

Alessandro Falaschi


Trasmissione dei Segnali e Sistemi di Telecomunicazione



Febbraio 2015
Edizione 1.3

Web Edition

ai miei genitori

<i>Titolo</i>	Trasmissione dei segnali e sistemi di telecomunicazione
<i>Autore</i>	Alessandro Falaschi - http://infocom.uniroma1.it/alef/wiki/
<i>Rilascio</i>	Edizione 1.3, Febbraio 2015
<i>ISBN</i>	978-88-91178-32-9
<i>Licenza</i>	 Creative Commons <i>Attribuzione - Non Commerciale - Condividi allo stesso modo</i> http://creativecommons.org/licenses/by-nc-sa/3.0/deed.it
<i>Sito Web</i>	http://infocom.uniroma1.it/alef/libro
<i>Strumenti editoriali</i>	<ul style="list-style-type: none"> . Lyx - http://www.lyx.org/ . Skencil - http://www.skencil.org/ . Inkscape - http://www.inkscape.org/ . Gimp - http://gimp.linux.it/www/ . Gnuplot - http://www.gnuplot.info/ . Octave - http://www.gnu.org/software/octave/
<i>Copertina</i>	Marco Sebastiani - http://www.marcosebastiani.it/
<i>Donazioni</i>	Se apprezzi questo lavoro ed hai voglia di supportarlo http://infocom.uniroma1.it/alef/wiki/Main/Donazioni
<i>Liberatoria</i>	L'eventuale inclusione non autorizzata di materiale protetto da copyright è da considerare transitoria, almeno fino a quando non saranno prodotte copie originali dello stesso. Ove possibile, sono forniti i riferimenti all'origine del materiale. L'autore si impegna alla rimozione immediata dei contenuti che saranno ritenuti lesivi dei diritti altrui.
<i>Dello stesso autore</i>	http://infocom.uniroma1.it/alef/wiki/Didattica/LoStratoApplicativodiInternet

Prefazione

Iniziai a autoprodurre questo testo come forma di rivincita rispetto alle fotocopie di appunti scritti a mano su cui da studente preparavo gli esami, e dopo qualche anno mi resi conto di essermi imbarcato in una impresa senza fine. Cionostante, ho sempre continuato a migliorare la qualità dello scritto e ad integrare nuovi argomenti, incoraggiato dal buon successo di lettori che riscuoteva durante la sua permanenza on-line, ed a distanza di più di quindici anni dalla partenza, annuncio il rilascio dell'edizione 1.3.

L'assoluta varietà degli argomenti affrontati, anziché spingere verso una trattazione superficiale e approssimata, pose fin da subito una sfida di completezza ed esattezza, affrontata cercando di mantenere un buon livello di rigore analitico, di alternare la teoria alle applicazioni, di bilanciare i concetti con gli esempi concreti. Ne è risultato un testo a più livelli, dove il corpo principale attribuisce la giusta rilevanza a concetti, principi, argomenti e tecniche, i cui aspetti più rigorosi sono approfonditi nell'ambito di numerose note ed appendici, mentre i risvolti pratici sono affrontati attraverso esempi ed esercizi.

Allo scopo di evitare inutili ripetizioni, si fa molto uso di rimandi e collegamenti che letteralmente *attraversano* l'intero testo e che consentono di *ricucire assieme* argomenti correlati ed interdipendenti, ma trattati in capitoli diversi, lasciando la libertà di seguire percorsi di lettura diversi. Inoltre, sono presenti *numerossime illustrazioni*, che mostrano sia gli schemi circuitali e simbolici dei dispositivi discussi, sia l'andamento delle curve di prestazione o di altre grandezze in funzione del tempo, della frequenza, o di parametri di sistema. Infine, per alcuni argomenti solo accennati si sono aggiunti riferimenti diretti a *Wikipedia*, invitando così il lettore all'approfondimento.

Giustamente ci si può chiedere: ma con tutti gli ottimi testi che già esistono su questi argomenti, che bisogno c'era di un ulteriore lavoro? A parte che quando iniziai a scrivere, alcuni testi in italiano ora disponibili non erano ancora usciti, ritengo che il mio lavoro abbia comunque prodotto un'opera che si distingue sotto diversi aspetti. Il più appariscente è probabilmente la *disponibilità gratuita* in formato elettronico, che ha di fatto reso il testo un riferimento comune a tutta la comunità linguistica italiana, e che ne permette la facile consultazione e navigabilità. Il secondo aspetto distintivo è la *varietà di argomenti* presenti, trattati in modo omogeneo e interdipendente, come difficilmente si riesce a fare in ambito universitario, a causa del livello di frammentazione didattica che lo affligge. Una terza considerazione riguarda l'elevata *qualità tipografica* per un testo autoprodotta, tanto più apprezzabile in quanto ottenuta con il solo utilizzo di strumenti *opensource*. Infine, l'aspetto forse più nascosto ma a mio avviso realmente qualificante, è l'attività di *revisione dinamica* a cui è sottoposto durante i periodi didattici, che lo rende materia in evoluzione continua.

Oltre alla pubblicazione web del testo in formato HTML, ho deciso di produrre anche la versione a *stampa on-demand* ed *ebook* tramite a *youcanprint.it*, acquistabile oltre che per posta, anche presso una delle più di mille librerie affiliate. Personalmente ritengo la versione fisica *insostituibile* per uno studio serio e ragionato, ma riconosco che anche la versione PDF *completamente navigabile* sia un formato di particolare valore intrinseco, e ne consento dunque il download a chi *acquista* una copia, o si rende disponibile a *supportarne* lo sviluppo tramite *donazione*.

Note di rilascio all'edizione 1.3

Breve riassunto delle novità rispetto alla versione dello scorso anno. Oltre al consueto lavoro di correzione, chiarimento, collegamento, re-impaginazione, in fondo al testo ora è presente un *indice analitico*, per compilare il quale ho svolto una revisione totale del testo, la cui sintassi è migliorata in svariati passaggi, così come i termini usati hanno guadagnato in coerenza e chiarezza. Quanto ai nuovi contenuti, mi sono tolto il gusto di approfondire gli aspetti relativi a:

- *equalizzazione* (§ 14.6), in particolare se attuata per via numerica: si parte dal metodo che minimizza l'errore quadratico medio (MMSE) e che produce un *filtro di Wiener*, con coefficienti calcolati risolvendo un sistema di equazioni, oppure mediante il metodo di *discesa del gradiente*, eventualmente realizzato in modo adattativo come LMS. Si accenna quindi al *decision feedback equalizer* o DFE, e poi si affronta la *maximum likelihood sequence detection* o MLSD, ...che si attua grazie all'*algoritmo di Viterbi*, a me ben noto per la sua applicazione nel riconoscimento vocale... ma quelli erano altri tempi!
- *comunicazioni radiomobili* (§ 15.3.4): si è fatta chiarezza sulla distinzione tra gli affievolimenti (*fading*) dovuti ai fenomeni su *grande* e su *piccola scala*, completata l'analisi delle condizioni per *fading piatto*, e migliorata l'esposizione relativa al *fading di Rayleigh*. E' stata aggiunta l'analisi relativa al fading *selettivo in frequenza* a cui è associato il concetto di *banda di coerenza*, e quindi affrontato il tema più legato all'effettivo movimento, che determina una *dispersione spettrale* ed una *variabilità temporale* del canale, quantificata dal concetto di *tempo di coerenza*;
- elementi di *statistica* (§ 7.5): si imposta il problema della *verifica di ipotesi*, mostrando come il criterio di *massima verosimiglianza* ne costituisca la naturale soluzione; lo stesso criterio è quindi applicato al problema della *stima di parametro*, in particolare di *media* e *varianza*, nel cui caso si ricavano le espressioni della varianza di stima;
- *trasformata di Fourier a tempo discreto* o DTFT (§ 4.3), assieme alla sua relazione con DFT e trasformata Z, raccontata per mezzo di nuove figure; contestualmente l'esposizione degli argomenti su campionamento, quantizzazione ed elaborazione numerica subisce varie migliorie, compresa la citazione della *somma di Poisson* (pag. 46) e la dimostrazione dell'*ortogonalità dei sinc* (§ 4.1.3);
- *filtro adattato* (§ 9.5), sembra incredibile ma ogni anno lo ritocco, stavolta si è evidenziato il legame con la decisione di *massima verosimiglianza*, aggiunto un paragrafo sul circuito *integrate and dump*, uno sul calcolo della P_e per *segnalazione unipolare* o *antipodale*, e uno sulla realizzazione del filtro adattato come un *correlatore*.

Sono poi stati fatti alcuni spostamenti di argomenti da un capitolo all'altro, sanando situazioni-limite frutto dello sviluppo graduale del testo: il capitolo della probabilità perde la quantizzazione, il calcolo della probabilità di errore, ed il filtro di ricezione ottimo, per cederli rispettivamente ai capitoli sul campionamento e sulle trasmissioni numeriche di banda base.

Alcuni capitoli, come l'11 e il 12 su *segnali modulati* e loro *prestazioni*, tra i primi ad essere scritti, hanno subito una revisione approfondita, guadagnandone in chiarezza. Il cap. 13 sulla *modulazione numerica* ha subito una profonda revisione, con l'aggiunta di diverse nuove figure, e la modifica dell'ordine di esposizione; la sezione sullo *spread spectrum* (§ 13.8) si è inoltre arricchita di un accenno al *frequency* e *time hopping*, e tra le "altre possibilità" sono citati anche i sistemi MIMO ed il *cognitive radio*. Il cap. 17 sulla teoria dell'informazione, relativamente recente, è stato oggetto di ampliamenti, avendo aggiunto un cenno al *soft decision decoding* ed una sotto-sezione sui *codici ciclici*, con la citazione dei codici BCH e di *Reed-Solomon*. Infine, nel cap. 18 si è migliorata l'esposizione della parte relativa alla *fonetica*, aggiungendo (e discutendo) la figura di una forma d'onda e del relativo *spetrogramma*, e si è approfondita l'esposizione dell'*analisi* LPC, aggiungendo il caso della stima del *periodo di pitch* mediante autocorrelazione.

Ci sono poi delle sviste (o meglio errori) che sono stati corretti, anche se uno si chiede come mai non siano stati segnalati prima da qualche lettore - ma per carità, *nessuno è perfetto!* Ad esempio, il prodotto scalare o *cross-energy* (pag. 31) si traduce come energia mutua, e non *incrociata!* Poi, la *distanza di Hamming* (pag. 98) conta il numero di bit differenti tra due codeword qualunque, e non la minima tra queste distanze! Inoltre, il precedente disegno della autocorrelazione (pag. 222) *non aveva* il massimo nell'origine! Poi, la densità spettrale di un segnale OFDM (pag. 327) *non è* la somma dei sinc quadri, ma il quadrato della somma dei sinc! Infine, è stata corretta l'espressione (§ 17.1.1.2) del legame tra frequenza binaria in uscita dal codificatore binario, e intensità informativa della sorgente.

Per ultimo, voglio condividere una illuminazione che ha finalmente fatto chiarezza su di un dubbio che mi tormentava fin da studente: perché se il filtro adattato deve avere risposta in frequenza coniugata a quello di trasmissione, si usa invece un filtro di ricezione passa-banda ideale? Perché nel caso di segnalazione a banda minima, il filtro adattato è appunto rettangolare!

Un sentito grazie a tutti coloro che mi hanno incoraggiato a continuare, ed *io continuerò!*

Alessandro Falaschi, 2015

Indice

Prefazione

Note di rilascio ii

1 Introduzione

1.1	Trasmissione dell'informazione	1
1.2	Trasmissioni numeriche	2
1.3	Segnali analogici, certi ed aleatori	5
1.3.1	Rappresentazione di segnali analogici	6
1.3.2	Rappresentazione di processi aleatori	6
1.3.3	Transito dei segnali attraverso sistemi fisici	6
1.4	Segnali numerici	7
1.5	Teoria delle probabilità	8
1.6	Sistemi di telecomunicazione	8
1.7	Segnali e sistemi	11
1.7.1	Caratteristiche dei sistemi	11
1.7.2	Caratteristiche dei segnali	12
1.7.3	Aspetti fisici delle grandezze energetiche	14

2 Serie di Fourier

2.1	Prerequisiti trigonometrici	15
2.1.1	Numeri complessi	15
2.1.2	Formule di Eulero	16
2.1.3	Fasori	16
2.2	Serie di Fourier	17
2.2.1	Segnali reali	18
2.2.1.1	Simmetria coniugata o Hermitiana	18
2.2.1.2	Interpretazione degli X_n come fasori	19
2.2.1.3	Serie trigonometrica	20
2.2.1.4	Serie di Fourier di un'onda rettangolare	20
2.2.2	Serie di Fourier troncata	22
2.3	Teorema di Parseval	23
2.4	Appendici	25
2.4.1	Algebra vettoriale	25
2.4.2	Esempi di Sviluppo in serie	27

3 Trasformata di Fourier

3.1	Definizione	29
3.2	Energia mutua e densità di energia	31

3.3	Prime proprietà della trasformata di Fourier	31
3.4	Impulso matematico	35
3.4.1	Trasformata di una costante	35
3.4.2	Trasformata di segnali periodici	35
3.4.3	Proprietà di <i>setacciamento</i>	36
3.5	Risposta impulsiva e convoluzione	36
3.5.1	Risposta impulsiva	37
3.5.2	Integrale di convoluzione	37
3.5.3	La risposta impulsiva come funzione memoria	38
3.5.4	Convoluzione con l'impulso traslato	39
3.6	Moltiplicazione in frequenza e nel tempo	39
3.6.1	Moltiplicazione in frequenza (<i>filtraggio</i>)	40
3.6.2	Moltiplicazione nel tempo (<i>modulazione e finestrazione</i>)	41
3.7	Derivazione ed integrazione nel tempo	42
3.8	Trasformata di segnali periodici	44
3.8.1	Treno di impulsi	44
3.8.2	Segnale periodico	45
3.8.3	Trasformata del treno di impulsi	45
3.8.4	Trasformata di segnale periodico	45
3.9	Appendici	46
3.9.1	Esercizio: quanti modi di calcolare la $\mathcal{F}\{x(t)\}$?	46
3.9.2	Sulla trasformata di una costante	47
3.9.3	Finestrazione e stima spettrale	48
3.9.4	Trasformata di un gradino	49
3.9.5	Sintesi delle proprietà della trasformata di Fourier	50
3.9.6	Trasformate di segnali	50

4 Campionamento, quantizzazione ed elaborazione numerica

4.1	Teorema del campionamento	51
4.1.1	Aliasing	53
4.1.2	Generalizzazione del filtro di restituzione	54
4.1.3	Ortogonalità delle funzioni sinc	54
4.1.4	Uso pratico	54
4.2	Quantizzazione	56
4.2.1	Quantizzazione lineare	56
4.2.1.1	SNR di quantizzazione	57
4.2.2	Quantizzazione logaritmica e codifica PCM	58
4.3	Trasformata di Fourier di sequenze	60
4.4	Trasformata discreta di Fourier	61
4.4.1	Relazione tra DTFT, DFT e trasformata zeta	64
4.4.2	Fast Fourier Transform	66
4.4.3	Relazione tra DFT e DCT	66
4.5	Filtraggio numerico via DFT	67
4.5.1	Convoluzione discreta	68
4.5.2	Convoluzione circolare	68
4.5.3	Convoluzione tra sequenze di durata finita via DFT	69
4.5.4	Convoluzione di segnali via DFT	69
4.6	Riassumendo	70
4.6.1	Le frequenze della DFT	70
4.6.2	Le ampiezze della DFT	71

5	Trasmissione dati in banda base	73	6	Reti di trasmissione a circuito	119
5.1	Trasmissione su canale numerico	73	6.1	Introduzione	119
5.1.1	Trasmissione numerica di banda base	73	6.1.1	Elementi della rete telefonica	119
5.1.2	Codifica di linea e segnale dati	74	6.1.2	La rete di accesso	120
5.1.2.1	Segnale binario e onda rettangolare	75	6.2	Multiplicazione	121
5.1.2.2	Effetto della limitazione in banda e ISI	75	6.2.1	Multiplicazione a divisione di tempo	121
5.1.2.3	Diagramma ad occhio	76	6.3	Rete plesiocrona	122
5.1.2.4	Trasmissione multilivello	77	6.3.1	Trama PCM	123
5.2	Generazione del segnale dati	78	6.3.2	Messaggi di segnalazione	124
5.2.1	Codici di linea a banda infinita	78	6.3.3	Sincronizzazione di centrale	125
5.2.2	Segnale dati limitato in banda	80	6.3.4	Multiplicazione asincrona e PDH	126
5.2.2.1	Requisiti per l'impulso di trasmissione	80	6.3.4.1	Bit stuffing	127
5.2.2.2	Condizioni di Nyquist	82	6.3.4.2	Add and Drop Multiplexer - ADM	127
5.2.2.3	Filtro a coseno rialzato	83	6.3.5	Sincronizzazione di rete	128
5.2.2.4	Codice di Gray	84	6.3.5.1	Elastic store	128
5.3	Errori nelle trasmissioni numeriche di banda base	86	6.4	Gerarchia digitale sincrona	128
5.3.1	Rumore bianco limitato in banda	86	6.5	Topologia di rete	132
5.3.2	Soglie di decisione	87	6.6	Rete in fibra ottica	133
5.3.3	Probabilità di errore per simbolo	87	6.6.1	Dispositivi SDH	133
5.3.3.1	Espressione della P_e per simbolo	89	6.6.2	Topologia ad anello	134
5.3.4	Diagramma ad occhio in presenza di rumore	91	6.6.2.1	Rete di trasporto	134
5.3.5	Uso del codice di Gray e P_e per bit	91	6.6.2.2	Rete di accesso in fibra	134
5.4	Gestione degli errori di trasmissione	93	6.6.3	Sistemi di protezione automatica	135
5.4.1	Controllo di errore	93	6.7	Instradamento	136
5.4.1.1	Errori su parole	94	6.8	Commutazione	137
5.4.2	Detezione di errore	95	6.8.1	Reti a divisione di spazio	137
5.4.2.1	Parità	95	6.8.2	Reti multistadio	138
5.4.2.2	Somma di controllo	95	6.8.3	Commutazione numerica a divisione di tempo	138
5.4.2.3	Codici polinomiali e CRC	96	6.8.3.1	Time Slot Interchanger	139
5.4.3	Correzione di errore e codifica di canale	98	6.8.3.2	Commutazione bidimensionale	139
5.4.3.1	Codice a blocchi	98	6.9	Appendici	140
5.4.3.2	Codice a ripetizione	99	6.9.1	Plain old telephone service (POTS)	140
5.4.3.3	Errori a burst ed interleaving	99	6.9.2	ISDN	142
5.5	Protocolli ARQ	100	6.9.3	Sistema di segnalazione numero 7	142
5.5.1	Send and wait	100	6.9.4	ADSL	144
5.5.2	Continuous RQ	102	6.9.5	TDM mediante modulazione di ampiezza degli impulsi	145
5.5.2.1	Go back N	102	6.10	Riferimenti	146
5.5.2.2	Selective repeat	102	7	Probabilità, processi, decisione statistica	147
5.5.2.3	Efficienza	103	7.1	Teoria delle probabilità	147
5.5.3	Controllo di flusso	104	7.1.1	Assiomi delle probabilità	148
5.5.3.1	Round trip time	104	7.1.2	Teoremi di base	148
5.5.3.2	Finestra scorrevole	104	7.1.3	Probabilità condizionali	148
5.5.3.3	Numero di sequenza	105	7.1.4	Teorema di Bayes	149
5.6	Sincronizzazione dati	108	7.1.5	Indipendenza statistica	150
5.6.1	Trasmissione asincrona	108	7.2	Variabili aleatorie	150
5.6.2	Trasmissione sincrona	110	7.2.1	Densità di probabilità e funzione di distribuzione	151
5.7	Ricevitore ottimo	112	7.2.2	Valore atteso, momento e momento centrato	152
5.8	Equalizzazione numerica	114	7.2.3	Variabile aleatoria a distribuzione uniforme	153
5.8.1	Equalizzazione zero forcing	115	7.2.4	Variabile aleatoria gaussiana e funzione $erfc\{.\}$	154
5.9	Codifica di carattere	116	7.2.5	Altre variabili aleatorie	155
5.9.1	Codifica UNICODE	116	7.3	Processi stazionari ed ergodici	155
			7.3.1	Media di insieme	156
			7.3.2	Media temporale	156
			7.3.3	Media temporale calcolata come media di insieme	157

7.3.4	Processo stazionario	157	9	Densità spettrale e filtraggio	217
7.3.5	Processo stazionario ed ergodico	158	9.1	Correlazione e covarianza	217
7.3.6	Riassumendo	159	9.1.1	Correlazione	218
7.3.7	Processo ad aleatorietà parametrica	159	9.1.2	Indipendenza statistica e incorrelazione	219
7.4	Trasformazione di v.a. e cambio di variabili	160	9.1.3	Correlazione di processo ergodico	220
7.4.1	Caso unidimensionale	160	9.1.4	Autocorrelazione e intercorrelazione	220
7.4.2	Caso multidimensionale	162	9.1.4.1	Proprietà dell'autocorrelazione	221
7.5	Elementi di statistica	164	9.2	Densità spettrale	222
7.5.1	Verifica di ipotesi	165	9.2.1	Teorema di Wiener	222
7.5.2	Funzione di verosimiglianza	165	9.2.2	Processo armonico	223
7.5.3	Decisione di massima verosimiglianza	165	9.2.3	Processo gaussiano bianco limitato in banda	223
7.5.4	Stima di parametro	165	9.2.4	Processo di segnale dati	224
7.5.5	Stima di massima verosimiglianza	166	9.3	Stima spettrale	225
7.6	Appendici	168	9.3.1	Periodogramma	225
7.6.1	Funzione caratteristica	168	9.4	Filtraggio di segnali e processi	226
7.6.2	Detezione di sinusoidi nel rumore	168	9.4.1	Densità spettrale in uscita da un filtro	226
8	Traffico, code e reti a pacchetto	171	9.4.2	Caratteristiche statistiche in uscita da un filtro	227
8.1	Distribuzione binomiale per popolazione finita	171	9.5	Filtro adattato	229
8.2	Distribuzione di Poisson	173	9.5.1	Segnalazione antipodale	233
8.2.1	Variabile aleatoria esponenziale negativa	174	9.5.2	Segnalazione ortogonale	233
8.3	Sistema di servizio orientato alla perdita	175	9.6	Caratteristiche dei sistemi fisici	234
8.3.1	Frequenza di arrivo e di servizio	176	9.7	Unità di elaborazione dei segnali	235
8.3.2	Intensità media di traffico	176	9.7.1	Prodotto	236
8.3.3	Probabilità di rifiuto	176	9.7.2	Somma	237
8.3.4	Efficienza di giunzione	178	9.7.3	Stima della autocorrelazione	238
8.3.5	Validità del modello	179	9.8	Filtri digitali	238
8.4	Sistemi di servizio orientati al ritardo	180	9.8.1	Filtro trasversale	238
8.4.1	Risultato di Little	181	9.8.2	Filtro trasversale del 1° ordine	239
8.4.2	Sistemi a coda infinita ed a server unico	181	9.8.3	Filtro a risposta impulsiva <i>infinita</i> del 1° ordine	240
8.4.3	Sistemi a coda finita e con più server	183	9.9	Filtri analogici	241
8.5	Reti per trasmissione dati	185	9.9.1	Filtro analogico ad un polo	241
8.5.1	Il pacchetto dati	185	9.9.2	Frequenza di taglio	242
8.5.2	Modo di trasferimento delle informazioni	186	9.9.3	Assenza di distorsione lineare	242
8.5.2.1	Schema di moltiplicazione	187	9.10	Appendici	242
8.5.2.2	Principio di commutazione	187	9.10.1	Coefficiente di correlazione	242
8.5.2.3	Architettura protocollare	190	9.10.2	Gaussiana multidimensionale ed indipendenza statistica	243
8.6	Appendici	193	9.10.3	Densità spettrale per onda PAM	243
8.6.1	La rete Internet	193	9.10.4	Potenza di un segnale dati	246
8.6.1.1	Storia	193	9.10.5	Autocorrelazione dell'uscita di un filtro	247
8.6.1.2	Le caratteristiche	194	9.10.6	Grafici di esempio	247
8.6.1.3	Gli indirizzi	194	10	Segnali modulati	249
8.6.1.4	TCP	197	10.1	Caratteristiche ed applicazioni	249
8.6.1.5	IP	200	10.1.1	Moltiplicazione a divisione di frequenza - FDM	249
8.6.1.6	Ethernet	203	10.1.1.1	Collegamenti punto-multipunto	250
8.6.1.7	Fast e Gigabit Ethernet	206	10.1.1.2	Accesso multiplo	250
8.6.2	Rete ATM	208	10.1.1.3	Collegamenti punto-punto	251
			10.1.2	Canale telefonico	251
			10.1.3	Antenne e lunghezza d'onda	252
			10.1.4	Banda di segnale	252
			10.1.5	Trasmissione a banda laterale unica	252
			10.2	Rappresentazione dei segnali modulati	253
			10.2.1	Inviluppo complesso	253

10.2.2	Modulazione di ampiezza e/o angolare	254	11.3.1.1	Ricevitore a PLL	283
10.2.3	Componenti analogiche di bassa frequenza	255	11.3.1.2	Ricevitore a discriminatore	284
10.2.4	Filtro di Hilbert	255	11.3.2	Densità spettrale di segnali a modulazione angolare	284
10.2.4.1	Estrazione delle componenti analogiche di bassa frequenza	256	11.3.2.1	Segnale modulante sinusoidale	285
10.2.5	Segnale analitico	257	11.3.3	Densità spettrale FM con processo aleatorio modulante	288
10.2.6	Densità spettrale di segnali passabanda	258	11.3.3.1	Indice di modulazione per processi	289
10.2.7	Schema delle trasformazioni	260	11.3.3.2	Modulazione a basso indice	289
10.3	Transito dei segnali modulati nei sistemi fisici	260	11.4	Appendici	290
10.3.1	Filtraggio	260	11.4.1	Calcolo della potenza di un segnale AM BLU	290
10.3.2	Intermodulazione tra componenti analogiche	261	11.4.1.1	Calcolo della potenza di segnali BLD-PI, PS, PPS	291
10.3.3	Equalizzazione di banda base	261	11.4.2	Ricostruzione della portante mediante quadratura	291
10.4	Condizioni per inviluppo complesso reale	262	11.4.3	Il mixer	291
10.4.1	Filtro passa banda ideale	262	11.4.4	Trasmissione televisiva	292
10.4.2	Simmetria coniugata attorno ad f_0	262	11.4.5	Modulazione FM a basso indice	295
10.5	Demodulazione delle componenti analogiche di b. f.	262	11.4.6	FM broadcast	295
10.6	Rappresentazione dei processi in banda traslata	263	12 Prestazioni delle trasmissioni modulate	297	
10.6.1	Conclusioni	265	12.1	Il rumore nei segnali modulati	297
10.6.2	Processo gaussiano bianco limitato in banda	266	12.1.1	Rapporto segnale-rumore e banda di rumore	297
10.7	Appendici	266	12.1.2	Demodulazione di un processo di rumore	298
10.7.1	Risposta impulsiva del filtro di Hilbert	266	12.2	Prestazioni delle trasmissioni AM	299
10.7.2	Trasformata di Hilbert di un segnale modulato	267	12.2.1	Potenza di segnale e di rumore dopo demodulazione ed SNR	299
10.7.3	Autocorrelazione di processi passabanda	267	12.2.1.1	Modulazione BLD-PS	299
11 Modulazione per segnali analogici	271		12.2.1.2	Modulazione BLU-PS	300
11.1	Modulazione di ampiezza - AM	271	12.2.1.3	Modulazione BLD-PI	300
11.1.1	Banda laterale doppia - BLD	272	12.3	Prestazioni delle trasmissioni FM	301
11.1.1.1	Portante soppressa - PS	272	12.3.1	Rumore dopo demodulazione FM	301
11.1.1.2	Portante intera - PI	272	12.3.2	Caso di basso rumore	302
11.1.1.3	Portante parzialmente soppressa - PPS	273	12.3.3	Caso di elevato rumore	305
11.1.1.4	Efficienza di PI-PPS	273	12.3.4	Enfasi e de-enfasi	306
11.1.2	Banda laterale unica - BLU	274	13 Modulazione numerica	307	
11.1.2.1	Generazione di segnali BLU	275	13.1	Modulazione di ampiezza	307
11.1.3	Banda laterale ridotta - BLR	275	13.1.1	Modulazione BPSK	307
11.1.4	Potenza di un segnale AM	275	13.1.2	Modulazione L-ASK	309
11.2	Demodulazione di ampiezza	275	13.1.3	Prestazioni di L-ASK	310
11.2.1	Demodulazione coerente o omodina	276	13.2	Modulazione di fase	311
11.2.1.1	Errori di fase e frequenza	276	13.2.1	Modulazione QPSK ed L-PSK	311
11.2.1.2	Phase Locked Loop - PLL	276	13.2.2	Prestazioni QPSK	313
11.2.2	Demodulazione incoerente in fase e quadratura	277	13.2.3	Prestazioni L-PSK	314
11.2.3	Demodulatore di inviluppo per AM-BLD-PI	278	13.3	Modulazione QAM	316
11.2.4	Demodulazione per segnali a banda laterale unica e ridotta	279	13.3.1	Prestazioni di QAM	317
11.2.5	Demodulatore eterodina	279	13.4	Codifica differenziale	319
11.2.5.1	Frequenze immagine	280	13.4.1	Modulazione DBPSK	319
11.2.5.2	Supereterodina	281	13.4.2	DQPSK	321
11.3	Modulazione angolare	281	13.4.3	OFDM differenziale	321
11.3.1	Ricezione di un segnale a modulazione angolare	283	13.5	Modulazione di frequenza L-FSK	322
			13.5.1	Modulazione FSK ortogonale	322
			13.6	Schema riassuntivo delle prestazioni	324
			13.7	Modulazione OFDM	324
			13.7.1	Rappresentazione nel tempo ed in frequenza	325
			13.7.2	Architettura di modulazione	328

13.7.3	Efficienza dell'OFDM	329	14.7.2	Ingresso aleatorio	370
13.7.4	Architettura di demodulazione	330	14.7.3	Effetto sulla modulazione FM	371
13.7.5	Prestazioni	330	14.8	Appendici	372
13.7.5.1	Calcolo della P_e per portante	331	14.8.1	Valutazione dell' SNR dovuto a diverse fonti di disturbo	372
13.7.5.2	Potenza di rumore per portante	331	14.8.2	Potenza assorbita da un bipolo	372
13.7.5.3	Prestazioni per portante	333	14.8.3	Condizioni per il massimo trasferimento di potenza	373
13.7.5.4	Caso di rumore bianco	334	14.8.4	Potenza ceduta ad un carico $Z_c(f) \neq Z_g^*(f)$	374
13.7.5.5	Confronto con la portante singola	334	14.8.5	Derivazione del tempo di ritardo di gruppo	375
13.7.6	Equalizzazione	335	15 Collegamenti e mezzi trasmissivi	377	
13.7.6.1	Codifica differenziale	335	15.1	Bilancio di collegamento	377
13.7.7	Sensibilità alla temporizzazione	335	15.2	Collegamenti in cavo	379
13.7.8	Ottimalità	335	15.2.1	Costanti distribuite, grandezze derivate, e condizioni generali	379
13.7.9	Codifica	336	15.2.2	Trasmissione in cavo	381
13.7.10	Portanti pilota	337	15.2.2.1	Casi limite	383
13.8	Sistemi a spettro espanso	337	15.2.3	Tipologie di cavi per le telecomunicazioni	384
13.8.1	Sequenze pseudo-casuali	338	15.2.3.1	Coppie simmetriche	384
13.8.2	Sequenza diretta	339	15.2.3.2	Cavo coassiale	385
13.8.2.1	Guadagno di processo	340	15.3	Collegamenti radio	387
13.8.2.2	Despreading	340	15.3.1	Trasduzione elettromagnetica	387
13.8.2.3	Prestazioni in presenza di rumore	341	15.3.2	Trasferimento energetico	388
13.8.2.4	Prestazioni in presenza di un tono interferente	341	15.3.3	Condizioni di propagazione e attenuazioni supplementari	389
13.8.2.5	Accesso multiplo	342	15.3.3.1	Condizioni di visibilità	389
13.8.3	Frequency Hopping	343	15.3.3.2	Diffusione e riflessione atmosferica	390
13.8.4	Time Hopping	343	15.3.3.3	Absorbimento atmosferico	391
13.9	Altre possibilità	343	15.3.3.4	Dimensionamento di un collegamento soggetto a pioggia	391
13.10	Appendici	344	15.3.3.5	Cammini multipli	392
13.10.1	Sincronizzazione	344	15.3.3.6	Collegamento in diversità	394
13.10.2	Ortogonalità tra simboli sinusoidali	345	15.3.4	Collegamenti radiomobili	394
14 Transito nei circuiti e distorsioni	349		15.3.4.1	Le componenti della attenuazione supplementare	395
14.1	Caratterizzazione dei circuiti	350	15.3.4.2	Path loss	396
14.2	Bipoli	351	15.3.4.3	Fading su larga scala e shadowing	396
14.2.1	Potenza assorbita da un bipolo	351	15.3.4.4	Fading su piccola scala	397
14.2.2	Connessione tra generatore e carico	352	15.3.4.5	Fading selettivo in frequenza	399
14.2.2.1	Potenza disponibile e massimo trasferimento di potenza	352	15.3.4.6	Dispersione spettrale e variabilità temporale	401
14.2.2.2	Assenza di distorsioni lineari	353	15.4	Collegamenti in fibra ottica	404
14.2.2.3	$Z_g(f)$ reale	353	15.4.1	Trasmissione ottica	404
14.3	Reti due porte	353	15.4.2	Bilancio di collegamento	407
14.3.1	Modello circuitale	353	15.4.3	Multiplicazione a divisione di lunghezza d'onda - WDM	410
14.3.2	Schema simbolico	354	15.4.4	Ridondanza e pericoli naturali	411
14.3.3	Trasferimento energetico	354	15.4.5	Sonet e SDH	411
14.4	Misure di potenza in decibel	357	15.4.6	Dalle fibre ottiche alle reti ottiche	411
14.5	Distorsione lineare	359	15.5	Appendici	412
14.5.1	Guadagno di potenza in dB	359	15.5.1	Collegamenti satellitari	412
14.5.2	Tempo di ritardo di gruppo	359	15.5.2	Allocazione delle frequenze radio	416
14.5.3	Effetto della distorsione lineare sui segnali	360			
14.5.3.1	Segnali di banda base	360			
14.5.3.2	Segnali modulati	361			
14.5.4	Calcolo dell' SNR	362			
14.6	Equalizzazione	362			
14.6.1	Equalizzazione numerica	363			
14.6.1.1	Equalizzatore MMSE	364			
14.6.1.2	Equalizzatore decision feedback	366			
14.6.1.3	Equalizzatore di massima verosimiglianza	366			
14.6.1.4	Prestazioni	368			
14.7	Distorsione di non linearità	369			
14.7.1	Ingresso sinusoidale	369			

16 Rumore termico	419	17.2.3	Capacità di canale discreto	451
16.1 Rumore nei bipoli passivi	419	17.2.4	Capacità di canale continuo	453
16.2 Rapporto segnale rumore dei generatori	420	17.3	Codici di canale	457
16.3 Rumore nelle reti due porte	420	17.3.1	Codice lineare a blocchi	459
16.3.1 Reti passive	421	17.3.1.1	Codice di Hamming	460
16.3.1.1 Rapporto SNR in uscita	422	17.3.1.2	Codici ciclici	462
16.3.1.2 Fattore di rumore per reti passive	422	17.3.2	Codice convoluzionale	464
16.3.2 Reti attive	422	18 Codifica di sorgente multimediale	469	
16.3.2.1 Fattore di rumore per reti attive	423	18.1	Codifica audio	469
16.3.3 Fattore di rumore per reti in cascata	424	18.1.1	Codifica di forma d'onda	470
16.4 Rumore nei ripetitori	427	18.1.1.1	DPCM o PCM Differenziale	470
16.4.1 Rumore termico accumulato	428	18.1.1.2	ADPCM o DPCM Adattivo	470
16.4.2 Compromesso tra rumore termico e di intermodulazione	429	18.1.1.3	Codifica per sottobande	473
17 Teoria dell'informazione e codifica	431	18.1.2	Codifica basata su modello	473
17.1 Codifica di sorgente	431	18.1.2.1	Produzione del segnale vocale	473
17.1.1 Codifica di sorgente discreta	432	18.1.2.2	Codifica a predizione lineare - LPC	476
17.1.1.1 Entropia	433	18.1.2.3	Predizione lineare ad eccitazione residuale - RELP	479
17.1.1.2 Intensità informativa e codifica binaria	434	18.1.2.4	Quantizzazione vettoriale	480
17.1.1.3 Codifica con lunghezza di parola variabile	435	18.1.2.5	Predizione lineare ad eccitazione codificata - CELP	482
17.1.1.4 Codifica per blocchi	438	18.1.3	Codifica psicoacustica	483
17.1.1.5 Sorgente con memoria	440	18.2	Codifica di immagine	486
17.1.1.6 Codifica per sorgenti con memoria	441	18.2.1	Dimensioni	486
17.1.1.7 Compressione basata su dizionario	442	18.2.2	Spazio dei colori	488
17.1.2 Codifica con perdite di sorgente continua	444	18.2.3	Formato GIF	490
17.1.2.1 Curva velocità-distorsione	444	18.2.4	Codifica JPEG	492
17.1.2.2 Entropia di sorgente continua	445	18.3	Codifica video	498
17.1.2.3 Sorgente continua con memoria	446	18.3.1	Standard video	503
17.2 Codifica di canale	447	18.3.1.1	H.261	503
17.2.1 Canale simmetrico binario e decisore Bayesiano	447	18.3.1.2	H.263	505
17.2.2 Informazione mutua media per sorgenti discrete	449	18.3.1.3	MPEG-1	507
		18.3.1.4	MPEG-2	508
		18.3.2	Contenitori	510
		18.3.2.1	Transport Stream	512
		Bibliografia		515
		Indice analitico		517

Capitolo 1

Introduzione

Svolgiamo innanzitutto una rassegna dei molteplici aspetti che intervengono nei sistemi di telecomunicazione, che oltre al semplice inoltro per via *elettrica* di un *messaggio informativo* da un luogo ad un altro, coinvolgono un discreto numero di apparati differenti e cooperanti, nel contesto di una organizzazione *in rete* dei dispositivi.

La trasmissione può riguardare un messaggio generato al tempo stesso della sua trasmissione, oppure esistente a priori. Il supporto fisico del messaggio, chiamato *segnale*, identifica due categorie molto generali: quella dei segnali *analogici*, e quella dei segnali *numerici*. Nel primo caso rientra ad esempio la voce umana, mentre esempi di segnali numerici sono i documenti conservati su di un computer.

1.1 Trasmissione dell'informazione

Sorgente, destinatario e canale L'origine del segnale da trasmettere è indicata (vedi Fig. 1.1) come *sorgente*, di tipo analogico o numerico per i due tipi di segnale. Ciò che giace tra sorgente e *destinatario* viene descritto da una entità astratta denominata *canale* di comunicazione, le cui caratteristiche condizionano i messaggi trasmessi.

Distorsioni e disturbi Il canale può ad esempio imporre una limitazione alla *banda di frequenze* del segnale in transito¹; cause fisiche ineliminabili producono inoltre, al lato ricevente, l'insorgere di un segnale di disturbo additivo, comunemente indicato con il termine di *rumore*, che causa la ricezione di un segnale diverso da quello stesso presente all'uscita del canale. Pertanto, ci si preoccupa di caratterizzare il canale in modo da scegliere i metodi di trasmissione più idonei a rendere minima l'alterazione sul messaggio trasmesso.

Rapporto segnale-rumore L'entità delle alterazioni subite dal messaggio viene spesso quantificata nei termini del *rapporto segnale rumore* (SNR o SIGNAL-TO-NOISE RATIO), che rappresenta un indice di qualità del collegamento stesso, e che per ora definiamo genericamente come il rapporto tra l'entità del segnale utile ricevuto e quello del rumore ad esso sovrapposto, indicato con n nella figura 1.1.

¹Approfondiremo nel seguito il senso di questa locuzione; per ora è sufficiente interpretarla in termini generici, ovvero di fedeltà della riproduzione al segnale originario.

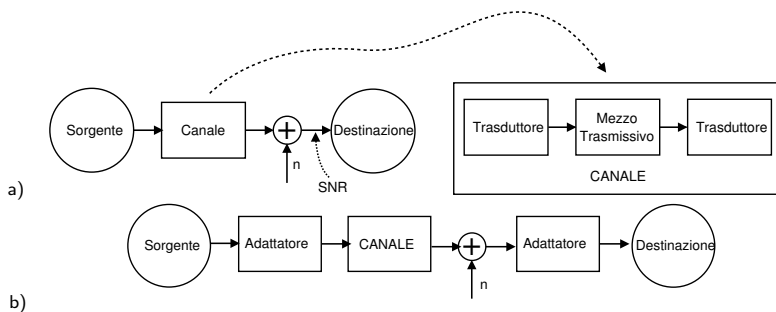
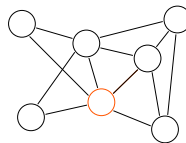


Figura 1.1: Elementi funzionali per la trasmissione dell'informazione

Trasmissione La Fig. 1.1a) evidenzia come il canale, nella realtà fisica, è costituito da un *mezzo trasmissivo* su cui si propaga un segnale di natura elettromagnetica, che viene convertito in tale forma da appositi *trasduttori* di trasmissione e ricezione². Considerando per il momento i trasduttori come facenti parte del canale stesso, proseguiamo l'analisi concentrandoci sugli ulteriori aspetti del processo di comunicazione.

Adattatori La figura 1.1b) evidenzia l'esistenza (in trasmissione, in ricezione od a entrambe le estremità) di dispositivi *adattatori*, che hanno lo scopo di ridurre od eliminare le cause di deterioramento del messaggio introdotte dalla trasmissione: si può ad esempio ricorrere ad *equalizzatori* per correggere la risposta in frequenza di un canale, ad *amplificatori* per contrastare l'attenuazione subita dal segnale, ovvero a *codificatori di linea* per rendere le caratteristiche del segnale idonee ad essere trasmesse sul canale a disposizione.

Rete La trasmissione lungo un canale in uso esclusivo alla coppia sorgente-destinazione è piuttosto raro; di solito i collegamenti sono condivisi tra più comunicazioni, ognuna con differente origine e destinatario. Il problema della condivisione delle risorse trasmissive, ed il coordinamento di queste attività, produce la necessità di analizzare in modo esplicito le *reti di telecomunicazione*, che entrano a far parte integrante dei sistemi di trasmissione dell'informazione.



Gli aspetti delle telecomunicazioni brevemente accennati sono immediatamente applicabili a segnali di natura *analogica*, in cui il segnale è definito per tutti gli istanti di tempo, ed assume valori qualsiasi. Nel caso invece in cui il segnale è definito solo per istanti di tempo discreti e valori discreti, si entra nell'ambito delle *trasmissioni numeriche*.

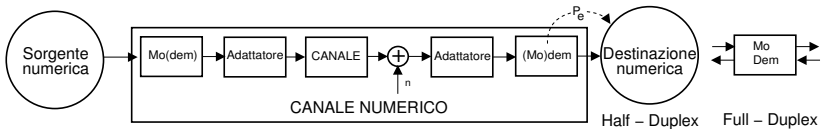
1.2 Trasmissioni numeriche

Sorgente e destinazione numerica il messaggio informativo in questo caso è di natura discreta, ossia ad intervalli di tempo regolari sono prodotti *simboli* ap-

²Un classico esempio di trasduttore è quello dell'antenna, nel caso di trasmissione radio.

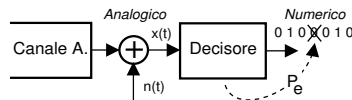
partenenti ad un insieme finito, come ad es. le lettere dell'alfabeto;

Modem i simboli che compongono un segnale *numerico* devono essere trasformati in un segnale analogico mediante l'utilizzo di dispositivi chiamati *modem*³, come rappresentato dalla figura seguente, in cui è evidenziato come per una trasmissione unidirezionale⁴ occorra solo *metà* delle funzioni del modem per entrambi i lati del collegamento, mentre nel caso di collegamento *full duplex* (in cui entrambi gli estremi possono essere contemporaneamente sorgente e destinazione) il modem opera allo stesso tempo nelle due direzioni.



Canale numerico la figura precedente suggerisce come sia possibile racchiudere tutto ciò che è compreso tra i due modem in un *unico* blocco, denominato *canale numerico*. Quest'ultimo è concettualizzato come una entità autonoma, e nel progetto di una sistema di comunicazione numerica è caratterizzato da un *fattore di qualità* individuato dalla ...

Probabilità di errore quando il modem ricevente (in figura indicato come *decisore*) produce un simbolo *diverso* da quello trasmesso si verifica un *errore*, e la frazione percentuale di questi eventi rispetto al totale costituisce la *probabilità di errore*. Gli errori sono causati dal *rumore additivo* presente in uscita dal canale analogico, e/o dalle alterazioni introdotte dallo stesso. Evidentemente, le *prestazioni* individuate dalla probabilità di errore sono strettamente legate a quelle individuate dal rapporto segnale-rumore per il canale analogico sottostante.



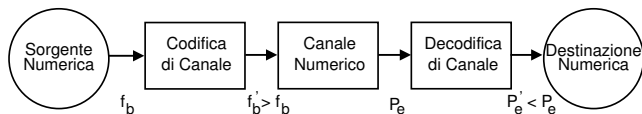
Codifica di canale Nelle trasmissioni numeriche, si può introdurre una *ridondanza* nella sequenza trasmessa, inviando più simboli di quanti non ne produca la sorgente, e quindi di fatto *umentando* il numero di simboli da trasmettere per unità di tempo; i simboli in più sono scelti in modo da essere in qualche modo *dipendenti* tra loro, e questa loro caratteristica rende possibile la riduzione della probabilità di errore di cui soffre il canale numerico. Infatti, grazie alla dipendenza (nota) tra i simboli trasmessi, il ricevitore è ora in grado di "accorgersi" che si è verificato un errore, in quanto la dipendenza prevista non è più rispettata; pertanto, il ricevitore può attuare delle contromisure. La ridondanza introdotta può essere così elevata da permettere la correzione di errori isolati⁵, oppure il ricevitore può semplicemente richiedere la *ritrasmissione* del simbolo errato. Le trasformazioni del segnale ora descritte prendono il nome di *codifica*

³La parola *Modem* è una contrazione delle due parole *modulatore-demodulatore*.

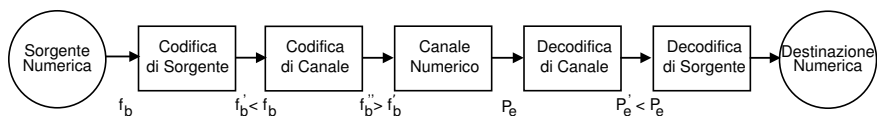
⁴Nelle trasmissioni unidirezionali, sorgente e destinazione non si scambiano i ruoli. La trasmissione stessa viene anche indicata con il termine di *half-duplex*.

⁵Si parla in questo caso di codifica FEC, ovvero di *Forward Error Correction*.

di canale, e devono essere “rimosse” all’uscita dello stesso da un processo inverso di *decodifica*.



Codifica di sorgente Possono essere introdotti due ulteriori blocchi, che operano una *codifica* (e relativa de-codifica) *di sorgente* sulla sequenza trasmessa, con uno scopo è per così dire “inverso” a quello della codifica di canale: infatti, la codifica di sorgente *rimuove le dipendenze* tra i simboli presenti nelle sequenze generate dalla sorgente, ottenendo di fatto un riduzione del numero di simboli da trasmettere per unità di tempo⁶. Un tipico esempio di codifica di sorgente è rappresentato dagli algoritmi di compressione esistenti per i file di computer (come i file *zippati*); in tal caso, il fattore di compressione ottenibile dipende dalla natura del file trattato, ed è tanto maggiore quanto più quest’ultimo presenta caratteristiche di ripetitività e quindi di predicibilità del suo contenuto. In altre parole, l’uscita di un codificatore di sorgente è una sequenza di simboli tendenzialmente *indipendenti* tra loro, nel senso che ogni forma di predicibilità di un simbolo a partire dai circostanti è stata rimossa.



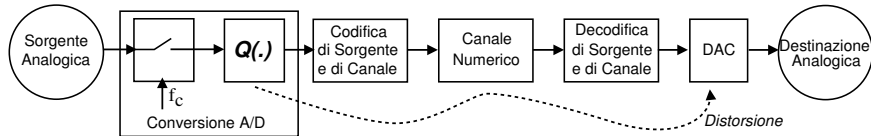
Campionamento e quantizzazione Poniamoci ora il problema di utilizzare un *canale numerico* per effettuare una *trasmissione analogica*. Il vantaggio di tale “contorsione” è da ricercarsi nel migliore comportamento delle trasmissioni numeriche rispetto ai disturbi, nonché alla loro *generalità*⁷. Per ottenere il risultato desiderato, occorre applicare alla sorgente analogica un procedimento di *campionamento*, prelevandone i valori ad istanti discreti, e quindi di *quantizzazione*, rappresentando tali valori mediante un insieme finito di simboli. Il risultato è una sequenza numerica che può essere di nuovo convertita nel segnale originario, utilizzando un dispositivo di conversione *digitale-analogica* (DAC) dal lato del ricevitore. Esempi pratici di quest’ultimo processo sono ben noti, come ad esempio nel caso dei CD audio.

Rumore di quantizzazione La riduzione dei valori campionati nell’ambito di un insieme finito produce una *ulteriore distorsione*, che può essere pensata sommarsi in ricezione al segnale originario, producendo una nuova fonte di degrado. L’entità del rumore dovuto alla quantizzazione è inversamente legata alla *risoluzione del quantizzatore*, ovvero alla capacità di differenziare tra valori di ingresso molto

⁶Pensiamo per similitudine ad un imballaggio, il cui contenuto è prima disposto in modo da occupare il minimo volume (codifica di sorgente), ed a cui viene poi aggiunto del materiale antiurto (codifica di canale).

⁷Nei collegamenti numerici, non occorre specializzare il metodo di trasmissione al mezzo a disposizione, anzi quest’ultimo è totalmente “mascherato” dal fornitore del collegamento numerico stesso.

vicini tra loro. In definitiva, la distorsione risulta tanto minore, quanto maggiore è il flusso informativo⁸ prodotto dal quantizzatore, espresso in bit/secondo.



Teoria velocità-distorsione CLAUDE SHANNON enunciò negli anni '50 una serie di teoremi, che sono la base dell'analisi dei sistemi di comunicazione. Tali aspetti sono affrontati al Cap. 17, e possono essere riassunti nei seguenti tre enunciati:

- un qualsiasi canale pone un limite al massimo flusso informativo che transita in esso. Il limite deriva dai vincoli che il canale impone sulla massima banda B del segnale in transito, sulla massima potenza di segnale S ricevuta, e sulla potenza di rumore N presente al ricevitore. Il massimo flusso di informazione in transito prende il nome di *capacità di canale* C , e può essere espresso come $C = B \log_2 \left(1 + \frac{S}{N}\right)$ bit/sec. In questi termini, la massima velocità di trasmissione in un canale può dipendere da una limitazione sulla banda, o sulla potenza, od essere causata da un eccessivo rumore: solo alterando uno di questi fattori (scegliendo un diverso canale), è possibile trasmettere a velocità più elevata;
- una qualsiasi sorgente produce un flusso informativo in bit/secondo tanto più elevato quanto minore è la distorsione introdotta dal processo di quantizzazione;
- considerando una coppia sorgente + canale, dato che il canale limita il massimo flusso informativo prodotto dalla sorgente, quest'ultima verrà necessariamente riprodotta con una distorsione tanto maggiore quanto minore è la capacità di canale. A meno di non impiegare più tempo per la trasmissione... oppure di cambiare canale (ad esempio, usandone uno con una banda maggiore).

1.3 Segnali analogici, certi ed aleatori

I segnali analogici, indicati con $s(t)$ (con le parentesi tonde), rappresentano *l'andamento nel tempo* di una grandezza fisica. Come esempio possiamo citare il *segnale vocale*, in cui un'onda trasversale di pressione-velocità è convertita in una tensione da un microfono. Oppure citare un *segnale di immagine*, che è bidimensionale, definito quindi su di un piano anziché nel tempo, rappresentato da una grandezza $S(x, y)$ che ne individua la luminanza, e scandito per linee generando un segnale temporale. Un segnale può anche presentare *valori complessi*⁹, e in tal caso assume contemporaneamente due diversi valori (parte reale e parte immaginaria, oppure modulo e fase).

E' importante distinguere tra i segnali cosiddetti *certi* e quelli *aleatori*. Un esempio di *segnale certo* può essere una cosinusoide di cui sia nota sia l'ampiezza che la fase, mentre un *segnale aleatorio* non è noto con esattezza prima che questo venga prodotto (ad esempio il rumore di un ruscello, o le notizie diramate da un telegiornale). L'insieme di tutti i segnali aleatori appartenenti ad una medesima classe viene indicato

⁸Vedremo infatti al § 4.2.1.1 che un aumento della risoluzione del processo di quantizzazione corrisponde ad un aumento del numero di bit necessari a rappresentare ogni valore (o campione) di segnale.

⁹Come vedremo al cap. 10, un segnale a valori complessi è il risultato di una particolare rappresentazione, detta *inviluppo complesso*, utile nell'analisi dei segnali modulati.

nel suo complesso come *processo aleatorio*, ed un segnale particolare di questo insieme come una sua *realizzazione*.

1.3.1 Rappresentazione di segnali analogici

Lo studio delle proprietà dei segnali si articola prendendo in considerazione per gli stessi rappresentazioni alternative, scelte in modo da poter valutare più agevolmente le alterazioni subite nel passaggio attraverso sistemi fisici. In particolare, sarà definito lo *sviluppo in serie di Fourier* per la rappresentazione dei segnali periodici, e quindi la *trasformata di Fourier* che descrive una classe più ampia di segnali.

L'analisi di Fourier consente di definire il concetto di *banda occupata* da un segnale, nonché di come la sua potenza e/o energia si *distribuisce in frequenza*; quest'ultimo andamento viene indicato con il termine di *spettro di densità di potenza* (o di *energia*).

1.3.2 Rappresentazione di processi aleatori

Anche nel caso in cui il segnale non è noto a priori, e dunque è impossibile calcolarne la trasformata di Fourier in forma chiusa, si può ugualmente giungere ad una rappresentazione che caratterizzi le realizzazioni del processo nei termini della distribuzione (statistica) in frequenza della potenza di segnale.

Ciò è possibile considerando la *funzione di autocorrelazione*, che esprime il grado di interdipendenza statistica tra i valori assunti in istanti diversi dalle realizzazioni del processo, e che costituisce un elemento unificante ai fini della *stima spettrale* dei segnali.

Osserveremo come processi molto correlati siano caratterizzati da una densità di potenza di tipo *colorato*, mentre processi scarsamente correlati saranno identificati da una densità di potenza di tipo *bianco*¹⁰.

1.3.3 Transito dei segnali attraverso sistemi fisici

Saranno forniti i metodi di valutazione dei *peggioramenti* indotti su di un segnale che transiti in un mezzo trasmissivo, sulla base di alcuni concetti chiave, e di come riuscire a *ridurli* od *eliminarli*.

Risposta impulsiva e convoluzione Per mezzo dell'integrale di *convoluzione* si esprime in forma chiusa l'uscita di un sistema (ad es. un circuito elettrico¹¹) in base alla conoscenza dell'ingresso, e di una particolare caratteristica del sistema, la *risposta impulsiva*. Quest'ultima rappresenta la sua uscita quando in ingresso è presente una particolare funzione analitica, detta *impulso matematico*.

Risposta in frequenza Operando nel dominio della frequenza, osserveremo come la trasformata di Fourier della risposta impulsiva rappresenti la *risposta in frequenza* della rete, ovvero l'uscita alle diverse frequenze quando l'ingresso ha uno spettro bianco.

¹⁰I termini *colorato* e *bianco* hanno origine da una similitudine con l'energia luminosa, per cui se la luce bianca indica l'indiscriminata presenza di tutte le lunghezze d'onda, così uno spettro bianco indica la presenza in egual misura di tutte le frequenze; viceversa, come una luce colorata dipende dal prevalere di determinate frequenze nella radiazione elettromagnetica, così uno spettro colorato indica la prevalenza di alcune frequenze su altre.

¹¹Quando un circuito elettrico ha la funzione di trasportare un segnale tra una coppia di morsetti ad un'altra, il circuito prende il nome di *rete due porte* o *quadripolo*.

Modulazione Nel caso in cui il segnale da trasmettere occupi una banda concentrata attorno ad una frequenza più o meno elevata (detta portante), come nel caso dei *segnali modulati*, si ricorre alla rappresentazione mediante le *componenti analogiche di bassa frequenza*. Il caso opposto, caratterizzato da una estensione frequenziale contigua alla frequenza zero, è detto di *banda base*. L'uso dei segnali modulati è obbligatorio, qualora questi debbano essere trasmessi su canali di tipo cosiddetto *passa-banda*.

Trasferimento energetico In conseguenza delle particolarità del mezzo trasmissivo, e delle condizioni di *adattamento di impedenza*, il segnale è ricevuto con una potenza ridotta, che non deve scendere sotto la *soglia di sensibilità* del ricevitore.

La trasmissione dei segnali mediante un sistema di comunicazione coinvolge diversi altri aspetti, che sono brevemente introdotti al § 1.6, nel contesto della caratterizzazione dei sistemi di telecomunicazione.

1.4 Segnali numerici

Sono indicati con la notazione $s[k]$ (con le parentesi quadre) o s_k (con un pedice), per evidenziare che il loro dominio è l'insieme dei numeri interi. Sono valide le stesse definizioni fornite al § 1.7.2 a riguardo dei segnali analogici, relativamente ai concetti di potenza, energia e periodicità, utilizzando qui delle sommatorie in luogo degli integrali.

Sequenze Un segnale viene chiamato numerico quando assume valori appartenenti ad un *insieme finito di simboli*; per questo motivo, la sua essenza è indicata anche come *sequenza simbolica*. Ad esempio, un testo scritto assume valori nell'ambito dei caratteri stampabili. Se si rappresenta ogni carattere con il suo numero ordinale, si ha allora una vera sequenza di numeri.

Segnali tempo-discreti Si può alternativamente rappresentare ogni carattere con un diverso valore di tensione, ottenendo un segnale analogico che è una rappresentazione a *più livelli* di tensione della sequenza originaria.

Frequenza di simbolo Il concetto di occupazione di banda, applicabile ai segnali analogici, è qui sostituito da quello di velocità di emissione, espressa in *simboli/secondo*, ed indicata come *frequenza di simbolo* o f_s . Una sequenza prodotta da una sorgente numerica si presta facilmente ad essere trasformata in un'altra, con un diverso alfabeto ed una differente frequenza di simbolo¹².

Frequenza binaria Qualora si desideri ottenere una *trasmissione binaria*, ossia rappresentabile come una sequenza di *zeri* ed *uni*, l'alfabeto di rappresentazione ha cardinalità pari a 2. In tal caso, ogni simbolo di una sequenza ad L livelli (es. 13) può essere posto in corrispondenza ad un gruppo di M elementi binari (o bit), con M pari al primo intero maggiore di $\log_2 L$ (es. 4). La grandezza M , pari al numero di bit/simbolo, moltiplicata per il numero di simboli a secondo f_s ,

¹²Per fissare le idee, consideriamo i simboli di una sequenza numerica $s[k]$ ad L valori: questi possono essere presi a gruppi di M , producendo simboli a velocità M volte inferiore, ma con L^M valori distinti. Se si dispone di un alfabeto di uscita ad H valori, i gruppi di M simboli L -ari originari possono essere rappresentati con gruppi di N simboli H -ari purché $L^M \leq H^N$. Es.: per codificare in binario ($H = 2$) simboli con $L = 26$ livelli, occorrono almeno $N = 5$ bit/simbolo, ottenendo così $2^5 = 32 > L = 26$.

permette di calcolare il flusso informativo in bit/secondo, che prende il nome di *frequenza binaria*: $f_b = M \cdot f_s$.

Campionamento Si è già accennato a come un segnale analogico possa essere rappresentato mediante i suoi valori campionati a frequenza di f_c *campioni/secondo* e quantizzati con un numero M di *bit/campione*, consentendo l'uso di un canale numerico. In tal caso, la sorgente numerica equivalente sarà caratterizzata da una velocità di trasmissione di f_b bit/secondo, pari al prodotto $f_c \cdot M$.

Modulazione numerica Qualora la risposta in frequenza del canale imponga un processo di modulazione, esistono tecniche *specifiche* per i segnali numerici, che traggono vantaggio dalla natura discreta del messaggio da trasmettere.

Trasmissione a pacchetto Un segnale numerico può avere origini delle più disparate, e non necessariamente essere il risultato di un processo di quantizzazione. Ad esempio, può trattarsi di un file da trasmettere tra due computer; in tal caso, si può suddividere la sequenza numerica in messaggi più piccoli (chiamati *pacchetti di dati*), numerarli consecutivamente, ed inviarli singolarmente attraverso la rete di interconnessione, anche impiegando percorsi differenti per ogni sotto-messaggio: sarà compito del lato ricevente ri-assemblare i singoli pacchetti nell'ordine originario. Il caso descritto è un tipo particolare di rete, detto *a commutazione di pacchetto*, di cui saranno esposti i principi di funzionamento e le metodologie di progetto di massima delle risorse, mirate all'ottenimento di *prestazioni* definite in termini di *ritardo medio* di trasmissione.

1.5 Teoria delle probabilità

Molti dei concetti utilizzati per trattare i processi aleatori, per definire la quantità di informazione di un messaggio, le prestazioni di un canale, il dimensionamento di reti di comunicazione, sono fondati sulla conoscenza della *teoria delle probabilità*, che verrà pertanto illustrata, almeno nei suoi concetti fondamentali, i quali saranno immediatamente applicati ai casi specifici che si verificano nei sistemi di telecomunicazione.

Teoria del traffico In particolare, è accennata l'applicazione della teoria delle probabilità al problema del dimensionamento di collegamenti che debbano trasportare più messaggi contemporaneamente, operando una moltiplicazione degli stessi su di un medesimo mezzo trasmissivo.

1.6 Sistemi di telecomunicazione

Introduciamo brevemente quattro diversi *punti di vista* in cui è possibile inquadrare le problematiche di comunicazione: gli aspetti *fisici*, di *elaborazione*, di *sistema*, di *rete*, e di *trasporto*.

Aspetti fisici Un canale di comunicazione, *dal punto di vista fisico*, si identifica con il mezzo trasmissivo, per la descrizione del quale si adotta frequentemente un modello circuitale. Elenchiamo i mezzi comunemente adottati:

Collegamenti radio: il segnale si propaga nello spazio libero come onda elettromagnetica sferica, e viene irradiato mediante antenne, che ne focalizzano la potenza lungo direzioni privilegiate. La trasmissione è resa possibile grazie al processo di modulazione;

Collegamenti in cavo: da quelli tra computer e stampante, a quelli su doppino (telefonia), a quelli in cavo coassiale (televisione, ethernet). Possono essere di tipo half o full duplex a seconda che i due estremi della comunicazione siano unidirezionali o bidirezionali;

Collegamenti in fibra ottica: sono realizzati facendo viaggiare energia luminosa attraverso una guida d'onda di materiale dielettrico. La tecnica è idonea alla trasmissione dei soli segnali numerici, dato che la sorgente luminosa in trasmissione viene accesa e spenta velocissimamente in corrispondenza dei bit (zero od uno) del messaggio.

Modello circuitale Il collegamento ed i trasduttori ad esso relativi, sono spesso realizzati ricorrendo ad un circuito elettrico equivalente, in modo da poterne descrivere il comportamento mediante strumenti analitici noti.

Aspetti sistemistici Da un punto di vista *di sistema*, il transito dei segnali attraverso sistemi fisici è analizzato in termini del peggioramento introdotto, che può essere catalogato nell'ambito di diverse categorie:

Distorsioni Si distinguono quelle cosiddette *lineari*, causate da una risposta in frequenza non ideale, dalle distorsioni *non lineari*, che causano invece una deformazione istantanea sulla forma d'onda in transito;

Non stazionarietà Questi fenomeni sono caratterizzati da una variazione nel tempo delle caratteristiche del canale, e ricorrono spesso nel caso di comunicazioni con mezzi mobili;

Attenuazione Un segnale in transito lungo un canale presenta in uscita una ampiezza inferiore a quella di ingresso. L'alterazione può aver luogo sia per cause fisiche intrinseche (lunghezza del collegamento, disadattamento di impedenze, tecnologia degli amplificatori), che in dipendenza di fatti contingenti (percorsi multipli, pioggia); in questo secondo caso, il fenomeno è trattato come l'esito di un processo aleatorio;

Portata Affinché possano essere soddisfatti i requisiti di qualità (ad esempio l'SNR) desiderati, risulta che la lunghezza del collegamento deve essere inferiore ad un massimo, in conseguenza dell'attenuazione del collegamento, della potenza trasmessa, e degli altri fatti contingenti;

Qualità del servizio Con questo termine sono indicate diverse grandezze, ognuna applicabile in un particolare contesto, e che rappresentano un indice di "bontà" del processo comunicativo. Tra queste grandezze possiamo citare il *rapporto segnale rumore* SNR e la *probabilità di errore* P_e , relative rispettivamente alle trasmissioni analogiche e numeriche; il *ritardo medio*, rilevante nel caso di trasmissioni a pacchetto; il *tempo di fuori servizio*, qualificante della affidabilità dei sottosistemi di comunicazione.

Rete Dal punto di vista della *rete di comunicazione*, la consegna del messaggio informativo alla destinazione deve tener conto di aspetti indicati come:

Commutazione La rete è costituita da un insieme di *nodi di commutazione*, interconnessi da collegamenti che vengono usati in modalità condivisa da molte comunicazioni contemporanee, e che sono *attraversati* dai messaggi in transito, che devono essere *smistati* verso la porta di uscita corretta;

Instradamento La determinazione del percorso dei messaggi nella rete, scelto tra i possibili percorsi alternativi che collegano la sorgente con la destinazione, prende il nome di *instradamento*;

Segnalazione Il coordinamento tra i nodi della rete avviene mediante lo scambio tra gli stessi di informazioni aggiuntive dette *di segnalazione*, che costituiscono un vero e proprio processo di comunicazione parallelo a quello prettamente informativo;

Protocollo Lo scambio dei messaggi tra le coppie di nodi della rete, od anche tra i nodi ed un organo di controllo centrale, avviene utilizzando particolari linguaggi, detti *protocolli* di segnalazione;

Mobilità Se i nodi della rete modificano la propria posizione nel tempo, o gli utenti desiderano usufruire degli stessi servizi indipendentemente dal punto di accesso alla rete, occorre individuare soluzioni specifiche, come ad esempio la registrazione ed il mantenimento dei propri dati presso una *unità di controllo*, e l'adozione di procedure di *autenticazione* che permettano di certificare l'identità degli utenti.

Elaborazione terminale Questa categoria comprende tutti gli aspetti legati alle trasformazioni operate sull'informazione ai due estremi del collegamento¹³. Tra questi è possibile riconoscere

Codifica di sorgente Le trasformazioni sul segnale da trasmettere, che permettono di impegnare la minor quantità di risorse trasmissive (ad esempio, banda), possono prevedere operazioni che tengono conto delle specifiche *caratteristiche del segnale* da trattare, come nel caso della codifica vocale, o della codifica video;

Codifica di canale In modo analogo, le trasformazioni necessarie a combattere gli errori nelle trasmissioni numeriche, possono tener conto delle caratteristiche statistiche *dei disturbi*;

Modulazione e formattazione Le operazioni necessarie alla trasmissione di un segnale radio, o di un segnale numerico, possono tener conto delle caratteristiche del *canale trasmissivo*, e adottare soluzioni che possono facilitare la realizzazione delle due funzioni precedenti.

¹³L'importanza e la specificità di tali trasformazioni assume un rilievo sempre maggiore con l'evoluzione (in termini di miniaturizzazione e potenza di calcolo) dei dispositivi di elaborazione, in special modo per ciò che riguarda le trasmissioni numeriche.

Trasporto Dal punto di vista del *trasporto* dell'informazione, sono rilevanti gli aspetti di:

Multiplazione Si tratta di raggruppare tra loro le singole comunicazioni in transito per un tratto in comune; con il risultato di migliorare sensibilmente l'efficienza della rete. Infatti, mediante la multiplazione si può garantire un elevato tasso di utilizzo delle risorse, che non giacciono mai inutilizzate proprio grazie al ri-uso continuo e multiplo delle stesse per parecchie comunicazioni in transito. Tecniche di multiplazione comunemente adottate sono la *multiplazione ...*

- di *tempo*, in cui lo stesso collegamento è utilizzato per più comunicazioni contemporanee in base ad uno schema di alternanza temporale;
- di *frequenza*, in cui le diverse comunicazioni occupano differenti regioni di frequenza, in uno stesso collegamento tra (ad esempio) due antenne;
- di *codice*, in cui diverse comunicazioni avvengono simultaneamente nella medesima banda di frequenza, adottando una particolare codifica che ne permette la separazione dal lato ricevente.

Controllo Riguarda la corretta consegna del messaggio al destinatario, e coinvolge la gestione degli errori di trasmissione di cui si è discusso a riguardo della codifica di canale, le problematiche di riassetto delle comunicazioni inoltrate in forma di pacchetti distinti, la gestione della segnalazione per ciò che riguarda l'adattamento dei protocolli di instradamento alle condizioni di carico della rete, ed il coordinamento delle sorgenti che desiderano trasmettere utilizzando il medesimo mezzo trasmissivo.

1.7 Segnali e sistemi

Vengono qui brevemente riassunte le definizioni ricorrenti nel caratterizzare gli elementi fondamentali nei cui termini sono descritti i sistemi di telecomunicazione.

In termini generali, un *sistema* è un gruppo di oggetti che interagiscono armoniosamente, e che sono combinati in modo da conseguire un obiettivo desiderato. Un sistema può essere parte (sottosistema) di un sistema più grande, e si può definire una intera gerarchia di sistemi, ognuno con il proprio dominio.

Un *segnale* è un evento che veicola un contenuto informativo. Nel nostro caso, possiamo interessarci alla *risposta* di un sistema ad un dato segnale. A volte, un sistema è descritto unicamente in termini della sua risposta a determinati segnali.

1.7.1 Caratteristiche dei sistemi

Idealizziamo ora un sistema come una trasformazione $\mathcal{T}[\cdot]$, tale che ad ogni segnale di ingresso $x(t)$ corrisponda una uscita $y(t)$: $\mathcal{T}[x(t)] = y(t)$. In base a tale formalismo, riportiamo alcune caratteristiche dei sistemi, che ne descrivono il comportamento in termini più generali.

Linearità Un sistema è *lineare* quando l'uscita associata ad una combinazione lineare di ingressi, è la combinazione lineare delle uscite previste per ogni singolo ingresso:

$$\mathcal{T}\left[\sum_i a_i x_i(t)\right] = \sum_i a_i \mathcal{T}[x_i(t)]$$

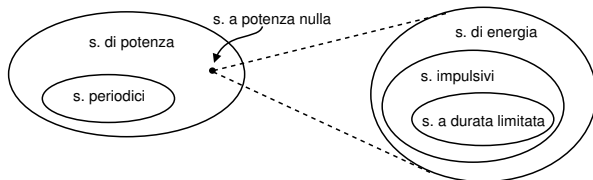


Figura 1.2: Visione insiemistica per le diverse classi di segnali

Al contrario, un legame ingresso-uscita *senza memoria*¹⁴ del tipo $y(t) = g(x(t))$, in cui $g(\cdot)$ è una generica funzione *non lineare*¹⁵... non è lineare!

Permanenza Un sistema è *permanente* (o stazionario) se l'uscita associata ad un ingresso traslato nel tempo, è la traslazione temporale dell'uscita che si avrebbe per lo stesso ingresso non traslato, ovvero: se $\mathcal{T}[x(t)] = y(t)$, allora $\mathcal{T}[x(t - \tau)] = y(t - \tau)$. Nel caso contrario, il sistema è detto tempo-variante, non stazionario, o non permanente.

Realizzabilità fisica E' detta anche *causalità*, poiché descrive l'impossibilità di osservare una uscita, prima di aver applicato un qualunque ingresso. Una definizione alternativa asserisce che i valori di uscita $y(t)$ ad un istante $t = t_0$, non possono dipendere da valori di ingresso $x(t)$ per istanti futuri $t > t_0$.

Stabilità è definita come la proprietà di fornire uscite limitate (in ampiezza) per ingressi limitati.

1.7.2 Caratteristiche dei segnali

Da un punto di vista analitico, un segnale è una funzione del tempo, del tipo descritto al §1.3, e per esso si possono operare le classificazioni:

Segnale di potenza Un segnale analogico può avere una estensione temporale limitata, oppure si può immaginare che si estenda da meno infinito a infinito. Nel secondo caso il segnale si dice di *potenza* se ne esiste (ed è diversa da zero) la media quadratica

$$0 < \mathcal{P}_s = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |s(t)|^2 dt < \infty$$

Un segnale di potenza è inoltre detto

Segnale periodico di periodo T , nel caso in cui si verifichi che

$$s(t) = s(t + T)$$

per qualsiasi valore di t , mentre si dice

¹⁴Un operatore si dice *senza memoria* quando ogni valore dell'uscita dipende da un unico valore di ingresso.

¹⁵Una funzione $y(x)$ è lineare quando il suo sviluppo in serie di potenze si arresta al primo ordine, ed è quindi esprimibile in forma $y = ax + b$ (equazione di una retta).

Segnale di energia un segnale di durata limitata o illimitata, se esiste il valore

$$0 < \mathcal{E}_s = \int_{-\infty}^{\infty} |s(t)|^2 dt < \infty$$

Perché ciò avvenga, occorre che $s(t)$ tenda a zero (per t che tende ad ∞) più velocemente (od in modo uguale) ad $\frac{1}{\sqrt{t}}$, e quindi $|s(t)|^2$ tenda a zero come (o più in fretta di) $\frac{1}{t}$.

In particolare, se un segnale ha *durata limitata*, ovvero è nullo per t al di fuori di un intervallo $[t_1, t_2]$ (vedi Fig. in basso), allora è anche di energia. Infine, viene detto

Segnale impulsivo un segnale di energia, che tende a zero come (o più velocemente di) $\frac{1}{t}$:

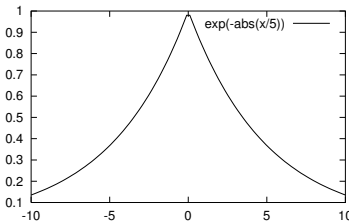
$$0 < \int_{-\infty}^{\infty} |s(t)| dt < \infty$$

E' il caso delle funzioni sommabili, per le quali $|s(t)|^2$ tende a zero come (o più di) $\frac{1}{t^2}$, e dunque di energia.

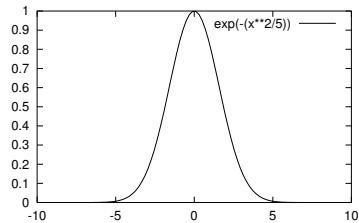
Riassumendo

- Un segnale *impulsivo* è di energia;
- Un segnale a *durata limitata* è impulsivo, e di energia;
- Un segnale *periodico* non è di energia, ma di potenza;

Esempi di segnali di energia

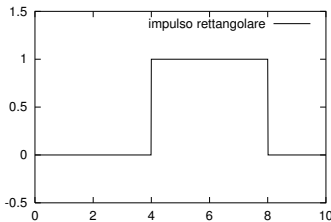


Impulso Esponenziale Bilatero

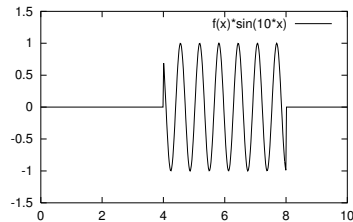


Impulso Gaussiano

Esempi di segnale a durata limitata



Impulso rettangolare tra 4 ed 8



Sinusoide troncata

Qualora il segnale sia associato a delle grandezze elettriche, allora i concetti di *Potenza* ed *Energia* hanno il correlato fisico illustrato di seguito.

1.7.3 Aspetti fisici delle grandezze energetiche

Potenza istantanea Se consideriamo una resistenza R , ed applichiamo ai suoi capi una tensione $v(t)$, in essa scorre una corrente $i(t) = \frac{v(t)}{R}$, e la *potenza ceduta* alla resistenza ad ogni istante t è pari a

$$p(t) = v(t) i(t)$$

che si misura in *Watt* (equivalente a Joule/secondo), e che rappresenta la *potenza istantanea* assorbita. Ricordando che $i(t) = \frac{v(t)}{R}$, si ottiene anche $p(t) = \frac{v^2(t)}{R} = i^2(t) R$.

Energia Se integriamo $p(t)$ su di un intervallo temporale T , si ottiene *l'energia complessiva* assorbita da R nell'intervallo T :

$$e_T(t) = \int_{t-T}^t p(\tau) d\tau \quad [\text{joule}]$$

Nello stesso intervallo T , la resistenza *assorbe* una potenza $p_T(t) = \overline{e_T(t)} = \frac{1}{T} e_T(t)$ [Watt], che costituisce una *media a breve termine* dell'energia assorbita nell'intervallo¹⁶.

Se un segnale $x(t)$ è *periodico* con periodo T (o $\frac{T}{n}$ con n intero), i valori di $\overline{e_T(t)} = p_T(t)$ coincidono con quelli calcolabili con T comunque grande. Se $R = 1 \Omega$, tali valori coincidono inoltre con le definizioni di potenza ed energia del segnale:

$$\text{Energia: } \mathcal{E}_x = \int_{-\frac{T}{2}}^{\frac{T}{2}} |x|^2(t) dt = e_T\left(\frac{T}{2}\right) \quad [\text{Volt}^2 \cdot \text{sec}] \text{ o } [\text{Ampere}^2 \cdot \text{sec}]$$

$$\text{Potenza: } \mathcal{P}_x = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |x|^2(t) dt = \frac{1}{T} e_T\left(\frac{T}{2}\right) = p_T\left(\frac{T}{2}\right) \quad [\text{Volt}^2] \text{ o } [\text{Ampere}^2]$$

Potenza dissipata Se la resistenza è diversa da 1Ω , le due quantità non coincidono più. Nelle misure fisiche in genere si ottiene la *potenza dissipata* sullo strumento di misura (o irradiata dall'antenna, o dagli altoparlanti) espressa in Watt. Per risalire alla *potenza/energia di segnale* delle grandezze elettriche presenti ai suoi capi (tensione o corrente) occorre dividere (o moltiplicare) la potenza in Watt per R . Ad esempio, una potenza assorbita \mathcal{P} di 10 Watt su 8Ω equivale ad una potenza di segnale $\mathcal{P} \cdot R = 80 (\text{Volt})^2$, ovvero di $\frac{\mathcal{P}}{R} = 1.25 (\text{Ampere})^2$.

Valore efficace Si indica allora come *valore efficace* quel livello di segnale continuo che produrrebbe lo stesso effetto energetico. Nell'esempio precedente, otteniamo: $V_{eff} = \sqrt{80} = 8.94 \text{ Volt}$; $I_{eff} = 1.118 \text{ Ampere}$. Infatti:

$\mathcal{P}_T(\text{segnale}) = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} V_{eff}^2 dt = \frac{T}{T} (8.94)^2 = 80 \text{ Volt}^2$ che su 8Ω dissipa $\frac{V^2}{R} = 10$ Watt.

¹⁶Anticipando una notazione che verrà usata nel corso del testo, il pedice T indica l'estensione temporale a cui è riferita la grandezza che presenta il pedice, mentre la soprilineatura di una grandezza che dipende dal tempo, indica una media temporale della grandezza stessa.

Capitolo 2

Serie di Fourier

Sono qui impostate alcune relazioni trigonometriche fondamentali, per poi definire la rappresentazione di segnali periodici mediante lo sviluppo in serie di Fourier. Sono quindi introdotte le proprietà di simmetria e di approssimazione, e sviluppati i concetti di ortogonalità. E' infine enunciato il teorema di Parseval e definito lo spettro di potenza. Il capitolo termina con richiami di algebra vettoriale.

2.1 Prerequisiti trigonometrici

2.1.1 Numeri complessi

Un numero complesso \underline{x} è costituito da una coppia di valori numerici a e b che ne rappresentano la parte reale e quella immaginaria:

$$\underline{x} = a + jb$$

E' spesso utile ricorrere ad una rappresentazione di \underline{x} nel piano complesso, che mette in luce l'espressione alternativa¹ di \underline{x} nei termini di modulo $|x|$ e fase φ :

$$\underline{x} = |x| e^{j\varphi}$$

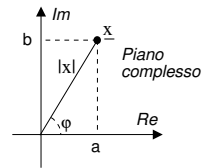
Queste due quantità si ottengono dalle parti reale ed immaginaria, mediante le relazioni

$$|x| = \sqrt{a^2 + b^2} \quad \text{e} \quad \varphi = \arctan \frac{b}{a}$$

mentre le relazioni inverse risultano

$$a = |x| \cos \varphi \quad \text{e} \quad b = |x| \sin \varphi$$

Per ogni numero complesso \underline{x} , è definito il suo coniugato \underline{x}^* come quel numero complesso con uguale parte reale, e parte immaginaria di segno opposto, ovvero uguale modulo, e fase cambiata di segno: $\underline{x}^* = a - jb = |x| e^{-j\varphi}$.



¹La rappresentazione in modulo e fase consente di calcolare il prodotto tra numeri complessi (es $\underline{x} = |x| e^{j\varphi}$ e $\underline{y} = |y| e^{j\theta}$) in modo semplice: $\underline{y} = \underline{x} \cdot \underline{y} = |x| |y| e^{j(\varphi+\theta)}$.

2.1.2 Formule di Eulero

L'esponenziale $e^{j\varphi}$ è un particolare numero complesso con modulo pari ad uno², e che quindi si scompone in parte reale ed immaginaria come

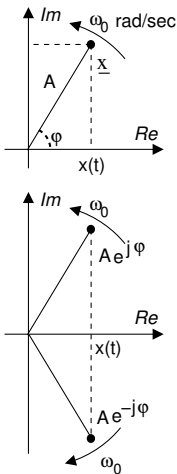
$$e^{\pm j\varphi} = \cos \varphi \pm j \sin \varphi$$

Da questa relazione sono derivabili le *formule di Eulero*, che esprimono le funzioni trigonometriche in termini di esponenziali complessi come

$$\cos \varphi = \frac{e^{j\varphi} + e^{-j\varphi}}{2} \quad \text{e} \quad \sin \varphi = \frac{e^{j\varphi} - e^{-j\varphi}}{2j}$$

e che possono tornare utili nel semplificare i calcoli, trasformando i prodotti tra funzioni trigonometriche in somme di angoli³.

2.1.3 Fasori



Un segnale del tipo $x(t) = A \cos(2\pi f_0 t + \varphi)$ è completamente rappresentato dal numero complesso $\underline{x} = A e^{j\varphi}$ detto *fasore*, la cui conoscenza permette di riottenere il segnale originario mediante la relazione $x(t) = \Re \{ \underline{x} \cdot e^{j2\pi f_0 t} \}$, che una volta sviluppata⁴ risulta infatti pari a

$$\begin{aligned} x(t) &= \Re \left\{ A \cdot e^{j(2\pi f_0 t + \varphi)} \right\} \\ &= A \cdot \Re \left\{ \cos(2\pi f_0 t + \varphi) + j \sin(2\pi f_0 t + \varphi) \right\} \\ &= A \cos(2\pi f_0 t + \varphi) \end{aligned}$$

Osserviamo che il risultato ottenuto può interpretarsi graficamente come l'aver impresso al fasore una rotazione di velocità angolare $\omega_0 = 2\pi f_0$ radianti/secondo in senso antiorario, ed aver proiettato il risultato sull'asse reale. In alternativa, possiamo esprimere il segnale di partenza anche come

$$x(t) = \frac{1}{2} \left\{ \underline{x} e^{j2\pi f_0 t} + \underline{x}^* e^{-j2\pi f_0 t} \right\} \quad (2.1)$$

Tale operazione coinvolge anche le *frequenze negative*, e corrisponde a tener conto anche di un secondo vettore rotante, che si muove ora in senso orario, che ha una parte immaginaria di segno sempre opposto al primo, e che è moltiplicato per il coniugato del fasore. Vedremo tra breve che l'ultima espressione fornita è esattamente quella della *serie di Fourier* per il caso in questione.

²L'espressione più generale e^γ con $\gamma = \alpha + j\varphi$ è ancora un numero complesso, di modulo e^α . Infatti $e^\gamma = e^{\alpha + j\varphi} = e^\alpha e^{j\varphi} = e^\alpha (\cos \varphi + j \sin \varphi)$.

³L'affermazione nasce dalla relazione $e^\alpha e^\beta = e^{\alpha + \beta}$. Ad esempio quindi, il prodotto $\cos \alpha \cdot \sin \beta$ diviene

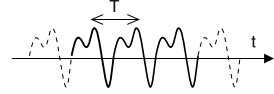
$$\begin{aligned} &= \frac{1}{4j} (e^{j\alpha} + e^{-j\alpha}) (e^{j\beta} - e^{-j\beta}) = \frac{1}{4j} [e^{j\alpha} e^{j\beta} - e^{j\alpha} e^{-j\beta} + e^{-j\alpha} e^{j\beta} - e^{-j\alpha} e^{-j\beta}] \\ &= \frac{1}{4j} [e^{j(\alpha + \beta)} - e^{-j(\alpha + \beta)} - e^{j(\alpha - \beta)} + e^{-j(\alpha - \beta)}] = \frac{1}{4j} [2j \sin(\alpha + \beta) - 2j \sin(\alpha - \beta)] \\ &= \frac{1}{2} [\sin(\alpha + \beta) - \sin(\alpha - \beta)] \end{aligned}$$

⁴Un modo alternativo di ottenere lo stesso risultato è quello di esprimere gli esponenziali complessi in termini trigonometrici, ottenendo $x(t) = \Re \{ |\underline{x}| (\cos \varphi + j \sin \varphi) [\cos(2\pi f_0 t) + j \sin(2\pi f_0 t)] \}$, e sviluppare il calcolo facendo uso delle relazioni $\cos \alpha \cos \beta = \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)]$ e $\sin \alpha \sin \beta = \frac{1}{2} [\cos(\alpha - \beta) - \cos(\alpha + \beta)]$, ma avremmo svolto più passaggi.

2.2 Serie di Fourier

Come anticipato a pag. 12, un segnale *periodico* $x(t)$ è un segnale di potenza, che assume ripetutamente gli stessi valori a distanza multipla di un intervallo temporale T denominato *periodo*, ovvero tale che

$$x(t) = x(t + T) \quad \forall t$$



L'inverso di T è detto *frequenza fondamentale* $F = \frac{1}{T}$ o *prima armonica* di $x(t)$, espressa in Hertz, dimensionalmente pari all'inverso di un tempo [sec^{-1}].

Per i segnali periodici esiste una forma di rappresentazione basata sulla conoscenza di una serie infinita di coefficienti complessi $\{X_n\}$ denominati *coefficienti di Fourier*, calcolabili a partire da un periodo del segnale come

$$X_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) e^{-j2\pi n F t} dt \quad (2.2)$$

e che permettono⁵ la *ricostruzione* di $x(t)$, sotto forma di una combinazione lineare di infinite funzioni esponenziali complesse $e^{j2\pi n F t}$, mediante l'espressione nota come *serie di Fourier*:

$$x(t) = \sum_{n=-\infty}^{\infty} X_n e^{j2\pi n F t} \quad (2.3)$$

Osserviamo che:

- La conoscenza di $\{X_n\}$ *equivale* a quella di $x(t)$ e *viceversa*, esistendo il modo di passare dall'una all'altra rappresentazione;
- Le funzioni della base di rappresentazione $e^{j2\pi n F t}$ sono funzioni trigonometriche a frequenza multipla (*n-esima*) della fondamentale, detta anche *n-esima armonica*⁶
- I termini $X_n e^{j2\pi n F t}$ sono chiamati *componenti armoniche* di $x(t)$ a frequenza $f = nF$;
- Il coefficiente $X_0 = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) dt$ rappresenta la componente continua (o *valor medio*) di $x(t)$;
- La serie di Fourier dà valori esatti in tutti i punti in cui $x(t)$ è continuo, mentre in corrispondenza di discontinuità di prima specie fornisce un valore pari alla media dei valori agli estremi, cosicché il valore dell'energia di un periodo è preservato;
- I coefficienti di Fourier X_n possono essere calcolati anche per un segnale di estensione finita T . Antitrasformando, il segnale diventa periodico!

⁵Condizione sufficiente ad assicurare la convergenza della serie (2.2) è che il segnale soddisfi alle condizioni di *Dirichlet* http://en.wikipedia.org/wiki/Dirichlet_conditions

⁶Questa terminologia richiama alla mente nozioni di teoria musicale, in cui gli armonici di una nota sono appunto note di frequenza multipla della prima. In particolare la seconda armonica corrisponde ad un intervallo di ottava, e la 4^a a due ottave. E la terza armonica? Prendendo ad esempio un *La* a 440 Hz, la terza armonica si trova a $440 * 3 = 1320$ Hz. Sapendo che ogni semitono della scala temperata corrisponde ad un rapporto di frequenza pari a $2^{\frac{1}{12}}$ rispetto al semitono precedente, proviamo a determinare il numero di semitoni N_s tra la terza armonica ed il *La* (fondamentale). Risulta allora $2^{\frac{N_s}{12}} = \frac{1320}{440} = 3 \rightarrow \frac{N_s}{12} = \log_2 3 \simeq 1.5849 \rightarrow N_s = 19$ semitoni, ovvero un intervallo di *tredicesima*, ovvero il *Mi* che viene dopo il *La* dell'ottava successiva, che si trova ad 880 Hz.

- Se poniamo $nF = f$ (con f variabile continua), possiamo interpretare le componenti armoniche come i valori campionati di una funzione (complessa) della frequenza: $X_n = \overline{X}(nF)$. Ad $\overline{X}(f)$ si dà il nome di *inviluppo dello spettro di ampiezza* di $x(t)$, che si ottiene estendendo la definizione dei coefficienti di Fourier: $\overline{X}(f) = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) e^{-j2\pi ft} dt$;
- I coefficienti X_n sono valori complessi. Al loro posto si possono usare, in alternativa:

$$\begin{cases} M_n &= |X_n| && \text{spettro di modulo} \\ \varphi_n &= \arctan \frac{\Im\{X_n\}}{\Re\{X_n\}} && \text{spettro di fase} \end{cases}$$

essendo

$$X_n = |X_n| e^{j\varphi_n} = \Re\{X_n\} + j\Im\{X_n\}$$

2.2.1 Segnali reali

2.2.1.1 Simmetria coniugata o Hermitiana

I coefficienti della serie di Fourier possono essere calcolati anche per segnali complessi; nel *caso particolare di $x(t)$ reale* i coefficienti di Fourier risultano godere della proprietà di *simmetria coniugata*, ovvero per essi vale l'eguaglianza

$$X_{-n} = X_n^*$$

che esprime come i coefficienti con indice n negativo abbiano una parte reale uguale a quella dei coefficienti con (uguale) indice positivo, e parte immaginaria cambiata di segno⁷. Ciò comporta una proprietà analoga per il modulo e la fase di $\{X_n\}$, e dunque possiamo scrivere:

$$x(t) \text{ Reale} \Leftrightarrow \begin{cases} \Re\{X_{-n}\} = \Re\{X_n\} \\ \Im\{X_{-n}\} = -\Im\{X_n\} \end{cases} ; \begin{cases} |X_{-n}| = |X_n| \\ \arg\{X_{-n}\} = -\arg\{X_n\} \end{cases}$$

Tali relazioni evidenziano che

Se $x(t)$ è reale, i coefficienti X_n risultano avere modulo pari e fase dispari, ovvero parte reale pari e parte immaginaria dispari.

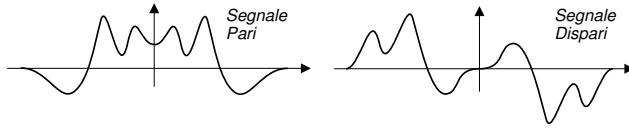
Un corollario di questo risultato è che⁸

Se $x(t)$ è reale pari, i coefficienti X_n sono reali (pari), mentre se $x(t)$ è reale dispari, gli X_n sono immaginari (dispari).

⁷La dimostrazione di questa proprietà si basa sull'osservazione che scomponendo l'esponenziale complesso che compare nella formula (2.2) come $e^{-j2\pi nFt} = \cos 2\pi nFt - j \sin 2\pi nFt$, ed essendo $x(t)$ reale, l'integrale che calcola gli X_n si suddivide in due termini, relativi al calcolo della parte reale e quella immaginaria, ovvero: $X_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \cos 2\pi nFt dt - \frac{j}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \sin 2\pi nFt dt$.

Essendo il coseno una funzione pari, il primo integrale fornisce gli stessi risultati per n cambiato di segno; il secondo integrale invece cambia segno con n , essendo il seno una funzione dispari.

⁸Con riferimento alla scomposizione del calcolo di X_n alla nota precedente, notiamo che se $x(t)$ è (reale) pari, allora $\Im\{X_n\} = 0$, in quanto $x(t) \sin 2\pi nFt dt$ è dispari, ed il suo integrale esteso ad un intervallo simmetrico rispetto all'origine è nullo. Se invece $x(t)$ è (reale) dispari, si ottiene $\Re\{X_n\} = 0$, per lo stesso motivo applicato al termine $x(t) \cos 2\pi nFt dt$.

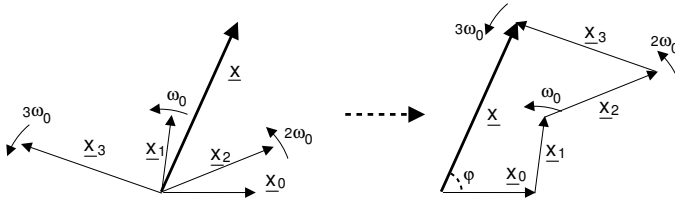


2.2.1.2 Interpretazione degli X_n come fasori

Confrontando la formula di ricostruzione

$$x(t) = \sum_{n=-\infty}^{\infty} X_n e^{j2\pi n F t}$$

con la (2.1) ricavata al § 2.1.3 per il caso di un coseno, e tenendo conto della proprietà di simmetria coniugata $X_{-n} = X_n^*$, si nota come un segnale reale possa essere pensato composto a partire da un insieme infinito di fasori X_n (di valore *doppio* di quello dei coefficienti X_n), ognuno rotante con una velocità angolare $\omega_n = 2\pi n F$ multipla della frequenza fondamentale.



Esercizio: calcoliamo i coefficienti dello sviluppo in serie di Fourier per il segnale $x(t) = A \cos(2\pi F t + \varphi)$. Esprimiamo innanzitutto l'integrale che fornisce i coefficienti, nei termini della formula di Eulero per il coseno:

$$\begin{aligned} X_n &= \frac{A}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \frac{e^{j2\pi F t} e^{j\varphi} + e^{-j2\pi F t} e^{-j\varphi}}{2} e^{-j2\pi n F t} dt \\ &= \frac{A}{2T} \left(e^{j\varphi} \int_{-\frac{T}{2}}^{\frac{T}{2}} e^{j2\pi F t} e^{-j2\pi n F t} dt + e^{-j\varphi} \int_{-\frac{T}{2}}^{\frac{T}{2}} e^{-j2\pi F t} e^{-j2\pi n F t} dt \right) \end{aligned}$$

in cui $F = \frac{1}{T}$, e consideriamo la funzione integranda $e^{\pm j2\pi F t} e^{-j2\pi n F t}$ per i diversi valori di n:

- per $n = 0$, osserviamo che $e^{-j2\pi n F t} \Big|_{n=0} = e^0 = 1$, e dunque $X_0 = 0$ in quanto

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} e^{j2\pi F t} dt = 0$$

poiché in un intervallo T entra esattamente un ciclo di (co)sinusoide a frequenza F , risultando in un valor medio nullo.

- per $n = 1$, si ha che $\int_{-\frac{T}{2}}^{\frac{T}{2}} e^{j2\pi F t} e^{-j2\pi F t} dt = \int_{-\frac{T}{2}}^{\frac{T}{2}} e^0 dt = T$, mentre $\int_{-\frac{T}{2}}^{\frac{T}{2}} e^{-j2\pi F t} e^{-j2\pi F t} dt = \int_{-\frac{T}{2}}^{\frac{T}{2}} e^{-j2\pi 2 F t} dt = 0$, dato che in un periodo T entrano due cicli esatti della funzione periodica integranda, ottenendo quindi

$$X_1 = \frac{A}{2} e^{j\varphi}$$

- per $n = -1$ valgono considerazioni analoghe, ottenendo in definitiva

$$X_{-1} = \frac{A}{2} e^{-j\varphi}$$

2.2.1.3 Serie trigonometrica

Nel caso in cui gli X_n abbiano simmetria coniugata, la formula di ricostruzione può scriversi

$$x(t) = X_0 + \sum_{n=1}^{\infty} \left\{ X_n e^{j2\pi n F t} + X_{-n} e^{-j2\pi n F t} \right\} = M_0 + \sum_{n=1}^{\infty} M_n 2 \cos(2\pi n F t + \varphi_n)$$

ovvero in forma di serie di coseni; si noti che X_0 è necessariamente reale, in quanto la fase deve risultare una funzione dispari della frequenza.

In modo simile, le proprietà relative alle parti reale ed immaginaria permettono di scrivere:

$$\begin{aligned} x(t) &= X_0 + \sum_{n=1}^{\infty} \left\{ (R_n + jI_n) e^{j2\pi n F t} + (R_n - jI_n) e^{-j2\pi n F t} \right\} \\ &= R_0 + \sum_{n=1}^{\infty} \{ 2R_n \cos(2\pi n F t) - 2I_n \sin(2\pi n F t) \} \end{aligned}$$

in cui

$$R_0 = M_0 = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) dt \quad \text{e} \quad \begin{cases} R_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \cos(2\pi n F t) dt \\ I_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \sin(2\pi n F t) dt \end{cases}$$

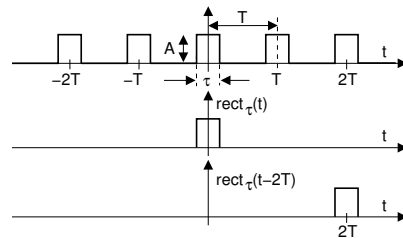
Pertanto, nel caso in cui $x(t)$ sia un segnale reale, la serie di Fourier si riduce ad uno sviluppo in termini di funzioni trigonometriche, ed in particolare ad una serie di soli coseni nel caso in cui $x(t)$ sia pari, oppure una serie di soli seni, nel caso in cui sia dispari.

2.2.1.4 Serie di Fourier di un'onda rettangolare

La figura a lato mostra un segnale *ad onda quadra* con un *duty cycle*⁹ del 33%, la cui espressione analitica può essere scritta come

$$x(t) = \sum_{n=-\infty}^{\infty} A \text{rect}_{\tau}(t - nT)$$

e per la quale si è adottata la notazione $\text{rect}_{\tau}(t)$ per rappresentare un impulso rettangolare di base τ ed altezza unitaria, centrato nell'origine dei tempi. L'argomento $(t - nT)$ indica una *traslazione* (o spostamento) del rettangolo *a destra* (ossia verso gli istanti *positivi*) di una quantità pari a nT , cosicché la sommatoria rappresenta appunto la replica dello stesso impulso rettangolare infinite volte in avanti ed all'indietro.



⁹Il DUTY CYCLE si traduce come *ciclo di impegno*, ed è definito come il rapporto percentuale per il quale il segnale è diverso da zero, ossia $\text{duty cycle} = \frac{\tau}{T} * 100 \%$.

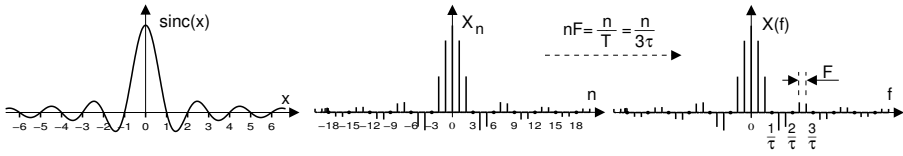


Figura 2.1: Funzione $\text{sinc}(x)$ e coefficienti di Fourier dell'onda quadra

Esercizio Il calcolo dei coefficienti di Fourier per il segnale in questione non presenta particolari difficoltà, ma l'esito si presta ad alcune utili considerazioni. Applicando infatti un risultato noto¹⁰, si ottiene

$$\begin{aligned} X_n &= \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) e^{-j2\pi n F t} dt = \frac{1}{T} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} A e^{-j2\pi n F t} dt = \\ &= \frac{A}{T} \left. \frac{e^{-j2\pi n F t}}{-j2\pi n F} \right|_{-\frac{\tau}{2}}^{\frac{\tau}{2}} = \frac{A}{\pi n F T} \frac{e^{j2\pi n F \frac{\tau}{2}} - e^{-j2\pi n F \frac{\tau}{2}}}{2j} = \end{aligned} \quad (2.4)$$

$$= A \frac{\tau}{T} \frac{\sin(\pi n F \tau)}{\pi n F \tau} = A \frac{\tau}{T} \text{sinc}(n F \tau) \quad (2.5)$$

Nella seconda uguaglianza, gli estremi di integrazione sono stati ristretti all'intervallo di effettiva esistenza del segnale, mentre la penultima uguaglianza si giustifica ricordando le formule di Eulero.

Definizione della funzione sinc Il risultato (2.5) ottenuto mostra come i coefficienti X_n della serie di Fourier per l'onda rettangolare risultano dipendere dai valori di $\frac{\sin(\pi n F \tau)}{\pi n F \tau}$ calcolati per n intero; tale espressione viene però rappresentata nei termini della funzione

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

che ricorrerà spesso nel testo, che è raffigurata nella parte di sinistra della Fig. 2.1, e che come si può notare passa da zero per valori interi dell'argomento x .

Nella parte centrale di Fig. 2.1 sono mostrati i valori $X_n = \text{sinc}(n F \tau)$, in cui si è posto $\tau = \frac{T}{3}$ (corrispondente al duty cycle del 33%), dando luogo a termini X_n nulli in corrispondenza degli indici $n = 3, 6, 9, \dots$. La parte destra di Fig. 2.1, infine, mostra ancora i coefficienti X_n , ma lungo una scala in Hertz, ottenuta considerando che $n F = \frac{n}{T}$ rappresenta la frequenza dell' n -esima armonica, e che la posizione $\tau = \frac{T}{3}$ adottata fornisce $n F = \frac{n}{T} = \frac{n}{3\tau}$.

Osserviamo ora che, mentre la spaziatura tra le armoniche è pari ad $F = \frac{1}{T}$ e dipende esclusivamente dal *periodo* della forma d'onda, gli zeri della funzione $\text{sinc}(n F \tau)$ occorrono a frequenze multiple di $\frac{1}{\tau}$. Per meglio comprendere le implicazioni di tali osservazioni, valutiamo come si modificano i valori X_n al variare di τ e di T .

Relazione tra i coefficienti della serie ed i parametri dell'onda quadra La parte in alto di Fig. 2.2 mostra quattro possibili modi di variare l'onda quadra di partenza: la colonna di sinistra rappresenta il caso in cui il periodo T si mantenga

¹⁰Sappiamo infatti che $\frac{\partial e^{f(x)}}{\partial x} = \frac{\partial f(x)}{\partial x} \cdot e^{f(x)}$, e quindi $\int_a^b \frac{\partial f(x)}{\partial x} \cdot e^{f(x)} dx = e^{f(x)} \Big|_a^b$

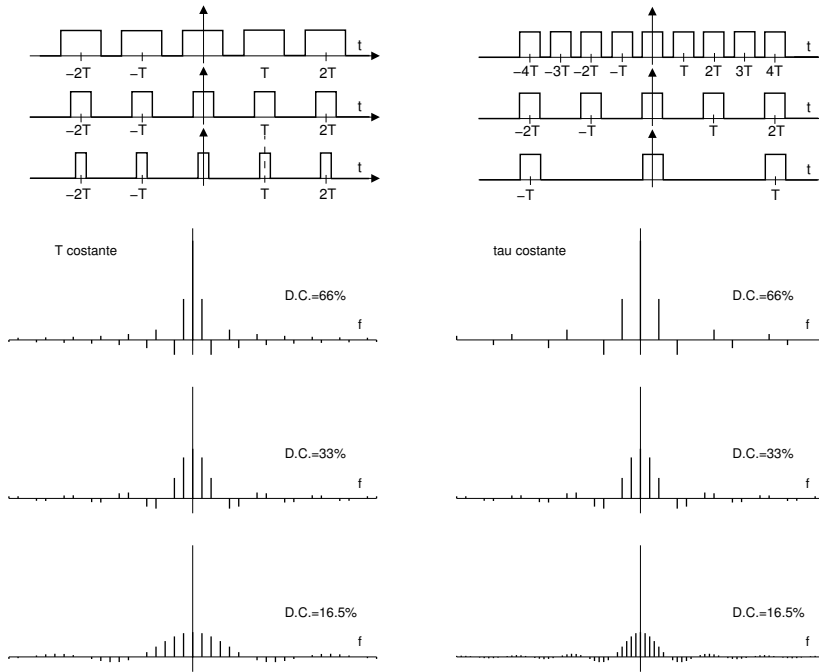


Figura 2.2: Modifiche allo spettro di ampiezza per variazioni della forma d'onda

costante, mentre la durata τ di un singolo ciclo *raddoppia* (prima riga) o si *dimezza* (terza riga), mentre la colonna di destra considera il caso in cui τ si mantenga invariato, ed il periodo T varii in modo da ottenere lo stesso duty cycle $\frac{\tau}{T}$ di sinistra, ovvero pari al 66% (prima riga) o 12,5% (terza riga).

La parte inferiore di Fig. 2.2 mostra le corrispettive variazioni per i valori dei coefficienti dello sviluppo in serie, calcolate facendo uso della (2.5). Al lato sinistro (per T costante) osserviamo che le armoniche mantengono la stessa spaziatura $\frac{1}{T}$, ma l'inviluppo *sinc* ($nF\tau$) si *contrae* ed *espande* rispettivamente. Il lato destro della figura mostra invece come questa volta rimane costante la velocità con cui gli X_n vanno a zero, mentre le armoniche si *diradano* (sopra) ed *infittiscono* (sotto) all'aumentare ed al diminuire di T rispettivamente. Infine, notiamo come al diminuire del duty cycle si assista in entrambi i casi ad una riduzione dell'ampiezza degli X_n , legata alla riduzione di potenza del segnale (vedi sezione 2.3).

2.2.2 Serie di Fourier troncata

Consideriamo un'onda quadra con duty-cycle del 50%

$$x(t) = \sum_{k=-\infty}^{\infty} \text{rect}_{\frac{T}{2}}(t - kT)$$

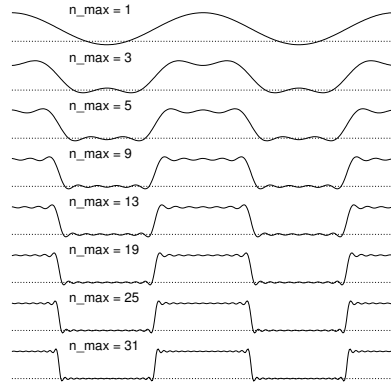
rappresentata mediante una serie troncata di Fourier in cui si considerano solo i coefficienti X_n con indice $-N \leq n \leq N$. Sappiamo che $X_n = \frac{\tau}{T} \frac{\sin(\pi n F \tau)}{\pi n F \tau}$ e, per $\tau = \frac{T}{2}$, si ottiene $X_n = \frac{1}{2} \frac{\sin(\frac{n\pi}{2})}{\frac{n\pi}{2}} = \frac{1}{2} \text{sinc}\left(\frac{n}{2}\right)$, che risulta diverso da zero solo con n dispari, e dunque:

$$X_0 = \frac{1}{2}; \quad X_n = \begin{cases} \frac{(-1)^{\frac{n-1}{2}}}{\pi n} & \text{con } n \text{ dispari} \\ 0 & \text{con } n \text{ pari} \end{cases}$$

Essendo inoltre $x(t)$ reale pari, sappiamo che può essere espresso come serie di coseni:

$$x(t) = X_0 + \sum_{n=1}^{\infty} 2X_n \cos(2\pi n F t)$$

Nella figura a fianco riportiamo il risultato ottenuto arrestando lo sviluppo in serie all'indice mostrato per ogni curva, e generando quindi il segnale



$$\hat{x}_N(t) = X_0 + \sum_{N=1}^{n_{Max}} 2X_n \cos(2\pi n F t)$$

Come osservabile, la ricostruzione è sempre più accurata, tranne che per le oscillazioni in prossimità della discontinuità, che prendono il nome di *fenomeno di Gibbs*.

Il caso mostrato è emblematico della inaccuratezza che si commette considerando contributi frequenziali ridotti rispetto a quelli propri della forma d'onda¹¹, a causa (ad esempio) di un *filtraggio* del segnale.

2.3 Teorema di Parseval

Stabilisce l'equivalenza di due rappresentazioni del segnale dal punto di vista energetico. La potenza, infatti, è calcolabile in modo simile in entrambi i domini del tempo e della frequenza, risultando

$$\mathcal{P}_x = \lim_{\Delta T \rightarrow \infty} \frac{1}{\Delta T} \int_{-\frac{\Delta T}{2}}^{\frac{\Delta T}{2}} |x(t)|^2 dt = \sum_{n=-\infty}^{\infty} |X_n|^2$$

¹¹Un risultato teorico, che qui citiamo solamente, mostra che l'errore quadratico di ricostruzione $\varepsilon = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} (x(t) - \hat{x}(t))^2 dt$ che è presente utilizzando solo le prime N armoniche è il minimo rispetto a quello ottenibile utilizzando un qualunque altro gruppo di N armoniche che non siano le prime.

Sviluppiamo i calcoli che danno luogo al risultato mostrato:

$$\begin{aligned}
 \mathcal{P}_x &= \lim_{\Delta T \rightarrow \infty} \frac{1}{\Delta T} \int_{-\frac{\Delta T}{2}}^{\frac{\Delta T}{2}} |x(t)|^2 dt = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |x(t)|^2 dt = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) x^*(t) dt = \\
 &= \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \left[\sum_n X_n e^{j2\pi n F t} \right] \left[\sum_m X_m^* e^{-j2\pi m F t} \right] dt = \\
 &= \sum_n \sum_m X_n X_m^* \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} e^{j2\pi(n-m)Ft} dt = \\
 &= \sum_{n=-\infty}^{\infty} X_n X_n^* = \sum_{n=-\infty}^{\infty} |X_n|^2 = \sum_{n=-\infty}^{\infty} M_n^2 = \sum_{n=-\infty}^{\infty} (R_n^2 + I_n^2)
 \end{aligned}$$

Ortogonalità degli esponenziali complessi Nei precedenti calcoli si è fatto uso del risultato

$$\frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} e^{j2\pi(n-m)Ft} dt = \begin{cases} 0 & \text{con } n \neq m \\ 1 & \text{con } n = m \end{cases} \quad (2.6)$$

che deriva dalla circostanza che la funzione integranda (per $n \neq m$) è esprimibile nei termini di seni e coseni con periodo uguale o sotto-multiplo di T , e quindi a valor medio nullo; per $n = m$ invece essa vale $e^0 = 1$, e dunque il risultato. Questo prende il nome di *proprietà di ortogonalità* degli esponenziali complessi, in base ai principi di algebra vettoriale forniti in appendice 2.4.1.

Spettro di potenza per segnali periodici In appendice (pag. 26) si mostra come l'integrale

$$\|x(t)\|^2 = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) x(t)^* dt$$

oltre a misurare la potenza del segnale periodico $x(t)$, ne misuri la norma quadratica da un punto di vista algebrico.

Tornando ad esaminare il risultato $\mathcal{P}_x = \sum_{n=-\infty}^{\infty} |X_n|^2$ espresso dal teorema di Parseval, notiamo che $|X_n|^2$ è la potenza di una singola componente armonica di $x(t)$:

$$\mathcal{P}_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} [X_n e^{j2\pi n F t}] [X_n^* e^{-j2\pi n F t}] dt = \frac{|X_n|^2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} dt = |X_n|^2$$

e quindi osserviamo che

La potenza totale \mathcal{P}_x di un segnale periodico $x(t)$ è pari alla somma delle potenze delle sue componenti armoniche $X_n e^{j2\pi n F t}$.

Si presti attenzione che il risultato è una diretta conseguenza dell'ortogonalità della base di rappresentazione: infatti, la potenza di una somma *non* è in generale pari alla somma delle potenze¹²; l'uguaglianza ha luogo solo nel caso di in cui gli addendi siano ortogonali.

¹²In generale risulta, con la notazione di prodotto scalare (\bar{a}, \bar{b}) tra vettori-segnali \bar{a} e \bar{b} introdotta al § 2.4.1: $(\bar{x} + \bar{y}, \bar{x} + \bar{y}) = (\bar{x}, \bar{x}) + (\bar{y}, \bar{y}) + (\bar{x}, \bar{y}) + (\bar{y}, \bar{x})$.

La successione $\{\mathcal{P}_n\} = \{\dots, |X_{-k}|^2, \dots, |X_0|^2, \dots, |X_k|^2, \dots\}$ rappresenta come la potenza totale si ripartisce tra le diverse armoniche a frequenza $f = nF$, e prende il nome di *spettro di potenza* del segnale $x(t)$.

Osserviamo che necessariamente i termini $\mathcal{P}_n = |X_n|^2$ risultano reali e positivi. Inoltre, se $x(t)$ è reale, risulta $|X_n|^2 = |X_{-n}^*|^2 = |X_{-n}|^2$, e quindi si ottiene $\mathcal{P}_n = \mathcal{P}_{-n}$; pertanto un segnale *reale* è caratterizzato da uno spettro di potenza *pari*.

Problema: si determini lo spettro di potenza di un'onda quadra. *Soluzione:* Essendo $X_n = \frac{1}{2} \text{sinc}\left(\frac{n}{2}\right)$, si ottiene $\{\mathcal{P}_n\} = \left\{|X_n|^2\right\} = \frac{1}{4} \left\{\text{sinc}^2\left(\frac{n}{2}\right)\right\}$.

2.4 Appendici

2.4.1 Algebra vettoriale

Spazio normato Un insieme di elementi viene detto *spazio lineare* (o *spazio vettoriale*), quando sono definite le operazioni di somma tra elementi e di moltiplicazione degli stessi per dei coefficienti, e queste operazioni danno come risultato ancora un elemento dell'insieme.

Lo *spazio prodotto interno* (o *spazio normato*) è quello spazio lineare, in cui è definito il *prodotto scalare* (\bar{x}, \bar{y}) tra generici vettori \bar{x} ed \bar{y} ¹³. In tal caso, si può definire la *norma* $\|\bar{x}\|$ di un vettore \bar{x} come

$$\|\bar{x}\| = \sqrt{(\bar{x}, \bar{x})}$$

Due vettori si dicono *ortogonali* se il loro prodotto scalare è nullo, ossia $(\bar{x}, \bar{y}) = 0$.

Un generico punto \bar{x} di uno spazio lineare può esprimersi come combinazione di vettori \bar{u}_i di una base di rappresentazione, con coefficienti x_i :

$$\bar{x} = \sum_i x_i \bar{u}_i$$

Se lo spazio è normato, e per i vettori della base risulta $(\bar{u}_i, \bar{u}_j) = 0$ per tutti gli $i \neq j$, allora la base è detta *ortogonale*, ed i coefficienti x_i si determinano per proiezione di \bar{x} lungo i vettori della base:

$$x_i = (\bar{x}, \bar{u}_i)$$

In tal caso, il prodotto scalare tra due vettori \bar{x} ed \bar{y} ha espressione¹⁴

$$(\bar{x}, \bar{y}) = \sum_i x_i y_i^* \|\bar{u}_i\|^2$$

Se $\|\bar{u}_i\|^2 = 1$, allora gli \bar{u}_i sono *unitari* e la base è detta *ortonormale*.

¹³Il prodotto scalare è un operatore che associa ad una coppia di vettori uno scalare. Indicando con (\bar{x}, \bar{y}) il prodotto scalare tra \bar{x} ed \bar{y} , tale operatore deve soddisfare alle seguenti tre proprietà:

- $(\bar{x}, \bar{y}) = (\bar{y}, \bar{x})$ - proprietà *commutativa*;
- $(a\bar{x} + b\bar{y}, \bar{z}) = a(\bar{x}, \bar{z}) + b(\bar{y}, \bar{z})$ - proprietà *distributiva*;
- $(\bar{x}, \bar{x}) \geq 0$ (con il segno uguale solo se $\bar{x} = 0$).

Qualora lo spazio normato sia anche *completo*, prende il nome di *spazio di Hilbert* (vedi http://it.wikipedia.org/wiki/Spazio_di_Hilbert). Il senso "naif" di *completo* è che *i punti ci sono tutti*, mentre quello più matematicamente forbito è che *tutte le successioni di Cauchy sono convergenti ad un elemento dello spazio*.

¹⁴E' facile verificare che il risultato ottenuto è direttamente applicabile allo spazio descritto dalla geometria euclidea, in cui gli u_i sono unitari ed orientati come gli assi cartesiani, ottenendo in definitiva

$$(\bar{x}, \bar{y}) = x_1 y_1 + x_2 y_2 + x_3 y_3$$

Spazio dei segnali periodici I concetti ora esposti sono immediatamente applicabili all'insieme dei segnali periodici di periodo T , una volta assimilati a elementi di uno spazio normato, per i quali viene definito un operatore di prodotto scalare tra due segnali $x(t)$ ed $y(t)$ come l'integrale

$$(x(t), y(t)) = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) y^*(t) dt$$

a cui corrisponde una *norma quadratica* immediatamente riconoscibile come la *potenza* del segnale:

$$\|x(t)\|^2 = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |x(t)|^2 dt$$

Considerando quindi il sottospazio dei segnali con periodo $T = \frac{1}{F}$ costituito dall'insieme $\{e^{j2\pi n F t}\}$, la validità di (2.6) di pag. 24 prova che gli $\{e^{j2\pi n F t}\}$ costituiscono una base ortonormale per i segnali periodici di periodo T ; in particolare si riconosce che l'espressione (2.2) di pag. 17 rappresenta la proiezione¹⁵ del segnale lungo i vettori della base, mentre la formula di ricostruzione (2.3) costituisce la rappresentazione del segnale nei termini delle sue componenti ortogonali.

Ri-definizione dei coefficienti di Fourier Moltiplicando il segnale periodico per $e^{-j2\pi m F t}$ ed eseguendo l'integrale tra due istanti t_1 e t_2 presi a distanza di un multiplo intero di periodi (ossia $t_2 - t_1 = kT$), si ottiene

$$\begin{aligned} \int_{t_1}^{t_2} x(t) e^{-j2\pi m F t} dt &= \int_{t_1}^{t_2} \left(\sum_{n=-\infty}^{\infty} X_n e^{j2\pi n F t} \right) e^{-j2\pi m F t} dt = \\ &= \sum_{n=-\infty}^{\infty} X_n \int_{t_1}^{t_2} e^{j2\pi(n-m)F t} dt = (t_2 - t_1) \cdot X_m \end{aligned}$$

in quanto per $n \neq m$ la funzione integranda ha valor medio nullo, dato che nell'intervallo (t_1, t_2) (dovunque collocato dell'asse dei tempi) presenta un numero intero di periodi. Pertanto, il calcolo dei coefficienti può ottenersi a partire da un qualunque intervallo esteso su un numero intero di periodi:

$$X_n = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} x(t) e^{-j2\pi n F t} dt$$

Osserviamo inoltre come l'espressione che permette il calcolo della lunghezza di un vettore

$$\|\bar{x}\| = \sqrt{\sum_i (x_i)^2}$$

non sia nient'altro che la riproposizione del teorema di Pitagora, che (su due dimensioni) asserisce l'uguaglianza dell'area del quadrato costruito sull'ipotenusa, con la somma delle aree dei quadrati



costruiti sui cateti. Infatti

¹⁵ Infatti, il prodotto scalare si calcola come il prodotto dei moduli, moltiplicato per l'angolo compreso tra i due: $(\bar{x}, \bar{y}) = |x| \cdot |y| \cdot \cos \theta$. Se il secondo vettore ha lunghezza unitaria, si ottiene la proiezione del primo nella direzione del secondo.

Disuguaglianza di Schwartz Consiste nel risultato

$$\left| \int_{-\infty}^{\infty} x(t) y^*(t) dt \right|^2 \leq \int_{-\infty}^{\infty} |x(t)|^2 dt \cdot \int_{-\infty}^{\infty} |y(t)|^2 dt$$

che a volte può tornare utile nei calcoli che coinvolgono segnali di energia. La dimostrazione si basa sull'identificare l'insieme di tali segnali come uno spazio normato, dotato di un operatore di prodotto scalare tra $x(t)$ ed $y(t)$ definito come

$$(x(t), y(t)) = \int_{-\infty}^{\infty} x(t) y^*(t) dt \quad (2.7)$$

Con tali posizioni, il risultato mostrato deriva da quello valido per un qualunque spazio vettoriale, che fa uso della disuguaglianza $|\cos \theta| \leq 1$, e che mostra che

$$(\bar{x}, \bar{y})^2 = (|x| \cdot |y| \cdot \cos \theta)^2 \leq |x|^2 \cdot |y|^2$$

Applicando quindi la definizione di prodotto scalare (2.7) ai due vettori-segnale $\bar{x} = x(t)$ e $\bar{y} = y(t)$, si ottiene il risultato espresso dalla disuguaglianza di Schwartz, in cui $\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} x(t) x^*(t) dt = (\bar{x}, \bar{x})$, e per la quale vale il segno di uguale se e solo se $x(t) = Ky(t)$, con K costante reale.

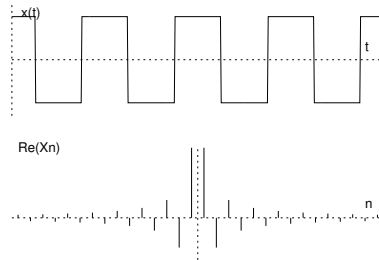
2.4.2 Esempi di Sviluppo in serie

Nello schema che segue, sono mostrate le ampiezze delle componenti armoniche X_n per alcuni segnali periodici di periodo T , di cui è fornita l'espressione nel tempo per $|t| < T/2$.

Onda quadra simmetrica

$$x(t) = \begin{cases} +1 & |t| < T/4 \\ -1 & T/4 \leq |t| < T/2 \end{cases}$$

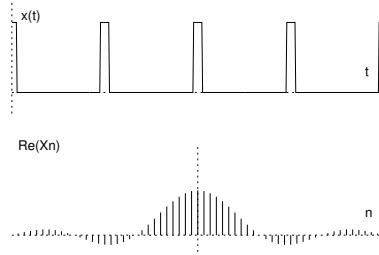
$$X_n = \begin{cases} \text{sinc}\left(\frac{n}{2}\right) & n \neq 0 \\ 0 & n = 0 \end{cases}$$



Treno di impulsi rettangolari

$$x(t) = \begin{cases} +1 & |t| < \tau/2 \\ 0 & \tau/2 \leq |t| < T/2 \end{cases}$$

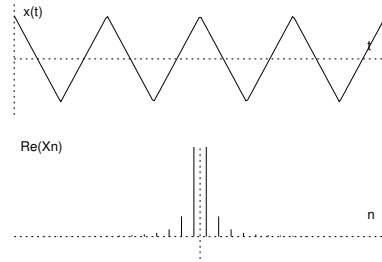
$$X_n = \frac{\tau}{T} \text{sinc}\left(\frac{n\tau}{T}\right)$$



Onda triangolare simmetrica

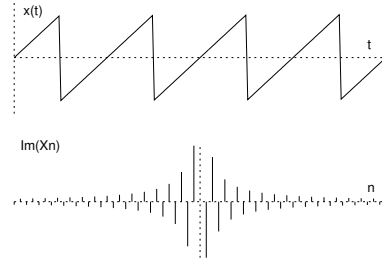
$$x(t) = 1 - 4\frac{|t|}{T} \quad |t| < T/2$$

$$X_n = \begin{cases} \text{sinc}^2\left(\frac{n}{2}\right) & n \neq 0 \\ 0 & n = 0 \end{cases}$$

**Dente di sega simmetrico**

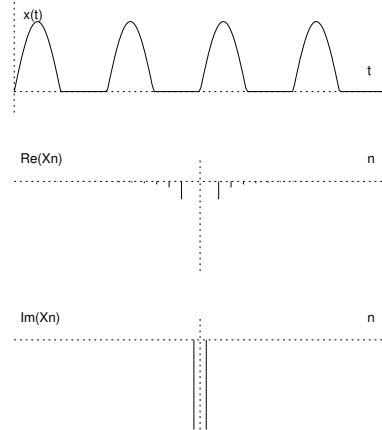
$$x(t) = 2\frac{t}{T} \quad |t| < T/2$$

$$X_n = \begin{cases} j\frac{(-1)^n}{n\pi} & n \neq 0 \\ 0 & n = 0 \end{cases}$$

**Rettificata a singola semionda**

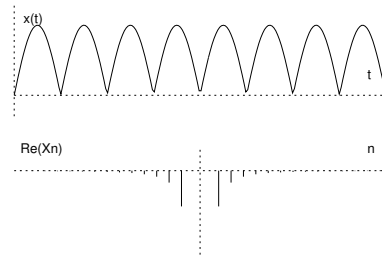
$$x(t) = \begin{cases} \sin \omega_0 t & 0 \leq t < T/2 \\ 0 & -T/2 \leq t < 0 \end{cases}$$

$$X_n = \begin{cases} \frac{1}{\pi(1-n^2)} & n \text{ pari} \\ -j\frac{1}{4} & n = \pm 1 \\ 0 & \text{altrimenti} \end{cases}$$

**Rettificata a onda intera**

$$x(t) = |\sin \omega_0 t|$$

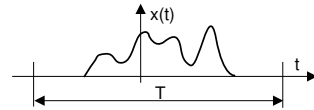
$$X_n = \begin{cases} \frac{2}{\pi(1-n^2)} & n \text{ pari} \\ 0 & \text{altrimenti} \end{cases}$$



Capitolo 3

Trasformata di Fourier

Abbiamo già osservato al § 2.2 che lo sviluppo in serie di Fourier può essere applicato ad un segnale limitato nel tempo, e che l'uso della formula di ricostruzione rende periodico il segnale originario. Se però facciamo tendere ad infinito il periodo *fittizio* T su cui sono calcolati i coefficienti X_n , le armoniche della serie di Fourier tendono ad infittirsi, fino ad arrivare ad una distanza infinitesima; allo stesso tempo, la periodizzazione del segnale ricostruito tende via via a scomparire.



3.1 Definizione

La trasformata di Fourier serve a rappresentare quei segnali per i quali non sussiste una struttura periodica, ed è un operatore funzionale che, applicato ad un segnale definito nel dominio del tempo, ne individua un altro nel dominio della variabile *continua* frequenza (a differenza della serie *discreta* di Fourier, idonea al caso in cui siano presenti *solo armoniche* della fondamentale). L'operazione di trasformazione è spesso indicata con la simbologia $X(f) = \mathcal{F}\{x(t)\}$, ed il segnale trasformato si indica con la stessa variabile di quello nel tempo, scritta in maiuscolo. La sua definizione formale dal punto di vista analitico è:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt$$

la cui esistenza è garantita per segnali $x(t)$ impulsivi (ovvero per i quali $\int_{-\infty}^{\infty} |x(t)| dt < \infty$, cioè assolutamente sommabili). Un segnale impulsivo è anche di energia, mentre non è sempre vero il viceversa. Spesso però, $X(f)$ esiste anche per segnali di energia; vedremo inoltre che può essere definita (grazie ad operazioni di passaggio al limite) anche per segnali di potenza periodici.

L'antitrasformata di Fourier $\mathcal{F}^{-1}\{\}$ è l'operatore analitico che svolge l'associazione inversa a $\mathcal{F}\{\}$, e che consente di ottenere, a partire da un segnale definito nel dominio della frequenza, quel segnale nel dominio del tempo la cui trasformata è il primo segnale. L'operazione di antitrasformazione è definita come

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df$$

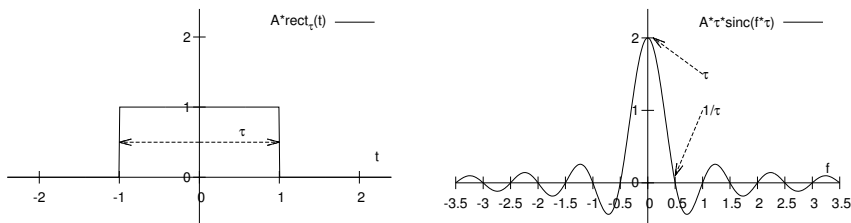


Figura 3.1: \mathcal{F} -trasformata di un rettangolo di base $\tau = 2$ ed ampiezza $A = 1$

e vale ovunque $x(t)$ sia continuo, mentre nelle discontinuità di prima specie fornisce il valor medio di $x(t)$. Il risultato della trasformata $X(f) = M(f) \exp^{j\varphi(f)}$ è anche detto *spettro di ampiezza complessa*, mentre $M(f)$ ed $\varphi(f)$ sono detti spettri di *modulo* e *fase*.

La formula di ricostruzione, se messa a confronto con la serie di Fourier, può essere pensata come una somma integrale di *infinite* componenti $X(f) df e^{j2\pi ft}$ di ampiezza (complessa) infinitesima, evidenziando come ora siano presenti *tutte* le frequenze e non solo le armoniche. Una seconda analogia con la serie di Fourier deriva dal considerare un segnale $x(t)$ di durata limitata T , e calcolare $X(f) = \mathcal{F}\{x(t)\}$ per $f = \frac{n}{T} = nF$. In tal caso, è facile verificare¹ che risulta

$$X(f = nF) = T \cdot X_n \quad (3.1)$$

con X_n pari all' n -esimo coefficiente di Fourier calcolato per $x(t)$ su quello stesso periodo.

Prima di procedere con le proprietà e le caratteristiche di questa trasformata, svolgiamo un semplice esercizio.

Trasformata di un rettangolo Disponendo del segnale $x(t) = A \text{rect}_\tau(t)$, se ne vuol calcolare lo spettro di ampiezza $X(f)$. Svolgendo i calcoli si ottiene:

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} A \text{rect}_\tau(t) e^{-j2\pi ft} dt = A \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} e^{-j2\pi ft} dt = A \left. \frac{e^{-j2\pi ft}}{-j2\pi f} \right|_{-\frac{\tau}{2}}^{\frac{\tau}{2}} = \\ &= \frac{A}{\pi f} \frac{e^{j2\pi f \frac{\tau}{2}} - e^{-j2\pi f \frac{\tau}{2}}}{2j} = A\tau \frac{\sin(\pi f \tau)}{\pi f \tau} = A\tau \cdot \text{sinc}(f\tau) \end{aligned}$$

Questo risultato, graficato in fig 3.1, ricorda quello già incontrato a pag. 20 per la serie di Fourier dell'onda quadra. Il noto andamento $\frac{\sin x}{x}$ rappresenta ora la distribuzione *continua* in frequenza dello spettro di ampiezza, ed il primo zero della curva si trova presso $f = \frac{1}{\tau}$, in modo del tutto simile al treno di impulsi rettangolari di base τ . Notiamo esplicitamente inoltre che, *aumentando* la durata del *rect*, lo spettro si *concentra*, addensandosi nella regione delle frequenze più basse; mentre al contrario, qualora il *rect* sia più breve, $X(f)$ si estende a regioni di frequenza più elevata.

¹

$$\begin{aligned} X(nF) &= \int_{-\infty}^{\infty} x(t) e^{-j2\pi nFt} dt = \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) e^{-j2\pi nFt} dt = \\ &= T \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) e^{-j2\pi nFt} dt = T \cdot X_n \end{aligned}$$

3.2 Energia mutua e densità di energia

Similmente al caso dei segnali periodici, individuiamo la relazione tra l'energia totale di un segnale, e come questa si *distribuisce* nel dominio della frequenza. In base alle considerazioni geometriche del § 2.4.1, definiamo il *prodotto scalare* tra segnali di energia $x(t)$ e $y(t)$ (detto anche *energia mutua*²) il valore

$$\mathcal{E}_{xy} = (\bar{x}, \bar{y}) = \int_{-\infty}^{\infty} x(t) y^*(t) dt \quad (3.2)$$

che, nel caso in cui $x(t) = y(t)$, coincide con l'energia \mathcal{E}_x di $x(t)$. Se entrambi $x(t)$ e $y(t)$ possiedono trasformata di Fourier possiamo scrivere:

$$\begin{aligned} \mathcal{E}_{xy} &= \int y^*(t) \left[\int X(f) e^{j2\pi ft} df \right] dt = \int X(f) \left[\int y^*(t) e^{j2\pi ft} dt \right] df \\ &= \int_{-\infty}^{\infty} X(f) Y^*(f) df \end{aligned}$$

Il risultato

$$\int_{-\infty}^{\infty} x(t) y^*(t) dt = \int_{-\infty}^{\infty} X(f) Y^*(f) df$$

esprime il *teorema di Parseval* per segnali di energia, ed implica che le trasformate di segnali ortogonali, sono anch'esse ortogonali. Ponendo ora $x(t) = y(t)$, si ottiene:

$$\mathcal{E}_x = (\bar{x}, \bar{x}) = \|x\|^2 = \int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df$$

Esaminando quest'ultima espressione, possiamo indicare

$$\mathcal{E}_x(f) = |X(f)|^2$$

come lo *spettro di densità di energia* di $x(t)$. Infatti, l'integrale $\int_{f_1}^{f_2} |X(f)|^2 df$ rappresenta il contributo all'energia totale \mathcal{E}_x di $x(t)$, limitatamente alla banda di frequenze comprese tra f_1 ed f_2 .

3.3 Prime proprietà della trasformata di Fourier

Simmetria coniugata Nel caso in cui $x(t)$ sia reale, risulta³

$$X(f) = X^*(-f)$$

e quindi la parte reale di $X(f)$ è *pari*, e quella immaginaria *dispari*, ossia modulo $|X(f)|$ pari e fase $\arg\{X(f)\}$ dispari; si applica inoltre il corollario di pag. 18.

²Nei testi anglofoni la (3.2) è indicata come *cross-energy*, tradotta letteralmente come *energia incrociata*, o meglio, *in comune*.

³Infatti $X^*(f) = \left[\int x(t) e^{-j2\pi ft} dt \right]^* = \int x^*(t) e^{j2\pi ft} dt = X(-f)$ dato che $x(t)$ è reale.

Dualità Trasformata ed antitrasformata differiscono solo per il segno. Ciò comporta che se sostituiamo alla variabile f del risultato $X(f)$ di una \mathcal{F} -trasformata, la variabile t , si ottiene una funzione del tempo $X(t)$ che, se nuovamente trasformata, fornisce ... il segnale originario $x(t)$, espresso come funzione della variabile f , cambiata di segno: $x(-f)$. Il concetto esposto, verificabile analiticamente con un pò di pazienza, si riassume come

$$\begin{aligned} \text{se } x(t) \xrightarrow{\mathcal{F}\{\}} X(f) \text{ allora sostituendo } f \text{ con } t &\rightarrow X(t) \xrightarrow{\mathcal{F}\{\}} x(-f) \\ \text{se } X(f) \xrightarrow{\mathcal{F}^{-1}\{\}} x(t) \text{ allora sostituendo } t \text{ con } f &\rightarrow x(f) \xrightarrow{\mathcal{F}^{-1}\{\}} X(-t) \end{aligned}$$

e consente l'uso di risultati ottenuti "in un senso" (ad es. da tempo a frequenza) per derivare senza calcoli i risultati nell'altro (da frequenza a tempo), o viceversa.

Esempio: Trasformata di un sinc(t) Supponiamo di voler \mathcal{F} -trasformare il segnale $x(t) = B \frac{\sin(\pi t B)}{\pi t B} = B \text{sinc}(tB)$: l'applicazione cieca dell'integrale che definisce la trasformata di Fourier al segnale $x(t)$ appare un'impresa ardua...

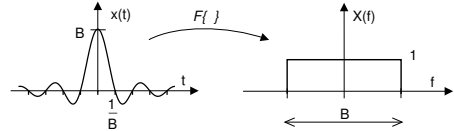
Allora, ricordando che

$$\mathcal{F}\{\text{rect}_\tau(t)\} = \tau \text{sinc}(f\tau)$$

scriviamo direttamente

$$\mathcal{F}\{B \cdot \text{sinc}(tB)\} = \text{rect}_B(f)$$

Pertanto la trasformata di un $\frac{\sin x}{x}$ nel tempo, è un rettangolo in frequenza.



Linearità Discende molto semplicemente dalla proprietà distributiva dell'integrale che definisce la trasformata. Pertanto:

$$\text{se } z(t) = ax(t) + by(t) \text{ allora } Z(f) = aX(f) + bY(f)$$

Valore medio e valore iniziale Subito verificabile una volta notato che la \mathcal{F} -trasformata, calcolata per $f = 0$, si riduce all'integrale di $x(t)$, e quindi al suo *valor medio*. Pertanto:

$$m_x = \int_{-\infty}^{\infty} x(t) dt = X(f = 0) \quad \text{e, per dualità} \quad x_0 = x(t = 0) = \int_{-\infty}^{\infty} X(f) df$$

dove l'ultima relazione esprime la proprietà del *valore iniziale*.

Esempio Come applicazione, troviamo subito che

$$\int_{-\infty}^{\infty} \text{sinc}(tB) dt = \frac{1}{B} \text{rect}_B(f = 0) = \frac{1}{B}$$

Traslazione nel tempo Si tratta di una proprietà molto semplice, e che ricorre frequentemente nei calcoli sui segnali. Manifesta la relazione esistente tra la trasformata dei segnali e quella degli stessi traslati, e si esprime con il predicato:

$$\begin{aligned} \text{se } z(t) &= x(t - T) \\ \text{allora } Z(f) &= X(f) e^{-j2\pi fT} \end{aligned}$$

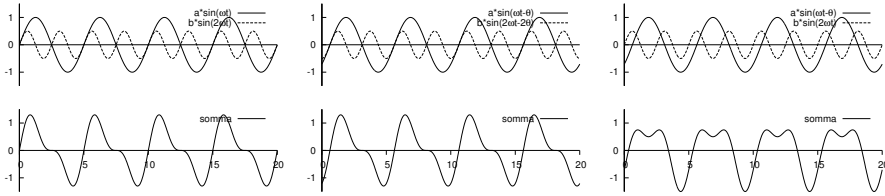


Figura 3.2: Confronto tra diversi spettri di fase

la cui dimostrazione è fornita sotto⁴.

Esempio La figura a lato esemplifica il risultato ottenuto nel caso in cui $x(t) = \text{rect}_\tau(t - T)$, mostrando come la traslazione temporale del rect determini per $x(t)$ uno spettro di modulo ancora pari a

$$X(f) = \mathcal{F}\{\text{rect}_\tau(t)\} = \tau \text{sinc}(f\tau)$$

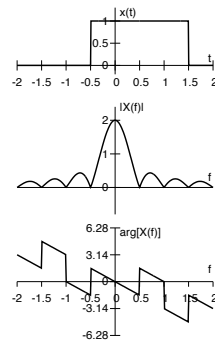
a cui si aggiunge un contributo di fase *lineare*

$$\varphi(f) = -2\pi fT$$

Nel caso in figura, si è posto $\tau = 2$ e $T = .5$, ottenendo in definitiva

$$Z(f) = X(f) e^{-j2\pi fT} = 2 \text{sinc}(2f) e^{-j\pi f}$$

Le discontinuità di fase osservabili hanno ampiezza π , ed hanno origine dai cambi di segno della funzione $\text{sinc}(f\tau) = |\text{sinc}(f\tau)| e^{-j\pi\phi(f)}$, in cui $\phi(f)$ vale ± 1 con periodo $1/\tau$.



Poniamo ora l'attenzione sul fatto che l'espressione $x(t - T)$ indica un *ritardo* del segnale $x(t)$ di una quantità pari a T .

Linearità di fase La circostanza che un *ritardo temporale* del segnale $x(t)$ si traduca in una alterazione *lineare* della fase⁵ della sua trasformata $X(f)$, comporta una conseguenza notevole anche nel passaggio da f a t , ossia:

Se si desidera che un segnale mantenga inalterata la sua forma d'onda, pur subendo una alterazione della propria trasformata, l'unica possibilità è quella di modificare solamente lo spettro di fase, con andamento lineare in frequenza.

Esempio Consideriamo un segnale periodico $x(t)$ costituito da due sole armoniche:

$$x(t) = a \sin(\omega t) + b \sin(2\omega t)$$

(avendo posto $2\pi F = \omega$), assieme alla sua versione ritardata

$$x(t - T) = a \sin(\omega(t - T)) + b \sin(2\omega(t - T)) = a \sin(\omega t - \omega T) + b \sin(2\omega t - 2\omega T)$$

Ponendo $\omega T = \theta$, otteniamo

$$x(t - T) = a \sin(\omega t - \theta) + b \sin(2\omega t - 2\theta)$$

⁴La dimostrazione si basa su di un semplice cambio di variabile: $Z(f) = \int x(t - T) e^{-j2\pi f t} dt = \int x(\theta) e^{-j2\pi f(T + \theta)} d\theta = e^{-j2\pi f T} \int x(\theta) e^{-j2\pi f \theta} d\theta = X(f) e^{-j2\pi f T}$

⁵La circostanza evidenziata fa sì che, nel caso in cui $X(f)$ presenti un andamento *lineare* della fase, si usa dire che è presente un *ritardo di fase*.

e verifichiamo che la seconda armonica subisce un ritardo di fase esattamente doppio.

In fig. 3.2 si è posto $a = 1$; $b = .5$; $\theta = \frac{\pi}{4}$ e $F = .2$, ed è mostrato sia il segnale somma originario, sia quello ottenuto considerando un contributo di fase lineare per le due armoniche. Verifichiamo che nel secondo caso, la forma d'onda è la stessa ottenibile per $T = 0$, in quanto le armoniche sono traslate del medesimo intervallo temporale. A destra invece, la fase della seconda armonica viene annullata, ottenendo dalla somma un segnale $a \sin(2\pi Ft - \theta) + b \sin(2\pi 2Ft)$. Come è evidente, in questo caso il risultato assume una forma completamente diversa⁶.

Traslazione in frequenza (Modulazione) E' la proprietà duale della precedente, e stabilisce che

$$\text{se } Z(f) = X(f - f_0) \quad \text{allora } z(t) = x(t) e^{j2\pi f_0 t}$$

la cui dimostrazione è del tutto analoga a quanto già visto. Da un punto di vista mnemonico, distinguiamo la traslazione temporale da quella in frequenza per il fatto che, nel primo caso, i *segnali* della traslazione e dell'esponenziale complesso sono *uguali*, e nel secondo, *opposti*.

Da un punto di vista pratico, può sorgere qualche perplessità per la comparsa di un segnale *complesso* nel tempo. Mostriamo però che anti-trasformando uno spettro ottenuto dalla somma di due traslazioni opposte, si ottiene un segnale reale:

$$\mathcal{F}^{-1}\{X(f - f_0) + X(f + f_0)\} = x(t) e^{j2\pi f_0 t} + x(t) e^{-j2\pi f_0 t} = 2x(t) \cos 2\pi f_0 t$$

Pertanto, lo sdoppiamento e la traslazione di $X(f)$ in $\pm f_0$ sono equivalenti ad un segnale cosinusoidale di frequenza f_0 , la cui ampiezza è modulata dal segnale $x(t) = \mathcal{F}^{-1}\{X(f)\}$. E' proprio per questo motivo, che la proprietà è detta di *modulazione* (vedi anche a pag. 41).

Coniugato Deriva direttamente⁷ dalla definizione di \mathcal{F} -trasformata:

$$\mathcal{F}\{x^*(t)\} = X^*(-f); \quad \mathcal{F}^{-1}\{X^*(f)\} = x^*(-t) \quad (3.3)$$

Nel caso di segnali reali, ritroviamo la proprietà di simmetria coniugata $X(f) = X^*(-f)$.

Cambiamento di scala Quantifica gli effetti sullo spettro di una variazione nella velocità di scorrimento del tempo (e viceversa). Possiamo ad esempio pensare come, riavvolgendo velocemente un nastro magnetico, si ascolta un segnale di durata più breve, e dal timbro più *acuto*. Questo fenomeno viene espresso analiticamente come:

$$\mathcal{F}\{x(at)\} = \frac{1}{|a|} X\left(\frac{f}{a}\right)$$

in cui scegliendo ad es. $a > 1$, si ottiene una *accelerazione* temporale ed una *allargamento* dello spettro (ed il contrario, con $a < 1$). La dimostrazione (per $a > 0$) è riportata alla nota⁸. Un corollario di questa proprietà, è che se $a = -1$, allora

$$\mathcal{F}\{x(-t)\} = X(-f)$$

⁶Nel seguito illustreremo che una conseguenza del risultato discusso, è la sensibilità delle trasmissioni numeriche alle distorsioni di fase.

⁷Infatti $\mathcal{F}\{x^*(t)\} = \int_{-\infty}^{\infty} x^*(t) e^{-j2\pi ft} dt = \int_{-\infty}^{\infty} [x(t) e^{j2\pi ft}]^* dt$.

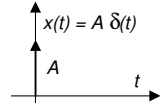
⁸ $\int x(at) e^{-j2\pi ft} dt = \frac{1}{a} \int x(at) e^{-j2\pi \frac{f}{a} at} d(at) = \frac{1}{a} \int X(\beta) e^{-j2\pi \frac{f}{a} \beta} d\beta = \frac{1}{a} X\left(\frac{f}{a}\right)$

3.4 Impulso matematico

Prima di esporre altre proprietà della trasformata di Fourier, occorre definire ed analizzare le proprietà della “funzione” *impulso matematico*, indicato con il simbolo $\delta()$ e chiamato anche *delta di Dirac*. Questo è definito come un segnale $\delta(t)$ che vale zero ovunque, tranne per $t = 0$ dove vale ∞ ; per contro, l'area di $\delta(t)$ è *unitaria*:

$$\delta(t) = \begin{cases} \infty & \text{con } t = 0 \\ 0 & \text{altrove} \end{cases} \quad \text{e} \quad \int_{-\infty}^{\infty} \delta(t) dt = 1$$

Da un punto di vista analitico, $\delta(t)$ non è una funzione, ma una *distribuzione*⁹, definita come il limite a cui tende una serie di funzioni, come discusso in appendice 3.9.2. E' prassi rappresentare graficamente $A \cdot \delta(f)$ come una freccia (vedi figura) con scritto accanto il valore dell'area A .



3.4.1 Trasformata di una costante

La trasformata di una costante è un impulso matematico, di area pari al valore della costante.

Questa proprietà è valida per entrambi i domini (f e t) di partenza, fornendo

$$\mathcal{F}\{A\} = A \cdot \delta(f) \quad \text{e} \quad \mathcal{F}^{-1}\{A\} = A \cdot \delta(t)$$

In appendice 3.9.2 sono svolte riflessioni che illustrano come interpretare questo risultato. Qui osserviamo semplicemente che la costante A può essere vista come il limite, per $\tau \rightarrow \infty$, di un segnale rettangolare:

$$A = \lim_{\tau \rightarrow \infty} A \text{rect}_{\tau}(t)$$

la cui trasformata per $\tau \rightarrow \infty$ risulta

$$\mathcal{F}\left\{\lim_{\tau \rightarrow \infty} A \text{rect}_{\tau}(t)\right\} = \lim_{\tau \rightarrow \infty} A \tau \text{sinc}(f\tau) = \begin{cases} \infty & \text{con } f = 0 \\ 0 & \text{altrove} \end{cases}$$

Ci troviamo pertanto nelle esatte circostanze che definiscono un impulso matematico, e resta da verificare che $\int_{-\infty}^{\infty} \tau \text{sinc}(f\tau) df = 1$: si può mostrare (pag. 32) che tale integrale vale uno per qualunque τ , e quindi possiamo scrivere $\mathcal{F}\{A\} = A \cdot \delta(f)$.

3.4.2 Trasformata di segnali periodici

Consideriamo un segnale periodico $x(t)$, del quale conosciamo lo sviluppo in serie

$$x(t) = \sum_{n=-\infty}^{\infty} X_n e^{j2\pi n F t}$$

Applicando la proprietà di linearità, il risultato per la trasformata di una costante, e ricordando la proprietà della traslazione in frequenza, troviamo¹⁰ che la \mathcal{F} -trasformata di $x(t)$ vale:

$$X(f) = \sum_{n=-\infty}^{\infty} X_n \delta(f - nF)$$

⁹Vedi ad es. [http://it.wikipedia.org/wiki/Distribuzione_\(matematica\)](http://it.wikipedia.org/wiki/Distribuzione_(matematica)) per approfondire

¹⁰ $X(f) = \mathcal{F}\left\{\sum_{n=-\infty}^{\infty} X_n e^{j2\pi n F t}\right\} = \sum_{n=-\infty}^{\infty} X_n \mathcal{F}\{1 \cdot e^{j2\pi n F t}\} = \sum_{n=-\infty}^{\infty} X_n \cdot \delta(f - nF)$

Lo spettro di ampiezza di un segnale periodico è quindi costituito da *impulsi matematici*, situati in corrispondenza delle frequenze armoniche, e di area pari ai rispettivi coefficienti della serie di Fourier.

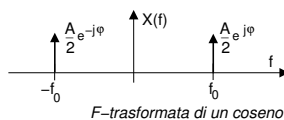
Un modo alternativo di calcolare la trasformata di segnali periodici è illustrato alla sezione 3.8.1.

Trasformata di un coseno Applichiamo il risultato trovato nel verso opposto, ossia per individuare le componenti armoniche, a partire dall'espressione della trasformata di Fourier. Nel caso di un coseno, che scriviamo

$$x(t) = A \cos(2\pi f_0 t + \varphi) = A \frac{e^{j(2\pi f_0 t + \varphi)} + e^{-j(2\pi f_0 t + \varphi)}}{2}$$

la \mathcal{F} -trasformata risulta:

$$\begin{aligned} X(f) &= \mathcal{F} \left\{ \frac{A}{2} (e^{j2\pi f_0 t} e^{j\varphi} + e^{-j2\pi f_0 t} e^{-j\varphi}) \right\} \\ &= \frac{A}{2} \{ e^{j\varphi} \delta(f - f_0) + e^{-j\varphi} \delta(f + f_0) \} \end{aligned}$$



in cui riconosciamo $X_1 = \frac{A}{2} e^{j\varphi}$ e $X_{-1} = \frac{A}{2} e^{-j\varphi}$ come mostrato in figura.

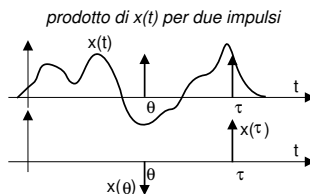
Potenza di un coseno Cogliamo l'occasione per calcolare la potenza di una sinusoidale. Applicando il teorema di Parseval si ottiene:

$$\mathcal{P}_x = |X_1|^2 + |X_{-1}|^2 = 2 \frac{A^2}{4} = \frac{A^2}{2}$$

3.4.3 Proprietà di setacciamento

Osserviamo innanzitutto che il *prodotto* di un segnale per un impulso unitario dà come risultato lo stesso impulso, con *area* pari al valore del segnale nell'istante in cui è centrato l'impulso:

$$x(t) \delta(t - \tau) = x(\tau) \delta(t - \tau) \quad (3.4)$$



Questa considerazione consente di scrivere il valore di $x(t)$ per un istante $t = \tau$, nella forma

$$x(\tau) = \int_{-\infty}^{\infty} x(t) \delta(t - \tau) dt \quad (3.5)$$

Quest'ultima proprietà è detta di *setacciamento* (in inglese, SIEVING) in quanto consiste nel passare (metaforicamente) al setaccio $x(t)$, che compare in entrambi i membri dell'espressione ottenuta, così come la farina compare *su entrambi i lati* del setaccio stesso.

3.5 Risposta impulsiva e convoluzione

Il titolo di questa sezione individua due concetti cardine nella descrizione dei sistemi fisici e delle relazione tra lo stimolo ad essi applicato, e l'effetto corrispondente.

3.5.1 Risposta impulsiva

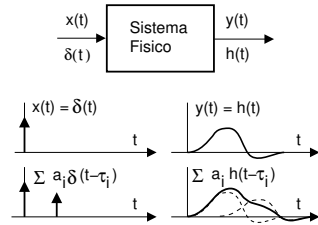
Consideriamo un sistema fisico (elettrico, meccanico, pneumatico...) che venga sollecitato (in un punto considerato come ingresso) da un segnale impulsivo $\delta(t)$, e consideriamo l'andamento temporale di una grandezza (meccanica, pneumatica, elettrica...) che possiamo considerare una uscita. Tale risultato prende il nome di *risposta impulsiva* (ossia all'impulso) e viene indicata con $h(t)$. L'andamento di $h(t)$ rappresenta la grandezza di uscita, osservata dopo che è passato un tempo pari a t da quando si è applicato in ingresso l'impulso $\delta(t)$.

Se il sistema è *lineare e permanente*¹¹, applicando un ingresso costituito da *più impulsi*, ognuno con differente area a_i e centrato ad un diverso istante τ_i , ovvero

$$x(t) = \sum_{i=1}^N a_i \delta(t - \tau_i) \quad (3.6)$$

si ottiene una uscita pari a

$$y(t) = \sum_{i=1}^N a_i h(t - \tau_i) \quad (3.7)$$



Si rifletta sul significato della sommatoria, con l'aiuto della figura a lato: ad un dato istante t , il valore dell'uscita $y(t)$ risulta dalla somma di N termini, ognuno pari al valore della risposta impulsiva calcolata con argomento pari al tempo trascorso tra l'istante di applicazione dell'*i-esimo* impulso e l'istante di osservazione.

3.5.2 Integrale di convoluzione

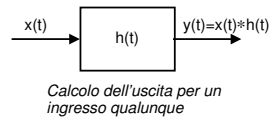
Consideriamo ancora lo stesso sistema fisico, al cui ingresso sia posto un generico segnale $x(t)$ che, grazie alla proprietà di sovrapposizione, rappresentiamo come scomposto in infiniti termini, ossia come somma integrale di impulsi centrati in τ (variabile) ed area $x(\tau) d\tau$ (infinitesima):

$$x(t) = \int_{-\infty}^{\infty} x(\tau) d\tau \delta(t - \tau)$$

Questa espressione, formalmente simile alla (3.6), è equivalente alla proprietà di sovrapposizione (3.5), dato che $\delta(t)$ è una funzione pari.

L'andamento della grandezza di uscita sarà il risultato della sovrapposizione di infinite risposte impulsive, ognuna relativa ad un diverso valore dell'ingresso:

$$y(t) = \int_{-\infty}^{\infty} x(\tau) h(t - \tau) d\tau$$



in cui $x(\tau) d\tau$ è l'area degli impulsi che costituiscono l'ingresso, e $h(t - \tau)$ è l'uscita all'istante t causata dall'impulso in ingresso centrato all'istante τ . Il risultato ottenuto, formalmente simile a (3.7), prende il nome di *integrale di convoluzione*, e viene indicato in forma simbolica

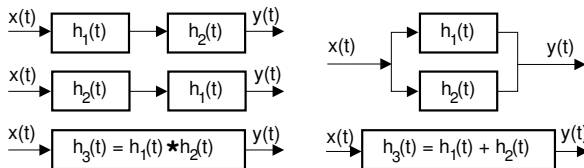
¹¹Il significato di questa classificazione si trova al Capitolo 1, a pag. 11.

da un asterisco (*), in modo che ci si possa riferire ad esso anche come *prodotto di convoluzione*, ossia $g(t) = x(t) * h(t)$.

Notiamo come $h(t)$ caratterizzi completamente il sistema fisico, in quanto permette di calcolarne l'uscita per un qualsiasi ingresso.

Proprietà commutativa Se un segnale con andamento $h(t)$ è posto in ingresso ad un sistema con risposta impulsiva $x(t)$, si ottiene ancora la stessa uscita, in quanto l'integrale di convoluzione è *commutativo*¹²:

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau) h(t - \tau) d\tau = \int_{-\infty}^{\infty} h(\tau) x(t - \tau) d\tau = h(t) * x(t)$$



Questa proprietà, assieme a quella di linearità, consente di stabilire le equivalenze mostrate in figura, dove si mostra come l'attraversamento *in serie* ed *in parallelo* di più sistemi lineari può essere

Risposta impulsiva equivalente per sistemi in serie e parallelo
riserondito all'attraversamento di un sistema equivalente, con risposta impulsiva pari rispettivamente alla convoluzione ed alla somma delle singole risposte impulsive.

3.5.3 La risposta impulsiva come funzione memoria

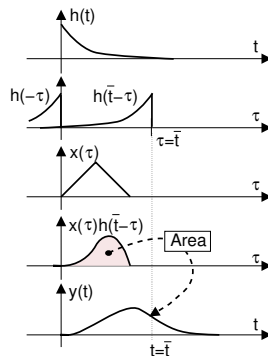
Diamo ora un'interpretazione grafica della convoluzione: poniamo che $h(t)$ sia un esponenziale decrescente ed $x(t)$ un segnale triangolare, come mostrato in figura, assieme alla funzione integranda che compare nel calcolo dell'uscita ad un generico istante $t = \bar{t}$.

L'andamento di $h(\bar{t} - \tau)$ con τ variabile indipendente, si ottiene ribaltando $h(t)$ rispetto all'origine dei tempi e traslandola (nel passato, quindi a destra) di \bar{t} . Il risultato dell'integrale di convoluzione, quando $t = \bar{t}$, è pari a

$$y(\bar{t}) = \int_{-\infty}^{\infty} x(\tau) h(\bar{t} - \tau) d\tau$$

ossia pari all'area del prodotto $x(\tau) h(\bar{t} - \tau)$, tratteggiata in figura; per altri valori di \bar{t} , il termine $h(\bar{t} - \tau)$ sarà traslato di una diversa quantità.

Il calcolo dell'area di $x(\tau) h(\bar{t} - \tau)$ ha il significato di sommare le risposte causate da tutti i valori di ingresso, ogni risposta presa con il rispettivo ritardo $\bar{t} - \tau$ tra gli istanti (passati) $\tau \leq \bar{t}$ di applicazione dei valori di ingresso, e l'istante \bar{t} di osservazione. Pertanto, i valori di $h(t)$ rappresentano la memoria, da parte del sistema fisico, degli ingressi precedenti.



¹²Infatti, adottando il cambio di variabile $t - \tau = \theta$, si ottiene $\int_{-\infty}^{\infty} x(\tau) h(t - \tau) d\tau = \int_{-\infty}^{\infty} x(t - \theta) h(\theta) d\theta$.

Estensione temporale della convoluzione In base alla costruzione grafica discussa, è facile verificare che se $x(t)$ ed $h(t)$ presentano una estensione temporale limitata, ovvero $x(t) \neq 0$ con $0 \leq t \leq T_x$ e $h(t) \neq 0$ con $0 \leq t \leq T_h$, allora il risultato $y(t) = x(t) * h(t)$ ha estensione compresa tra $t = 0$ e $t = T_x + T_h$, ossia presenta una durata pari alla somma delle durate.

3.5.4 Convoluzione con l'impulso traslato

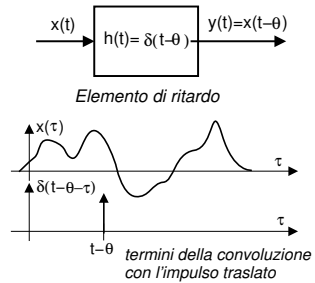
Consideriamo un sistema fisico che operi un semplice ritardo θ sui segnali in ingresso: in tal caso risulterà $h(t) = \delta(t - \theta)$, ovvero: la risposta all'impulso è un impulso ritardato.

Per calcolare l'uscita, che sappiamo essere pari a $y(t) = x(t - \theta)$, possiamo ricorrere all'integrale di convoluzione, ottenendo

$$\begin{aligned} y(t) &= x(t) * h(t) = x(t) * \delta(t - \theta) = \\ &= \int_{-\infty}^{\infty} x(\tau) \delta(t - \theta - \tau) d\tau = x(t - \theta) \end{aligned}$$

Questo risultato ci permette di enunciare un principio generale, che verrà utilizzato di frequente, e che recita:

La convoluzione tra un segnale $x(t)$ ed un impulso matematico $\delta(t - \theta)$ centrato ad un istante θ provoca la traslazione di $x(t)$ all'istante in cui è centrato l'impulso.



3.6 Moltiplicazione in frequenza e nel tempo

La descrizione di un sistema fisico per mezzo della sua risposta impulsiva è di fondamentale utilità soprattutto per merito della seguente proprietà della trasformata di Fourier:

La \mathcal{F} -trasformata della convoluzione tra due segnali è pari al prodotto delle trasformate dei segnali:

$$\mathcal{F}\{x(t) * y(t)\} = X(f)Y(f)$$

La dimostrazione è riportata alla nota¹³. Sussiste inoltre anche la proprietà *duale*, ovvero ad un *prodotto nel tempo* corrisponde una *convoluzione in frequenza*, che si scrive

$$\mathcal{F}\{x(t) \cdot y(t)\} = X(f) * Y(f)$$

In Fig. 3.3 è mostrato come l'ultima relazione individui un *isomorfismo* tra spazi di segnale. Nel seguito, trattiamo delle conseguenze e dei risvolti legati alla coppia di proprietà duali ora introdotte, iniziando dalla prima.

¹³
$$\begin{aligned} Z(f) &= \mathcal{F}\{x(t) * y(t)\} = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x(\tau) y(t - \tau) d\tau \right] e^{-j2\pi ft} dt = \\ &= \int_{-\infty}^{\infty} x(\tau) \left[\int_{-\infty}^{\infty} y(t - \tau) e^{-j2\pi ft} dt \right] d\tau = \int_{-\infty}^{\infty} x(\tau) Y(f) e^{-j2\pi f\tau} d\tau = \\ &= Y(f) \int_{-\infty}^{\infty} x(\tau) e^{-j2\pi f\tau} d\tau = Y(f) \cdot X(f) \end{aligned}$$

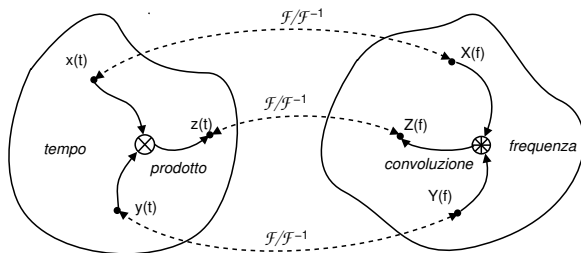


Figura 3.3: Isomorfismo tra gli spazi di segnale nel tempo e nella frequenza

3.6.1 Moltiplicazione in frequenza (*filtraggio*)

Questa proprietà consente una diversa modalità di calcolo dell'uscita da un sistema fisico, che può essere infatti ricavata operando nel dominio della frequenza, in base all'espressione

$$Y(f) = \mathcal{F}\{x(t) * h(t)\} = X(f)H(f)$$

e quindi valutando $y(t) = \mathcal{F}^{-1}\{Y(f)\}$. La trasformata della risposta impulsiva $H(f) = \mathcal{F}\{h(t)\}$ prende il nome di *risposta in frequenza*, per il motivo esposto di seguito, assieme ad un paio di esempi di applicazione di questa proprietà a casi già noti al lettore. Approfondimenti sulle operazioni di filtraggio posso essere trovati al cap. 9, da affrontare dopo lo studio di processi ergodici al § 7.3.

Risposta in frequenza Ponendo in ingresso al sistema un segnale esponenziale complesso $x(t) = e^{j2\pi f_0 t}$, in cui è presente l'unica frequenza f_0 (infatti $X(f) = \delta(f - f_0)$), la proprietà del prodotto per un impulso permette di valutare una uscita $Y(f) = H(f)\delta(f - f_0) = H(f_0)\delta(f - f_0)$, ossia un impulso centrato in f_0 e di area complessa $H(f_0)$, da cui

$$y(t) = H(f_0)e^{j2\pi f_0 t}$$

Quindi, il segnale in ingresso si ripropone in uscita, alterato in modulo e fase in base al valore che $H(f)$ assume alla frequenza f_0 : per questo motivo $H(f)$ è detta *risposta in frequenza* del sistema.

Autovettori e misura di $H(f)$ Ricordando come in algebra lineare, applicando una trasformazione lineare ad un proprio autovettore, si ottiene l'autovettore stesso, moltiplicato per il rispettivo autovalore, osserviamo che la stessa definizione è ora perfettamente applicabile alle funzioni esponenziali complesse $e^{j2\pi f_0 t}$, che risultano essere gli autovettori (o autofunzioni) di un sistema con risposta in frequenza $H(f)$, ed alle quali risulta associato l'autovalore $H(f_0)$. Questa particolarità consente di misurare $H(f)$ alle diverse frequenze, come illustrato a pag. 235, semplicemente osservando l'uscita del sistema, quando in ingresso è posto un segnale sinusoidale.

Sistema passa tutto Poniamo di avere $H(f) = 1$, e che quindi risulti $h(t) = \delta(t)$. In questo caso le componenti di $X(f)$ alle diverse frequenze non subiscono nessuna alterazione, ottenendo

$$y(t) = \mathcal{F}^{-1}\{Y(f)\} = \mathcal{F}^{-1}\{X(f)\} = x(t)$$

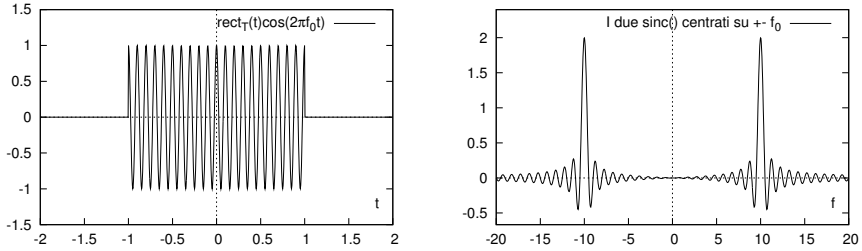


Figura 3.4: Trasformata di un coseno finestrato con $T = 2$, $f_0 = 10$

ed il sistema viene detto di tipo *passa tutto*. Per verifica, scriviamo l'integrale di convoluzione, che risulta $y(t) = \int_{-\infty}^{\infty} x(\tau) \delta(t - \tau) d\tau = x(t)$: ritroviamo quindi la proprietà di setacciamento.

Ritardo Se invece $H(f) = e^{-j2\pi f\theta}$, pari cioè ad un esponenziale complesso, il sistema equivale ad un elemento di ritardo, riproducendo in uscita l'ingresso presentatosi θ istanti prima. Infatti risulta:

$$y(t) = \mathcal{F}^{-1}\{Y(f)\} = \mathcal{F}^{-1}\{X(f)e^{-j2\pi f\theta}\} = x(t - \theta)$$

D'altra parte, scrivendo l'integrale di convoluzione, e ricordando che $h(t) = \mathcal{F}^{-1}\{e^{-j2\pi f\theta}\} = \delta(t - \theta)$, avremmo ottenuto $y(t) = \int_{-\infty}^{\infty} x(\tau) \delta(t - \theta - \tau) d\tau = x(t - \theta)$, ritrovando la proprietà della convoluzione per un impulso traslato. Un sistema siffatto è indicato come *canale perfetto* a pag. 349, in quanto privo di distorsioni lineari (vedi § 14.5).

Sistemi in cascata Ponendo l'uscita di un primo sistema fisico $y(t) = x(t) * h(t)$ in ingresso ad un secondo filtro con risposta impulsiva $g(t)$, si ottiene un risultato $z(t) = y(t) * g(t)$ la cui trasformata di Fourier si può calcolare come $Z(f) = X(f)H(f)G(f)$, dato che (vedi pag. 38) la cascata dei due sistemi fisici è equivalente ad un terzo sistema con risposta impulsiva $h'(t) = h(t) * g(t)$, e questa convoluzione temporale è equivalente al prodotto tra le rispettive risposte in frequenza.

3.6.2 Moltiplicazione nel tempo (*modulazione e finestatura*)

La relazione

$$Z(f) = \mathcal{F}\{x(t)y(t)\} = X(f) * Y(f)$$

ci permette di investigare le conseguenze frequenziali del prodotto temporale di due segnali.

Esempio Prendiamo il caso in cui $z(t) = A \text{rect}_T(t) \cos 2\pi f_0 t$, ovvero pari alla forma d'onda graficata a sinistra della Fig. 3.4. Applicando i risultati noti e la proprietà di traslazione in frequenza, risulta:

$$\begin{aligned} Z(f) &= \frac{A}{2} \mathcal{F}\left\{\text{rect}_T(t) \left(e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}\right)\right\} \\ &= \frac{AT}{2} (\text{sinc}[(f - f_0)T] + \text{sinc}[(f + f_0)T]) \end{aligned}$$

in cui $\mathcal{F}\{rect_T(t)\} = T \text{sinc}(fT)$ si è traslato in $\pm f_0$.

Il risultato dell'esempio, mostrato a destra in fig. 3.4, coincide con quello previsto: l'espressione di $Z(f)$ infatti è anche pari alla convoluzione tra $\mathcal{F}\{rect_T(t)\}$, ed i due impulsi traslati $\mathcal{F}\{\cos 2\pi f_0 t\} = \frac{1}{2}(\delta(f - f_0) - \delta(f + f_0))$: determinando quindi la replica dello spettro del $rect$, traslata alla frequenza del coseno.

Modulazione L'esempio ci permette di motivare il termine *modulazione* associato a questa proprietà. L'ampiezza del coseno risulta infatti *modulata* dal rettangolo. La *modulazione di ampiezza* (AM) dei radio ricevitori casalinghi si riferisce esattamente a questo processo, svolto allo scopo di condividere tra più emittenti la banda prevista per le trasmissioni, assegnando a ciascuna di esse una diversa frequenza portante f_0 su cui trasmettere. Ma questo è l'argomento del cap. 10.

Finestratura Dalla figura 3.4 si può anche arguire come, per T crescente, $Z(f)$ tenda sempre più ad assomigliare ad una coppia di impulsi, ossia al risultato noto per un un coseno di durata *infinita*. Qualora si consideri invece solo un *breve intervallo* di un segnale, il suo spettro si modifica, a seguito della convoluzione in frequenza con la trasformata della *finestra* di analisi. L'operazione di estrazione di una porzione di segnale di durata limitata, a partire da un segnale comunque esteso, è indicata come una operazione di *finestratura* (WINDOWING). In appendice (3.9.3) sono svolte considerazioni relative alla scelta di una finestra rettangolare o con *altro andamento*.

3.7 Derivazione ed integrazione nel tempo

Le ultime due proprietà riguardano un risultato di applicazione meno frequente, ma talvolta utile. Si ottiene infatti che le operazioni di derivata ed integrale di un segnale possono essere realizzate mediante il passaggio dello stesso attraverso un sistema fisico, dato che derivata ed integrale nel tempo sono equivalenti a prodotti in frequenza, e quindi realizzabili come convoluzione del segnale per una appropriata risposta impulsiva.

Derivazione nel tempo E' equivalente a moltiplicare lo spettro per $j2\pi f$:

$$\mathcal{F}\left\{\frac{d}{dt}x(t)\right\} = j2\pi f \cdot X(f)$$

e più in generale $\mathcal{F}\left\{\frac{d^n}{dt^n}x(t)\right\} = (j2\pi f)^n \cdot X(f)$. Per segnali di energia, la dimostrazione è svolta nella nota¹⁴. L'andamento dello spettro originario $X(f)$ risulta *esaltato* alle frequenze più elevate, in quanto il suo modulo è moltiplicato per $2\pi|f|$. La fase, invece, subisce un incremento di $\frac{\pi}{2}$ a tutte le frequenze (il numero immaginario puro $j2\pi f = 2\pi f e^{j\frac{\pi}{2}}$ ha fase $\frac{\pi}{2}$).

Esempio Calcolare $Y(f) = \mathcal{F}\{y(t)\}$ con

$$y(t) = \frac{d}{dt}x(t) \quad \text{e} \quad x(t) = \cos 2\pi f_1 t + \cos 2\pi f_2 t$$

¹⁴ $\mathcal{F}\left\{\frac{dx(t)}{dt}\right\} = \int_{-\infty}^{\infty} \frac{dx(t)}{dt} e^{-j2\pi f t} dt = x(t) e^{-j2\pi f t} \Big|_{-\infty}^{\infty} + j2\pi f \int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt = j2\pi f X(f)$ in quanto $x(t)$ che compare nel primo termine dell'integrale per parti, essendo di energia, tende a zero per $t \rightarrow \infty$.

Valutare quindi $y(t) = \mathcal{F}^{-1}\{Y(f)\}$ con $f_1 = 10$, $f_2 = 100$.

Si ottiene:

$$X(f) = \frac{1}{2} (\delta(f - f_1) + \delta(f + f_1) + \delta(f - f_2) + \delta(f + f_2))$$

Dato che $f \cdot \delta(f - a) = a \cdot \delta(f - a)$, risulta:

$$Y(f) = j2\pi \frac{1}{2} [f_1 (\delta(f - f_1) - \delta(f + f_1)) + f_2 (\delta(f - f_2) - \delta(f + f_2))]$$

Considerando ora che $j2\pi \frac{1}{2} = -\frac{2\pi}{2j}$, si ottiene $y(t) = -2\pi f_1 \sin \omega_1 t - 2\pi f_2 \sin \omega_2 t$ e quindi, per $f_1 = 10$ e $f_2 = 100$, si ha

$$y(t) = -2\pi [10 \sin \omega_1 t + 100 \sin \omega_2 t]$$

Integrazione nel tempo E' equivalente a dividere lo spettro per $j2\pi f$:

$$\mathcal{F} \left\{ \int_{-\infty}^t x(\theta) d\theta \right\} = \frac{X(f)}{j2\pi f}$$

Tale risultato è diretta conseguenza del precedente, in virtù dei legami tra integrale e derivata. Infatti, $\int_{-\infty}^t x(\theta) d\theta$ è una funzione di t , che compare nel limite superiore di integrazione, e la sua derivata è proprio $x(t)$.

In questo caso, le basse frequenze del segnale originario sono esaltate seguendo un andamento $1/2\pi |f|$, mentre la fase subisce una variazione (un ritardo) costante pari a $-\frac{\pi}{2}$. Notiamo come questa circostanza determini una singolarità per $f = 0$ nel caso in cui $X(0) \neq 0$, e come mostrato a pag. 32, ciò corrisponde ad un segnale $x(t)$ che sottende una area non nulla, e quindi $y(t) = \int_{-\infty}^t x(\theta) d\theta$ non si azzerava per $t \rightarrow \infty$: pertanto $y(t)$ non è di energia, e la sua trasformata non è garantita.¹⁵

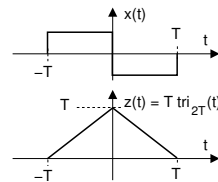
Esempio: Trasformata di un triangolo.

Consideriamo un segnale

$$x(t) = \text{rect}_T \left(t + \frac{T}{2} \right) - \text{rect}_T \left(t - \frac{T}{2} \right)$$

ed il suo integrale

$$z(t) = \int_{-\infty}^t x(\theta) d\theta = T \text{tri}_{2T}(t)$$



entrambi rappresentati nella figura precedente: $z(t)$ è nullo fino a $t < -T$, cresce linearmente fino a $t = 0$, e quindi il contributo all'integrale dato dall'area del *rect* negativo torna ad annullarne il valore.

Per calcolare la \mathcal{F} -trasformata di $z(t)$, calcoliamo prima quella di $x(t)$, e poi applichiamo la proprietà dell'integrazione. Applicando la proprietà di traslazione nel tempo, scriviamo

$$\begin{aligned} X(f) &= T \cdot \text{sinc}(fT) \cdot e^{+j2\pi f \frac{T}{2}} - T \cdot \text{sinc}(fT) \cdot e^{-j2\pi f \frac{T}{2}} = \\ &= T \cdot \frac{\sin(\pi f T)}{\pi f T} \cdot 2j \sin \pi f T = j2T \frac{\sin^2(\pi f T)}{\pi f T} \end{aligned}$$

Dividendo quindi per $j2\pi f$ si ottiene

$$Z(f) = \frac{j2T}{j2\pi f} \frac{\sin^2(\pi f T)}{\pi f T} \frac{T}{T} = \left(T \frac{\sin(\pi f T)}{\pi f T} \right)^2 = (T \text{sinc}(fT))^2$$

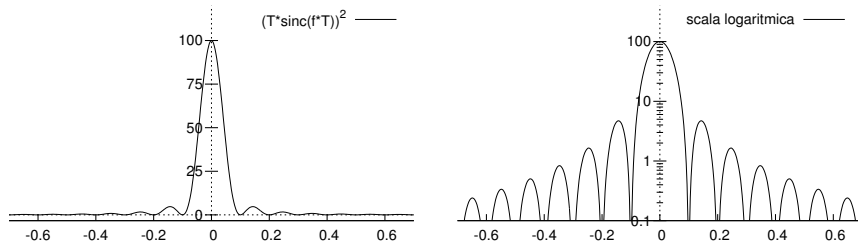


Figura 3.5: Andamento di $(T \operatorname{sinc}(fT))^2$ in scala lineare e logaritmica; $T = 10$.

il cui andamento è mostrato in figura 3.5. Da questo risultato ne consegue infine che $\mathcal{F}\{tri_{2T}(t)\} = T \operatorname{sinc}^2(fT)$, come riportato al § 3.9.6.

Densità di energia di $\operatorname{rect}_T(t)$ Lo stesso risultato mostrato nell'esempio, può essere ottenuto per altra via, notando che il triangolo è il risultato della convoluzione di due rettangoli:

$$z(t) = T \cdot tri_{2T}(t) = \operatorname{rect}_T(t) * \operatorname{rect}_T(t) \quad (3.8)$$

Come verifica, si ripercorra la costruzione grafica riportata alla sezione 3.5.3. E' quindi ora sufficiente applicare la proprietà del prodotto in frequenza, per ottenere:

$$Z(f) = \mathcal{F}\{T \cdot tri_{2T}(t)\} = [\mathcal{F}\{\operatorname{rect}_T(t)\}]^2 = [T \operatorname{sinc}(fT)]^2 \quad (3.9)$$

Il risultato fornito da (3.9), è anche pari alla densità di energia $\mathcal{E}_y(f)$ di un segnale rettangolare $y(t) = \operatorname{rect}_T(t)$: infatti per il teorema di Parseval, si ha $\mathcal{E}_y(f) = Y(f)Y^*(f)$, in cui $Y(f) = \mathcal{F}\{\operatorname{rect}_T(t)\} = T \operatorname{sinc}(fT)$, e pertanto

$$\mathcal{E}_y(f) = [T \operatorname{sinc}(fT)]^2$$

3.8 Trasformata di segnali periodici

Presentiamo ora un diverso modo di ottenere lo spettro di un segnale periodico, che in sostanza fornisce gli stessi risultati previsti dalla serie di Fourier, seguendo però un metodo diverso, che si basa sulla definizione di una particolare forma d'onda (ideale), nota come

3.8.1 Treno di impulsi

E' costituito da una serie infinita di impulsi matematici distanziati di un periodo T , si esprime analiticamente come

$$\pi_T(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT)$$

e si rivelerà di utilizzo frequente nei contesti del campionamento e delle trasmissioni numeriche.

¹⁵In realtà si può dimostrare che anche nel caso in cui $X(0) \neq 0$, è comunque possibile giungere ad un risultato, qualora si ricorra all'impulso matematico, ottenendo $\mathcal{F}\left\{\int_{-\infty}^t x(\theta) d\theta\right\} = \frac{X(f)}{j2\pi f} + \frac{\delta(f)}{2} X(0)$.

3.8.2 Segnale periodico

Consideriamo un segnale periodico di periodo T espresso come

$$x(t) = \sum_{m=-\infty}^{\infty} g(t - mT) \quad (3.10)$$

di cui $g(t)$ costituisce un periodo: la concatenazione di infinite repliche di $g(t)$, spaziate di un periodo T l'una dall'altra, riproduce il segnale periodico originario. Sfruttando la proprietà di convoluzione con l'impulso traslato, la stessa somma può essere scritta come

$$x(t) = \sum_{m=-\infty}^{\infty} g(t) * \delta(t - mT) = g(t) * \sum_{m=-\infty}^{\infty} \delta(t - mT) = g(t) * \pi_T(t)$$

dove nel secondo passaggio si è sfruttata la linearità della convoluzione. Ricordando ora la proprietà della moltiplicazione in frequenza, troviamo $X(f) = G(f) \cdot \mathcal{F}\{\pi_T(t)\}$; ci accingiamo allora a determinare $\mathcal{F}\{\pi_T(t)\}$, ossia la ...

3.8.3 Trasformata del treno di impulsi

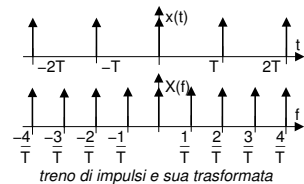
L'approccio che conviene seguire è di pensare a $\pi_T(t)$ come ad un segnale periodico, e svilupparlo in serie di Fourier. I coefficienti si calcolano allora come:

$$\begin{aligned} \Pi_n &= \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \left[\sum_{m=-\infty}^{\infty} \delta(t - mT) \right] e^{-j2\pi n F t} dt \\ &= \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \delta(t) e^{-j2\pi n F t} dt = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} 1 \cdot \delta(t) dt = \frac{1}{T} \end{aligned}$$

in quanto, tra tutti gli impulsi della sommatoria, ne resta solo uno, quello centrato in zero, dato che gli altri sono tutti esterni ai limiti di integrazione, mentre la seconda uguaglianza tiene conto della (3.4); pertanto, tutti i coefficienti risultano avere lo stesso valore, pari ad $\frac{1}{T}$, e possiamo dunque scrivere

$$\mathcal{F}\{\pi_T(t)\} = \mathcal{F}\left\{ \sum_{n=-\infty}^{\infty} \Pi_n e^{j2\pi n F t} \right\} = \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) = \frac{1}{T} \pi_{\frac{1}{T}}(f)$$

ottenendo il risultato cercato: $\mathcal{F}\{\pi_T(t)\} = \frac{1}{T} \pi_{\frac{1}{T}}(f)$. Quindi, la trasformata di un treno di impulsi è a sua volta un treno di impulsi, di periodo inverso a quello originario.



3.8.4 Trasformata di segnale periodico

Siamo finalmente in grado di esprimere la trasformata di un segnale periodico come il prodotto tra la \mathcal{F} -trasformata di un suo periodo ed un treno di impulsi in frequenza:

$$X(f) = G(f) \cdot \frac{1}{T} \pi_{\frac{1}{T}}(f) \quad (3.11)$$

Esempio Riprendendo in considerazione il caso dell'onda quadra affrontato al § 2.2.1.4, non è difficile riconoscere come, ponendo $g(t) = \text{Arect}_\tau(t)$, e corrispondentemente $G(f) = \text{Arsinc}(f\tau)$, il prodotto di $G(f)$ per il treno di impulsi $\frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(f - nF)$ (con $F = \frac{1}{T}$) fornisce il risultato già incontrato:

$$X(f) = A \frac{\tau}{T} \sum_{n=-\infty}^{\infty} \text{sinc}(nF\tau) \delta(f - nF)$$

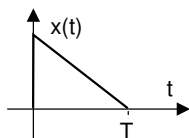
Somma di Poisson Il risultato ottenuto ci può aiutare a mettere in luce una uguaglianza indicata con il nome di *somma di Poisson*¹⁶, definita come la possibilità di esprimere un somma infinita realizzata sulla base di una funzione nel tempo, nei termini di una somma infinita realizzata sulla base di una funzione della frequenza, che è la trasformata di quella nel tempo. Nel caso in esame, riscriviamo la (3.11) come $X(f) = \sum_n \frac{1}{T} G\left(\frac{n}{T}\right) \delta\left(f - \frac{n}{T}\right)$; antitrasformando entrambi i membri si ottiene

$$x(t) = \sum_{m=-\infty}^{\infty} g(t - mT) = \sum_{n=-\infty}^{\infty} \frac{1}{T} G\left(\frac{n}{T}\right) e^{j2\pi \frac{n}{T} t}$$

che riconosciamo corrispondere all'espansione in serie di Fourier del segnale periodico $x(t)$, non appena constatato come i termini $\frac{1}{T} G\left(\frac{n}{T}\right)$ altro non siano che i suoi coefficienti Fourier, come d'altra parte risulta anche dalla (3.1).

3.9 Appendici

3.9.1 Esercizio: quanti modi di calcolare la $\mathcal{F}\{x(t)\}$?



Sia dato il segnale

$$x(t) = \begin{cases} 1 - \frac{t}{T} & \text{con } 0 \leq t \leq T \\ 0 & \text{altrimenti} \end{cases}$$

mostrato in figura. Descrivere quanti più modi possibili di calcolarne lo spettro di densità di energia $\mathcal{E}_x(f)$.

1. Si calcola $X(f) = \mathcal{F}\{x(t)\} = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt$ e quindi $\mathcal{E}_x(f) = |X(f)|^2$;
2. Anticipando un risultato del § 9.2.1, è possibile calcolare $\mathcal{R}_x(f) = \int_{-\infty}^{\infty} x(t) x(t + \tau) dt$, e quindi $\mathcal{E}_x(f) = \mathcal{F}\{\mathcal{R}_x(f)\}$;
3. Notando che $x(t) = y(t) \cdot z(t)$ con $y(t) = \text{tri}_{2T}(t)$ e $z(t) = \text{rect}_T(t - \frac{T}{2})$, possiamo scrivere $X(f) = Y(f) * Z(f)$, e quindi procedere come in 1);
4. Notiamo che la derivata¹⁷ di $x(t)$ vale $g(t) = \frac{d}{dt} x(t) = \delta(t) - \frac{1}{T} \text{rect}_T(t - \frac{T}{2})$; questo ci permette di calcolare $G(f)$ come $G(f) = \mathcal{F}\{g(t)\} = 1 - \text{sinc}(fT)$. Otteniamo quindi $X(f) = \frac{G(f)}{j2\pi f}$, e procediamo come in 1).

¹⁶Per un approfondimento si veda ad es. http://it.wikipedia.org/wiki/Formula_di_sommazione_di_Poisson.

¹⁷La derivata di una discontinuità di prima specie è pari ad un impulso matematico, di area uguale all'altezza della discontinuità. Infatti l'integrale dell'impulso $\int_{-\infty}^t \delta(\theta) d\theta$ è proprio un gradino. Questa considerazione consente di risolvere in modo semplice le trasformate di segnali in cui è presente una discontinuità.

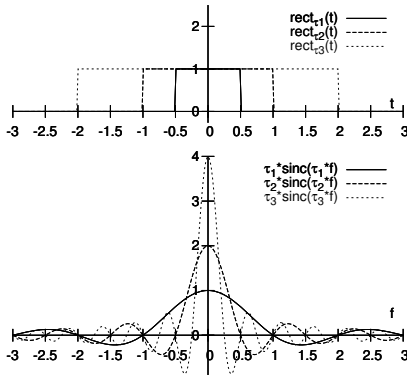
3.9.2 Sulla trasformata di una costante

Svolgiamo alcune considerazioni relative al risultato mostrato a pag. 35, illustrando come l'impulso $\delta(\cdot)$ permetta di rappresentare particolari situazioni. Consideriamo pertanto il segnale costante $x(t) = A$, che trattiamo come un segnale periodico con periodo T tendente ad ∞ ¹⁸, ed esprimiamo $x(t)$ nei termini dei coefficienti di Fourier.

L'integrale $X_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} A e^{-j2\pi n F t} dt$ per $T \rightarrow \infty$ fornisce zero per tutti gli n tranne

$$\text{che per } n = 0, \text{ e quindi si ottiene } X_n = \begin{cases} A & \text{con } n = 0 \\ 0 & \text{con } n \neq 0 \end{cases}.$$

In alternativa, pensiamo la costante come il limite a cui tende un'onda quadra con duty-cycle $\frac{\tau}{T}$ al tendere di τ a T : lo spettro di ampiezza è stato calcolato al capitolo 2, pag. 20, presenta righe alle armoniche $f = \frac{n}{T}$, mentre gli X_n con andamento sinc($nF\tau$) si azzerano alle frequenze $f = \frac{n}{\tau}$. Se $\tau \rightarrow T$, gli zeri annullano tutte le armoniche tranne X_0 , il cui valore $A\frac{\tau}{T}$ tende ora ad A .



Qualora invece si desideri calcolare la *trasformata* di Fourier anziché la *serie*, applicando la definizione $X(f) = \int_{-\infty}^{\infty} A e^{-j2\pi f t} dt$ si ottiene $X(f) = 0$ ovunque, tranne che in $f = 0$ dove $X(0) = \infty$. Convienne allora ricorrere ad un'operazione di passaggio al limite, e pensare il segnale $x(t) = A$ come il risultato dell'allargamento progressivo di un $rect_{\tau}(t)$, cioè come $x(t) = \lim_{\tau \rightarrow \infty} A rect_{\tau}(t)$. La figura a lato mostra come, considerando valori τ_i via via più grandi, si ottenga una trasformata $X_{\tau_i}(f) = A\tau_i \text{sinc}(f\tau_i)$ sempre più *alta e stretta*.

Notiamo ora che l'energia di $x_{\tau}(t) = A rect_{\tau}(t)$ vale $\mathcal{E}_{x_{\tau}} = \int_{-\infty}^{\infty} |x_{\tau}(t)|^2 dt = A^2\tau$; per il teorema di Parseval, l'energia coincide nei domini di tempo e frequenza, e quindi risulta

$$\mathcal{E}_{x_{\tau}} = \int_{-\infty}^{\infty} |X_{\tau}(f)|^2 df = A^2\tau$$

Al tendere di τ ad ∞ , l'energia diviene infinita, mentre la potenza vale

$$\mathcal{P}_x = \lim_{\tau \rightarrow \infty} \frac{\mathcal{E}_{x_{\tau}}}{\tau} = \lim_{\tau \rightarrow \infty} \int_{-\infty}^{\infty} \frac{|X_{\tau}(f)|^2}{\tau} df = A^2$$

L'espressione $\lim_{\tau \rightarrow \infty} \frac{|X_{\tau}(f)|^2}{\tau}$ rappresenta dunque¹⁹ lo *spettro di densità di potenza* $\mathcal{P}_x(f)$ della costante A , che finalmente scriviamo come $\mathcal{P}_x(f) = A^2\delta(f)$, in cui $\delta(f)$ è la funzione *impulso matematico* introdotta in 3.4. In tal modo infatti, è facile verificare che risulta $\mathcal{P}_x = \int_{-\infty}^{\infty} A^2\delta(f) df = A^2$, e $\mathcal{P}_x(f) = \begin{cases} \infty & \text{con } f = 0 \\ 0 & \text{altrove} \end{cases}$. In altre parole, il formalismo dell'impulso matematico rende possibile trattare questo caso, dove la potenza (finita) è tutta concentrata ad una unica frequenza ($f = 0$) dando luogo ad una densità infinita.

¹⁸In effetti, un segnale costante è un segnale periodico, con periodo T qualsiasi.

¹⁹vedi anche §9.3.1 a pag. 225

3.9.3 Finestratura e stima spettrale

Applichiamo ora la teoria svolta al § 3.6.2 per speculare sul problema dell'analisi spettrale effettuata a partire da un solo segmento temporale del segnale.

Il calcolo dello spettro di $y(t) = x(t)w(t)$ fornisce, come noto, il valore $Y(f) = X(f) * W(f)$. Quindi, il vero spettro $X(f)$ di $x(t)$ non può essere conosciuto, se non tramite l'effetto della convoluzione con quello $W(f)$ della funzione finestra $w(t)$. Già a pagina 41 si è fatto notare come, se $x(t) = A \cos 2\pi f_0 t$ e $w(t) = \text{rect}_T(t)$, si ottiene che $W(f) = T \text{sinc}(fT)$, pertanto

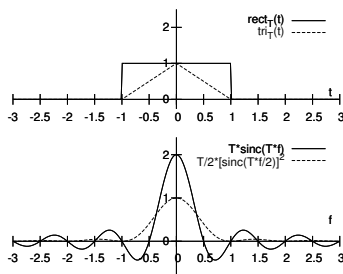
$$\mathcal{F}\{x(t) \cdot w(t)\} = \frac{AT}{2} (\text{sinc}[(f - f_0)T] + \text{sinc}[(f + f_0)T])$$

che è tanto più diverso dalle due linee spettrali del coseno (vedi Fig. 3.4), quanto più è piccolo T .

Valutiamo ora gli effetti derivanti dall'uso di una funzione finestra diversa da quella rettangolare. Se ad esempio si sceglie di adottare una finestra triangolare di eguale durata T , il risultato mostrato a pag. 43 dalle (3.8) e (3.9) permette di ottenere

$$W(f) = \mathcal{F}\{w(t) = \text{tri}_T(t)\} = \frac{T}{2} \left[\text{sinc}\left(\frac{fT}{2}\right) \right]^2$$

Come può essere verificato dalla figura a fianco, la finestra triangolare esibisce un andamento nel tempo *più dolce* rispetto al $\text{rect}(t)$, e ciò si riflette in una maggiore concentrazione della sua trasformata alle frequenze più basse. Infatti, $W(f)$ ha ora un *lobo principale* di estensione doppia (il primo zero si trova ad $f = \frac{2}{T}$ anziché ad $\frac{1}{T}$ come per il rect), e le *code laterali* decrescono *più rapidamente*, andando a zero come $\frac{1}{f^2}$, mentre l'ampiezza risulta dimezzata.



L'andamento del *lobo principale* e delle *code* di $W(f)$ si riflette nell'andamento della trasformata qualora il segnale originario contenga, ad esempio, più di una frequenza: per la linearità della trasformata, il risultato sarà la replica di $W(f)$ centrata alle frequenze presenti.

La Fig. 3.6 mostra la trasformata di una finestra rettangolare di base $\tau = 2, 0.5,$ e 0.25 secondi, contenente la somma di due cosinusoidi di frequenza $f_0 = 10$ e $f_1 = 15$ Hz. Osserviamo come al diminuire del prodotto $(f_1 - f_0) \cdot \tau$, le due trasformate $W(f)$ *interagiscono*, fino ad esibire un andamento complessivo in cui non è più possibile distinguere la presenza di due diversi toni. Il fenomeno illustrato avviene tanto prima, quanto più il lobo principale di $W(f)$ è esteso; pertanto, l'uso di una finestra triangolare peggiora la situazione. D'altro canto, se la frequenza dei due coseni è sufficientemente diversa, la rapidità di azzeramento delle code di una finestra triangolare consente di ottenere una stima spettrale più vicina alla reale composizione del segnale, evitando di mostrare *artefatti* conseguenza dalle lunghe code presenti nel caso di finestra rettangolare.

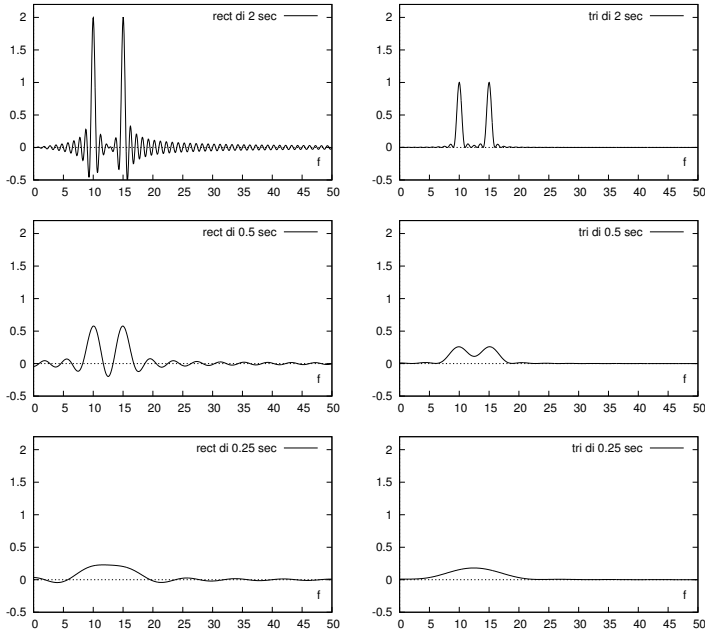


Figura 3.6: Trasformata di due toni a 10 e 15 Hz, con finestra temporale $rect_T$ e tri_T di durata 2, .5 e .25 secondi

Considerazioni analoghe a quelle ora svolte, possono essere intraprese per diverse scelte²⁰ di funzione finestra, in dipendenza dal particolare obiettivo della stima spettrale.

3.9.4 Trasformata di un gradino

Definiamo la funzione gradino come $g(t) = \begin{cases} 1 & \text{per } t > 0 \\ \frac{1}{2} & \text{per } t = 0 \\ 0 & \text{per } t < 0 \end{cases}$ che, fornendo $\int_{-\infty}^{\infty} |g(t)| dt =$

∞ , non dovrebbe avere una trasformata. Come già visto per la costante, il calcolo può essere condotto a termine come limite a cui tende la trasformata di una diversa funzione, il cui limite tende al gradino. Scegliamo quindi $g_\alpha(t) = e^{-\alpha t}$ per $t > 0$, per la quale risulta $\lim_{\alpha \rightarrow 0} g_\alpha(t) = g(t)$, e troviamo

$$G_\alpha(f) = \int_0^\infty e^{-\alpha t} e^{-j2\pi f t} dt = \frac{e^{-(\alpha + j2\pi f)t}}{-(\alpha + j2\pi f)} \Big|_0^\infty = \frac{1}{\alpha + j2\pi f} = \frac{\alpha - j2\pi f}{\alpha^2 + (2\pi f)^2}$$

²⁰Nel tempo sono state definite un elevato numero di finestre temporali, ognuna migliore sotto certi aspetti, e peggiore sotto altri. Consultando Wikipedia http://en.wikipedia.org/wiki/Window_function, possiamo elencare le finestre di *Hamming*, *Hann*, *Cosine*, *Lanczos*, *Bartlett*, *Gauss*, *Blackman*, *Kaiser*, *Nuttall*, *Bessel*, *Dolph-Chebyshev*, *Exponential*, *Tukey*...

Si può mostrare che

$$\lim_{\alpha \rightarrow 0} \Re \{G_\alpha(f)\} = \lim_{\alpha \rightarrow 0} \frac{\alpha}{\alpha^2 + (2\pi f)^2} = \frac{1}{2} \delta(f)$$

mentre risulta

$$\lim_{\alpha \rightarrow 0} \Im \{G_\alpha(f)\} = \lim_{\alpha \rightarrow 0} \frac{-j2\pi f}{\alpha^2 + (2\pi f)^2} = \frac{1}{j2\pi f}$$

ottenendo²¹ in definitiva

$$G(f) = \mathcal{F}\{g(t)\} = \frac{1}{2} \left(\delta(f) - \frac{j}{\pi f} \right)$$

3.9.5 Sintesi delle proprietà della trasformata di Fourier

Proprietà	$z(t)$	$Z(f) = \mathcal{F}\{z(t)\}$
Linearità	$ax(t) + by(t)$	$aX(f) + bY(f)$
Coniugato	$x^*(t)$	$X^*(-f)$
Cambiamento di scala	$x(at)$	$\frac{1}{ a } X\left(\frac{f}{a}\right)$
Ritardo	$x(t - T)$	$X(f) e^{-j2\pi f T}$
Traslazione in frequenza	$x(t) e^{j2\pi f_0 t}$	$X(f - f_0)$
Modulazione di ampiezza	$x(t) \cos 2\pi f t$	$\frac{1}{2} X(f - f_0) + \frac{1}{2} X(f + f_0)$
Prodotto in frequenza	$\int_{-\infty}^{\infty} x(\tau) y(t - \tau) d\tau$	$X(f) Y(f)$
Prodotto nel tempo	$x(t) y(t)$	$\int_{-\infty}^{\infty} X(\sigma) Y(f - \sigma) d\sigma$
Dualità	$X(t)$	$x(-f)$
Simmetria coniugata	$x(t)$ reale	$X(f) = X^*(-f)$
Derivazione	$\frac{d}{dt} x(t)$	$j2\pi f \cdot X(f)$
Integrazione	$\int_{-\infty}^t x(\theta) d\theta$	$\frac{X(f)}{j2\pi f}$

3.9.6 Trasformate di segnali

$x(t)$	$X(f)$	\mathcal{P}/\mathcal{E}	$\mathcal{P}(f)/\mathcal{E}(f)$	Pot/En
$\cos(2\pi f_0 t + \varphi)$	$\frac{1}{2} e^{j\varphi} \delta(f - f_0) + \frac{1}{2} e^{-j\varphi} \delta(f + f_0)$	$\frac{1}{2}$	$\frac{1}{4} \delta(f - f_0) + \frac{1}{4} \delta(f + f_0)$	P
A	$A \cdot \delta(f)$	A^2	$A^2 \cdot \delta(f)$	P
$A \cdot \text{rect}_\tau(t)$	$A \cdot \tau \text{sinc}(f\tau)$	$A^2 \cdot \tau$	$A^2 \cdot \tau^2 \text{sinc}^2(f\tau)$	E
$A \cdot \text{tri}_{2\tau}(t)$	$A \cdot \tau \text{sinc}^2(f\tau)$	$A^2 \cdot \frac{2}{3} \cdot \tau$	$A^2 \cdot \tau^2 \text{sinc}^4(f\tau)$	E
$e^{-\beta t}, t \geq 0$	$\frac{1}{\beta + j2\pi f}$		$\frac{1}{\beta^2 + 4(\pi f)^2}$	E
$e^{-\beta t }$	$\frac{2\beta}{\beta^2 + 4(\pi f)^2}$		$\frac{4\beta^2}{\beta^4 + 8(\pi\beta f)^2 + 16(\pi f)^4}$	E
$e^{-\alpha(\beta t)^2}$	$\frac{1}{\beta} e^{-\alpha \left(\frac{f}{\beta}\right)^2}$		$\frac{1}{\beta^2} e^{-2\alpha \left(\frac{f}{\beta}\right)^2}$	E

²¹In realtà per $f = 0$ si ottiene che $\Im \{G_\alpha(f)\} = 0$, e lo stesso vale per $G(f)$, ossia $\Im \{G(f=0)\} = 0$.

Capitolo 4

Campionamento, quantizzazione ed elaborazione numerica

Sono esposti i metodi che consentono di rappresentare un segnale per mezzo dei suoi *campioni*¹ temporali, permettendone la rappresentazione in forma numerica: i campioni infatti, sono una sequenza di *numeri*² e quindi dei *bit*. Le operazioni descritte ai precedenti capitoli che si svolgono sul segnale $x(t)$, possono quindi essere eseguite direttamente sulla sua rappresentazione numerica, dando luogo alle tecniche indicate come *elaborazione numerica* dei segnali, di cui anche si fornisce qualche accenno, relativo alle rappresentazione frequenziale di sequenze, ed al suo impiego nelle operazioni di filtraggio.

4.1 Teorema del campionamento

Esprime la possibilità di ricostruire un segnale limitato in banda a partire dai suoi campioni:

Un segnale con spettro nullo a frequenze maggiori di W è completamente descritto dai suoi valori osservati ad intervalli temporali regolari $t_n = nT_c$, con n intero e periodo di campionamento $T_c \leq \frac{1}{2W}$; da questi è quindi possibile risalire ai suoi valori per qualunque altro istante.

La frequenza $f_{c_{min}} = \frac{1}{T_{c_{Max}}} = 2W$, chiamata *frequenza di Nyquist*³, corrisponde alla *minima* frequenza con cui occorre campionare un segnale $x(t)$ limitato in banda, ed è pari al doppio della massima frequenza presente. Se questa condizione è rispettata, il segnale originario può essere ricostruito ricorrendo ad una formula di interpolazione⁴ (detta *cardinale*⁵) che esprimiamo come:

¹Il termine *campione* rappresenta il valore di un segnale ad un determinato istante, e considerato come sinonimo di esemplare, o esempio, ovvero *sample* in inglese; da non confondere con *champion*, o primatista!

²*Digits* in inglese, da cui il termine *digitale* come sinonimo di *numerico*.

³In realtà questo teorema è stato derivato indipendentemente, in tempi diversi, anche da altri scienziati, come *Whittaker*, *Kotelnikov* e *Shannon*.

⁴Con *interpolazione* si intende la stima dei valori del segnale, per gli istanti compresi tra due campioni noti.

⁵In realtà la formula (4.1) non è l'unica possibile, come vedremo al § 4.1.2

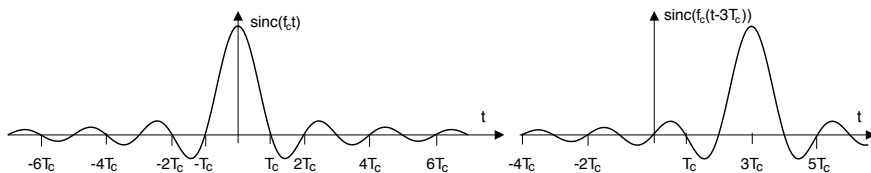


Figura 4.1: La funzione $\text{sinc}(f_c t)$ centrata in $t = 0$ e traslata in $t = 3T_c$

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT_c) \cdot \text{sinc}(f_c(t - nT_c)) \quad (4.1)$$

in cui la funzione $\text{sinc}(f_c t) = \frac{\sin \pi f_c t}{\pi f_c t}$, detta per questo motivo *seno cardinale* ed introdotta a pag. 21, svolge appunto il ruolo di funzione interpolatrice, ed è mostrata in Fig. 4.1, assieme ad una sua replica traslata.

Per dimostrare la validità di (4.1) ed approfondirne il significato, adottiamo lo schema simbolico mostrato nella parte sinistra di fig. 4.2, che (come mostreremo) realizza le operazioni della formula di ricostruzione. Innanzitutto, viene generato un segnale $x^\bullet(t)$ costituito da impulsi (vedi § 3.8.1) con area pari ai campioni di segnale, ossia $x^\bullet(t) = \sum_{n=-\infty}^{\infty} x(nT_c) \delta(t - nT_c)$, mediante moltiplicazione di $x(t)$ (limitato in banda tra $-W$ e W) per un treno di impulsi con periodo $T_c \leq \frac{1}{2W}$: questa operazione corrisponde a porci nelle esatte condizioni del teorema del campionamento.

Ricordando ora il risultato ottenuto al § 3.8.3 per la trasformata di un treno di impulsi, possiamo ottenere lo spettro di ampiezza di $X^\bullet(f) = \mathcal{F}\{x^\bullet(t)\}$, mostrato in basso a destra di fig. 4.2, come

$$\begin{aligned} X^\bullet(f) &= \mathcal{F}\{x(t) \cdot \pi T_c \delta(t)\} = X(f) * \frac{1}{T_c} \Pi_{\frac{1}{T_c}}(f) = X(f) * \frac{1}{T_c} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_c}\right) = \\ &= f_c \cdot \sum_{n=-\infty}^{\infty} X(f) * \delta(f - n f_c) = f_c \cdot \sum_{n=-\infty}^{\infty} X(f - n f_c) \end{aligned} \quad (4.2)$$

dove il penultimo passaggio scambia l'integrale (di convoluzione) di una somma con una somma di integrali, e l'ultimo passaggio tiene conto della proprietà di convoluzione con un impulso: $X^\bullet(f)$ è quindi costituito da infinite repliche di $X(f)$, centrate a multipli della frequenza di campionamento $f_c = \frac{1}{T_c}$. Ma il filtro passa-basso ideale $H(f)$ (chiamato anche con il nome di *filtro di restituzione*) limita la banda di $X^\bullet(f)$ al solo intervallo di frequenze $(-\frac{f_c}{2}, \frac{f_c}{2})$, e dunque lascia passare solo una delle repliche spettrali, ovvero produce un segnale $y(t)$ con spettro di ampiezza $Y(f) = H(f) X^\bullet(f) = T_c f_c X(f) = X(f)$, che quindi è perfettamente equivalente ad $x(t)$ originario, ricostruito sulla base dei suoi soli campioni $x(nT_c)$.

L'ultimo passo di questa dimostrazione evidenzia come lo schema di fig. 4.2 in effetti esegua proprio l'operazione definita dalla formula di ricostruzione (4.1). Infatti, $y(t)$ è il risultato della convoluzione tra $x^\bullet(t)$ e $h(t) = \mathcal{F}^{-1}\{H(f)\} = \mathcal{F}^{-1}\{T_c \text{rect}_{f_c}(f)\} = \text{sinc}(f_c t)$, e dunque ogni impulso di cui è composto $x^\bullet(t)$, quando posto in ingresso al filtro di restituzione, ne trasla la risposta impulsiva $h(t) = \text{sinc}(f_c t)$ all'istante nT_c in

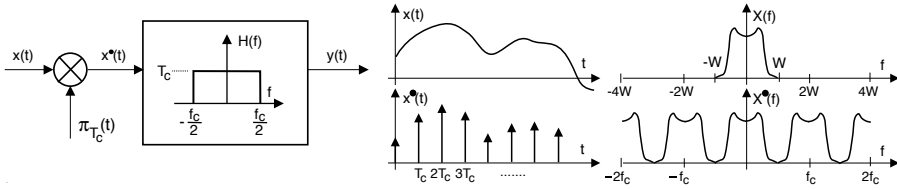
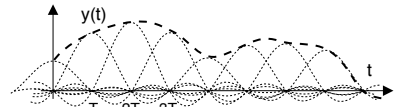


Figura 4.2: Circuito di campionamento e restituzione; segnali coinvolti e relativi spettri

cui è centrato l'impulso. In formule, appunto:

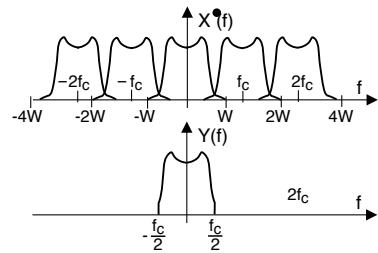
$$\begin{aligned}
 y(t) &= \left[x(t) \cdot \sum_n \delta(t - nT_c) \right] * h(t) = \sum_n x(nT_c) \delta(t - nT_c) * \text{sinc}(f_c t) = \\
 &= \sum_n x(nT_c) \text{sinc}(f_c(t - nT_c))
 \end{aligned}$$

L'effetto del risultato ottenuto è rappresentato nella figura a fianco, che esemplifica come il teorema del campionamento definisca essenzialmente una formula di *interpolazione*: i valori del segnale ricostruito hanno l'esatto valore dei campioni di segnale negli istanti di campionamento, mentre negli istanti intermedi il valore si forma come *somma* di tutte le "code" dei *sinc* adiacenti.

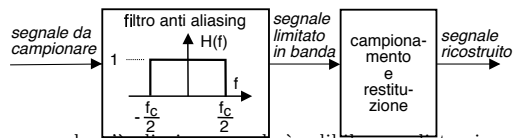


4.1.1 Aliasing

Questo termine ha origine dalla parola inglese⁶ *alias* (copia, clone) e sta ad indicare il fenomeno che si produce nell'applicare il teorema del campionamento quando i requisiti non sono soddisfatti, e cioè quando la frequenza di campionamento è inferiore alla frequenza di Nyquist, ossia $f_c = \frac{1}{T_c} < 2W$ (ovvero $T_c > \frac{1}{2W}$). In questo caso la (4.2) indica come le repliche spettrali che compongono $X^*(f)$ siano più ravvicinate, e si sovrappongono, come rappresentato dalla figura a lato: l'aliasing è infatti indicato anche come *fold-over*, o *ripiegamento*. Quando questo avviene, il filtro passa-basso di restituzione non è più in grado di estrarre la replica centrata in $f = 0$, e dunque alla sua uscita è presente un segnale $y(t)$ che si differenzia da $x(t)$ in particolar modo per i contenuti energetici nella regione delle frequenze più elevate⁷.



Il fenomeno dell'aliasing può insorgere, oltre che nel caso in cui si commetta il banale errore di adottare $f_c < 2W$, anche a causa di una imperfetta limitazione in banda del segnale da campionare, che infatti viene sempre preventivamente filtrato, in modo di assicurarsi che



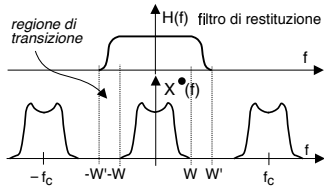
⁶In realtà *alias* è di origine latina !!!

⁷In un segnale audio, ad esempio, ci si accorge che c'è aliasing quando si udivile una distorsione (rumore) congiuntamente ai passaggi con maggior contenuto di alte frequenze.

non contenga componenti a frequenze maggiori della metà di quella di campionamento.

4.1.2 Generalizzazione del filtro di restituzione

Altri problemi possono essere causati dal filtro di restituzione $H(f)$, che difficilmente si riesce a realizzare ideale, e che invece presenta una *regione di transizione* tra banda passante e banda soppressa, di estensione non nulla.



In questo caso occorre *sovracampionare* ad una frequenza $f_c = 2W' > 2W$, in modo che le repliche spettrali siano più distanziate tra loro, e quindi il filtro di ricostruzione possa ancora isolare l'unica replica di $X^\bullet(f)$ in banda base, come si osserva in figura. Notiamo ora che ciò avviene anche se ora la risposta impulsiva del filtro di restituzione non presenta più l'andamento di un $\text{sinc}(f_c t)$, e dunque la formula di interpolazione non ha più l'espressione *cardinale* fornita nella 4.1: ciò significa che l'operazione di campionamento e restituzione può essere realizzata in forme anche molto diverse tra loro.

4.1.3 Ortogonalità delle funzioni sinc

Si può dimostrare⁸ che le funzioni *sinc* costituiscono una base di rappresentazione ortogonale, in quanto

$$\int_{-\infty}^{\infty} \text{sinc}(f_c(t - kT_c)) \text{sinc}(f_c(t - hT_c)) dt = \begin{cases} 0 & \text{se } h \neq k \\ T_c & \text{se } h = k \end{cases} \quad (4.3)$$

Pertanto, il valore dell'energia di un segnale limitato in banda è calcolabile a partire dai suoi campioni, e vale:

$$\begin{aligned} \mathcal{E}_x &= \int_{-\infty}^{\infty} x(t) x^*(t) dt = \sum_k \sum_h x_k x_h^* \int_{-\infty}^{\infty} \text{sinc}(f_c(t - kT_c)) \text{sinc}(f_c(t - hT_c)) dt \\ &= \sum_k \sum_h x_k x_h^* T_c \delta(h, k) = T_c \sum_k |x_k|^2 \end{aligned}$$

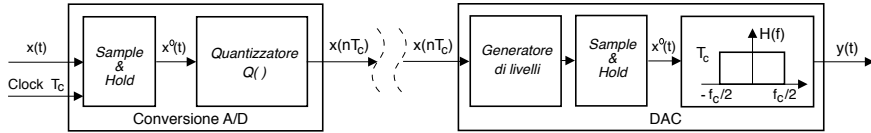
4.1.4 Uso pratico

Lo schema proposto in Fig. 4.2 aveva il solo scopo di visualizzare gli aspetti teorico-matematici del teorema del campionamento.

⁸Applicando il teorema di Parseval (§ 3.2) e la proprietà di traslazione temporale, la (4.3) può essere riscritta come

$$\int_{-\infty}^{\infty} T_c \text{rect}_{f_c}(f) e^{-j2\pi f k T_c} T_c \text{rect}_{f_c}(f) e^{+j2\pi f h T_c} df = (T_c)^2 \int_{-f_c/2}^{f_c/2} e^{-j2\pi f \frac{k-h}{2W}} df$$

in cui l'esponenziale complesso sotto integrale compie un numero *intero* di oscillazioni a media nulla per $f \in [-1/2T_c, 1/2T_c]$ se $k \neq h$, e dunque in tal caso l'integrale è nullo; al contrario, l'esponenziale vale 1 se $k = h$, ed il suo integrale vale f_c , determinando così il risultato mostrato.



Campionamento e restituzione nel mondo reale

Per coglierne i lati pratici, innanzitutto consideriamo che *non viene generato il segnale $X^\bullet(f)$* , dato che i campioni di $x(t)$ sono in realtà ottenuti mediante un circuito *Sample and Hold* (s&H, ovvero *campiona e mantieni*) che produce una uscita costante (per un tempo $\tau \leq T_c$), pari al valore assunto dall'ingresso negli istanti di campionamento $t = nT_c$; l'uscita del s&H viene quindi *quantizzata* (vedi il § 4.2), ovvero misurata e convertita in un valore numerico, dal dispositivo $Q(\cdot)$.

I valori così ottenuti possono essere memorizzati, oppure trasmessi. Per ricostruire il segnale originario si adotta un DAC (*Digital to Analog Converter*, ossia Convertitore Digitale-Analogico) che può essere idealizzato nei tre componenti mostrati in figura, ossia un dispositivo che per ogni diverso valore numerico genera un segnale di ampiezza pari ad uno dei livelli di quantizzazione, un s&H ed un filtro passa-basso di restituzione⁹.

Osserviamo ora come il s&H si limiti ad *approssimare* il segnale $x^\bullet(t)$, realizzando al suo posto il segnale $x^\circ(t)$, mediante un treno di impulsi rettangolari modulati in ampiezza, in accordo allo schema di principio disegnato al suo interno. Pertanto, il filtro di ricostruzione *non è alimentato da $x^\bullet(t)$* , ma dal segnale $x^\circ(t) = \sum_n x(nT_c) \cdot \text{rect}_\tau(t - nT_c)$. Per determinare quale sia in questo caso l'uscita del filtro di restituzione $H(f)$, ne descriviamo l'ingresso $x^\circ(t)$ nella forma:

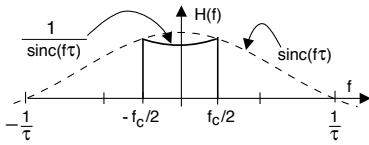
$$\begin{aligned} x^\circ(t) &= \sum_n x(nT_c) \cdot \text{rect}_\tau(t) * \delta(t - nT_c) = \\ &= \text{rect}_\tau(t) * \sum_n x(nT_c) \cdot \delta(t - nT_c) = \text{rect}_\tau(t) * x^\bullet(t) \end{aligned}$$

e dunque

$$X^\circ(f) = \tau \text{sinc}(f\tau) \cdot X^\bullet(f) \quad (4.4)$$

Osserviamo quindi che usare rettangoli di base $\tau < T_c$ al posto degli impulsi, equivale a moltiplicare $X^\bullet(f)$ per un involuppo di tipo $\frac{\sin x}{x}$ che, seppur con $\tau \ll T_c$ non causa grossi inconvenienti (gli zeri posti ad $\frac{1}{\tau}$ si allontanano dall'origine e $\frac{\sin x}{x}$ vicino ad $x = 0$ è quasi costante), per τ prossimo a T_c produce una alterazione dell'ampiezza della replica in banda base.

⁹In effetti, il DAC necessita di un segnale di temporizzazione, sincronizzato con f_c . Questo segnale può essere trasmesso separatamente, o essere ri-generato localmente a partire dalla stima della velocità alla quale sono ricevuti i valori $x(nT_c)$.

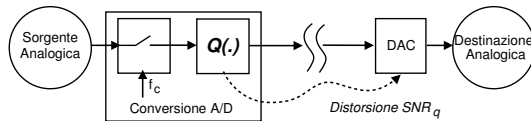
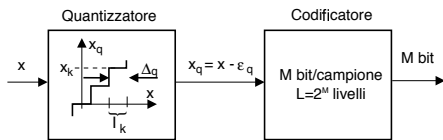


In tal caso (τ è noto) il filtro di ricostruzione può essere realizzato in modo da avere un andamento *inverso* a quello del $\frac{\sin x}{x}$, e tale che $H(f) \cdot \tau \text{sinc}(f\tau) = \text{costante}$. Infatti, questo accorgimento prende il nome di *sinc correction*.

All'appendice 6.9.5 è illustrato un metodo di MULTIPLAZIONE di più segnali campionati in una unica trasmissione.

4.2 Quantizzazione

Senza entrare nei dettagli di come sia effettivamente realizzato il circuito¹⁰ di un quantizzatore, limitiamoci a definirne la funzione come quella di suddividere la dinamica Δ_x dei possibili valori x in ingresso in $L = 2^M$ intervalli¹¹ di ampiezza $\Delta_q = \frac{\Delta_x}{L}$, ed associare ad ogni campione di segnale x che ricade entro l'intervallo I_k , lo stesso valore (o *livello*) $x_q = x_k$ che corrisponde al centro dell'intervallo I_k , commettendo quindi un errore $\epsilon_q = x - x_q$, come rappresentato in figura, ed il cui effetto si ripercuote in fase di restituzione D/A, come descritto nella figura che segue.



Quindi, ogni valore quantizzato x_q viene rappresentato per mezzo di un *codice binario* c_q con un numero fisso M di bit, in modo da ottenere una frequenza

binaria $f_b = M \cdot f_c$ bit/secondo. Lo stesso codice, quando ricevuto dal convertitore D/A, produrrà la generazione di un campione di valore pari ad x_q . Al § 18.1 l'argomento della *codifica di sorgente* verrà ripreso, e applicato al caso specifico del segnale vocale.

Procediamo ora con una prima analisi di due diversi modi di scegliere il codice da associare ad ogni livello.

4.2.1 Quantizzazione lineare

Il primo caso è quello in cui ad ogni livello k , con $k = 0, 1, \dots, L - 1$, si associa la rappresentazione binaria stessa di k (es.: se $k = 3$, $c_q = 11$): dato che anche gli intervalli I_k sono tutti di eguale ampiezza, la caratteristica ingresso-uscita del quantizzatore è uno *zig zag* in diagonale, come mostrato nella parte destra di fig. 4.3; inoltre, realizzando l'intervallo centrale *a cavallo* di $x = 0$, si limita il massimo errore a $\pm \frac{\Delta_q}{2}$, ma il questo caso il numero di intervalli (e di livelli) risulta dispari, e pari ad $L - 1$. Notiamo ora che, in accordo alla teoria velocità-distorsione (§ 17.1.2.1), più si sceglie un numero M di bit/campione piccolo (riducendo quindi la velocità binaria), e maggiore sarà la distorsione, che in questa sede identifichiamo nel rapporto SNR_q tra la potenza \mathcal{P}_x del segnale non quantizzato, e quella \mathcal{P}_e del rumore di quantizzazione.

¹⁰Vedi ad es. http://sms.unipv.it/misure/ME/Conversione_A-D_Slides.pdf

¹¹Come noto, per individuare uno tra L possibili oggetti occorrono $M = \log_2 \lceil L \rceil$ bit, e scegliendo L come una potenza di due, si ottiene il risultato esatto. Ad esempio, scegliendo $L = 64$, si ottiene $M = 6$.

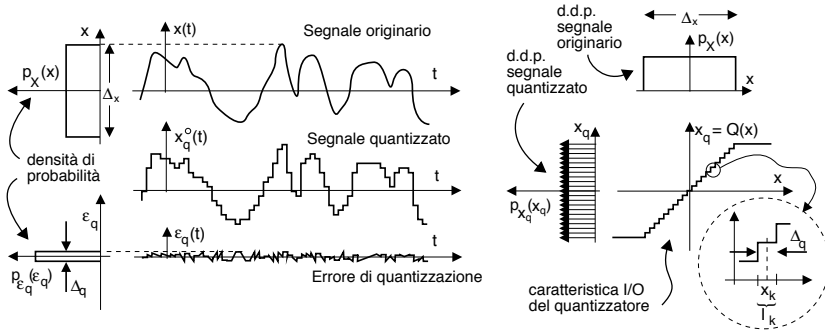


Figura 4.3: Processo di quantizzazione per segnali a distribuzione di ampiezza uniforme

4.2.1.1 SNR di quantizzazione

Individuiamo la relazione che lega SNR_q al numero di bit/campione M , determinando innanzitutto una espressione per la potenza del segnale \mathcal{P}_x , e quindi per quella del rumore \mathcal{P}_ϵ , con l'aiuto del grafico mostrato in fig. 4.3, e adottando l'ipotesi semplificativa che i valori in ingresso al campionatore abbiano origine da un processo ergodico a media nulla, e che soprattutto siano rappresentati da una v.a. con densità di probabilità uniforme $p_x(x) = \frac{1}{\Delta_x} \text{rect}_{\Delta_x}(x)$. In queste ipotesi, la potenza dei campioni è pari alla varianza della v.a. x , e risulta (vedi § 7.2.3)

$$\mathcal{P}_x = \sigma_x^2 = \frac{\Delta_x^2}{12}$$

Per quanto riguarda la potenza del segnale di errore, consideriamo che in fase di restituzione il campione quantizzato al livello k sarà riprodotto con il valore $x_q = x_k = k\Delta_q - \frac{\Delta_x}{2}$, commettendo l'errore $\epsilon_q = x - x_q$, di entità limitata entro l'intervallo $\pm \frac{\Delta_q}{2}$. Di nuovo, supponiamo che anche ϵ_q sia una v.a. uniformemente distribuita tra $\pm \frac{\Delta_q}{2}$, ed indipendente¹² da x_k ; pertanto, la potenza della componente di errore è anche qui pari alla varianza, e cioè

$$\mathcal{P}_\epsilon = \sigma_\epsilon^2 = \frac{\Delta_q^2}{12} = \frac{1}{12} \left(\frac{\Delta_x}{L-1} \right)^2$$

Siamo finalmente in grado di valutare l' SNR di quantizzazione:

$$SNR_q = \frac{\mathcal{P}_x}{\mathcal{P}_\epsilon} = \frac{\Delta_x^2}{12} 12 \left(\frac{L-1}{\Delta_x} \right)^2 = (L-1)^2 \cong L^2$$

in cui l'ultima approssimazione ha validità nel caso evidente in cui $L \gg 1$. Il risultato mostra che l' SNR_q cresce in modo quadratico con l'aumentare dei livelli, ovvero se L raddoppia SNR_q quadruplica. Ricorrendo alla notazione in decibel¹³ per l' SNR ,

¹²Questa ed altre ipotesi adottate sono palesemente non vere, ma permettono di giungere ad un risultato abbastanza semplice, e che può essere molto utile nel dimensionamento di massima degli apparati.

¹³Una discussione relativa alla misura delle grandezze in decibel, è fornita al § 14.4. Qui ci limitiamo ad usare i dB come misura relativa di un rapporto, ossia

$$SNR_q \text{ (dB)} = 10 \log_{10} \frac{\mathcal{P}_x}{\mathcal{P}_\epsilon} = 10 \log_{10} \mathcal{P}_x - 10 \log_{10} \mathcal{P}_\epsilon = \mathcal{P}_x \text{ (dBV}^2) - \mathcal{P}_\epsilon \text{ (dBV}^2)$$

otteniamo il risultato $SNR_q(L)|_{dB} = 10 \log_{10} L^2 = 20 \log_{10} L$ e, ricordando che $L = 2^M$, si ottiene

$$SNR_q(M)|_{dB} = M \cdot 20 \log_{10} 2 \simeq 6 \cdot M \quad \text{dB} \quad (4.5)$$

dato che $\log_{10} 2 \simeq 0.3$. In modo simile, valutiamo il miglioramento in dB ottenibile aumentando di uno il numero di bit per ogni campione, ovvero raddoppiando il numero di livelli: $SNR_q(2L)|_{dB} = 20 \log_{10} 2L = 20 \log_{10} L + 20 \log_{10} 2 \cong SNR_q(L)|_{dB} + 6 \text{dB}$. Pertanto ogni bit in più, provoca un miglioramento di 6 dB per l' SNR_q .

Consideriamo ora cosa accade se il segnale in ingresso x ha una dinamica *minore* di quanto previsto: in tal caso σ_x^2 si riduce, mentre $\sigma_\epsilon^2 = \frac{1}{12} \left(\frac{\Delta_x}{L-1} \right)^2$ non cambia, e dunque SNR_q peggiora, come se avessimo ridotto i livelli. Ma la tecnica illustrata di seguito, è capace di mantenere un SNR_q elevato anche con bassi livelli di segnale.

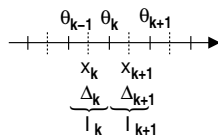
4.2.2 Quantizzazione logaritmica e codifica PCM

La tecnica che ora illustriamo nasce sia come risposta al problema appena evidenziato, ossia del peggioramento di SNR_q in presenza di bassi segnali, sia in considerazione del fatto che in realtà il processo x non ha distribuzione di ampiezza uniforme come ipotizzato sopra, rendendo il risultato (4.5) di ridotta utilità pratica.

Si può mostrare¹⁴ che, per una generica $p_X(x)$, il quantizzatore *ottimo* (che rende minimo SNR_q) non suddivide la dinamica Δ_x in $L - 1$ intervalli I_k di uguale estensione $\Delta_q = \frac{\Delta_x}{L-1}$, ma utilizza intervalli di estensione *variabile* $\Delta_q(k)$ diversa per ogni k , ed in particolare di estensione *ridotta* nelle regioni per le quali $p_X(x)$ è *più grande*. In questo modo, ai valori x più probabili (e quindi più frequenti) corrispondono intervalli di estensione ridotta, producendo in corrispondenza degli stessi un errore di quantizzazione ridotto, sia in termini di ampiezza che di potenza $\sigma_q^2(k)$, riducendone così il contributo alla potenza totale di errore \mathcal{P}_ϵ . Infatti \mathcal{P}_ϵ si ottiene ora come valore atteso, pesando i contributi $\sigma_q^2(k)$ con le rispettive probabilità:

$$\mathcal{P}_\epsilon = E_k \{ \sigma_q^2(k) \} = \sum_{k=1}^L p_k \sigma_q^2(k)$$

in cui $p_k = \int_{I_k} p_X(x) dx$ è la probabilità che $x \in I_k$. Il modo ottimo di disporre i confini (θ_{k-1}, θ_k) che delimitano I_k , in modo da rendere minima \mathcal{P}_ϵ , è noto come algoritmo di LLOYD-MAX¹⁵, ed ottiene il risultato di determinare errori più grandi per intervalli rari, ed errori più piccoli per intervalli più probabili.



in cui le grandezze espresse in dBV^2 rappresentano potenze di segnale di tensione, in unità logaritmiche.

¹⁴Vedi ad es. [http://en.wikipedia.org/wiki/Quantization_\(signal_processing\)](http://en.wikipedia.org/wiki/Quantization_(signal_processing))

¹⁵Il metodo è iterativo, ed inizia suddividendo l'intervallo Δ_x in modo uniforme. Per ogni iterazione:

- si determinano i valori quantizzati x_k (detti *centroidi*) come $x_k = E \{ x \in I_k \} = \int_{I_k} x \cdot p_X(x/k) dx = \frac{\int_{I_k} x p_X(x) dx}{p_k}$. In tal modo, i valori x_k si spostano (internamente a I_k) verso la regione in cui $p_X(x)$ ha un valore più elevato, ovvero dove la v.a. si addensa;
- si ri-calcolano i confini di decisione θ_k come $\theta_k = \frac{x_k + x_{k+1}}{2}$, seguendo lo spostamento degli x_k .

Le iterazioni si arrestano quando non si riscontrano cambiamenti apprezzabili.

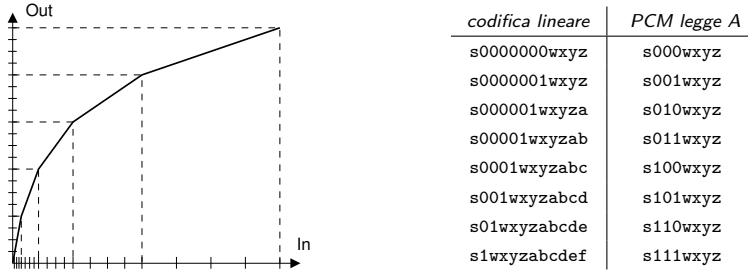


Figura 4.4: Caratteristica 1/0 di un quantizzatore logaritmico e relativa tabella di codifica

Nella quantizzazione del segnale vocale, anche se è arbitrario identificare con esattezza una $p_X(x)$, si verifica strumentalmente che quest'ultima è addensata nelle regioni con valori più piccoli. Per questo motivo, la legge di quantizzazione che si è adottata per ottenere gli 8 bit a campione utilizzati nel PCM segue un andamento logaritmico¹⁶, dimezzando progressivamente la pendenza della caratteristica di ingresso-uscita del quantizzatore.

La fig. 4.4 mostra un esempio di tale realizzazione (per i soli valori positivi), evidenziando come il risultato possa essere approssimato individuando (a partire dall'origine) regioni di valori di ingresso con una ampiezza che di volta in volta raddoppia, e suddividendo ogni regione in un uguale numero di intervalli equispaziati. La caratteristica non lineare è realizzabile per via completamente numerica: si attua un campionamento con legge lineare, con un numero di livelli molto maggiore del necessario (es con $4096 = 2^{12}$ livelli). Il numero di bit pari a zero nella parte più significativa del campione individua la regione dei valori di ingresso, mentre i bit rimanenti sono *shiftati* a destra, per mantenere costante il numero di intervalli per regione, ottenendo in definitiva una rappresentazione in *virgola mobile* del valore del campione.

Esempio La tabella a destra di fig. 4.4 esemplifica il processo di conversione PCM *legge A*, in cui la codifica lineare è di 12 cifre, mentre gli 8 bit della codifica PCM si suddividono in un bit di segno, tre di esponente e quattro di mantissa. Il MSB della codifica lineare rappresenta il segno (s), e resta inalterato. Il numero di zeri più significativi individua in quale delle 8 regioni cade il valore di ingresso, rappresentato da tre bit nel codice PCM. Infine delle restanti cifre di ingresso, se ne conservano solo le 4 più significative. Ad esempio, il valore 10000001010, con quattro cifre significative, diviene 1-000-1010 (come risulta dalla prima riga della tabella), mentre 0001100110101 diventa 0-101-100, come riporta la sesta riga.

I risultati del *mapping* ora esposto possono essere inseriti in una ROM come coppie di *ingresso-uscita*, ovvero come coppie *uscita-ingresso* per il dispositivo che effettua il procedimento inverso.

Le prossime sezioni non tengono più conto delle questioni relative alla quantizzazione dei campioni, anche se nella realizzazione di dispositivi reali può rivestire interesse di progetto.

¹⁶L'andamento esatto della curva segue uno di due standard, denominati legge μ (per USA e Giappone) e legge *A* (per gli altri), lievemente diverse nella definizione, ma sostanzialmente equivalenti.

4.3 Trasformata di Fourier di sequenze

Avendo illustrato come sia sufficiente la conoscenza dei soli campioni temporali $x(nT_c)$ per descrivere completamente un segnale continuo e limitato in banda $x(t)$, e come alla sequenza tempo-discreta dei suoi campioni x_n corrisponda una periodizzazione in frequenza $X^\bullet(f)$, notiamo come ciò sia in qualche modo *speculare* alla proprietà dei segnali periodici nel tempo, di godere di una rappresentazione equivalente nel dominio della frequenza, costituita dalla sequenza dei coefficienti X_n noti come serie di Fourier. L'analogia è più stringente di quanto non possa apparire, dato che è assolutamente lecito ed esatto usare l'espressione della serie di Fourier (2.3) nella direzione *opposta*, ossia per ottenere lo spettro periodico di ampiezza $X^\bullet(f)$ a partire dalla sequenza dei campioni temporali x_n :

$$X^\bullet(f) = \sum_{n=-\infty}^{\infty} x_n e^{-j2\pi f n T_c} \quad (4.6)$$

definendo così una *trasformata di Fourier a tempo discreto*¹⁷ o **DTFT**, che produce una $X^\bullet(f)$ periodica in frequenza di periodo $f_c = \frac{1}{T_c}$ ¹⁸, in cui T_c è il periodo di campionamento con cui sono prelevati i valori x_n ¹⁹. Alla (4.6) è associata una *anti-trasformazione*, in grado di valutare i campioni temporali x_n partendo dalla conoscenza di un periodo di $X^\bullet(f)$, definita come

$$x_n = \frac{1}{f_c} \int_{-\frac{f_c}{2}}^{\frac{f_c}{2}} X^\bullet(f) e^{j2\pi f n T_c} df \quad (4.7)$$

e che è del tutto analoga (a parte il segno) alla (2.2) che calcola i coefficienti della serie di Fourier.

Molte delle proprietà già note per la serie e la trasformata di Fourier sono valide anche in questo caso, come ad esempio

- simmetria coniugata per $X^\bullet(f)$ nel caso di x_n reale;
- linearità: $ax_n + by_n \iff aX^\bullet(f) + bY^\bullet(f)$;
- ritardo: $x_{n-k} \iff X^\bullet(f) e^{-j2\pi k f T_c}$;
- modulazione: $x_n e^{j2\pi n f_0 T_c} \iff X^\bullet(f - f_0)$;
- convoluzione: $z_n = x_n * y_n = \sum_{k=-\infty}^{\infty} x_k y_{n-k} \iff Z^\bullet(f) = X^\bullet(f) \cdot Y^\bullet(f)$;
- prodotto: $z_n = x_n \cdot y_n \iff Z^\bullet(f) = \frac{1}{f_c} \int_{-\frac{f_c}{2}}^{\frac{f_c}{2}} X^\bullet(\sigma) Y^\bullet(f - \sigma) d\sigma$.

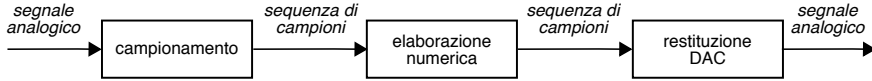
Tutto ciò permette effettuare operazioni sui segnali come analisi spettrale e filtraggio, senza dover svolgere calcoli analitici come integrali e trasformate, bensì operando direttamente sui campioni di segnale mediante appositi programmi di calcolo numerico, eseguiti su dispositivi ottimizzati a tale scopo²⁰, e quindi effettuare il processo di conversione D/A per riottenere un risultato tempo-continuo.

¹⁷Condizione sufficiente per la convergenza della serie (??) è che risulti $\sum_{n=-\infty}^{\infty} |x_n| < \infty$, in quanto $|X^\bullet(f)| = \left| \sum_{n=-\infty}^{\infty} x_n e^{-j2\pi f n T_c} \right| \leq \sum_{n=-\infty}^{\infty} |x_n|$.

¹⁸Infatti se applichiamo la (4.6) per calcolare $X^\bullet(f + f_c)$ si ottiene $\sum_{n=-\infty}^{\infty} x_n e^{-j2\pi(f+f_c)nT_c} = \sum_{n=-\infty}^{\infty} x_n e^{-j2\pi f n T_c} e^{-j2\pi f_c n T_c} = X^\bullet(f)$ dato che, essendo $f_c = \frac{1}{T_c}$, risulta $e^{-j2\pi f_c n T_c} = e^{-j2\pi n} = 1$ per qualsiasi n .

¹⁹Proprio come ai coefficienti della serie di Fourier per segnali periodici, intervallati di F Hz, corrisponde un segnale periodico nel tempo, di periodo $T = \frac{1}{F}$.

²⁰I chip progettati appositamente per svolgere calcoli di elaborazione numerica del segnale sono detti DSP (*Digital Signal Processor*).



Resta comunque il fatto che nelle (4.6) e (4.7) tuttora compaiono una somma di infiniti termini ed un integrale di funzione continua, mentre per effettuare le operazioni di elaborazione numerica possono essere usate solo sequenze di durata finita e somme. Per questo motivo affrontiamo la sezione successiva, che illustra come ciò possa essere risolto eseguendo il campionamento anche in frequenza, e limitando i segnali a durate finite.

4.4 Trasformata discreta di Fourier

Disponendo di una sequenza di lunghezza *finita* composta da N valori x_n , $n = 0, 1, \dots, N-1$, corrispondenti a campioni di un segnale $x(t)$ prelevati con ritmo $f_c = \frac{1}{T_c}$, si indica come DISCRETE FOURIER TRANSFORM (**DFT**) la nuova sequenza²¹

$$X_m = \sum_{n=0}^{N-1} x_n e^{-j2\pi \frac{m}{N} n} \quad (4.8)$$

univocamente definita per $m = 0, 1, \dots, N-1$, e che costituisce una approssimazione²² del campionamento in frequenza della trasformata $X(f) = \mathcal{F}\{x(t)\}$, calcolata per $f = \frac{m}{N} f_c$, e moltiplicata per f_c :

$$X_m \simeq f_c X \left(f = \frac{m}{N} f_c \right) \quad (4.9)$$

Notiamo subito che la (4.8) è valida per qualsiasi m , ed ha un andamento periodico con periodo N , a cui corrisponde la frequenza $f_c = \frac{1}{T_c}$, in accordo con la periodicità

²¹La (4.8) può essere fatta discendere dalla (4.6) vincolando f ad assumere i soli valori discreti $f = \frac{m}{N} \frac{1}{T_c}$, e limitando l'indice della sommatoria ad un insieme finito di campioni.

²²Una prima fonte di approssimazione deriva dall'operazione di finestrata legata all'uso di un numero finito di campioni, operando quindi su $x_w(t) = x(t) w(t_c)$ anziché su $x(t)$. Per analizzare le altre fonti di approssimazione, iniziamo a scrivere l'espressione di $X_w(f) = \mathcal{F}\{x_w(t)\}$ per $f = \frac{m}{N} f_c$:

$$\begin{aligned} X_w \left(f = \frac{m}{N} f_c \right) &= \int_0^{(N-1)T_c} x(t) e^{-j2\pi \frac{m}{N} f_c t} dt \\ &\simeq \sum_{n=0}^{N-1} x_n \cdot \int_0^{(N-1)T_c} \text{sinc}(f_c(t - nT_c)) e^{-j2\pi \frac{m}{N} f_c t} dt \end{aligned}$$

in cui la seconda eguaglianza utilizza l'interpolazione cardinale $x(t) = \sum_{n=-\infty}^{\infty} x_n \cdot \text{sinc}(f_c(t - nT_c))$ fornita dalla (4.1), ed introduce una seconda fonte di approssimazione legata all'intervallo *finito* di variazione per m (infatti, benché l'integrale abbia estensione limitata, i valori di $x(t)$ che cadono entro tale estensione, dovrebbero dipendere da *tutti* i suoi campioni). L'ultimo integrale è a sua volta una approssimazione (a causa degli estremi di integrazione limitati, e peggiore per i sinc centrati in prossimità dei confini della finestra) della trasformata (calcolata in $f = \frac{m}{N} f_c$) di $\text{sinc}(f_c(t - nT_c))$, pari quest'ultima a $T_c \text{rect}_{f_c}(f) e^{-j2\pi f n T_c}$, che quando valutata per $f = \frac{m}{N} f_c$, fornisce il risultato

$$X_w \left(f = \frac{m}{N} f_c \right) \simeq T_c \sum_{n=0}^{N-1} x_n e^{-j2\pi \frac{m}{N} n}$$

per valori $|m| \leq \frac{N}{2}$, a causa della estensione limitata (in frequenza) di $\text{rect}_{f_c}(f)$. E' però facile verificare che $X_w(\frac{m}{N} f_c)$ è periodica in m con periodo N , cosicché i valori assunti per $m = \frac{N}{2} + 1, \frac{N}{2} + 2, \dots$ sono uguali a quelli per $m = -\frac{N}{2} + 1, -\frac{N}{2} + 2, \dots$

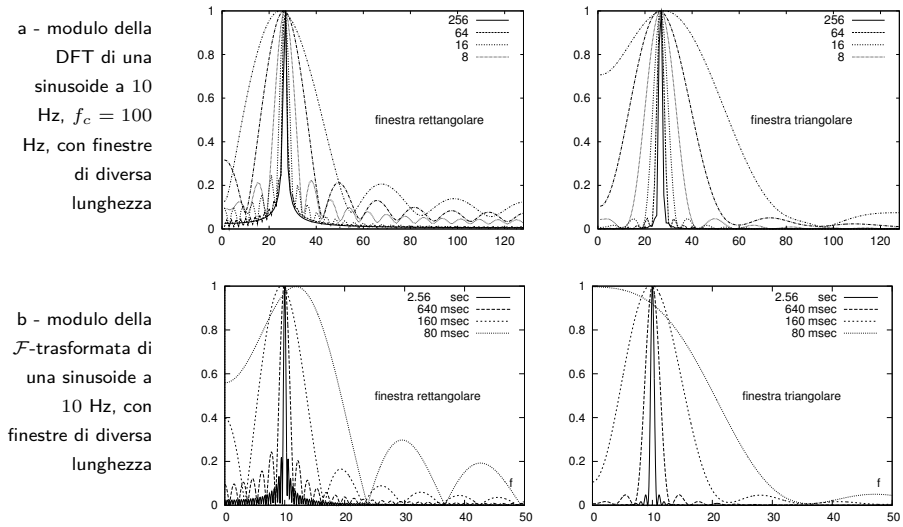


Figura 4.5: Confronto tra DFT ed \mathcal{F} -trasformata con uguale estensione temporale

in frequenza evidenziata per $X^\bullet(f)$ (vedi (4.6) e fig. 4.2); per questo motivo, qualora il segnale originario $x(t)$ contenga componenti a frequenze maggiori di $\frac{f_c}{2}$, gli X_m con indici prossimi ad $\frac{N}{2}$ presenteranno errore di aliasing²³. Notiamo inoltre come i valori X_m siano tutti relativi a frequenze ≥ 0 , ma nel caso di una sequenza x_n reale la proprietà di simmetria coniugata, associata alla periodicità in frequenza, rende il risultato interessante solo per indici $m \leq \frac{N}{2}$, dato che successivamente si trovano valori coniugati a quelli della prima metà. Notiamo infine che la (4.8) può essere espressa in forma matriciale: ad esempio, per $N = 4$ si ottiene

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{-j\frac{\pi}{2}} & e^{-j\pi} & e^{-j\frac{3\pi}{2}} \\ 1 & e^{-j\pi} & e^{-j2\pi} & e^{-j3\pi} \\ 1 & e^{-j\frac{3\pi}{2}} & e^{-j3\pi} & e^{-j\frac{9\pi}{2}} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (4.10)$$

da cui notiamo la proprietà di *simmetria* per la matrice dei coefficienti.

Esempio Allo scopo di concretizzare le differenze tra la trasformata di Fourier ed i valori forniti dalla DFT, in fig. 4.5-a sono riportati i valori $|X_n|$ per la DFT di una sinusoida a 10 Hz, adottando due diverse finestre di analisi (vedi § 3.9.3), prelevando alla medesima frequenza di campionamento (100 Hz) un numero variabile di campioni (mostrato in figura), e ponendo i rimanenti a zero, per calcolare in tutti i casi la medesima DFT a 256 punti²⁴. Il risultato è quindi confrontato (fig. 4.5-b) con quello ottenibile per via analitica calcolando

²³Come osservato al § 4.1.1, lo spettro $X^\bullet(f)$ di un segnale campionato a frequenza f_c è costituito dalle repliche del segnale originario, distanziate di multipli di f_c : $X^\bullet(f) = \sum_{n=-\infty}^{\infty} X(f - nf_c)$, e coincide con $X(f)$ per $-f_c/2 < f < f_c/2$, se $X(f)$ è limitata in banda tra $\pm W$ ed $f_c \geq 2W$. Al contrario, se $f_c < 2W$, allora le repliche $X(f - nf_c)$ si sovrappongono, e la (4.9) si riscrive come $X_m \simeq f_c X^\bullet(f = \frac{m}{N} f_c)$.

²⁴Il metodo esposto di porre a zero i campioni fino al raggiungimento di una potenza di due, in modo da utilizzare per il calcolo una FFT, è detto ZERO PADDING. Il calcolo della DFT su di un

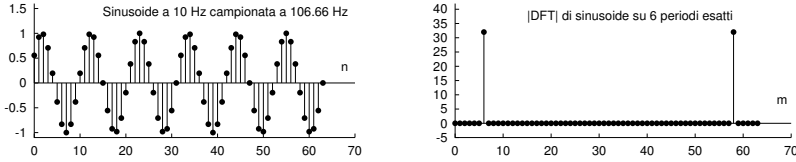


Figura 4.6: Sinusoide campionata su di un numero intero di periodi e relativo modulo di DFT

la \mathcal{F} -trasformata dello stesso segnale, adottando le medesime finestre temporali, di durata uguale al primo caso. Le curve ottenute nel caso di 80 msec (e 8 campioni!) dipendono da meno di un periodo di segnale, e perciò presentano una componente continua. Aumentando la durata della finestra, l'approssimazione di calcolare una $\mathcal{F}\{\}$ mediante la DFT migliora, anche se persiste un ridotto potere di risoluzione spettrale.

Osservazione Probabilmente a questo punto più di un lettore può trovarsi stupito di non avere incontrato due linee spettrali, come sarebbe stato lecito aspettarsi per una sinusoide. La spiegazione è che ciò può accadere, a patto che il numero di campioni N su cui si effettua la DFT sia un multiplo intero k del numero di campioni $M = T/T_c$ che ricadono entro uno stesso periodo T della sinusoide (con la ovvia condizione che sia $M > 2$ per rispettare il vincolo $f_c > 2/T$); in questo caso infatti, l'applicazione della IDFT (4.11) produce una sequenza ancora periodica. Nel caso in esame, essendo $T = 100$ msec e scegliendo $N = 64$ e $k = 6$, il vincolo $N = kM = kTf_c$ permette di ottenere $f_c = \frac{N}{kT} = \frac{64}{6 \cdot 10^{-1}} = 106, \bar{6}$ Hz, ovvero $M = Tf_c = 10^{-1} \cdot 106, \bar{6} = 10, \bar{6}$ campioni/periodo. La fig. 4.6 mostra questo risultato, evidenziando come la riga spettrale si manifesti per $m = 6$, ossia alla frequenza $f = \frac{m}{N} f_c = \frac{6}{64} 106,6 = 10$ Hz, mentre la riga presente in $m = 58$ è in realtà il ripiegamento periodico di quella a -10 Hz.

Il passaggio dai campioni nel tempo x_n a quelli in frequenza X_m è invertibile²⁵, ricorrendo alla INVERSE DISCRETE FOURIER TRANSFORM (**IDFT**)

$$x_n = \frac{1}{N} \sum_{m=0}^{N-1} X_m e^{j2\pi \frac{m}{N} n} \quad (4.11)$$

che per n esterno a $[0, N - 1]$ continua a valere, ed assume valori periodici, coerentemente a quanto accade per lo sviluppo in serie di Fourier. Infatti il legame tra DFT e serie di Fourier è molto stretto, in quanto i valori X_m rappresentano una approssimazione²⁶ dei rispettivi coefficienti della serie di Fourier $X_m^{SF} = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x_T(t) e^{-j2\pi \frac{m}{T} t} dt$,

numero di punti pari ai campioni di segnale disponibili, non avrebbe dato luogo all'effetto finestra, ma avrebbe fornito in tutti i casi andamenti simili a quello osservabile per 256 punti. Infine, notiamo che nelle figure sono mostrati solo i primi 128 valori, essendo i rimanenti speculari.

²⁵Sostituendo infatti la (4.8) nella (4.11), otteniamo

$$x_n = \frac{1}{N} \sum_{m=0}^{N-1} \left(\sum_{k=0}^{N-1} x_k e^{-j2\pi \frac{m}{N} k} \right) e^{j2\pi \frac{m}{N} n} = \frac{1}{N} \sum_{k=0}^{N-1} x_k \sum_{m=0}^{N-1} e^{j2\pi \frac{m}{N} (k-n)}$$

ma, dato che $\sum_{m=0}^{N-1} e^{j2\pi \frac{m}{N} (k-n)} = \begin{cases} N & \text{se } k = n \\ 0 & \text{altrimenti} \end{cases}$, allora nella sommatoria esterna sopravvive solo il termine x_n , dimostrando l'uguaglianza.

²⁶La relazione (4.12) si dimostra combinando le relazioni (3.1) e (4.9): $X_n \simeq f_c X\left(\frac{n}{N} f_c\right) = f_c X\left(\frac{n}{NT_c}\right) = f_c X\left(\frac{n}{T}\right) = f_c X(nF) = f_c T X_n^{SF} = \frac{1}{T_c} T X_n^{SF} = N X_n^{SF}$

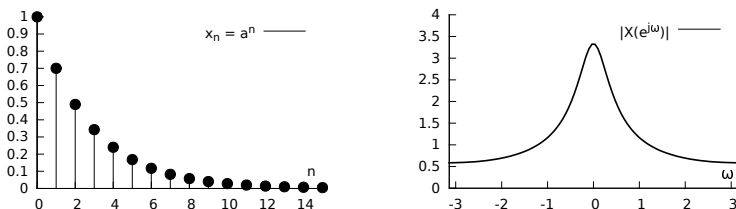


Figura 4.7: Sequenza $x_n = a^n$ e modulo della relativa trasformata di Fourier a tempo discreto

calcolati a partire da un segmento $x_T(t)$ estratto da $x(t)$, e moltiplicati per N :

$$X_n \simeq N \cdot X_n^{SF} \quad (4.12)$$

Per approfondire i risvolti di questo risultato, affrontiamo la sezione successiva.

4.4.1 Relazione tra DTFT, DFT e trasformata zeta

Così come per i segnali analogici sussiste una relazione (vedi pag. 241) tra la trasformata di FOURIER e quella di LAPLACE, così nel contesto delle sequenze, esistono legami tra DTFT e *trasformata zeta*, definita quest'ultima come

$$X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n} \quad (4.13)$$

che, nel caso in cui converga per $|z| = 1$, può essere fatta corrispondere alla DTFT (4.6) semplicemente ponendo $z = e^{j\omega}$, ovvero calcolando $X(z)$ sul *cerchio unitario*:

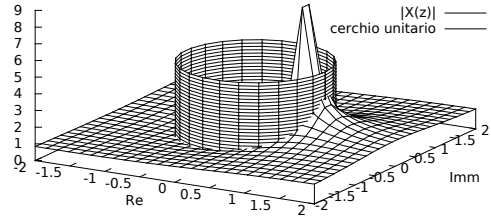
$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x_n e^{-j\omega n} = X(z)|_{z=e^{j\omega}} = X^\bullet(f)|_{f=\frac{\omega}{2\pi T_c}} \quad (4.14)$$

Infatti, nelle consuete condizioni in cui gli x_n sono i campioni di un segnale $x(t)$ limitato in banda tra $\pm W$ e prelevati con ritmo $f_c \geq 2W$, la (4.14) effettivamente *coincide* (per $-\pi \leq \omega < \pi$) con la $X^\bullet(f)$ (eq. (4.6), per $-\frac{f_c}{2} \leq f \leq \frac{f_c}{2}$) in cui si ponga $f = \frac{\omega}{2\pi T_c}$, mettendo cioè in corrispondenza le frequenze $\pm \frac{f_c}{2}$ di $X^\bullet(f)$ con le pulsazioni $\pm \pi$ di $X(e^{j\omega})$. Al di fuori di tale intervallo, $X(e^{j\omega})$ è periodica in ω con periodo 2π , analogamente a ciò che risulta (con periodo f_c) per la trasformata di Fourier $X^\bullet(f)$ di sequenze; se invece x_n è *sottocampionata*, ossia $f_c < 2W$, anche $X(e^{j\omega})$ è affetta da aliasing, così come avviene per $X^\bullet(f)$.

Esempio Consideriamo la sequenza $x_n = \begin{cases} a^n & \text{se } n \geq 0 \\ 0 & \text{altrimenti} \end{cases}$ il cui andamento per $a = 0.7$ è mostrato in fig. 4.7, la cui trasformata zeta $X(z) = \sum_{n=0}^{\infty} a^n z^{-n}$ risulta pari a²⁷ $X(z) = \frac{1}{1-az^{-1}}$, ed il cui modulo, dopo aver scritto la variabile complessa z come $z = x + jy$, è espresso come $|X(z)| = 1/\sqrt{\left(\frac{x^2-ax+y^2}{x^2+y^2}\right)^2 + \left(\frac{ay}{x^2+y^2}\right)^2}$. Facendo ora variare z nell'inter-

²⁷Il risultato si ottiene ricordando che $\sum_{n=0}^{\infty} \alpha^n = \frac{1}{1-\alpha}$ qualora $|\alpha| < 1$.

vallo $[-2 - j2, 2 + j2]$, si ottiene per il modulo di $X(z)$ l'andamento mostrato nella figura a lato, in cui è anche raffigurato un cilindro di raggio unitario, la cui intersezione con $|X(z)|$ individua l'andamento di $|X(e^{j\omega})| = 1/\sqrt{1+a^2-2a\cos\omega}$, ossia la DTFT della sequenza a^n , che a sua volta è mostrata in fig. 4.7 per $-\pi < \omega < \pi$.



Se la $X(e^{j\omega})$, ottenuta per una sequenza x_n *aperiodica* nel tempo, è campionata in N punti equispaziati e disposti sul cerchio unitario, ossia ponendo $\omega = 2\pi \frac{m}{N}$ con $m = 0, 1, \dots, N - 1$, si ottiene allora una sequenza periodica in frequenza²⁸

$$\tilde{X}_m = \sum_{n=-\infty}^{\infty} x_n e^{-j2\pi \frac{m}{N} n} = X(e^{j\omega}) \Big|_{\omega=2\pi \frac{m}{N}} = X^*(f) \Big|_{f=f_c \frac{m}{N}} \quad (4.15)$$

che può coincidere con la sequenza X_m ottenuta calcolando la DFT (4.8) di una sequenza x_n , qualora questa abbia una durata limitata $\leq N$. D'altro canto, è possibile applicare la IDFT (4.11) ad un periodo della sequenza \tilde{X}_m , ed ottenere quindi una nuova sequenza di valori nel tempo, anch'essa periodica di periodo N , espressa come

$$\tilde{x}_n = \frac{1}{N} \sum_{m=0}^{N-1} \tilde{X}_m e^{j2\pi \frac{m}{N} n} \quad (4.16)$$

Infatti, i valori \tilde{x}_n dipendono da quelli $x_n = x(t)|_{t=nT_c}$ del segnale originario $x(t)$, campionato agli istanti $t = nT_c$, mediante la relazione²⁹

$$\tilde{x}_n = \sum_{r=-\infty}^{\infty} x_{n+rN} \quad (4.17)$$

e quindi i primi N valori di \tilde{x}_n coincidono con i campioni di $x(t)$ solo se quest'ultimo ha durata limitata, con estensione minore di NT_c , ossia se N è sufficientemente elevato in modo che NT_c copra tutta la durata di $x(t)$, e la (4.13) si riconduca alla somma di un numero finito di termini. D'altra parte, se $x(t)$ ha durata maggiore di NT_c , ovvero $X(z)$ è stata campionata su di un numero di campioni troppo ristretto, allora l'applicazione della IDFT (4.16) ad $\tilde{X}(k)$ provoca il fenomeno di *aliasing temporale*.

²⁸ Infatti, $e^{-j2\pi \frac{m+N}{N} n} = e^{-j2\pi \frac{m}{N} n} e^{-j2\pi n}$, ed il secondo termine vale 1 per qualsiasi n . Indichiamo qui ed al prossimo §, una sequenza periodica mediante la tilde $\tilde{\cdot}$.

²⁹ Infatti, sostituendo la (4.15) in (4.16), otteniamo

$$\tilde{x}_n = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{h=-\infty}^{\infty} x_h e^{-j2\pi \frac{k}{N} h} e^{j2\pi \frac{k}{N} n}$$

che, scambiando l'ordine delle sommatorie, riscriviamo come

$$\tilde{x}_n = \sum_{h=-\infty}^{\infty} x_h \left(\frac{1}{N} \sum_{k=0}^{N-1} e^{-j2\pi \frac{k}{N} (h-n)} \right)$$

Dato che $\frac{1}{N} \sum_{k=0}^{N-1} e^{-j2\pi \frac{k}{N} (h-n)} = \begin{cases} 1 & \text{se } h = n + rN \\ 0 & \text{altrimenti} \end{cases}$, con r intero, si ottiene il risultato (4.17).

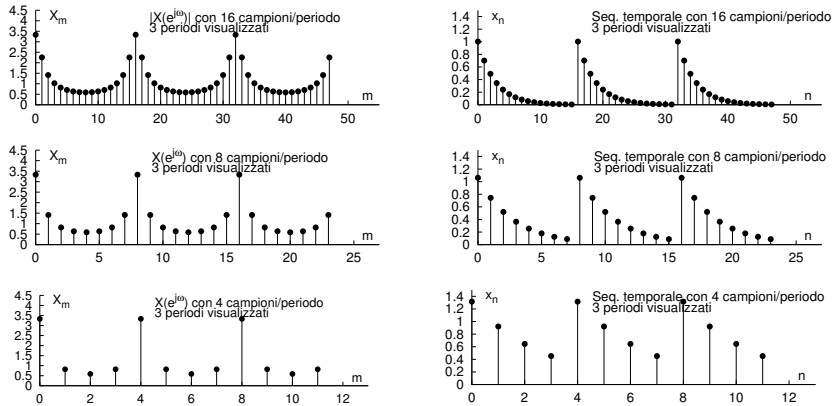


Figura 4.8: Aliasing temporale al diminuire della risoluzione del campionamento in frequenza

Esempio Nella parte sinistra di fig. 4.8 viene mostrato il modulo della sequenza X_m ottenuta campionando la $X(e^{j\omega})$ dell'esempio precedente, utilizzando 16, 8 o 4 campioni/periodo. Nella parte destra della stessa figura sono quindi rappresentate le corrispondenti sequenze x_n ottenute mediante IDFT. Si può notare che, mentre con 16 campioni/periodo la ricostruzione della sequenza $x_n = a^n$ è piuttosto fedele, con 8 campioni si inizia a verificare il fenomeno di aliasing temporale, che diviene ancor più evidente per 4 campioni/periodo.

4.4.2 Fast Fourier Transform

La sigla FFT descrive una classe di algoritmi di calcolo della DFT e della sua inversa, caratterizzati dall'uso di un numero molto ridotto di operazioni, rendendo grazie a ciò l'elaborazione numerica dei segnali computazionalmente praticabile. Analizziamo innanzitutto come il calcolo di ognuno degli N termini X_m della (4.8), considerando i valori $e^{-j2\pi \frac{m}{N} n}$ precalcolati (vedi (4.10)), richiede N moltiplicazioni complesse (equivalenti ognuna ad 4 moltiplicazioni e 2 somme reali) ed $N - 1$ somme complesse (ognuna pari a 2 somme reali): pertanto una DFT richiede $N(N(4 + 2) + 2(N - 1)) = N(8N - 2) \simeq 8N^2$ operazioni. Al contrario, gli algoritmi FFT più efficienti riducono il numero di operazioni ad $8N \log_2 N$: ad esempio, ponendo $N = 1024$, si ottiene un miglioramento di $2^3(2^{10})^2/2^3 2^{10} \cdot 10 = 2^{10}/10 \simeq 100$ volte! Queste prestazioni sono legate all'adozione di un valore di N che sia una potenza di due (ossia $N = 2^M$ con M intero), ma successivamente sono stati individuati metodi³⁰ che permettono una efficienza di calcolo comparabile anche per finestre di analisi di lunghezza qualsiasi.

4.4.3 Relazione tra DFT e DCT

Anche per la DFT risulta valida la proprietà di simmetria coniugata (§ 3.3) e quindi, se i valori della sequenza x_n di lunghezza N che compare nella (4.8) sono reali anziché complessi, allora i coefficienti di DFT X_m presentano parte reale pari e parte immaginaria dispari. In particolare, se immaginiamo di *estendere* la lunghezza della

³⁰Vedi ad es. http://it.wikipedia.org/wiki/Trasformata_di_Fourier_veloce

sequenza a $2N$ punti, ottenuti ribaltando sugli indici negativi la sequenza di partenza come $x_{-n} = x_n$ (vedi prima riga di fig. 4.9), allora siamo nelle condizioni di sequenza *reale pari*, che determina una trasformata solo reale (e pari), con parte immaginaria nulla.

Per arrivare a definire la DISCRETE COSINE TRANSFORM (DCT) si calcola una DFT *bilatera* sulla sequenza lunga $2N$ ottenuta trasladando quella descritta in modo da renderla effettivamente pari (seconda riga di fig. 4.9). Considerando che per segnali reali pari le componenti sinusoidali della base della DFT non danno contributi al risultato³¹, e adottando un nuovo cambio di variabile, si ottiene in definitiva la formula di calcolo della DCT come

$$X_m = \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) m \right] \quad (4.18)$$

a cui è associata la trasformazione inversa IDCT

$$x_n = \frac{1}{2} X_0 + \sum_{m=1}^{N-1} X_m \cos \left[\frac{\pi}{N} \left(m + \frac{1}{2} \right) n \right] \quad (4.19)$$

La DCT verrà usata in questo testo nell'ambito della compressione di immagini (§ 18.2.4): infatti i valori di luminanza dei pixel in cui si scompone una immagine, sono tutti valori reali.

4.5 Filtraggio numerico via DFT

La definizione di DFT illustrata al § 4.4 ben si presta a calcolare il risultato relativo ad un integrale di convoluzione, a patto di seguire alcune accortezze.

³¹Scriviamo la (4.8) come

$$\begin{aligned} X_m &= \sum_{n'=-N+1/2}^{N-1/2} x_{n'-1/2} e^{-j2\pi \frac{n'}{2N} m} = \\ &= \sum_{n'=-N+1/2}^{N-1/2} x_{n'-1/2} \cos \left(2\pi \frac{n'}{2N} m \right) - j \sum_{m'=-N+1/2}^{N-1/2} x_{n'-1/2} \sin \left(2\pi \frac{n'}{2N} m \right) = \\ &= 2 \sum_{n'=1/2}^{N-1/2} x_{n'-1/2} \cos \left(2\pi \frac{n'}{2N} m \right) = 2 \sum_{n=0}^{N-1} x_n \cos \left(2\pi \frac{n+1/2}{2N} m \right) = \\ &= 2 \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) m \right] \end{aligned}$$

in cui $x_{n'}$ è quella disegnata per seconda in fig. 4.9. La quarta eguaglianza tiene conto del fatto che il termine immaginario si annulla, in quanto sommatoria bilatera di una funzione dispari (ottenuta come prodotto di $x_{n'-1/2}$ pari e $\sin \left(2\pi \frac{n'}{2N} m \right)$ dispari), e del fatto che essendo i termini coseno pari, la sommatoria può essere ristretta ai soli indici positivi, raddoppiati. La penultima eguaglianza rappresenta il semplice cambio di variabile $n = n' - 1/2$, mentre l'ultima è (a parte il fattore 2) la definizione della DCT data in (4.18).

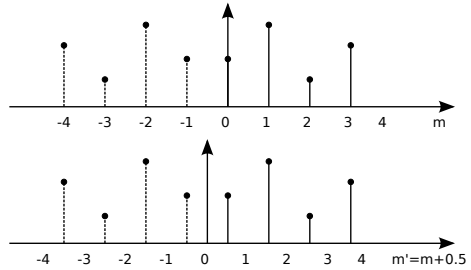


Figura 4.9: Estensione pari di sequenza reale

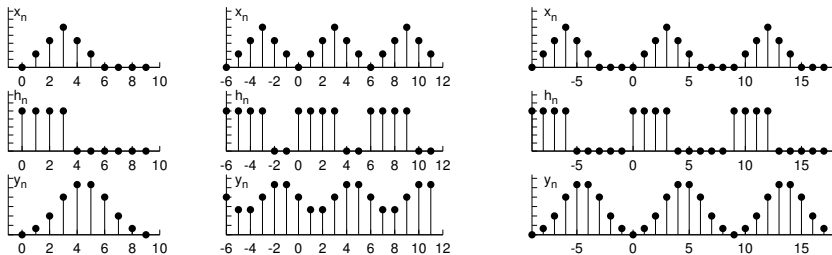


Figura 4.10: Confronto tra convoluzione discreta, circolare, e via DFT

4.5.1 Convoluzione discreta

Dati due segnali $x(t)$ e $h(t)$ limitati in banda tra $-W$ e W , anche il risultato della convoluzione $y(t) = x(t) * h(t)$ è limitato in banda, ed i suoi campioni $y_n = y(nT_c)$ (con $T_c > \frac{1}{2W}$) possono essere calcolati³² a partire da quelli di $x(t)$ e $h(t)$, come

$$y_n = \sum_{k=-\infty}^{\infty} x_k h_{n-k}$$

Nel caso in cui le sequenze x_n e h_n abbiano durata finita e pari rispettivamente a N ed M campioni, si otterrà una sequenza y_n di durata pari a $N + M - 1$ campioni.

Esempio: il lato sinistro di fig. (4.10) mostra una sequenza triangolare ed una rettangolare con un numero di elementi non nulli rispettivamente pari a 5 e 4, la cui convoluzione si estende su 8 valori.

4.5.2 Convoluzione circolare

Date due sequenze x_n ed h_n di durata finita N , il prodotto $Y_m = X_m H_m$ delle rispettive DFT $X_m = \sum_{n=0}^{N-1} x_n e^{-j2\pi \frac{m}{N} n}$ ed $H_m = \sum_{n=0}^{N-1} h_n e^{-j2\pi \frac{m}{N} n}$ possiede antitrasformata $\tilde{y}_n = \mathcal{DFT}^{-1}\{Y_m\}$ periodica in n di periodo N , e pari a

$$\tilde{y}_n = \sum_{p=0}^{N-1} \tilde{x}_p \tilde{h}_{n-p} \tag{4.20}$$

in cui \tilde{x}_n e \tilde{h}_n sono le sequenze periodiche di periodo N ottenute replicando infinitamente le sequenze originali x_n ed h_n (33). La convoluzione (4.20) è detta *circolare*

³²Il risultato può essere ottenuto esprimendo l'integrale $x(t) * h(t)$ nei termini dei campioni di $x(t)$ e $h(t)$, e sfruttando la proprietà di ortogonalità (vedi § 4.1.3) di sinc (.).

³³Infatti, ad x_n ed h_n corrispondono le DFT periodiche \tilde{X}_m ed \tilde{H}_m , che hanno per antitrasformata \tilde{x}_n ed \tilde{h}_n . Il prodotto $\tilde{X}_m \tilde{H}_m$, espresso in termini di \tilde{x}_n ed \tilde{h}_n , risulta pari a $\tilde{Y}_m = \tilde{X}_m \tilde{H}_m = \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \tilde{x}_p \tilde{h}_q e^{-j2\pi \frac{m}{N} (p+q)}$, ed applicando a questo la (4.11), otteniamo:

$$\tilde{y}_n = \frac{1}{N} \sum_{m=0}^{N-1} \tilde{Y}_m e^{j2\pi \frac{m}{N} n} = \frac{1}{N} \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} \tilde{x}_p \tilde{h}_q \left(\sum_{m=0}^{N-1} e^{j2\pi \frac{m}{N} (n-p-q)} \right)$$

Dato che $\sum_{m=0}^{N-1} e^{j2\pi \frac{m}{N} (n-p-q)} = \begin{cases} N & \text{se } q = (n-p) + lN \\ 0 & \text{altrimenti} \end{cases}$, con l intero, risulta allora $\tilde{y}_n = \sum_{p=0}^{N-1} \tilde{x}_p \tilde{h}_{n-p}$, come espresso dalla (4.20).

perché è possibile immaginare le sequenze x_n ed h_n *incollate* su due cilindri concentrici, e la somma svolta sui prodotti degli elementi coincidenti. Ogni valore di p corrisponde ad una diversa rotazione relativa (con angolo multiplo di $2\pi/N$) dei cilindri, ed il campione di h_n che era allineato ad x_{N-1} *rientra* dall'altro lato, per corrispondere ad x_0 .

La parte centrale di fig. (4.10) mostra il risultato della periodizzazione delle stesse sequenze dell'esempio precedente, assieme al risultato della convoluzione circolare tra esse, considerando un periodo $N = 6$ per entrambe: come evidente, il risultato è abbastanza differente da quello ottenuto nel caso della convoluzione discreta.

4.5.3 Convoluzione tra sequenze di durata finita via DFT

Sappiamo che la convoluzione produce un risultato di durata pari alla somma delle durate degli operandi; come anticipato, nel caso di due sequenze x_n ed h_n di durata N ed M , il risultato della convoluzione discreta $y_n = \sum_{k=0}^{N-1} x_k h_{n-k}$ produce valori non nulli per indici $n = 0, 1, \dots, N + M - 1$. Pertanto, perché la (4.20) produca lo stesso effetto di una convoluzione discreta, occorre costruire delle sequenze x'_n e h'_n di lunghezza almeno pari ad $N + M - 1$, ottenute a partire dai valori di x_n ed h_n , a cui si aggiungono $M - 1$ ed $N - 1$ valori nulli, rispettivamente. In tal modo, il prodotto $X'_m H'_m$ tra le DFT ad $N + M - 1$ punti di queste due nuove sequenze, può essere antitrasformato, per fornire il risultato corretto.

La parte destra di fig. (4.10) mostra come estendendo con zeri le due sequenze di esempio fino ad una stessa lunghezza $N = 9$, il risultato della loro convoluzione circolare anch'esso di periodo N ora fornisca con esattezza lo stesso risultato ottenuto calcolando direttamente la convoluzione discreta.

4.5.4 Convoluzione di segnali via DFT

Due segnali $x(t)$ e $h(t)$ limitati in banda non possono, a rigore, essere limitati nel tempo; viceversa, una finestra di segnale non può, a rigore, essere rappresentata dai suoi campioni. Infatti, l'effetto della convoluzione in frequenza tra la trasformata della finestra (nominalmente illimitata in banda) e lo spettro del segnale, produce una dispersione frequenziale di quest'ultimo.

Ciononostante, disponendo di un numero di campioni sufficientemente elevato, si può assumere che la trasformata della finestra si attenui, fino a rendersi trascurabile, oltre ad una certa frequenza. Inoltre, l'adozione di una frequenza di campionamento più elevata, provoca un allontanamento delle repliche spettrali del segnale campionato. In queste due ipotesi, è lecito ritenere l'elaborazione condotta sui campioni di segnale equivalente a quella da svolgere sul segnale originario.

Consideriamo ora il caso di operare su campioni prelevati alla frequenza opportuna, e di voler determinare la risposta di un filtro caratterizzato dalla propria h_n di durata finita M , ad un ingresso x_n di durata indefinita. Per applicare i risultati fin qui descritti, occorre suddividere la sequenza x_n in segmenti x_n^q di lunghezza L

$$x_n^q = \begin{cases} x_n & \text{per } qL \leq n < (q+1)L \\ 0 & \text{altrove} \end{cases}$$

in modo da poter scrivere $x_n = \sum_{q=-\infty}^{\infty} x_n^q$, operare una successione di convoluzioni discrete $y_n^q = x_n^q * h_n$, ed infine ottenere $y_n = x_n * h_n = \sum_{q=-\infty}^{\infty} y_n^q$ per la linearità della convoluzione.

Osserviamo ora che ognuno dei termini y_n^q risulta di estensione $N = M + L - 1$ punti, e può essere calcolato mediante la DFT inversa ad N punti del prodotto $X_m^q H_m^q$ tra le DFT ad N punti delle versioni *allungate con zero* (ZERO PADDED) di x_n^q ed h_n^q .

Infine, notiamo che l'estensione temporale $N = M + L - 1$ dei termini y_n^q è maggiore di quella dei segmenti originali x_n^q , di lunghezza L : pertanto la sequenza y_n si ottiene sommando ai primi $M - 1$ valori di ognuna delle y_n^q , gli ultimi $M - 1$ valori risultanti dalle operazioni precedenti. Per questo motivo, il metodo prende il nome di OVERLAP AND ADD.

4.6 Riassumendo

La DFT (4.8) e la IDFT (4.11) costituiscono una coppia di relazioni invertibili che permettono di passare da una sequenza complessa di lunghezza N ad un'altra di pari lunghezza. Ma:

- calcolando la DFT (4.8) su di una finestra di N campioni x_n di un segnale $x(t)$ limitato in una banda $W < f_c/2$, si ottengono delle stime X_m dei campioni della sua trasformata di Fourier $X(f)$ per $f = \frac{m}{N}f_c$, ossia $X_m \simeq f_c X(\frac{m}{N}f_c)$ con $m = 0, 1, \dots, N - 1$;
- calcolando la IDFT (4.11) degli X_m si ri-ottengono i campioni di $x(t)$ di partenza;
- sia gli X_m che gli x_n sono in realtà sequenze periodiche di periodo N ;
- i calcoli indicati dalle (4.8) e (4.11) sono in realtà svolti mediante un diverso algoritmo, chiamato *Fast Fourier Transform* o FFT, che ha il vantaggio di richiedere una complessità $O(N \log_2 N)$ ridotta rispetto a quella della DFT, che è $O(N^2)$.

Se la sommatoria (4.8) venisse applicata, anziché ad un numero finito N di campioni x_n , ad un loro numero infinito, allora

- si otterrebbe una sequenza periodica \tilde{X}_m (4.15) di periodo N , corrispondente al campionamento dello spettro periodico $X^\bullet(f)$;
- l'applicazione della IDFT a $\tilde{X}(k)$ produrrebbe una sequenza periodica \tilde{x}_n (4.16), coincidente con la sequenza originaria x_n solo nel caso in cui questa fosse di durata finita, minore o uguale ad N ;
- segmentando un segnale $x(t)$ in sotto-intervalli disgiunti, si può eseguire la convoluzione tra $x(t)$ ed una $h(t)$ di durata finita (§ 4.5.4), operando esclusivamente nel dominio digitale, e sommando tra loro le IDFT dei prodotti tra la DFT dei campioni di $h(t)$ e le DFT dei segmenti di $x(t)$.

L'interpretazione dei valori che risultano dalla applicazione della DFT su dei campioni di segnale, come stima della trasformata di Fourier del segnale, deve tenere conto oltre che delle fonti di approssimazioni evidenziate nella nota 22, anche dei corretti valori da assegnare alla scala delle frequenze e delle ampiezze, ossia:

4.6.1 Le frequenze della DFT

Occorre tener presente il valore della frequenza di campionamento, e se il segnale di partenza $x(t)$ è reale, della periodicità degli X_m . Infatti i valori X_m per $m = 0, 1, \dots, N - 1$ corrispondono ai campioni di $X^\bullet(f)$ per $f = \frac{m}{N}f_c$, ma se $x(t)$ è reale,

$X^\bullet(f)$ oltre ad essere periodico presenta simmetria coniugata, e dunque per valori $f > \frac{f_c}{2}$, ossia per $m = \frac{N}{2}$, $X^\bullet(f)$ assume valori *speculari* a quelli risultanti per $f < \frac{f_c}{2}$. Per fissare le idee, procediamo con un esempio: se $N = 512$ i primi 256 valori (da 0 a 255, ossia per $m = 0, 1, 2, \dots, N/2 - 1$) sono da mettere in corrispondenza con quelli di $X(f)$ con $f = 0, \frac{1}{N}f_c, \frac{2}{N}f_c, \dots, \frac{N/2-1}{N}f_c$; mentre i restanti 256 valori (da 256 a 511, ossia per $m = N/2, N/2+1, \dots, N-1$, e corrispondenti a $f = \frac{1}{2}f_c, \frac{N/2+1}{N}f_c, \dots, \frac{N-1}{N}f_c$) esibiscono un comportamento speculare a quello dei precedenti.

4.6.2 Le ampiezze della DFT

Come espresso dalla (4.12), i valori X_m rappresentano una approssimazione dei coefficienti della serie di Fourier calcolati sulla finestra temporale da cui provengono i campioni di segnale, e moltiplicati per il numero di campioni utilizzati nel calcolo: $X_m \simeq N \cdot X_m^{SF}$. Pertanto, i valori ottenuti dalla DFT devono essere normalizzati, dividendoli per N .

Capitolo 5

Trasmissione dati in banda base

Dopo che i segnali sono stati convertiti in sequenze numeriche mediante campionamento (Cap. 4), e queste ultime convertite in cifre binarie mediante quantizzazione (vedi § 4.2.1.1), non esiste alcuna differenza formale tra segnali analogici-numerizzati, e dati nativamente numerici, come documenti di computer. Pertanto, in questo capitolo si affrontano gli argomenti legati alla trasmissione dei dati, comprendendo in questo termine entrambi i casi.

5.1 Trasmissione su canale numerico

Al primo capitolo (§ 1.2) abbiamo evidenziato come tra sorgente e destinazione di una trasmissione numerica si possa idealizzare la presenza di un *canale numerico*, che racchiude i dispositivi idonei a svolgere diverse operazioni, in modo da permettere la trasmissione di informazioni numeriche mediante un segnale analogico (indicato come *segnale dati*), trasmesso su canale analogico. Se il canale presenta una risposta in frequenza di tipo *passa-banda*, il segnale dati dovrà presentare le stesse caratteristiche, ed al capitolo 13 saranno illustrati i principi di funzionamento dei dispositivi *modem* necessari a generare tali segnali. Nel caso in cui, invece, il canale analogico sia da considerare *passa-basso*, il *modem* che genera il segnale dati è indicato come *codificatore di linea*.

5.1.1 Trasmissione numerica di banda base

La figura 5.1 rappresenta uno schema generale di canale numerico, con evidenziati i principali elementi che lo compongono. Notiamo innanzitutto come *il ritmo* con cui le

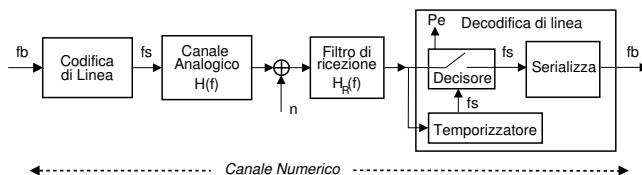
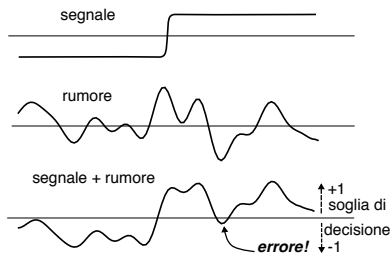


Figura 5.1: Elementi costitutivi di un canale numerico

informazioni numeriche sono emesse dal *codificatore di linea* (§ 5.1.2) sia descritto da una *frequenza di simbolo*¹ f_s potenzialmente inferiore al ritmo binario f_b che è proprio del messaggio². Come vedremo al § 5.1.2.4, il caso tipico per cui deve essere $f_s < f_b$ si verifica quando il canale analogico presenta una risposta in frequenza $H(f)$ limitata in banda, al punto da dover ridurre la banda del segnale trasmesso, fino a farla rientrare in quella del canale, in modo da poter ri-ottenere in uscita dal canale lo stesso segnale trasmesso³. Dal lato ricevente del canale numerico si trova il *decodificatore* di linea, che *ricostruisce* la sequenza delle informazioni trasmesse, prima *campionando* (con ritmo f_s) il segnale dati ricevuto, e quindi *decidendo* quale simbolo fosse rappresentato, in base al confronto tra il valore osservato e delle *soglie di decisione*.

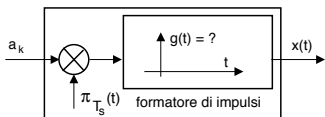
La presenza di un processo di *rumore additivo* $n(t)$ fa sì che il segnale preso in esame dal decisore possa superare la (o le) soglia di decisione, determinando un *errore* (con probabilità P_e) a riguardo di quale simbolo sia stato trasmesso. La valutazione di tale probabilità viene svolta al § ??, ma anticipiamo subito che P_e è tanto maggiore quanto più è grande la potenza (e dunque l'ampiezza) del rumore in ingresso al decisore, che può essere minimizzata dal filtro $H_R(f)$ di ricezione (vedi § 12.1.1); inoltre, $H_R(f)$ può realizzare un *filtro adattato* (§ 9.5), oppure ancora un dispositivo di *equalizzazione* (§ 14.6).



Lo schema di fig. 5.1 mostra inoltre come il decodificatore di linea svolga anche una funzione di *temporizzazione*, per consentire al dispositivo decisore di operare al passo con il ritmo f_s dei simboli in arrivo; al § 5.6 sono descritte alcune tecniche per affrontare questo aspetto. Infine, nel caso in cui ogni simbolo sia rappresentativo di più di un bit (vedi § 5.1.2.4), è evidenziata la presenza di un *serializzatore* che provvede ad emettere uno dopo l'altro i bit corrispondenti ai diversi simboli ricevuti.

5.1.2 Codifica di linea e segnale dati

Consideriamo una sorgente discreta che produce, ad una frequenza f_s , simboli a_k corrispondenti a valori numerici: possiamo allora descrivere il *segnale dati* $x(t)$ uscente dal *codificatore di linea* come un segnale analogico



$$x(t) = \sum_k a_k \cdot g(t - kT_s) \quad (5.1)$$

Costruzione del SEGNALE DATI

nella figura a lato⁴, in cui $\pi T_s(t)$ è un treno di impulsi (§ 3.8.1) con periodo $T_s = 1/f_s$.

e idealizzare il codificatore stesso come illustrato

¹Possiamo per ora pensare i *simboli* come se fossero le lettere dell'alfabeto di un messaggio scritto, ad ognuna delle quali si fa corrispondere un valore (o *livello*) di segnale.

² T_s è il periodo di simbolo, ed il suo inverso $f_s = 1/T_s$ è detto *frequenza di simbolo* (o *baud-rate*), detta anche *frequenza di segnalazione*, e si misura in simboli/secondo, detti appunto *baud*, in memoria di ÉMILE BAUDOT, vedi http://it.wikipedia.org/wiki/Codice_Baudot.

³Se non fosse preso questo provvedimento, e si trasmettesse un segnale con una occupazione spettrale maggiore della banda passante del canale, nel segnale ricevuto verrebbero a mancare alcune componenti frequenziali, e di conseguenza la forma d'onda del segnale risulterebbe modificata, causando così il fenomeno di *interferenza tra simboli* (vedi § 5.1.2.2).

⁴Nel caso in cui a_k assuma valori discreti in un alfabeto ad L livelli, codificati con $M = \lceil \log_2 L \rceil$ cifre binarie (*bit*) (il simbolo $\lceil \cdot \rceil$ rappresenta l'intero superiore), la trasmissione convoglia una *frequenza binaria* di valore pari a $f_b \left[\frac{\text{bit}}{\text{secondo}} \right] = M \left[\frac{\text{bit}}{\text{simbolo}} \right] \cdot f_s \left[\frac{\text{simboli}}{\text{secondo}} \right]$.

Notiamo come tale figura sia del tutto simile al s&H introdotto al § 4.1.4, tranne che ora $g(t)$ è generico. Come sarà approfondito al § 9.2.4, si può mostrare che (sotto condizioni abbastanza generali) lo spettro di densità di potenza di $x(t)$ può essere espresso come

$$\mathcal{P}_x(f) = \sigma_A^2 \frac{\mathcal{E}_G(f)}{T_s} \quad (5.2)$$

in cui $\mathcal{E}_G(f)$ rappresenta lo spettro di densità di energia di $g(t)$, che quindi determina l'andamento di $\mathcal{P}_x(f)$, mentre σ_A^2 è la varianza dei valori a_k , e dunque ne rispecchia la dinamica.

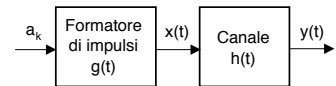
Onda PAM Il segnale dati (5.1) è a volte indicato come *onda* PAM, forzando la semantica di *Pulse Amplitude Modulation* per indicare la modulazione delle ampiezze degli impulsi che realizzano un segnale periodico nella forma di eq. (3.10) (pag. 45) in base ai valori a_k . Un ulteriore punto di vista è illustrato al § 6.9.5.

5.1.2.1 Segnale binario e onda rettangolare

Come applicazione della (5.2) osserviamo che, nel caso in cui si effettui una trasmissione binaria (ossia $a_k = \{0, 1\}$ ed $f_s = f_b$) e si scelga un impulso dati rettangolare $g(t) = \text{rect}_\tau(t - \frac{\tau}{2})$ (con $\tau \leq T_b$), allora la densità spettrale di $x(t)$ presenta⁵ un andamento $\text{sinc}^2(f\tau)$, con il primo zero per $f = \frac{1}{\tau}$, ed occupazione di banda (approssimativamente) pari ad alcuni multipli di tale valore⁶.

5.1.2.2 Effetto della limitazione in banda e ISI

Qualora il segnale dati $x(t)$ attraversi un canale con risposta impulsiva $h(t)$, in uscita si presenta un nuovo segnale



$$y(t) = \sum_k a_k \cdot g'(t - kT_s) \quad \text{in cui} \quad g'(t) = g(t) * h(t)$$

⁵Che questo sia il caso, può essere verificato per alcuni segnali che abbiamo studiato o studieremo:

1. Segnale campionato. In questo caso $a_k = s(kT_c)$ sono i campioni di segnale, ed abbiamo visto che $x^\circ(t)$ ha spettro periodico in frequenza, con un inviluppo di ampiezza dato da $\text{sinc}(f\tau)$;
2. Segnale periodico. Ponendo $a_k = \pm 1$ si genera un'onda quadra, il cui spettro è a righe con lo stesso inviluppo indicato al punto precedente;
3. Segnale dati. Se a_k sono variabili aleatorie statisticamente indipendenti, al § 9.2.4 si dimostrerà che $X(f)$ è di tipo continuo, con inviluppo ancora pari a $\text{sinc}(f\tau)$.

⁶Nella tabella che segue è riportata l'occupazione di banda necessaria a contenere 10 lobi del $\text{sinc}(f\tau) = \frac{1}{\tau} \mathcal{F}\{\text{rect}_\tau(t)\}$, per τ pari al periodo di simbolo T_s , in modo da dare un'idea delle specifiche necessarie al canale: osserviamo allora che il rettangolo può andare bene a basse velocità di trasmissione, infatti già per 10 Msimboli/sec, velocità di una LAN (*Local Area Network*, ossia una rete "locale" tra computer in uno stesso edificio), occorrono 100 MHz di banda.

f_s	apparato	T_s	$1/T_s$	$10/T_s$
$2.4 \cdot 10^3$	Modem (anni '80)	$4.2 \cdot 10^{-3}$	$2.4 \cdot 10^3$	24 KHz
$28.8 \cdot 10^3$	Modem (anni '90)	$3.5 \cdot 10^{-5}$	$28.8 \cdot 10^3$	288 KHz
$10 \cdot 10^6$	LAN	10^{-7}	10^7	100 MHz

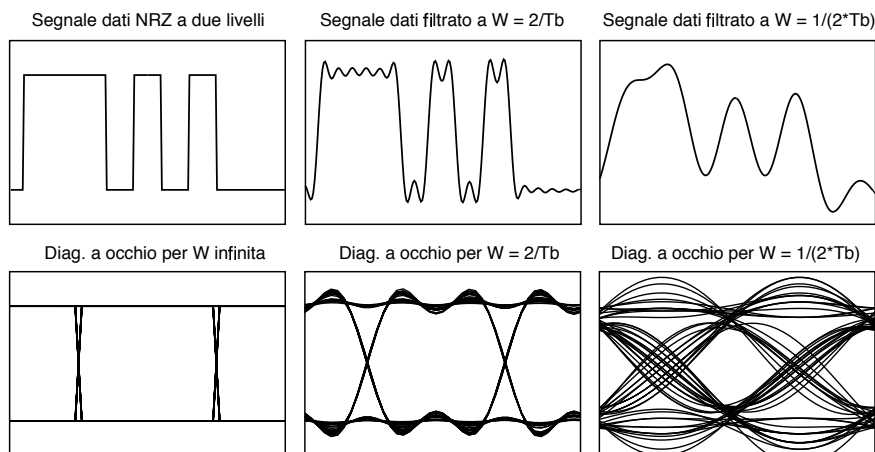


Figura 5.2: Segnale dati a banda infinita e limitata, e relativo diagramma ad occhio

come mostrato alla nota⁷. L'effetto della convoluzione tra $g(t)$ ed $h(t)$, è quello di *disperdere* nel tempo la forma d'onda $g(t)$, che quindi anche se delimitata entro un periodo di simbolo, *invade* gli intervalli temporali riservati ai simboli adiacenti, dando luogo al fenomeno della *interferenza intersimbolica* (ISI, *InterSymbolic Interference*).

5.1.2.3 Diagramma ad occhio

In Fig. 5.2 è riportato un esempio di segnale dati binario ad onda rettangolare con base pari a T_b , a cui è applicata una limitazione di banda mediante un filtro passa-basso ideale. Nella riga superiore è mostrato l'andamento del segnale, mentre alla riga inferiore si riporta un grafico noto come *diagramma ad occhio* (EYE DIAGRAM)⁸, che si ottiene visualizzando il segnale dati mediante un oscilloscopio con la base dei tempi sincronizzata al periodo di simbolo. Come si può vedere, in presenza di una limitazione di banda, il valore del segnale dati in corrispondenza di un determinato simbolo viene a dipendere anche dal valore dei simboli circostanti.

Riservandoci di riprendere l'argomento nel seguito, notiamo che il problema *non* si presenta qualora:

- la frequenza di simbolo sia molto inferiore alla banda del canale, *ovvero*
- la risposta impulsiva $h(t)$ abbia estensione temporale molto inferiore a T_s .

Nel seguito della sezione, assumiamo vere queste ipotesi.

$$\begin{aligned}
 {}^7 y(t) &= \left[\sum_k a_k \cdot g(t - kT_s) \right] * h(t) = \left[g(t) * \sum_k a_k \cdot \delta(t - kT_s) \right] * h(t) = \\
 &= g(t) * h(t) * \sum_k a_k \cdot \delta(t - kT_s) = g'(t) * \sum_k a_k \cdot \delta(t - kT_s) = \\
 &= \sum_k a_k \cdot g'(t - kT_s)
 \end{aligned}$$

⁸Il nome deriva dalla forma del diagramma, che in corrispondenza del *centro* degli impulsi mostra una apertura simile per l'appunto ad un occhio, e la cui analogia apparirà più evidente in seguito all'adozione di un impulso $g(t)$ a banda limitata (figura 5.5), ed in presenza di rumore (figura 5.7).

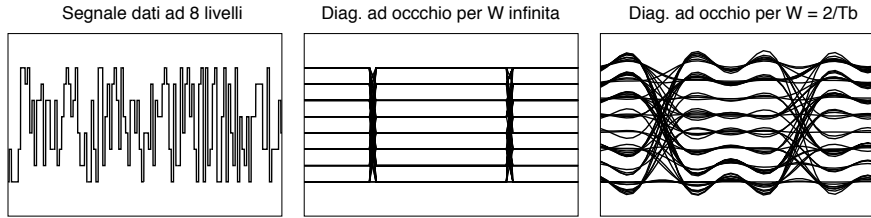
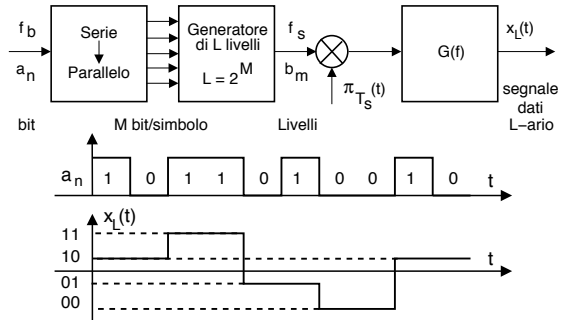


Figura 5.3: Segnale dati ad 8 livelli e diagramma ad occhio per banda infinita o limitata

5.1.2.4 Trasmissione multilivello

Nel caso in cui la banda a disposizione per la trasmissione sia insufficiente, una soluzione di semplice attuazione è quella di ricorrere ad una trasmissione non più *binaria*, ma che impieghi L possibili diversi simboli, o *livelli*⁹. A tale scopo, occorre raggruppare M bits del messaggio a_n (che arrivano a velocità f_b bits/sec) in una unica parola binaria.

Scegliendo $M = \log_2 L$, occorrono $T_s = MT_b$ secondi per accumulare M bit, ed emettere uno tra $L = 2^M$ possibili valori b_m , usati quindi come ampiezze degli impulsi (generati a ritmo T_s) necessari a produrre un segnale dati¹⁰ $x(t)$ ad L livelli, e caratterizzato da una frequenza di simbolo $f_s = 1/T_s = 1/MT_b = f_b/M$. La figura a lato mostra il segnale dati binario ad onda rettangolare, e quello che invece si ottiene con $L = 4$.



L'aumento del periodo di simbolo T_s corrisponde ad un aumento della durata di $g(t)$, e di conseguenza ad una *contrazione* di $G(f)$, determinando quindi una eguale riduzione della banda occupata da $x(t)$, come si verifica tenendo conto della (5.2).

Pertanto, l'occupazione di banda del segnale dati può essere ridotta a piacere, semplicemente aumentando il numero M di bit raggruppati in una singola parola. D'altra parte al § 5.3 si mostra come, a meno che la potenza del segnale non venga aumentata, si assisterà ad un peggioramento della probabilità di errore del ricevitore, in quanto a parità di potenza e quindi di ampiezza, i livelli risultano ora ravvicinati. Questo fenomeno è raffigurato mediante l'esempio di figura 5.3 dove a sinistra è mostrato un segnale dati ad 8 livelli, al centro il diagramma ad occhio corrispondente, ed a destra una versione del segnale limitata in banda.

⁹Proseguiamo l'esposizione riferendoci direttamente al termine *livelli*, indicando con questo la scelta tra L possibili valori di ampiezza per il segnale trasmesso.

¹⁰In ricezione si opererà il processo inverso, ripristinando la codifica binaria originaria a cui il codificatore ha associato il valore L -ario ricevuto, e *serializzando* tale parola ad M bit, in modo da ri-ottenere la sequenza binaria di partenza.

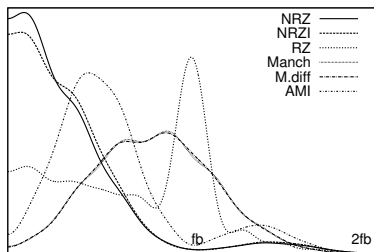
5.2 Generazione del segnale dati

Facendo di nuovo riferimento allo schema di generazione del segnale dati proposto al § 5.1.2, svolgiamo le considerazioni relative alla scelta della forma d'onda elementare $g(t)$.

5.2.1 Codici di linea a banda infinita

Come anticipato, lo spettro di densità di potenza $\mathcal{P}_X(f) = \sigma_A^2 \frac{\mathcal{E}_G(f)}{T_s}$ di un segnale dati ha andamento che dipende direttamente da quello dello spettro di densità di energia $\mathcal{E}_G(f)$ della risposta impulsiva $g(t)$ usata nel formatore di impulsi, e dunque nel caso in cui $g(t) = \text{rect}_\tau(t)$ si ottiene che $\mathcal{P}_X(f)$ ha andamento di tipo $\text{sinc}^2(f\tau)$, che come noto si estingue come $1/f^2$, con il primo zero per $f = 1/\tau$. Nel caso in cui si operi a bassa velocità (ossia con τ sufficientemente grande), si può considerare come se il canale fosse a *banda infinita*, e quindi capace di riprodurre il segnale inalterato.

La figura seguente mostra lo spettro di densità di energia $\mathcal{E}_G(f)$ corrispondente alla scelta di alcune delle forme d'onda $g(t)$ discusse nel seguito, e mostrato per frequenze



fino al doppio di f_b . Gli andamenti riportati in figura sono ottenuti generando i valori (0 o 1) per 200 simboli binari a_k in modo pseudo-casuale¹¹, campionando il segnale dati (5.1) con 16 campioni per periodo di bit, e calcolando una FFT smussata sui campioni così ottenuti. Ogni particolare $g(t)$ dà origine alla definizione di un *codice di linea* corrispondente, usato nella pratica per trasmettere informazioni di natura binaria. Elenchiamo di

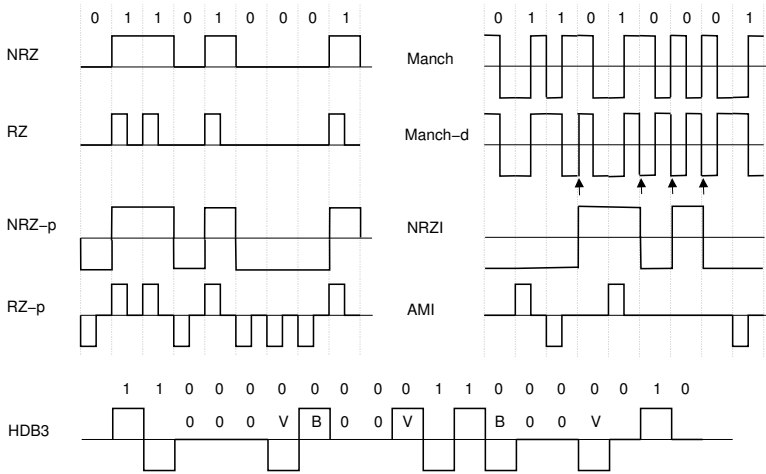
seguito caratteristiche e proprietà di tali segnali, aiutandoci con gli esempi riportati in figura.

Codici unipolari Sono associati a segnali sbilanciati, e presentano un valore nullo o diverso da zero per i due livelli logici 0 ed 1.

NRZ - No Return to Zero La sigla NRZ individua il fatto che il segnale “non torna a zero” per tutto il periodo di bit, essendo $g(t) = \text{rect}_{T_b}(t)$; pertanto ha spettro di tipo $\text{sinc}(fT_b)$, con il primo zero a $f = 1/T_b$, e presenta una componente continua. Rimane costante per dati costanti e ciò complica la sincronizzazione del clock del ricevitore, data l’assenza in questo caso di transizioni; la presenza di uno zero ad $f = 1/T_b$ aggrava inoltre la situazione.

RZ - Return to Zero In questo caso l’impulso $g(t)$ ha durata pari a $T_b/2$, il segnale presenta (a parità di ampiezza) minore energia di NRZ, mentre lo spettro presenta una componente pronunciata esattamente a frequenza f_b , agevolando la sincronizzazione sul bit ma occupando una banda maggiore. Il segnale si mantiene però costante per lunghe sequenze di *zeri*.

¹¹La non perfetta indipendenza statistica dei simboli prodotti dal generatore di numeri casuali di un computer si può riflettere su di una ridotta generalità del risultato mostrato, che tuttavia rispecchia molto bene i casi reali.



Codici bipolari Usano segnali bilanciati, e sono ricevuti mediante uno stadio di ingresso differenziale, riducendo la sensibilità al rumore. In funzione del tipo di codice, è possibile garantire l'assenza di una componente continua nel segnale.

NRZ polare, RZ polare Realizzano l'impulso con polarità negativa quando associato ad un bit pari a zero, e presentano media nulla solo se i valori 0 ed 1 sono equiprobabili. RZ polare non è mai costante, facilitando il compito della sincronizzazione.

Manchester o Diphase NRZ Realizza una codifica *di fase*, in quanto usa un impulso RZ a piena dinamica, in salita od in discesa, corrispondentemente ai bit 1 e 0. Per questo motivo il segnale risulta sempre a media nulla. L'occupazione spettrale è intermedia tra il caso NRZ ed RZ, dato che la durata media dell'impulso può essere T_b o $T_b/2$. L'uso del codice Manchester è prescritto dallo standard IEEE 802.3 per le LAN a bus con contesa di accesso CDMA/CD (vedi § 8.6.1.6).

Alternate Mark Inversion (AMI) Codifica gli 1 con polarità alternate, mediante un impulso $g(t)$ rettangolare di estensione T_b o $T_b/2$, e gli zeri con assenza di segnale, garantendo assenza di valore medio. Qualora la logica sia invertita (impulsi per gli zeri, e silenzio per gli uni) il codice prende il nome di *pseudo-ternario*. Da un punto di vista spettrale, l'AMI esibisce una occupazione di banda¹² ridotta rispetto a RZ, per via dei periodi silenti corrispondenti agli zeri. Se il periodo silente è prolungato, l'assenza di transizioni può compromettere la sincronizzazione di bit, e per questo motivo sono stati definiti ulteriori codici derivati, come ad esempio

HDB3 E' utilizzato per trasmettere il segnale PCM a 2 Mbps (vedi § 6.3.1), e l'acronimo significa *High-Density Bipolar-3-zeroes*. Come per AMI, rappresenta gli uni con polarità alternate, ma rimpiazza le sequenze di quattro zeri consecutivi forzando

¹²La densità spettrale mostrata in figura è relativa all'uso di una $g(t)$ di tipo RZ.

una *violazione* della regola sull'ultimo bit dei quattro, in modo che il ricevitore, rilevando la violazione, è in grado di riportare il bit a zero. Dato però che la presenza della violazione creerebbe la comparsa di una componente continua nel segnale, sono inseriti anche dei bit di *bilanciamento*, per rimuovere quest'ultima. I bit di bilanciamento si collocano al posto del primo dei quattro zeri, e la loro polarità è scelta in modo che la sequenza delle violazioni abbia una polarità alternata; in definitiva, dopo la prima violazione, si usa sempre anche il bit di bilanciamento.

Codici differenziali Sono ancora di tipo bipolare, e la forma d'onda non è più legata al solo valore dei bit, ma anche alla loro relazione temporale (vedi anche § 13.4). Ciò permette di risolvere l'ambiguità presente qualora si scambino le polarità degli estremi del collegamento.

NRZI Deriva dall'NRZ, e la **I** sta per *Inverted*. Ora il livello del segnale permane nello stesso stato per i bit pari ad uno, e cambia stato per i bit pari a zero. L'assenza di valor medio è legata alla statistica che descrive le sequenze di uni e dunque non può essere garantita, mentre permangono i problemi legati alla sincronizzazione. La ridotta occupazione spettrale lo rende però interessante.

Manchester Differenziale Usa un impulso RZ a piena dinamica come per Manchester, ma la polarità risulta invertita rispetto al bit precedente se il nuovo bit è uno, mentre è mantenuta nel caso arrivi un zero (in corrispondenza delle frecce); pertanto, in presenza degli uni non si verifica transizione al confine tra i periodi di bit. Questa soluzione è utilizzata nel contesto dello standard IEEE 802.5 per LAN *Token Ring*. L'occupazione spettrale è simile a quella osservabile per la codifica Manchester.

5.2.2 Segnale dati limitato in banda

Analizziamo i requisiti necessari a realizzare la trasmissione di un segnale dati che impegni una banda strettamente limitata, evidenziando le condizioni da realizzare, e descrivendo le alternative di progetto.

5.2.2.1 Requisiti per l'impulso di trasmissione

Limitazione di banda Abbiamo già osservato (vedi § 5.1.2) che (sotto opportune condizioni) la densità di potenza del segnale dati (vedi 5.1) risulta pari a $\mathcal{P}_x(f) = \sigma_A^2 \frac{|G(f)|^2}{T}$, e che la sua trasmissione inalterata è possibile solo disponendo di un canale che presenti una risposta in frequenza di tipo *passa-tutto* (ossia con modulo costante e fase lineare) nella banda di frequenze occupata da $G(f)$.

Nel caso in cui, ad es., si adotti $g(t) = \text{rect}_\tau(t)$, allora $G(f) = \tau \text{sinc}(f\tau)$, con $\tau \leq T_s$; pertanto, il primo passaggio per zero è a frequenza $\frac{1}{\tau} \geq \frac{1}{T_s}$, e la densità di potenza $\mathcal{P}_x(f)$ può essere trascurata solo dopo qualche multiplo di tale valore.

Se invece il canale sopprime contenuti frequenziali non trascurabili dal segnale dati (5.1), allora gli impulsi $g(t)$ *si deformano* (vedi § 5.1.2.2) e non sono più rettangolari; in particolare, possono estendersi per una durata maggiore di T_s , causando problemi di *interferenza tra simboli*¹³ (ISI).

¹³Come mostrato al § 5.1.2.2, il segnale dati filtrato è basato su impulsi $g'(t) = g(t) * h(t)$, con una durata pari alla somma delle durate di $g(t)$ e $h(t)$. Pertanto, l'impulso $a_n g(t - nT_s)$ si estende

Limitazione nel tempo Il problema della limitazione di banda potrebbe essere risolto adottando, in linea di principio, un impulso elementare di tipo $g(t) = \text{sinc}\left(\frac{t}{T_s}\right)$, che ha trasformata $G(f) = T_s \cdot \text{rect}\frac{1}{T_s}(f)$ strettamente limitata in banda, con frequenza massima $W = \frac{1}{2T_s}$, e che non subisce alterazioni purché il canale abbia un comportamento “passa tutto” in tale ristretto intervallo di frequenze. Notiamo che questa $g(t)$ passa da zero per $t = nT_s$, e pertanto non provoca interferenza tra i simboli collocati agli istanti nT_s , come verificabile notando che in tal caso l'espressione (5.1) risulta del tutto simile alla (4.1) relativa alla ricostruzione di una segnale campionato; ora però *non siamo interessati* al valore del segnale negli istanti che intercorrono tra quelli in cui sono centrati i simboli, mentre desideriamo unicamente recuperare i singoli valori originali, che troviamo *in modo esatto* agli istanti $t = nT_s$.

Lo svantaggio di adottare una forma d'onda $g(t)$ limitata in frequenza, è che il suo andamento è illimitato nel tempo, e dunque $g(t)$ può essere realizzata solo in modo approssimato¹⁴.

Limitazione di precisione Abbiamo appena mostrato come, adottando una $g(t) = \text{sinc}\left(\frac{t}{T_s}\right)$, si evita l'interferenza tra simboli, purché i campioni vengano prelevati *esattamente* agli istanti nT_s (15). Infatti, al di fuori di tali istanti il valore del segnale dipende dal valore delle code degli impulsi $g(t)$ centrati sugli altri simboli.

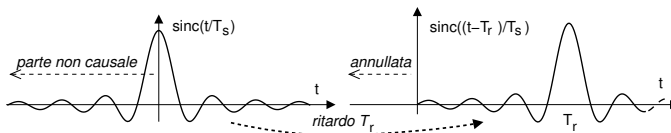
L'orologio (*clock*) del ricevitore, però, non ha una precisione infinita, e gli istanti di campionamento saranno affetti da errori di fase. Pertanto, è interessante ricercare una soluzione per $g(t)$ che, anche in presenza di errori di precisione nella determinazione degli istanti di campionamento, dia luogo ad errori quanto più ridotti possibile.

Riepilogando: vorremmo soddisfare contemporaneamente le esigenze:

1. Occupare una banda contenuta
2. Ricorrere ad un filtro poco complesso
3. Ridurre la sensibilità agli errori di campionamento

a valori di $t > (n+1)T_s$, e quindi $x((n+1)T_s) = a_{n+1}g(0) + a_n g(T_s)$, introducendo un errore pari a $a_n g(T_s)$, detto appunto *interferenza tra simboli*.

¹⁴Un sistema fisico che debba realizzare una risposta impulsiva $g(t) = \text{sinc}\left(\frac{t}{T_s}\right)$, non può presentare $g(t) \neq 0$ per $t < 0$: questo equivarrebbe infatti ad un sistema in grado di produrre una uscita *prima ancora* che sia applicato un segnale al suo ingresso. Se $g(t)$ ha estensione temporale illimitata, occorre ricorrere ad una versione ritardata e limitata $g'(t) = \begin{cases} g(t - T_R) & \text{con } t \geq 0 \\ 0 & \text{altrimenti} \end{cases}$. Se $T_R \gg T_s$, l'entità dell'approssimazione è accettabile, ed equivale ad un semplice ritardo pari a T_R ; d'altro canto, quanto maggiore è la durata della risposta impulsiva, tanto più difficile (ossia costosa) risulta la realizzazione del filtro relativo.



¹⁵Al contrario, se $g(t) = \text{rect}_{T_s}(t)$, il campionamento può avvenire ovunque nell'ambito del periodo di simbolo, ma si torna al caso di elevata occupazione di banda.

Per i punti 2 e 3, è sufficiente adottare $g(t)$ di tipo rettangolare, generando un segnale dati del tipo $x(t) = \sum_k a_k \cdot \text{rect}_\tau(t - kT_s)$, che ha lo svantaggio di occupare una banda infinita, e quindi la sua ricezione *inalterata* è possibile solo per canali ideali.

Prima di esporre una soluzione di compromesso a tutti e 3 i problemi, consideriamo che la trasmissione di $x(t)$ su di un canale non ideale $H(f)$ determina un effetto di cui al § 5.1.2.2 abbiamo tenuto conto considerando di ricevere un segnale dati caratterizzato da un impulso $g'(t) = g(t) * h(t)$. Pertanto, nel caso in cui $H(f)$ sia noto a priori, i risultati che troveremo (validi per $g'(t)$) individueranno in realtà un formatore di impulsi con $G(f) = \frac{G'(f)}{H(f)}$.

5.2.2.2 Condizioni di Nyquist

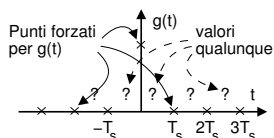
Individuiamo con questo termine i requisiti formali che devono essere soddisfatti dagli impulsi usati per generare il segnale dati.

Nel tempo Torniamo a riferirci alla (5.1) per osservare che, affinché $x(t)|_{t=nT_s}$ dipenda dal solo valore a_n e non dagli altri a_k con $k \neq n$, deve risultare

$$g(t) = \begin{cases} 1 & \text{se } t = 0 \\ 0 & \text{se } t = mT_s \text{ con } m \neq 0 \\ \forall & \text{altrove} \end{cases} \quad (5.3)$$

e cioè $g(t)$ deve passare da zero in tutti gli istanti multipli di T_s , tranne che per $t = 0$ dove deve valere 1, mentre per valori di t intermedi può assumere qualunque valore. In tal caso infatti dalla (5.1) si ottiene:

$$x(nT_s) = \sum_k a_k \cdot g(nT_s - kT_s) = \sum_k a_k \cdot g((n - k)T_s) = a_n$$



Le condizioni (5.3) prendono il nome di *condizioni di Nyquist per l'assenza di interferenza intersimbolo* (ISI, *Inter Symbol Interference*) nel dominio del tempo. Se una forma d'onda $g(t)$ soddisfa tali condizioni, allora viene detta *impulso di Nyquist*¹⁶.

In frequenza Dalle condizioni di Nyquist *nel tempo* (5.3) se ne derivano altre *in frequenza*, mediante i seguenti passaggi. Moltiplicando $g(t)$ per un treno di impulsi $\pi_{T_s}(t) = \sum_k \delta(t - kT_s)$, si ottiene

$$g(t) \cdot \pi_{T_s}(t) = \delta(t)$$

dato che $g(nT_s) = 0$ e $g(0) = 1$. Trasformando (vedi § 3.8.3) si ottiene:

$$1 = G(f) * \frac{1}{T_s} \cdot \Pi_{\frac{1}{T_s}}(f) = G(f) * \frac{1}{T_s} \cdot \sum_k \delta\left(f - k\frac{1}{T_s}\right)$$

Indicando con $f_s = \frac{1}{T_s}$ la frequenza di simbolo, ed eseguendo le convoluzioni tra $G(f)$ e gli impulsi centrati in $f = kf_s$, risulta infine

$$\sum_k G(f - kf_s) = T_s \quad (5.4)$$

¹⁶ Ad esempio, l'impulso rettangolare è di Nyquist, in quanto $\text{rect}_{T_s}(t) = \begin{cases} 1 & \text{se } t < \frac{T_s}{2} \\ 0 & \text{se } t = kT_s \end{cases}$.

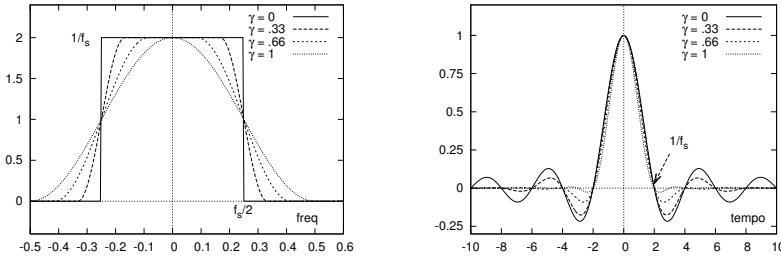
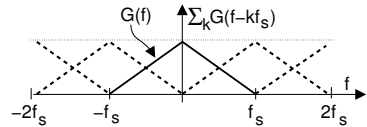


Figura 5.4: Filtro a coseno rialzato e impulso di Nyquist per $f_s = 0.5$, variando γ

che rappresenta la condizione *in frequenza* per l'assenza di interferenza intersimbolo. Il risultato ottenuto si interpreta considerando che una qualunque $G(f)$ va bene purché, se sommata con le sue repliche traslate di multipli di f_s , dia luogo ad una costante, ovvero $G(f)$ manifesta simmetria dispari rispetto ad $f_s/2$. In tal caso $G(f)$ può essere indicato come *filtro di Nyquist*. Notiamo che, seppure $G(f)$ possa essere qualsiasi, anche non limitata in banda, il nostro interesse è appunto per le $G(f)$ limitate in banda, come quella triangolare dell'esempio a lato.



5.2.2.3 Filtro a coseno rialzato

Individua una famiglia parametrica di filtri di Nyquist limitati in banda, e viene detto a *coseno rialzato* in quanto è composto da 2 archi di coseno, raccordati da una retta (vedi Fig. 5.4). La banda occupata ha espressione

$$B = \frac{f_s}{2} (1 + \gamma) \quad (5.5)$$

in cui γ è chiamato *coefficiente di roll-off*¹⁷, è compreso tra 0 e 1, e rappresenta un indice di dispersione del ramo di coseno. La banda di $G(f)$ varia quindi da un minimo (per $\gamma = 0$) pari a $B = f_s/2$, con $G(f)$ rettangolare, ad un massimo (per $\gamma = 1$) pari a $B = f_s$, con $G(f)$ che è proprio un periodo di coseno.

Il caso di *banda minima* si ottiene per $\gamma = 0$, che corrisponde a $G(f) = \frac{1}{f_s} \text{rect}_{f_s}(f)$, ovvero ad una $g(t) = \text{sinc}(f_s t)$, come già discusso a pag. 81 al § sulla limitazione nel tempo. Occupare una banda inferiore a quella minima non è possibile, perché in tal caso non sarebbero verificate le condizioni di Nyquist in frequenza, in quanto nella $\sum_k G(f - kf_s)$ resterebbero dei “buchi”.

Abbiamo già osservato alla nota (14) a pagina 81 come la realizzazione di $G(f)$ a *banda minima* sia complessa, e possa introdurre notevole ISI in presenza di errori negli istanti di campionamento. La situazione però migliora decisamente usando $\gamma > 0$, con γ via via più grande. Se $\gamma \neq 0$, per $g(t)$ si può ottenere l'espressione

$$g(t) = \text{sinc}(tf_s) \cdot \frac{\cos \gamma \pi t f_s}{1 - (2\gamma t f_s)^2} \quad (5.6)$$

¹⁷Il termine ROLL-OFF può essere tradotto come “rotola fuori”.

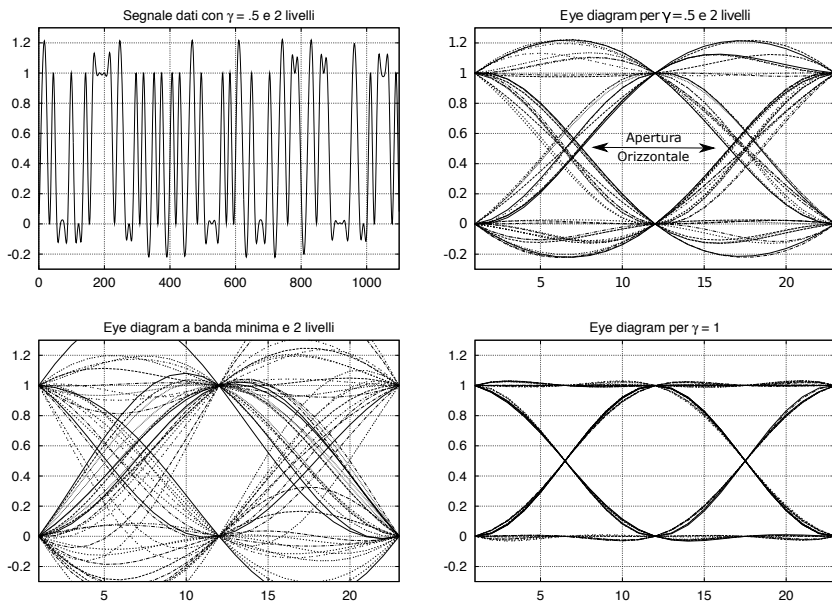


Figura 5.5: Segnale dati ed eye diagram per diversi valori di roll-off

a cui corrisponde una forma d'onda simile al $\frac{\sin(x)}{x}$, ma che va a zero molto più rapidamente, come verificabile osservando la parte destra di Fig. 5.4. Pertanto, se $\gamma \rightarrow 1$ ogni singola onda $g(t - kT_s)$ estenderà la sua influenza ad un numero di impulsi limitrofi *molto ridotto* rispetto al caso $\gamma = 0$, in quanto le oscillazioni sono molto più smorzate, e dunque il termine di errore di ampiezza in presenza di un errore di istante di campionamento è ridotto, dato che dipende da un minor numero di impulsi limitrofi.

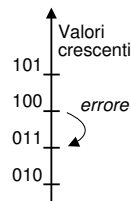
La fig 5.5 mostra l'andamento del segnale dati con $g(t)$ fornito dalla (5.6), calcolata per $\gamma = 0.5$, e per a_k a due valori, pari a 0 e 1. Notiamo che, al difuori degli istanti caratteristici $t = kT_s$ il segnale può assumere valori arbitrari, anche esterni alla dinamica degli a_k .

La rappresentazione fornita dal diagramma ad occhio, sempre mostrata in fig 5.5, permette di valutare meglio la precisione di temporizzazione che è necessaria ad evitare ISI, e che è pari a metà della *apertura orizzontale* dell'occhio. Gli ultimi due diagrammi permettono il confronto con i casi rispettivamente a banda minima e massima (per $\gamma = 1$), evidenziando l'influenza del roll-off.

5.2.2.4 Codice di Gray

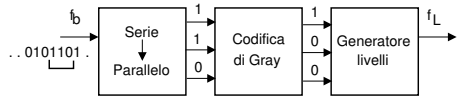
Si tratta di un accorgimento che consente di ridurre l'effetto degli errori commessi dal decisore (vedi § 5.3) in presenza di trasmissione multilivello (§ 5.1.2.4).

Per fissare le idee, supponiamo che il valore *analogico* dei livelli presenti nel segnale dati sia in corrispondenza lineare con il valore *numerico* dell'uscita del convertitore serie-parallelo (vedi fig. a pagina 77), ovvero che i livelli siano "numerati in binario", e livelli



contigui rappresentino configurazioni di bit “in sequenza naturale”, come mostrato in figura. Allora, se trasmettiamo ad esempio il livello associato a 100, ed il decisore commette l’errore¹⁸ di ritenere di aver ricevuto il livello contiguo, che rappresenta la sequenza 011, abbiamo tutti e tre i bit sbagliati!

Il codice di Gray consiste quindi in una tabella di conversione che sostituisce ai bit uscenti dal convertitore serie-parallelo, una diversa configurazione di bit. Possiamo immaginare l’operazione come quella di un accesso a memoria, in cui la parola originaria costituisce l’indirizzo, per mezzo del quale si individua la parola codificata da trasmettere al suo posto. La conversione è biunivoca (a partire dal codice si risale alla parola originaria), e le parole del codice di Gray hanno la proprietà di rappresentare i livelli di segnale contigui come configurazioni di bit che differiscono solo in una cifra binaria (ossia in un bit).



Con riferimento alla tabella, osserviamo che (ad esempio) per trasmettere la sequenza 110 di ingresso, si usa il livello numero 100, ossia il quarto (partendo da zero), lo stesso dell’esempio precedente; se il decisore sbaglia e ritiene di aver ricevuto il terzo livello (011, stesso errore precedente), a questo il decodificatore di Gray associa la sequenza 010, che infatti differisce dall’originale per un solo bit (il primo).

Ingresso binario	Uscita (livello)
100	111
101	110
111	101
110	100
010	011
011	010
001	001
000	000

Il procedimento illustrato, in presenza di un errore sul simbolo, produce un solo bit errato. Ciò comporta che la probabilità di osservare un bit errato è pari a $P_e^b = P_e^s/M$, con M pari al numero di bit/simbolo. Infatti

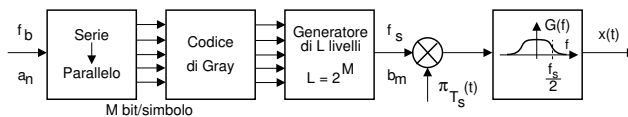
$$P_e^b = \frac{\text{N.Bit errati}}{\text{N.Bit totali}} = \frac{\text{N.Simboli errati}}{M \cdot \text{N.Simboli}} = P_e^s \frac{1}{M}$$

Riassumendo La figura che segue mostra la sequenza generale delle operazioni da intraprendere per generare un segnale dati multilivello, con codifica di Gray, e con caratteristica a coseno rialzato. Notiamo dunque che mentre alla sequenza binaria a_n compete una velocità di f_b bit/secondo, la sequenza multilivello b_m possiede invece un ritmo di

$$f_s = \frac{f_b}{M} = \frac{f_b}{\log_2 L}$$

simboli/secondo, ed il segnale dati risultante $x(t)$ occupa una banda a frequenze positive (vedi eq. 5.5)

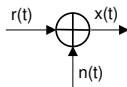
$$B = \frac{f_s (1 + \gamma)}{2} = \frac{f_b (1 + \gamma)}{2 \log_2 L} \tag{5.7}$$



¹⁸L’errore è causato dal rumore additivo che, sommandosi al segnale ricevuto (vedi Fig. 5.1), ne modifica il valore, fornendo al decisore valori diversi da quelli trasmessi.

5.3 Errori nelle trasmissioni numeriche di banda base

Fin qui abbiamo trascurato di prendere in considerazione gli effetti del rumore additivo presente in ricezione, a cui si è pur accennato al § 5.1.1, e che causa la ricezione di un segnale $x(t) = r(t) + n(t)$, e cioè oltre al segnale effettivamente ricevuto $r(t)$, nel ricevitore entra anche un diverso segnale $n(t)$ sommato ad $r(t)$, indicato come disturbo, o rumore (*noise*¹⁹), la cui forma d'onda è una *realizzazione* di un processo aleatorio stazionario ed ergodico (vedi § 7.3), la cui densità di probabilità del primo ordine è la “famosa” *gaussiana* (vedi § 7.2.4), ed il cui spettro di densità di potenza è *bianco*, ossia costante in frequenza. Nel caso in cui siano presenti più cause di disturbo, anche localizzate in punti diversi del collegamento, si fa in modo (vedi pag. 350) di ricondurle tutte ad un'unica fonte di rumore (equivalente) in ingresso al decisore. Come appare dalla figura a pag. 74, l'effetto del rumore è quello di causare degli *errori* nelle decisioni sui livelli, e quindi sui simboli e sui bit ricevuti. Intraprendiamo quindi l'analisi per valutare la *frequenza* di questi errori, ovvero la loro probabilità (vedi cap. 7).

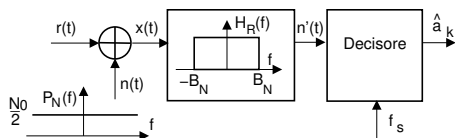


5.3.1 Rumore bianco limitato in banda

Come anticipato, il disturbo $n(t)$ è la realizzazione di un processo gaussiano ergodico, a valor medio nullo, e con spettro di densità di potenza *bianco* o costante

$$\mathcal{P}_N(f) = \frac{N_0}{2}$$

e viene caratterizzato dalla sua varianza σ_N^2 , pari alla potenza²⁰ \mathcal{P}_N di una sua realizzazione qualsiasi (per l'ergodicità): tale processo è anche indicato come *Additive White Gaussian Noise* o AWGN.

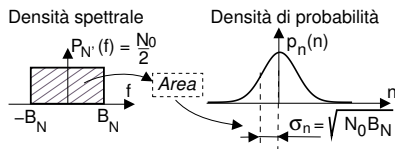


Allo scopo di limitare \mathcal{P}_N alla minima possibile, in ingresso al ricevitore è posto un filtro passa-basso ideale²¹ con risposta in frequenza $H_R(f)$ limitata in una banda $\pm B_N$ (detta *banda di rumore*, vedi § 12.1.1), tale da lasciar

passare le componenti frequenziali del segnale $r(t)$ per intero, e limitare al tempo stesso la banda (e dunque la potenza) di $\mathcal{P}_N(f)$ al minimo.

Pertanto, la potenza del rumore in uscita da $H_R(f)$ risulta²² pari a

$$\begin{aligned} \mathcal{P}_{N'} &= \sigma_{N'}^2 = \int_{-\infty}^{\infty} \mathcal{P}_N(f) |H_R(f)|^2 df = \\ &= \int_{-B_N}^{B_N} \frac{N_0}{2} df = N_0 B_N \end{aligned}$$



e questo valore si riflette (in virtù della ergodicità di $n(t)$) nella dinamica della v.a. di rumore n , come esemplificato in figura.

¹⁹Vedi il cap. 16 per analizzare la sua origine.

²⁰Al cap. 16 si illustra come in realtà $\mathcal{P}_N(f)$ non è costante per qualsiasi valore di f fino ad infinito, ma occupa una banda indeterminata ma limitata: altrimenti, avrebbe una potenza infinita.

²¹Si veda il § 5.7 per la scelta di un diverso filtro di ricezione, individuato applicando le condizioni per la minimizzazione della probabilità di errore.

²²Per i dettagli relativi al filtraggio di processi, ci si può riferire al § 9.4.1.

5.3.2 Soglie di decisione

Proseguiamo ora l'analisi indicando il segnale ricevuto nella forma

$$r(t) = \sum_k a[k] \cdot g(t - kT_s) \quad (5.8)$$

in cui $g(t)$ è un impulso di Nyquist (5.3), in modo da poter assumere assenza di ISI, ed i simboli $a[k]$ sono elementi di una sequenza aleatoria, ognuno scelto tra L possibili valori a_i , $i = 1, 2, \dots, L-1$, distribuiti entro un intervallo con dinamica pari a $a_L - a_1 = \Delta$.

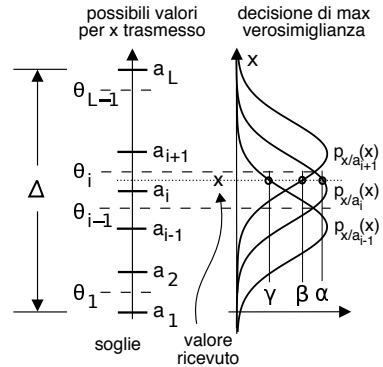
Agli istanti multipli del periodo di simbolo $t = kT_s = k/f_s$, il decisore esamina il valore del segnale $x(t) = r(t) + n(t)$, ed anziché ritrovare i valori $a[k]$ trasmessi, osserva la realizzazione di una variabile aleatoria gaussiana, con media pari al valore trasmesso $a[k] = a_i$, e varianza $\sigma_N^2 = N_0 B_N$. Quindi, per stabilire quale valore sia stato (probabilmente) trasmesso per il simbolo k -esimo, effettua una decisione di massima verosimiglianza o MV (vedi § 7.5.3) confrontando tra loro le densità di probabilità condizionate alle diverse ipotesi che sia stato trasmesso il simbolo a_i :

$$P_{X/a_i}(x) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left\{-\frac{(x - a_i)^2}{2\sigma_N^2}\right\} \quad (5.9)$$

e scegliendo per l' \hat{a}_i tale che $P_{X/\hat{a}_i}(x)$ è la più grande.

Nel caso in cui i valori a_i siano equispaziati, il criterio di massima verosimiglianza equivale (vedi figura) a definire $L-1$ soglie di decisione θ_i , $i = 1, 2, \dots, L-1$, poste a metà tra i valori a_i ed a_{i+1} ²³, e decidere per il valore a_i se il segnale ricevuto x cade all'interno dell'intervallo compreso tra θ_{i-1} e θ_i ⁽²⁴⁾, dato che ciò (nel caso in figura) corrisponde ad imporre

$$\alpha = P_{X/a_i}(x) > \beta = P_{X/a_{i+1}}(x) > \gamma = P_{X/a_{i-1}}(x)$$



5.3.3 Probabilità di errore per simbolo

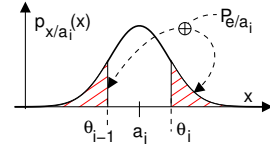
Dopo aver applicato il criterio di massima verosimiglianza su esposto, il decisore può aver commesso un errore nel caso in cui, a causa di un valore particolarmente elevato di rumore, x oltrepassi la soglia di decisione, ovvero se il campione di rumore osservato all'istante k è (in modulo) più grande di $\alpha = |\theta_i - a_i| = \frac{\Delta}{2(L-1)}$. La probabilità di errore si dice in questo caso *condizionata* alla trasmissione di a_i , e vale

$$P_{e/a_i} = 2 \int_{\theta_i}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left\{-\frac{(x - a_i)^2}{2\sigma_N^2}\right\} dx = P_\alpha$$

²³Questa proprietà di equidistanza tra le soglie, deriva dalla simmetria pari della d.d.p. gaussiana rispetto al suo valor medio: in generale, le soglie sono poste in modo da rendere eguali le probabilità di errore di falso allarme e di perdita, vedi § 7.6.2.

²⁴Chiaramente, tutti i valori minori di θ_1 provocano la decisione a favore di a_1 , e quelli maggiori di θ_{L-1} indicano la probabile trasmissione di a_L .

che chiameremo P_α , e che rappresenta (vedi figura) la somma delle aree tratteggiate. Lo stesso valore P_α è valido per tutti gli indici i compresi tra 2 ed $L-1$, mentre per a_1 ed a_L la probabilità di errore è dimezzata perché l'errore si verifica solo su di una soglia di decisione: $P_{e/a_1} = P_{e/a_L} = \frac{1}{2}P_\alpha$.



Applicando il cambiamento di variabile descritto al § 7.2.4, troviamo che $P_\alpha = \text{erfc} \left\{ \frac{\theta_i - a_i}{\sqrt{2}\sigma_N} \right\}$; sostituendo a $\theta_i - a_i$ la sua ampiezza espressa in funzione della dinamica di segnale Δ , troviamo

$$P_\alpha = \text{erfc} \left\{ \frac{\Delta}{2\sqrt{2}\sigma_N(L-1)} \right\} \quad (5.10)$$

Per arrivare all'espressione della probabilità di errore incondizionata²⁵ occorre eseguire una operazione di *valore atteso* (§ 7.2.2) rispetto a tutti gli indici i , con $i = 1, 2, \dots, L$, ovvero pesare le diverse probabilità di errore condizionate per le rispettive probabilità degli eventi condizionanti. Nel caso in cui i valori a_i siano *equiprobabili*, con probabilità $Pr(a_i) = \frac{1}{L}$, si ottiene:

$$P_e = E_{a_i} \{P_{e/a_i}\} = \sum_{i=1}^L Pr(a_i) P_{e/a_i} = \frac{1}{L} \left[(L-2)P_\alpha + 2\frac{1}{2}P_\alpha \right] = \left(1 - \frac{1}{L} \right) P_\alpha$$

in cui si è tenuto conto della diversa probabilità condizionate per i livelli intermedi e per i due agli estremi. Il risultato ottenuto, benché già idoneo a valutare la P_e , può essere ulteriormente elaborato per ottenere espressioni più adatte ai progetti di dimensionamento.

Pe come funzione di Eb/No I passaggi che seguono sono volti ad esprimere l'eq. (5.10) in modo che l'argomento di $\text{erfc}\{\}$ dipenda dalla grandezza E_b/N_0 , che come stiamo per verificare, riassume in sé i valori dei *parametri intrinseci* del collegamento, ossia potenza ricevuta \mathcal{P}_R , densità di rumore N_0 , e frequenza binaria f_b . Al contrario, la grandezza $SNR = \frac{\mathcal{P}_R}{\mathcal{P}_N}$, generalmente indicativa della qualità del segnale ricevuto dipenderebbe, oltre che dai parametri suddetti, anche da altre grandezze che invece rappresentano specifiche scelte per i *parametri di trasmissione*, come γ e L , il cui contributo vorremmo invece mantenere separato²⁶.

Definizione di Eb/No e suo contributo all'SNR La grandezza E_b rappresenta l'*energia per bit*²⁷ e la sua definizione

$$E_b = \mathcal{P}_R T_b = \frac{\mathcal{P}_R}{f_b}$$

²⁵Che non dipende cioè da quale simbolo sia stato trasmesso.

²⁶Infatti, la potenza di rumore $\mathcal{P}_N = N_0 B_N$, da cui SNR dipende, a sua volta è funzione dalla banda del filtro di ricezione, che come abbiamo visto è posta pari alla massima frequenza presente in $r(t)$, e quindi (prendendo in considerazione un codice di linea $g(t)$ a coseno rialzato, vedi § 5.2.2.3) pari a

$$B_N = \frac{f_s}{2} (1 + \gamma) = \frac{f_b}{2 \log_2 L} (1 + \gamma)$$

dipendendo quindi anch'essa da L e γ , oltre che dalla f_b . Pertanto, al variare di L e γ , varia anche SNR .

²⁷si rifletta sulla circostanza che la potenza è una energia per unità di tempo.

mostra come essa riassume in sé i parametri di sistema *potenza di segnale e velocità binaria*, mentre invece non dipende dai *parametri di trasmissione* L e γ . Anche N_0 costituisce un *parametro di sistema*, rappresentando una grandezza su cui non è possibile intervenire.

Allo scopo quindi di mantenere separati tra loro i diversi elementi che determinano la probabilità di errore, e di porre nella giusta evidenza le quantità E_b ed N_0 , esprimiamo dunque le potenze \mathcal{P}_N e \mathcal{P}_R in funzione di $T_b = 1/f_b$:

$$\mathcal{P}_N = N_0 B_N = \frac{N_0 (1 + \gamma)}{T_b 2 \log_2 L} \quad \text{e} \quad \mathcal{P}_R = \frac{E_b}{T_b} \quad (5.11)$$

in modo da ottenere

$$SNR = \frac{\mathcal{P}_R}{\mathcal{P}_N} = \frac{E_b T_b 2 \log_2 L}{T_b N_0 (1 + \gamma)} = \frac{E_b 2 \log_2 L}{N_0 (1 + \gamma)} \quad (5.12)$$

Legame tra Pr e la dinamica del segnale Prima di arrivare all'espressione di P_α (eq. 5.10) in funzione di E_b/N_0 , occorre esprimere \mathcal{P}_R in funzione dei parametri di trasmissione L e γ , nonché della dinamica Δ . Sotto le ipotesi in cui:

- si adotti un impulso di Nyquist a coseno rialzato con roll-off γ ;
- i simboli $a[k]$ siano statisticamente indipendenti ed a valori a_i equiprobabili;
- tali valori siano distribuiti uniformemente su L livelli, con dinamica $a_L - a_1 = \Delta$;

al § 9.10.4 si ottiene²⁸

$$\mathcal{P}_R = \frac{\Delta^2 L + 1}{12} \frac{1}{L - 1} \left(1 - \frac{\gamma}{4}\right) \quad (5.13)$$

Se in particolare risulta $L = 2$ e $\gamma = 0$ (come nel caso di trasmissione binaria a banda minima) allora si ha $\mathcal{P}_R = \frac{\Delta^2}{4}$. Ma per essere utilizzata ai nostri fini, la (5.13) deve prima essere invertita, in modo da esprimere Δ in funzione di \mathcal{P}_R :

$$\Delta = \sqrt{12 \frac{L - 1}{L + 1} \frac{\mathcal{P}_R}{(1 - \gamma/4)}} \quad (5.14)$$

5.3.3.1 Espressione della P_e per simbolo

Non resta ora che inserire la (5.14) nella espressione di P_α (eq. 5.10), ricordare che $\sigma_N^2 = \mathcal{P}_N$, e tenere conto della (5.12), in modo da ottenere²⁹

$$P_e = \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0} \frac{3 \log_2 L}{(L^2 - 1)(1 + \gamma)(1 - \frac{\gamma}{4})}} \right\} \quad (5.15)$$

che è graficata alla Fig 5.6, in funzione di $\frac{E_b}{N_0} \Big|_{dB}$, per tre condizioni operative. In particolare, notiamo che per $L = 2$ e $\gamma = 0$ si ottiene:

$$P_e = \frac{1}{2} \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0}} \right\} \quad (5.16)$$

²⁸Anche se il risultato sarà dimostrato al § 9.10.4, merita comunque un commento: osserviamo che \mathcal{P}_R diminuisce all'aumentare di γ (si stringe infatti l'impulso nel tempo); inoltre \mathcal{P}_R diminuisce al crescere di L , in quanto nel caso di più di 2 livelli, la forma d'onda assume valori molto vari all'interno della dinamica di segnale, mentre con $L = 2$ ha valori molto più *estremi*.

²⁹Per chiarezza sviluppiamo i passaggi, piuttosto banali anche se non ovvi:

Le scelte progettuali (γ e L) diverse da $L = 2$ e $\gamma = 0$ determinano immancabilmente un peggioramento della P_e , ma vengono intraprese per soddisfare esigenze di risparmio di banda (aumentando L)³⁰, e per ridurre i termini di interferenza intersimbolica (aumentando γ). Due domande riassuntive:

- perché P_e peggiora se aumento i livelli ?
Risposta (³¹).
- perché P_e peggiora se aumento γ ?
Risposta (³²).

Compromesso banda - potenza Osservando la fig. 5.6 notiamo che al crescere di L , e dunque occupando una banda minore, si può ottenere la stessa P_e solo a patto di aumentare E_b/N_0 , ovvero (a parità di f_b) aumentando la potenza trasmessa: questo è un aspetto di un risultato più generale della *teoria dell'informazione*. Claude Shannon ha infatti dimostrato (vedi pag. 454) che è *possibile trasmettere senza errori* (ricorrendo a tecniche di codifica di canale oltremodo sofisticate) purché la velocità di trasmissione f_b non ecceda la *capacità di canale*, definita come

$$C = B \log_2 \left(1 + \frac{\mathcal{P}_R}{N_0 B} \right)$$

in cui B è la banda del canale, \mathcal{P}_R la potenza ricevuta, e $N_0 B$ la potenza del rumore. Un secondo canale, con B *minore*, dispone di una minore capacità, in quanto $\log_2(\cdot)$

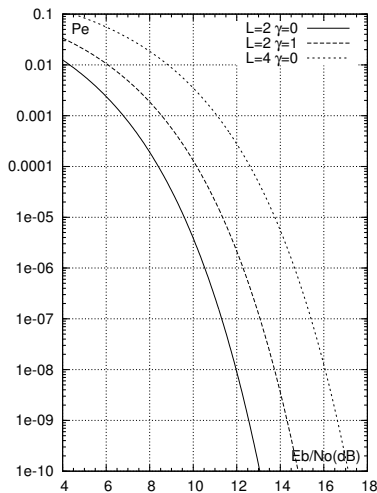


Figura 5.6: Andamento di P_e vs. E_b/N_0

$$\begin{aligned} P_e &= \left(1 - \frac{1}{L}\right) P_\alpha = \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \frac{\Delta}{2\sqrt{2}\sigma_N(L-1)} \right\} \\ &= \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{12} \frac{L-1}{L+1} \frac{\mathcal{P}_R}{(1-\gamma/4)} \frac{1}{2\sqrt{2}\mathcal{P}_N(L-1)} \right\} \\ &= \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ 2\sqrt{3} \frac{L-1}{L+1} \frac{1}{(1-\gamma/4)} \frac{1}{2\sqrt{2}} \sqrt{\frac{\mathcal{P}_R}{\mathcal{P}_N} \frac{1}{(L-1)}} \right\} \\ &= \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2}} \frac{L-1}{L+1} \frac{1}{(L-1)^2} \frac{1}{(1-\gamma/4)} \operatorname{SNR} \right\} \\ &= \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2}} \frac{1}{L^2-1} \frac{1}{(1-\gamma/4)} \frac{E_b}{N_0} \frac{2\log_2 L}{1+\gamma} \right\} \\ &= \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0} \frac{3\log_2 L}{(L^2-1)(1+\gamma)(1-\frac{\gamma}{4})}} \right\} \end{aligned}$$

³⁰ Aumentando L , l'argomento di (5.15) diminuisce, in quanto $(L^2 - 1)$ cresce più velocemente di $\log_2 L$.

³¹ Perché a parità di \mathcal{P}_R gli intervalli di decisione sono più ravvicinati, e le "code" della gaussiana sottendono un'area maggiore.

³² Perché occorre aumentare la banda del filtro di ricezione e dunque far entrare più rumore. D'altra parte questo peggioramento è compensato dalla riduzione dell'ISI.

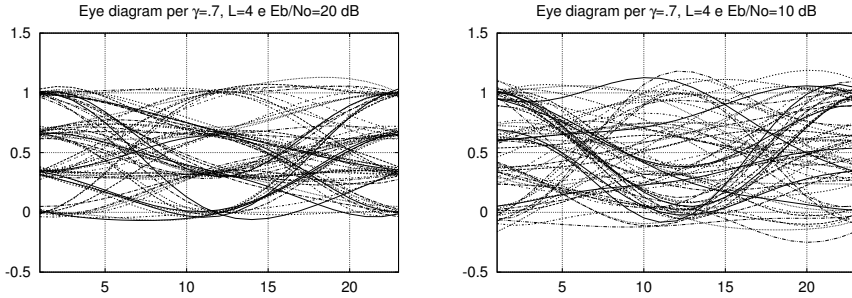


Figura 5.7: Diagramma ad occhio con E_b/N_0 pari a 20 e 10 dB, $\gamma = .7$, $L = 4$

crece più lentamente di quanto non decresca B ; pertanto, per mantenere la stessa capacità, è necessario trasmettere con una maggiore potenza di segnale \mathcal{P}_R . E' questo il motivo per cui, nel caso in cui vi siano limitazioni di potenza ma non di banda, come ad esempio nelle *comunicazioni satellitari*, conviene occupare la maggior banda possibile, mantenendo $L = 2$, in modo da risparmiare potenza³³.

Compromesso velocità - distorsione Abbiamo appena fatto notare come, riducendo la banda occupata dal segnale dati, occorra aumentare la potenza ricevuta per mantenere la stessa P_e . Viceversa, osserviamo anche che l'*aumento* della velocità binaria (e quindi dell'occupazione di banda) dovuta all'adozione di strategie per il controllo di errore (vedi § 5.4 e § 17.2), consente di *ridurre* la probabilità di errore stessa, e dunque migliorare la *fedeltà* del flusso binario, anche a parità di potenza ricevuta.

5.3.4 Diagramma ad occhio in presenza di rumore

Si tratta dello stesso tipo di grafico già descritto a pag. 84, e che ora ci aiuta a valutare in modo visivo la qualità di una trasmissione numerica. In fig. 5.7 sono riportati i grafici per un segnale dati a 4 livelli, in presenza di due diversi livelli di potenza di rumore: notiamo che al peggiorare del rapporto $\frac{E_b}{N_0}$, la zona priva di traiettorie (*l'occhio*) riduce la sua estensione verticale (*tende a chiudersi*). Pertanto, disponendo di un segnale numerico di qualità sconosciuta, questa può essere valutata in modo approssimato, qualora si disponga di un oscilloscopio, esaminando il *grado di apertura dell'occhio*.

5.3.5 Uso del codice di Gray e P_e per bit

La probabilità di errore (5.15) si riferisce all'evento di decidere per la ricezione del simbolo a_i quando invece è stato trasmesso a_{i-1} o a_{i+1} ⁽³⁴⁾, mentre ora determiniamo la probabilità che sia errato uno qualunque dei bit ottenibili dopo la *serializzazione* (vedi nota 10 a pag. 77) della codifica associata al simbolo ricevuto. Abbiamo già illustrato al § 5.2.2.4 come esista un modo particolare (il codice di Gray) di assegnare

³³Per contro, volendo mantenere immutata la qualità, e dunque la P_e , l'*aumento* dell'occupazione di banda prodotta dal controllo di errore (vedi § 5.4), consente di *ridurre* la potenza di segnale!

³⁴La probabilità di un errore legato al *salto* di due o più livelli θ è così piccola da potersi trascurare.

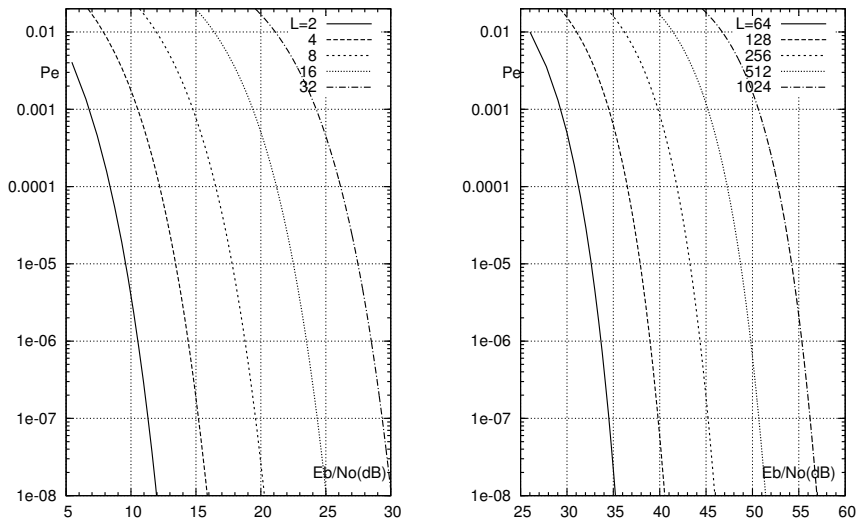


Figura 5.8: Probabilità di errore sul bit per trasmissione multilivello di banda base con codifica di Gray

codifiche binarie ai simboli a_i , al fine di associare ai simboli relativi a valori (livelli) contigui, parole differenti per un solo bit.

In presenza della codifica di Gray, pertanto, ogni simbolo errato contiene un solo bit errato, e quindi l'evento di errore *sul bit* si verifica quando il simbolo a cui appartiene è errato, e il bit è quello errato, ovvero: $Pr\{bit\ errato\} = Pr\{simbolo\ errato\} \cdot Pr\{bit\ errato/simbolo\ errato\} = P_e \cdot \frac{1}{\log_2 L}$. Ad esempio, con $L = 256$ livelli, la P_e sul bit si riduce di $\log_2 L = 8$ volte. L'espressione della P_e per bit nel caso si adotti una codifica di Gray, diviene quindi:

$$P_e^{bit} = \frac{1}{\log_2 L} \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0} \frac{3 \log_2 L}{(L^2 - 1)(1 + \gamma) \left(1 - \frac{\gamma}{4}\right)}} \right\} \quad (5.17)$$

Le curve in fig. 5.8 mostrano il valore di P_e^{bit} così determinato, per $\gamma = 0$, in funzione di $\frac{E_b}{N_0}$ espresso in dB, per diversi valori di L . Valori di $\gamma \neq 0$ determinano un peggioramento³⁵ di $10 \log_{10} (1 + \gamma) \left(1 - \frac{\gamma}{4}\right)$, che deve essere compensato da un eguale incremento in dB di $\frac{E_b}{N_0}$, per ottenere la stessa P_e .

Dimensionamento di una trasmissione numerica Una metodologia operativa di progetto può basarsi sull'imporre un determinato valore di P_e , a partire dal quale

- in base alle curve di fig. 5.8, si individuano i valori di E_b/N_0 (in dB) necessari, per diverse scelte di L ;

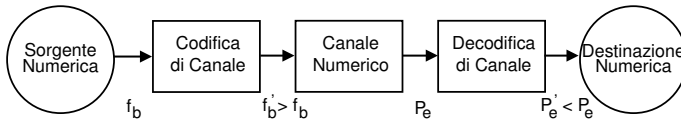
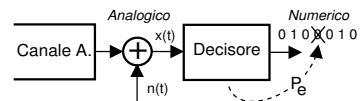
³⁵Di non grande entità: per $\gamma = 1$ il peggioramento risulta di 1.761 dB.

- nota la banda disponibile e la velocità f_b , si ottiene il valore di L , individuando così la curva appropriata, nell'ipotesi di adottare $\gamma = 0$;
- noto il livello di rumore N_0 , si determina E_b ;
- note le esigenze di precisione nella temporizzazione, si impone un valore del roll-off γ , e conseguentemente si aumenta il valore di E_b ;
- si determina la minima potenza che è necessario ricevere, come $W_{R_{min}} = E_b \cdot f_b$.

A pag. 105 viene proposto un esercizio riassuntivo, che comprende anche alcuni concetti presentati nella sezione seguente.

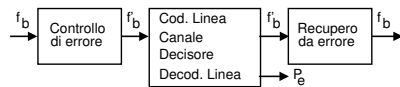
5.4 Gestione degli errori di trasmissione

La figura a lato, già proposta a pag. 3, ricorda ancora una volta come le conseguenze prodotte dagli errori del decisore, siano alcuni bit *diversi* da quelli trasmessi. Al § 5.3 abbiamo svolto il calcolo della probabilità di errore P_e dovuta a rumore additivo, mentre al § 5.2.2.3 è stato illustrato come anche una non perfetta temporizzazione, od una alterazione dell'impulso $g(t)$, possano causare errori di decisione dovuti all'ISI. Sempre al primo capitolo è proposta anche la figura replicata appresso, che illustra come la P_e possa essere *ridotta* adottando tecniche di *codifica di canale* e *controllo di errore*, che sono discusse di seguito.



5.4.1 Controllo di errore

Con questo termine si individuano le strategie atte a *proteggere* le informazioni da trasmettere, aumentando il numero effettivo dei bit inviati (che passano da f_b a $f'_b > f_b$ bit/secondo, vedi figura), in modo che i bit aggiunti siano dipendenti dagli altri, permettendo la gestione degli eventuali errori di trasmissione.



FEC e ARQ La quantità di bit aggiunti può permettere di *accorgersi* della presenza di errori di trasmissione, oppure di *correggere* detti errori. Se il sistema di trasmissione viene considerato unidirezionale³⁶, lo scopo del controllo di errore è quello di permettere al ricevitore di correggere gli errori che si sono verificati; questo approccio è anche indicato come *Forward Error Correction* (FEC, o correzione di errore *in avanti*), e si è sviluppato nel contesto funzionale della *codifica di canale* (vedi § 5.4.3).

³⁶O perché lo è effettivamente, oppure perché la trasmissione coinvolge informazioni generate in tempo reale, non memorizzate in trasmissione, e consumate immediatamente in ricezione (ad es. un segnale numerico ottenuto mediante campionamento e quantizzazione), per le quali non ha senso richiedere la trasmissione.

Se al contrario è presente un canale di comunicazione a ritroso, e non sussistono rigidi vincoli temporali sul massimo ritardo tra trasmissione e ricezione corretta, l'effetto di un errore *rivelato* ma non correggibile immediatamente può essere annullato chiedendo la ritrasmissione del dato errato, dando luogo alle strategie di *Automatic Repeat reQuest*³⁷ o ARQ (ossia richiesta automatica di ripetizione) (vedi § 5.5), che danno origine alla definizione dei *protocolli a finestra*, evoluti storicamente nel contesto della trasmissione dati.

Compromesso velocità - distorsione A pag. 91 sono svolte delle considerazioni su come l'adozione di tecniche di controllo di errore consenta di *contrattare* una migliore qualità del segnale ricevuto, a prezzo di una maggiore velocità di trasmissione (e quindi banda occupata) del segnale, mentre al § 17.2 è svolta la trattazione analitica relativa a questo fenomeno.

5.4.1.1 Errori su parole

I metodi di controllo di errore operano tipicamente su gruppi di n bit, ognuno dei quali può rivelarsi sbagliato con probabilità $P_e = p$; se gli eventi di errore sono statisticamente indipendenti (§ 7.1.5), la probabilità di *un* bit errato *su* n è pari a np , cioè n volte la probabilità di un bit errato³⁸.

D'altro canto, la probabilità di i bit errati su n è fornita dalla distribuzione Binomiale (vedi § 8.1) $P(i, n) = \binom{n}{i} p^i (1-p)^{n-i}$ che, se $p \ll 1$, può essere approssimata come

$$P(i, n) \approx \frac{n(n-1) \cdots (n-i+1)}{i!} p^i \quad (5.18)$$

che è molto inferiore³⁹ alla probabilità np di *un singolo* errore su n . Infine, sempre per $p \ll 1$, risulta che $P(i+1, n) \ll P(i, n)$, e quindi si può considerare la probabilità di ricevere i o *più* bit errati su n , praticamente uguale a quella di osservare solo i errori.

All'aumentare di p e di n , le approssimazioni non sono più valide; in tal caso però, il sistema di trasmissione è praticamente inusabile, e dunque ci si trova sempre nelle condizioni di p sufficientemente piccolo da permettere le approssimazioni.

L'esposizione prosegue illustrando al § 5.4.2 tre tecniche utilizzate comunemente per la detezione di errore, e formalizzando quindi (al § 5.4.3) il problema della codifica di canale, mediante la descrizione di alcune soluzioni che la realizzano. Infine, al § 5.5 sono descritti i protocolli ARQ adottati nel contesto della trasmissione dati.

³⁷L'aggettivo *automatic* si riferisce al fatto che spesso la gestione della ritrasmissione avviene a carico di uno strato protocollare di livello *inferiore* a quello che effettivamente consuma il messaggio, che in definitiva neanche si avvede della presenza del meccanismo di ritrasmissione.

³⁸La probabilità che solo il primo bit su n sia sbagliato è pari a $p_1 = p \cdot (1-p)^{n-1} \simeq p$ se $p \ll 1$ ed n non è troppo grande; lo stesso risultato si ottiene anche per gli altri $n-1$ casi possibili di un bit sbagliato e gli altri corretti, cosicché la probabilità di un solo *generico* bit sbagliato assomma quelle di tutti i casi.

³⁹Ad esempio, la probabilità di due bit errati può essere approssimata come $\frac{1}{2} n(n-1) p^2$. Infatti, la distribuzione Binomiale fornisce $P(2, n) = \binom{n}{2} p^2 (1-p)^{n-2}$, in cui $\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}$ e, se $p \ll 1$ ed n non è troppo elevato, $(1-p)^{n-2} \simeq 1$. All'aumentare di p e di n , l'approssimazione non è più valida, e la probabilità di più di un bit errato può risultare maggiore di quella di un solo bit errato.

5.4.2 Detezione di errore

In questo caso ci si limita a mettere il ricevitore in grado di accorgersi della presenza di errori, senza poterli correggere. Le parole errate possono essere scartate, oppure re-inviate su richiesta.

5.4.2.1 Parità

Viene comunemente usata nell'ambito della trasmissione asincrona e sincrona orientata al carattere, per rivelare errori sul bit, e consiste nell'aggiungere alla parola da trasmettere un ulteriore bit, in modo che in totale ci sia un numero *pari* di uni⁴⁰, applicando quindi una regola di parità *pari* (EVEN). Il caso opposto, ossia l'aggiunta di un bit in modo da rendere *dispari* il numero di uni, prende nome di parità ODD.

In entrambi i casi⁴¹, quando il ricevitore raggruppa i bit pervenuti, esegue un *controllo* detto appunto *di parità*, semplicemente contando il numero di uni, ed accorgendosi così se nella parola si sia verificato un errore (uno zero divenuto uno o viceversa). In tal caso, il ricevitore invierà all'altro estremo del collegamento una richiesta di ritrasmissione del gruppo di bit errati. Se invece si fosse verificato un errore che coinvolge due bit della parola, questo passerebbe inosservato, in quanto la parità prescritta verrebbe mantenuta. Infatti, la *distanza di Hamming* (vedi pag. 98) tra codeword ottenute aggiungendo il bit di parità è pari a due⁴².

Esempio: indicando la probabilità di errore sul bit con p (es 10^{-3}) ed applicando la (5.18) si ottiene che la probabilità di $i = 2$ errori in n bit (es $n = 10$) vale $\frac{1}{2}n(n-1)p^2$ (es $4.5 \cdot 10^{-5}$), e rappresenta il *tasso residuo* di errore su parola legato all'uso di un bit di parità; la trasmissione di $n - 1$ bit non protetti sarebbe invece stata esposta ad una probabilità np (es $0.9 \cdot 10^{-2}$) di un bit errato su n .

Il calcolo del bit di parità *pari* può essere effettuato svolgendo la *somma modulo due*⁴³ di tutti i bit che compongono la parola (ovvero complementando il risultato, nel caso di parità *dispari*).

5.4.2.2 Somma di controllo

Quando il messaggio è composto da M diverse parole di N bit, la probabilità che almeno una di queste sia errata aumenta in modo circa proporzionale ad M , in base ad un ragionamento del tutto analogo a quello della nota 38.

Per aumentare le capacità di rivelazione del controllo di parità applicato sulle singole parole (indicato ora come parità *di riga*, o *trasversale*), si aggiunge al gruppo di M parole una ulteriore parola (detta *somma di controllo*), i cui bit si ottengono applicando il controllo di parità a tutti i bit "omologhi" delle M parole incolonnate, generando così una parità *di colonna* (o *longitudinale*), come esemplificato in figura.

0000010 0	}	parità di riga (ODD)		
0101000 1				
1000110 0				
0100000 0				
0101101 1				
1000000 0				
1100011 1				
0000011 1				
1000001 1			←	parità di colonna (EVEN)

⁴⁰Ad esempio, alla sequenza 001001 verrà aggiunto uno 0, mentre a 010101 si aggiungerà ancora un 1, perché altrimenti gli uni complessivi sarebbero stati 3, che è dispari.

⁴¹Il ricevitore deve comunque essere al corrente del fatto se la parità sia ODD o EVEN !

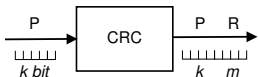
⁴²Considerando parole di 3 bit, le codeword (di 4 bit, in cui l'ultimo è una parità pari) risultano: (0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111). E' facile constatare che ognuna di esse differisce da tutte le altre per due bit.

⁴³La somma modulo due è equivalente alla operazione di OR esclusivo, viene a volte indicata con il simbolo \oplus , e corrisponde alle definizioni: $0 \oplus 0 = 0$, $0 \oplus 1 = 1$, $1 \oplus 0 = 1$, $1 \oplus 1 = 0$.

A volte, si preferisce calcolare la somma di controllo mediante una operazione di somma *modulo uno*⁴⁴, direttamente realizzabile in software in modo veloce. In tal caso, il ricevitore calcola una nuova somma di controllo longitudinale, includendo anche la somma di controllo originaria: in assenza di errori, il risultato deve fornire zero.

5.4.2.3 Codici polinomiali e CRC

L'utilizzo di una somma di controllo può produrre risultati scadenti nel caso di distribuzioni temporali dei bit errati particolarmente sfavorevoli, mentre l'uso del *Cyclic Redundancy Check* (o CRC) garantisce prestazioni *più uniformi*. Questo metodo di controllo di errore consiste nell'aggiungere ad una parola



P di k bit che si desidera trasmettere, un gruppo R di $m < k$ ulteriori bit *di protezione*, calcolati a partire dai primi k , e tali da permettere la deteazione di

eventuali errori; in questo senso, i codici polinomiali sono classificabili come *codici a blocchi* (§ 5.4.3.1).

L'aggettivo *polinomiale* trae origine dalla associazione tra un numero binario B di $n + 1$ bit, indicati con $b_i, i = 0, 1, \dots, n$, ed un polinomio⁴⁵ $B(x)$ a coefficienti binari nella variabile x , di grado n , con espressione

$$B(x) = b_n x^n + b_{n-1} x^{n-1} + \dots + b_1 x^1 + b_0$$

Un *codice polinomiale* è definito a partire da un *polinomio generatore* $G(x)$ di grado m , i cui coefficienti binari identificano una parola $G = g_m g_{m-1} \dots g_1 g_0$ di $m + 1$ bit.

Indicando ora con P la sequenza dei k bit p_i da proteggere, *aggiungiamo* a destra di questi un gruppo di m bit pari a zero, ottenendo una nuova parola $P \cdot 2^m$ lunga $k + m$ bit, che quindi dividiamo per G (mediante aritmetica modulo due⁴⁶), ottenendo un quoziente Q , ed un resto R con al massimo m bit. Pertanto, possiamo scrivere

$$\frac{P \cdot 2^m}{G} = Q \oplus \frac{R}{G}$$

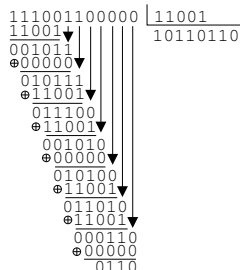
Le sequenze Q ed R costituiscono rispettivamente i coefficienti dei polinomi quoziente $Q(x)$ e resto $R(x)$, ottenibili dalla divisione di $P(x) \cdot 2^m$ per $G(x)$. Gli m bit R

⁴⁴La somma modulo uno è l'equivalente binario dell'operazione di somma (decimale) tradizionale, comprese quindi le operazioni di riporto verso le cifre più elevate. Il riporto finale viene poi nuovamente sommato al risultato della somma.

⁴⁵L'insieme di tutti i polinomi di grado minore od uguale ad n costituisce un particolare spazio algebrico, per il quale è possibile dimostrare una serie di proprietà, la cui verifica trascende dallo scopo di questo testo, e che consentono di stabilire le capacità del codice di rivelare gli errori.

⁴⁶Per fissare le idee, consideriamo $k = 8$ bit a da proteggere, pari a $P = 11100110$, $m = 4$ bit di CRC, ed un generatore $G = 11001$. La sequenza $P \cdot 2^m$ risulta pari a $11100110\ 0000$, e la divisione modulo 2 tra P e G fornisce un quoziente $Q = 10110110$ (che viene ignorato) ed un resto R pari a 0110 . Pertanto, viene trasmessa la sequenza $T = P \cdot 2^m \oplus R = 11100110\ 0110$ con $k + m = 12$ bit.

La divisione modulo 2 si realizza come mostrato nella figura a lato, calcolando un OR-ESCLUSIVO \oplus bit-a-bit tra i bit più significativi del divisore P , o del resto parziale, e quelli del dividendo G . Per ognuna di queste operazioni, si *abbassano* una o più cifre binarie del divisore, in modo che il resto parziale abbia *la stessa lunghezza* del divisore; in tal caso, si ottiene un bit di risultato di quoziente pari ad uno, mentre se il resto parziale è troppo corto, per ogni bit abbassato in più si ottiene un bit di resto parziale pari a zero. Quando tutti i bit del divisore sono stati usati, l'ultima operazione \oplus fornisce il resto R cercato.



del resto sono quindi utilizzati come *parola di protezione*, in modo da esprimere la sequenza T da trasmettere come $T = P \cdot 2^m \oplus R$ di $k + m$ bit, ovvero con i k bit più elevati pari a P e gli m bit in coda pari ad R . Il ricevitore effettua anch'esso una divisione, stavolta tra T e G , che in assenza di errori produce un *resto nullo*

$$\frac{T}{G} = \frac{P \cdot 2^m \oplus R}{G} = \frac{P \cdot 2^m}{G} \oplus \frac{R}{G} = Q \oplus \frac{R}{G} \oplus \frac{R}{G} = Q$$

in quanto sommando in aritmetica modulo due un numero per se stesso, si ottiene un risultato nullo. Pertanto, se $T/G = Q$ con resto nullo, la parola P è riottenuta semplicemente shiftando T a destra di m posizioni.

Nel caso invece in cui si siano verificati errori, indichiamo con E la sequenza binaria di errore, di lunghezza $k + m$ bit, ognuno dei quali è pari ad uno se in quella posizione si verifica errore, o zero in caso contrario, in modo da rappresentare il segnale ricevuto R come $R = T \oplus E$. Se $E \neq 0$ la divisione operata al ricevitore ora fornisce

$$\frac{R}{G} = \frac{T \oplus E}{G} = \frac{T}{G} \oplus \frac{E}{G} = Q \oplus \frac{E}{G}$$

e quindi si verifica la presenza di un resto pari ad E e diverso da zero, che indica appunto la presenza di errori, tranne nei casi in cui E risulti perfettamente divisibile per G , evento con bassa probabilità se G è scelto opportunamente.

Per applicare il metodo, sia il trasmettitore che il ricevitore devono utilizzare lo stesso generatore $G(x)$, per il quale esistono diverse scelte standardizzate⁴⁷. Si può dimostrare che le operazioni discusse permettono di rivelare

- tutti gli errori singoli;
- tutti gli errori doppi se G possiede almeno tre uni;
- qualsiasi numero dispari di errori, se $G(x)$ contiene il fattore $(x + 1)$;
- tutti gli errori *a burst* che si estendono per k o meno bit;
- la maggior parte degli errori impulsivi più estesi.

Calcolo del CRC Gli m bit aggiunti prendono appunto il nome di CYCLIC REDUNDANCY CHECK (o *controllo a ridondanza ciclica*) in virtù del particolare modo in cui questi possono essere calcolati: la parte del procedimento descritto che appare più complessa, ossia la divisione modulo due, è invece quella che ha maggiormente contribuito al successo del metodo, dato che è realizzabile a livello circuitale in modo relativamente semplice⁴⁸.

Si tratta infatti di utilizzare un registro a scorrimento controreazionato (vedi figura seguente), in cui sono immessi ad uno ad uno i k bit da proteggere, seguiti da m zeri consecutivi. Per ogni valore immesso, quelli già presenti *scorrono* a destra nel registro, ed il bit che *trabocca* alimenta gli OR esclusivi presenti nel registro, in corrispondenza degli uni di $G(x)$, tranne che per il bit corrispondente al termine più significativo; l'esempio in figura rappresenta il caso di $G(x) = x^5 + x^4 + x^2 + 1$.

⁴⁷Ecco quattro scelte utilizzate nei sistemi di trasmissione:

CRC-12 $G(x) = x^{12} + x^{11} + x^3 + x^2 + x + 1$

CRC-16 $G(x) = x^{16} + x^{15} + x^2 + 1$

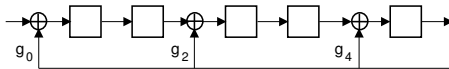
CRC-CCITT $G(x) = x^{16} + x^{12} + x^5 + 1$

CRC-32 $G(x) = x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x + 1$

Come discusso, un polinomio di ordine k genera un CRC di k bit; pertanto il CRC-12, che è usato per caratteri a 6 bit, genera 12 bit di CRC, mentre CRC-16 e CRC-CCITT, utilizzati in America ed in Europa rispettivamente per caratteri ad 8 bit, producono 16 bit di CRC. In alcuni standard di trasmissione sincrona punto-punto, è previsto l'uso di CRC-32.

⁴⁸Vedi ad es. http://en.wikipedia.org/wiki/Computation_of_cyclic_redundancy_checks

Al termine dell'inserimento di $k + m$ bit, lo stato del registro a scorrimento costituisce proprio il resto R , da usare come CRC.



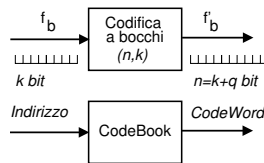
5.4.3 Correzione di errore e codifica di canale

Introduciamo ora alcune tecniche di controllo di errore, indicate come *codifica di canale*, utilizzabili anche per la semplice rivelazione di errore, ma più spesso finalizzate al tentativo di rendere il ricevitore in grado di *correggere* gli errori sopraggiunti nella trasmissione. L'argomento viene ripreso al § 17.3.

5.4.3.1 Codice a blocchi

La *codifica a blocchi* opera su blocchi di dati disgiunti, e per ogni gruppo di k bit in ingresso, distinto dagli altri gruppi di k , determina un gruppo di q bit di protezione, dipendenti dai k in ingresso, e che sono inviati assieme ai k originali per un totale di $n = n + q$ bit.

L'operazione compiuta dal codificatore può essere pensata o realizzata come l'accesso ad una tabella memoria (denominata CODEBOOK = *dizionario*), dove i k bit da codificare rappresentano un *indice* che individua 2^k differenti righe, in cui si trovano le *parole di codice* (CODEWORD), costituite ognuna da $n = k + q$ bit. Un codice siffatto è detto *codice* (n, k) .



Ridondanza Una misura del grado di protezione offerto dal codificatore di canale è rappresentata dalla *ridondanza* ρ , definita come il rapporto tra il numero di bit di protezione e quelli di informazione, e che è espressa in termini percentuali come

$$\rho = \frac{q}{k} \cdot 100 \%$$

Ad esempio, in una trasmissione con ridondanza del 50%, per ogni due bit di informazione ne viene inserito (da qualche parte) uno di protezione.

Distanza di Hamming L'entità della ridondanza introdotta migliora la capacità del codice di correggere uno o più errori, in quanto non tutte le 2^n possibili codeword (di n bit) sono utilizzate, ma solamente 2^k tra esse: le configurazioni assenti dal codebook possono comunque presentarsi in ricezione, a causa di errori di trasmissione, ma non essendo previste dal codebook, è possibile rilevare l'errore ed eventualmente correggerlo. Per una determinata coppia di codeword c_i, c_j , la loro *distanza di Hamming* $d_H(c_i, c_j)$ è definita come il numero di bit in cui differiscono⁴⁹, e le capacità di rivelazione e correzione di un codice sono quantificate dalla *minima* d_H tra tutte le possibile coppie di codeword $d_m = \min_{i \neq j} d_H(c_i, c_j)$, indicata come *distanza del codice*. Si dimostra infatti che un codice di canale è in grado di correggere $\frac{d_m - 1}{2}$ errori per parola, e di rivelare $d_m - 1$ errori per parola⁵⁰. Ad un codice a blocchi (n, k) può corrispondere una distanza $d_m \leq n - k + 1$.

⁴⁹Valutabile eseguendo il loro or esclusivo $c_i \oplus c_j$ e contando il numero di *uni* del risultato.

⁵⁰Ad esempio, se $d = 3$, la presenza di un solo bit errato fa sì che la sequenza ricevuta abbia un solo bit di differenza rispetto alla codeword trasmessa, ed al minimo due bit di differenza rispetto

Efficienza L'efficienza del codice è misurata dal *tasso di codifica* (CODE RATE)

$$R_c = \frac{k}{n} < 1$$

che rappresenta la frazione di bit informativi sul totale di quelli trasmessi, e che consente di scrivere la velocità di uscita dal codificatore come

$$f'_b = \frac{f_b}{R_c} > f_b$$

Ad esempio, in una trasmissione con un tasso di codifica pari a 0.5, il numero di bit uscenti (per unità di tempo) dal codificatore di canale è il doppio del numero dei bit entranti.

L'argomento dei codici a blocchi è molto vasto⁵¹, e fornisce molteplici soluzioni, la cui trattazione esauriente trascende dallo scopo attuale del testo; gli stessi tre casi di controllo di errore già trattati (parità⁵², somma di controllo e CRC) possono essere inquadrati nel contesto dei *codici a blocchi*. Qui ci limitiamo ad un esempio di codice a correzione molto elementare, il codice *a ripetizione*.

5.4.3.2 Codice a ripetizione

Un esempio di codice a blocchi molto semplice e con proprietà correttive è il codice a ripetizione $n : 1$, che per ogni bit in arrivo ne produce n identici in uscita, in modo che (se gli errori sono indipendenti) il decisore possa correggere l'errore in base ad una "votazione a maggioranza" (*majority voting*). Ponendo ad esempio $n = 3$, definiamo il codice a ripetizione $3 : 1$, composto da due codeword, 000 ed 111, per le quali risulta $d_m = 3$: pertanto, il codice è in grado di correggere un errore e rivelarne due⁵³. L'esercizio a pag. 106 mostra come, se la probabilità di errore per un bit P_e è molto piccola, l'applicazione della decisione a maggioranza permetta di ridurre il valore a $P_e^{dec} \simeq 3P_e^2$. Infine, per il codice a ripetizione si ottiene esattamente $d_m = n - k + 1$, ed essendo $k = 1$, risulta $d_m = n$, a discapito di un tasso di codifica mediocre, e pari a $R_c = 1/n$.

5.4.3.3 Errori a burst ed interleaving

Nonostante le capacità di correzione dei codici a blocchi possano sembrare adeguate a ridurre il valore di probabilità di errore, gli errori possono presentarsi in maniera non indipendente, ma concentrati in un breve intervallo di tempo: questa circostanza è indicata come *caso degli errori a pacchetto* (od a BURST=*scoppio*). In tal caso, si usa

a tutte le altre codeword, permettendo al ricevitore di correggere l'errore. Se invece sono presenti due errori, la procedura di correzione porterebbe a scegliere una codeword errata. Occorre quindi decidere a priori se utilizzare il codice a fini di correzione oppure di detezione, e nel secondo caso, lo stesso codice con $d = 3$ può rivelare fino a due errori per parola.

⁵¹Senza pretesa di esaustività, possiamo annoverare l'esistenza dei codici di *Hamming*, di *Hamard*, *BCH*, *Reed-Solomon*, *Reed-Muller*, di *Golay*, di *Gallager*, *turbo*, a *cancellazione*, a *fontana*, *punturati*...

⁵²Il concetto di parità può essere ulteriormente esteso, se ognuno dei k bit aggiunti è calcolato applicando la regola della parità ad un sottoinsieme degli m bit di ingresso, con sottoinsiemi eventualmente sovrapposti. Un codice del genere prende il nome di codice di *Hamming*, descritto al § 17.3.1.1.

⁵³Poniamo di dover trasmettere 0110. La sequenza diventa 000 111 111 000 e quindi, a causa di errori, ricevo 000 101 110 100. Votando a maggioranza, ricostruisco la sequenza corretta 0 1 1 0.

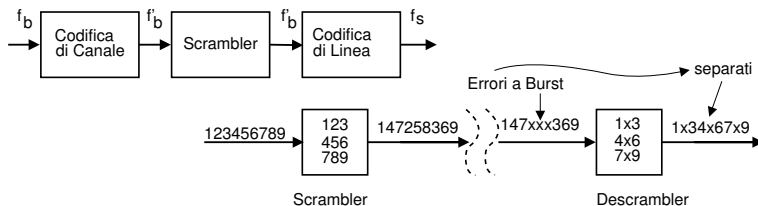


Figura 5.9: Schema di principio della trasmissione a dati intercalati

ricorrere alla tecnica nota come SCRAMBLING⁵⁴ o INTERLEAVING⁵⁵, attuabile a patto di accettare un ritardo di trasmissione. Si tratta infatti di modificare l'ordine dei dati inviati, in modo che gli errori che avvengono su bit *vicini* si riflettano in errori su bit... *lontani* e quindi appartenenti a codeword differenti, come illustrato in fig. 5.9. Ovviamente, occorre prevedere un processo inverso (*descrambling* o *deinterleaving*) all'altro capo del collegamento. E' appena il caso di notare che lo scrambler (similmente al codice di Gray) *non altera* il numero dei bit trasmessi.

5.5 Protocolli ARQ

Le trasmissioni ARQ prevedono l'esistenza di un canale di ritorno, mediante il quale chiedere la ri-trasmissione delle trame ricevute con errori⁵⁶; pertanto i dati anche se già trasmessi, devono essere temporaneamente memorizzati al trasmettitore, per rispondere alle eventuali richieste del ricevitore. Viene illustrato per primo un metodo molto semplice, ma potenzialmente inefficiente. Adottando invece buffer (detti *finestre*) di ricezione e trasmissione di dimensioni opportune, si riesce a conseguire una efficienza maggiore.

5.5.1 Send and wait

Viene trasmessa una trama alla volta, e si attende un riscontro (*ACKnowledgment*) di corretta ricezione prima di trasmettere la seguente. Nel caso in cui il ricevitore rilevi un errore, si genera invece un riscontro negativo (NACK), che causa la ritrasmissione della trama trasmessa in precedenza. Se il NACK giunge illeggibile, il trasmettitore attende fino allo scadere di un *allarme a tempo* (TIMEOUT) e quindi ritrasmette comunque l'ultimo dato inviato.

In figura 5.10 è riportata una tipica sequenza di passaggi, in cui (a) la trama $N + 2$ è ricevuta con errore, causando un primo NACK; quindi (b) è l'ACK($N + 3$) ad arrivare errato, causando lo scadere del timeout, e la ritrasmissione della trama $N + 3$. Notiamo

⁵⁴Letteralmente: arrampicamento, ma anche "arruffamento", vedi *scrambled eggs*, le uova strapazzate dell'*english breakfast*.

⁵⁵LEAVE = *foglia, sfogliare, rastrellare*, ed il termine potrebbe essere tradotto come *intercalamento*.

⁵⁶Queste tecniche hanno origine a scopo di controllo degli errori nei collegamenti punto-punto per i quali si osserva una *probabilità di errore* non trascurabile. Successivamente, sono stati utilizzati nelle reti a pacchetto, in cui è possibile la *perdita totale* dei pacchetti in transito. Per questo le implementazioni attuali dei ARQ, specie se applicati da un estremo all'altro di una rete, privilegiano l'uso di *timeout* piuttosto che quello di riscontri negativi.

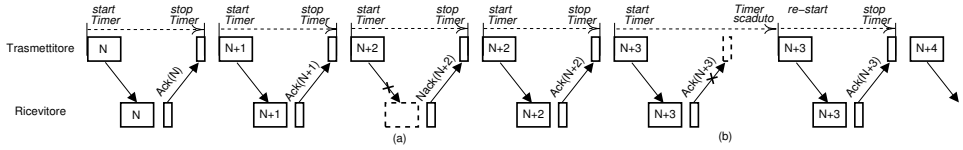


Figura 5.10: Richieste di ritrasmissione per un protocollo send and wait

che le trame devono essere etichettate con un numero di sequenza, in modo da permettere al ricevitore, nel caso (b), di riconoscere la trama come duplicata, e scartarla (l'ACK è inviato comunque per permettere la risincronizzazione del trasmettitore).

Utilizzo del collegamento Considerando l'intervallo di tempo t_T che intercorre tra la trasmissione di due trame consecutive, la trasmissione vera e propria dura solamente t_{Tx} istanti, dopodiché occorre attendere $2 \cdot t_p$ istanti (t_p è il tempo di *propagazione*) prima di ricevere l'ACK. Trascurando gli altri tempi (di trasmissione dell'ACK, e di elaborazione delle trame), si definisce una efficienza di utilizzo

$$U = \frac{t_{Tx}}{t_T} \simeq \frac{t_{Tx}}{t_{Tx} + 2 \cdot t_p} = \frac{1}{1 + 2 \cdot t_p/t_{Tx}} = \frac{1}{1 + 2 \cdot a}$$

in cui il parametro a che determina il risultato, può assumere valori molto diversi, in funzione della velocità di trasmissione e della lunghezza del collegamento.

Esempio Una serie di trame di $N = 1000$ bit viene trasmessa utilizzando un protocollo *send-and-wait*, su tre diversi collegamenti:

- un cavo ritorto di 1 km,
- una linea dedicata di 200 km,
- un collegamento satellitare di 50000 km.

Sapendo che la velocità di propagazione è di $2 \cdot 10^8$ m/sec per i casi (a) e (b), e di $3 \cdot 10^8$ m/sec per il caso (c), determinare l'efficienza di utilizzo $U = \frac{1}{1+2 \cdot a}$, per le due possibili velocità di trasmissione f_b di 1 kbps ed 1 Mbps. □

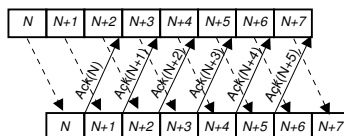
Il tempo necessario alla trasmissione $t_{Tx} = \frac{N}{f_b}$, risulta pari ad 1 sec ed 1 msec alle velocità di 10^3 e 10^6 bps rispettivamente. Il tempo di propagazione $t_p = \frac{\text{spazio}}{\text{velocità}}$ risulta pari a $5 \cdot 10^{-6}$ sec, $1 \cdot 10^{-3}$ sec e 0.167 sec nei tre casi (a), (b), e (c) rispettivamente. Pertanto:

- si ottiene $a = \frac{t_p}{t_{Tx}} = 5 \cdot 10^{-6}$ e $a = 5 \cdot 10^{-3}$ per le velocità di 1 kbps ed 1 Mbps rispettivamente, e quindi per entrambe $U \simeq 1$;
- per $f_b = 1$ kbps si ottiene $a = 10^{-3}$ e quindi $U \simeq 1$, per $f_b = 1$ Mbps risulta $a = 1$ e quindi $U = 0.33$;
- per le velocità di 1 kbps ed 1 Mbps si ottiene $a = 0.167$ ed $a = 167$ rispettivamente, a cui corrispondono efficienze pari a $U = 0.75$ e $U = 0.003$.

Sulla base del risultato dell'esempio notiamo che, considerando fissa la dimensione di trama, le prestazioni di un collegamento nei confronti di un protocollo ARQ possono essere caratterizzate, oltre che dal parametro a , anche dal cosiddetto *Prodotto Banda-Ritardo PBR* $= f_b \cdot t_p$, che infatti nei sei casi in esame vale $5 \cdot 10^{-3}$, 5, 1, 10^3 , 160, $1.6 \cdot 10^5$. Pertanto, abbiamo dimostrato come la trasmissione *send and wait* possa essere idonea per basse velocità e/o collegamenti brevi, in virtù della sua semplicità realizzativa; in caso contrario, è opportuno ricorrere ad uno dei metodi seguenti.

5.5.2 Continuous RQ

A differenza del protocollo *send-and-wait*, ora il trasmettitore invia le trame ininterrottamente, senza attendere la ricezione degli ACK. In presenza di trame ricevute correttamente, il ricevitore riscontra positivamente le stesse, consentendo al trasmettitore di liberare i buffer di trasmissione.

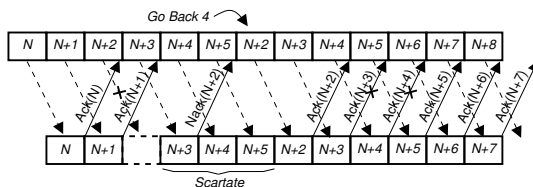


In presenza di trame ricevute con errori, la quantità di memoria tampone utilizzata al ricevitore determina la scelta di due possibili strategie di richiesta di ritrasmissione, denominate *go-back-N* e *selective-repeat*.

5.5.2.1 Go back N

In questo caso, il ricevitore dispone di una sola posizione di memoria, dove trattiene la trama appena ricevuta, per il tempo necessario al controllo di errore. In presenza di un errore di ricezione della trama $N + i$, rilevato⁵⁷ dopo la corretta ricezione di $N + i + 1$, il ricevitore invia un NACK($N + i$), chiedendo con ciò al trasmettitore di *andare indietro*, ed inizia a scartare tutte le trame con numeri maggiori di $N + i$, finché non riceve la $N + i$, e riprende le normali operazioni.

Se, trascorso un timeout, la $N + i$ non è arrivata, si invia di nuovo un NACK($N + i$). Nel caso in cui invece si corrompa un ACK, le operazioni continuano regolarmente, e l'ACK successivo agisce da riscontro positivo anche per le trame per le quali non si sono ricevuti riscontri. Il trasmettitore deve quindi mantenere memorizzate tutte le trame trasmesse e non ancora riscontrate, fino ad un numero massimo, raggiunto il quale la trasmissione si arresta.



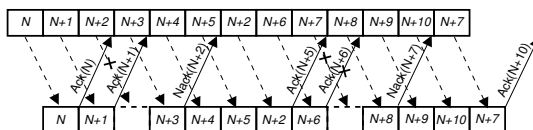
Una variante del metodo, idonea al caso in cui fenomeni di congestione di rete possano determinare la perdita dei NACK, prevede l'uso di un timer al trasmettitore, per re-inviare le trame non riscontrate.

Una variante del metodo, idonea al caso in cui fenomeni di congestione di rete possano determinare la perdita dei NACK, prevede l'uso di un timer al trasmettitore, per re-inviare le trame non riscontrate.

5.5.2.2 Selective repeat

L'origine di questo nome deriva dal fatto che non è più richiesto al trasmettitore di *tornare indietro* completamente, ma è sufficiente ritrasmettere solamente la trama che ha dato origine ad errore in ricezione, grazie alla capacità del ricevitore di memorizzare temporaneamente più trame, anche se ricevute fuori sequenza.

Come si nota in figura, a seguito della ritrasmissione della trama $N + 2$ per cui si è ricevuto il NACK, il trasmettitore continua a trasmettere nuove trame, fino al numero massimo previsto; in assenza di ulteriori ACK, un timer determina la ritrasmissione delle trame non riscontrate.



⁵⁷Sottolineiamo nuovamente l'importanza dei numeri di sequenza, che permettono al ricevitore di capire il numero della trama corrotta, grazie alla discontinuità dei numeri stessi.

Quando al ricevitore perviene la trama $N + 2$, questo emette un $\text{ACK}(N + 5)$, consentendo al trasmettitore di rilasciare tutta la memoria occupata dalle trame in sospenso, e di proseguire la trasmissione. La perdita di uno o più ACK è gestita allo stesso modo che per *go-back-N*, così come per ogni NACK inviato si inizializza un timer, allo scadere del quale ed in assenza di nuove trame ricevute, il NACK è re-inviato.

Dal punto di vista del ricevitore, questo è più complicato che nel caso *go-back-N*, dato che adesso occorre riordinare le trame ricevute, che possono arrivare sequenziate con un ordine differente da quello naturale. Per questo motivo, anche il ricevitore deve predisporre delle memorie temporanee dove salvare le trame correttamente arrivate, successivamente a quella che invece conteneva errori, e di cui si attende la ritrasmissione.

5.5.2.3 Efficienza

Indichiamo con p la probabilità di dover chiedere la ritrasmissione di una trama⁵⁸ per la quale si sono rilevati errori di trasmissione, e con m il numero totale di trasmissioni necessarie alla sua corretta ricezione. Osserviamo quindi che m è una variabile aleatoria, descritta dalle probabilità

$$\begin{aligned} p_M(1) &= Pr(m=1) = 1-p \\ p_M(2) &= p(1-p) \\ p_M(3) &= p^2(1-p) \\ &\vdots \\ p_M(m) &= p^{m-1}(1-p) \end{aligned}$$

in quanto la trama può essere ritrasmessa più volte, finché non si rilevano più errori. Pertanto, il numero medio di trasmissioni per una stessa trama è pari a

$$\begin{aligned} \bar{m} &= E\{m\} = \sum_{m=1}^{\infty} mp_M(m) = \sum_{m=1}^{\infty} mp^{m-1}(1-p) = \\ &= (1-p) \sum_{n=0}^{\infty} (n+1) \cdot p^n = (1-p) \frac{1}{(1-p)^2} = \frac{1}{1-p} \end{aligned}$$

in cui alla quarta eguaglianza si è posto $n = m - 1$, ed alla quinta si è tenuto conto della relazione espressa alla nota 27 di pag. 182, e del fatto che

$$\sum_{k=0}^{\infty} k\alpha^{k-1} = \sum_{k=1}^{\infty} k\alpha^{k-1} = \sum_{k=0}^{\infty} (k+1)\alpha^k$$

Quindi, per trasmettere una frequenza binaria di f_b bps (comprensivi di CRC e *overhead* dei numeri di sequenza), occorre disporre di un canale di capacità $f_b/(1-p)$ bps⁵⁹.

⁵⁸Nel caso in cui l'integrità della trama sia protetta da un codice a blocchi (n, k) con $d_H = l + 1$, la probabilità che la trama contenga più di l errori e che quindi venga accettata dal ricevitore anche se errata, vale approssimativamente $P(l+1, n) = \binom{n}{l+1} p^l$ (vedi formula (5.18)). Dato che il ricevitore accetta le trame che *non hanno* errori, oppure che hanno più di l errori, la probabilità che venga richiesta una ritrasmissione risulta $p = 1 - P(0, n) - P(l+1, n)$. Considerando ora che un buon codice deve fornire $P(l+1, n) \ll P(0, n)$, si ottiene $p \simeq 1 - P(0, n) = 1 - (1 - P_e)^n \simeq nP_e$, in cui P_e è la probabilità di errore sul bit (dato che $(1 - P_e)^n \simeq 1 - nP_e$ se $nP_e \ll 1$).

⁵⁹Dato che p aumenta con n (vedi pag. 94), l'efficienza del protocollo ARQ *peggiora* con l'aumentare della dimensione delle trame. Questo risultato determina l'esigenza di ricercare una soluzione di

Questo risultato approssimato si applica ad un protocollo di tipo *selective repeat*, e trascurando gli errori sul canale a ritroso.

5.5.3 Controllo di flusso

Si è illustrato che nei protocolli ARQ il trasmettitore, dopo un po che non riceve nuovi ACK, cessa a sua volta di inviare trame, dato che esaurisce la memoria temporanea in cui memorizzare le trame già inviate ed in attesa di riscontro. Nel caso in cui il ricevitore non sia in grado di smaltire in tempo i dati ricevuti, può scegliere di sospendere temporaneamente l'invio di riscontri, con il risultato di rallentare la velocità di invio dei dati. Questo meccanismo prende il nome di *controllo di flusso*, per l'evidente analogia idraulica, in cui una condotta viene ristretta al fine di ridurre il flusso di liquido in transito.

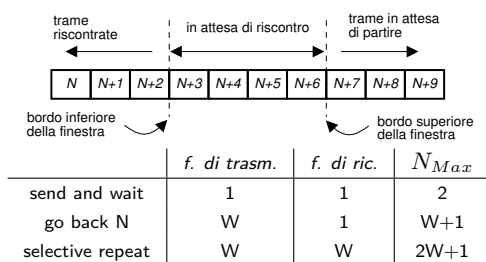
Dato che il ritardo tra la sospensione dell'invio degli ACK, e l'interruzione dell'invio di trame, dipende dalla dimensione delle memorie temporanee, e che questa dimensione incide allo stesso tempo anche sulla efficienza di utilizzo temporale del collegamento in condizioni di ricezione a piena velocità, svolgiamo alcune riflessioni sull'argomento.

5.5.3.1 Round trip time

E' il tempo che intercorre tra l'inizio della trasmissione di una trama, e l'arrivo del relativo ACK. La sua valutazione spesso si avvale della ipotesi di poter trascurare il tempo di trasmissione dell'ACK, e quindi si ottiene $RTT = t_{Tx} + 2t_p$. Se la trasmissione avviene a velocità f_b , allora in un tempo pari a RTT possono essere trasmessi $N_{ba} = f_b \cdot RTT$ bit, che possono essere pensati come il numero di *bit in aria*⁶⁰. Se la memoria temporanea del trasmettitore ha dimensioni $W \geq N_{ba}$, allora la trasmissione (senza errori) può avvenire senza soluzione di continuità, impegnando costantemente il collegamento.

5.5.3.2 Finestra scorrevole

La quantità massima di dati W che è possibile trasmettere senza ricevere un riscontro è indicata come *finestra di trasmissione* per il motivo che ora illustriamo. In figura



si mostra un gruppo di trame oggetto di una trasmissione; quelle già trasmesse ed in attesa di riscontro (da $N + 3$ a $N + 6$ in figura) sono racchiuse tra due confini, i *bordi* della finestra. Ogni volta che ne viene trasmessa una, il bordo *superiore* della finestra è spostato a destra, *allargandola*; ogni volta

compromesso, dato che l'incidenza dell'overhead sulla dimensione complessiva della trama invece *si riduce* all'aumentare di n .

⁶⁰L'espressione "*bit in aria*" trae spunto dalla metafora di una coppia di giocolieri, posti ai due estremi di una piazza, che si lanciano una serie di clave. Il primo ne lancia in continuazione, e quando iniziano ad arrivare al secondo, questi le rilancia verso il primo. Nel momento in cui la clava partita per prima torna nelle mani del primo giocoliere, un certo numero di clave sono sospese a mezz'aria, e corrispondono approssimativamente al numero di bit trasmessi in un tempo di pari durata, con una frequenza pari al ritmo di lancio delle clave, e non ancora riscontrati.

che si riceve un riscontro, è il bordo *inferiore* ad essere spostato a destra, *restringendo* così la finestra. In definitiva, il termine finestra trae origine dal fatto che, allo spostarsi dei bordi inferiore e superiore, la finestra *si apre e si chiude*.

La condizione di *massima apertura* della finestra identifica la quantità di memoria necessaria al trasmettitore per realizzare un protocollo ARQ, che quindi può essere ri-classificato in questi termini come indicato dalla tabella precedente, dove la colonna *finestra di ricezione* indica anche i requisiti di memoria al lato ricevente⁶¹. Notiamo infatti che mentre per *send-and-wait* è sufficiente la memoria di una sola trama, per *go-back-N* il trasmettitore deve ricordare fino ad un massimo di W trame in attesa di riscontro, e per *selective repeat* anche il ricevitore ha lo stesso vincolo, allo scopo di riordinare le trame ricevute fuori sequenza.

5.5.3.3 Numero di sequenza

Dato che non possono essere inviate più trame della dimensione della finestra, la loro numerazione può avvenire in forma ciclica, ossia utilizzando un contatore modulo N_{Max} , come indicato alla tabella precedente. Ad esempio, per *send-and-wait* è sufficiente un contatore binario (0–1) perché, nel caso in cui l'ACK sia corrotto, il ricevitore possa riconoscere la trama ricevuta come duplicata anziché nuova; un ragionamento simile⁶² determina la necessità di usare $W + 1$ numeri (0– W) nel caso *go-back-N*, e $2W + 1$ numeri (0– $2W$) nel caso *selective repeat*.

L'uso di un numero di bit ridotto per indicare il numero di sequenza, permette di limitare la dimensione dell'*overhead* di trama; ad esempio, con una finestra di dimensione 7, l'uso di go-back-N richiede 8 diversi numeri di sequenza, che quindi possono essere codificati utilizzando 3 bit.

Appendici

Esercizio Un sistema di trasmissione basato sul campionamento e sulla trasmissione numerica è rappresentato in figura 5.11. Il canale riportato all'estremità destra è considerato ideale entro una banda $\pm B = \pm 31.5$ KHz, purché la potenza al suo ingresso non superi il valore $\mathcal{P}_y^{Max} = 1$ Volt²; in tal caso la potenza in uscita risulta $\mathcal{P}_{y'} = 0.01 \cdot \mathcal{P}_y$. Al segnale ricevuto è sovrapposto un rumore additivo gaussiano bianco stazionario ergodico a media nulla, con spettro di densità di potenza $\mathcal{P}_N(f) = \frac{N_0}{2} = 4.61 \cdot 10^{-14}$ Volt²/Hz, e limitato nella banda $\pm B$.

- 1) Se $G(f)$ è a coseno rialzato con $\gamma = .5$, determinare la massima frequenza di simbolo $f_s = \frac{1}{T_s}$.
- 2) Desiderando una $P_e = P_e^c$ per la sequenza $\{c'\}$ pari a $P_e = 10^{-4}$, determinare il massimo numero di livelli/simbolo L .

⁶¹La ricezione di una sequenza di trame corrette, determina l'avanzamento alternato dei due bordi della finestra al ricevitore, che è inizialmente vuota, quindi contiene solo la trama ricevuta (avanza bordo superiore), e quindi è di nuovo svuotata, non appena viene trasmesso l'ACK (ed avanza il bordo inferiore). In presenza di errori, il bordo inferiore non avanza, ma resta fermo sulla trama ricevuta con errori, e di cui si attende la ritrasmissione. Mentre il trasmettitore continua ad inviare trame, il ricevitore le memorizza e fa avanzare il bordo superiore, finché non siano state ricevute tutte quelle trasmissibili senza riscontro, e pari alla dimensione massima della finestra in trasmissione.

⁶²Se il trasmettitore invia tutte le W trame, ma tutti gli ACK sono corrotti, allora la $(W + 1)$ -esima trama trasmessa è un duplicato della prima, ritrasmessa per time-out, ed il ricevitore può accorgersene solo se la trama reca un numero differente da quello della prima.

Per il caso *selective repeat*, vale un ragionamento simile, ma che per le differenze nella definizione del protocollo, porta ad un risultato diverso.

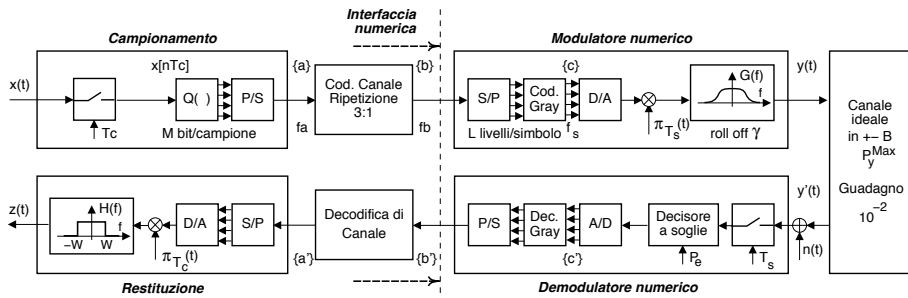


Figura 5.11: Sistema di trasmissione a cui si riferisce l'esercizio

- 3) Indicare la frequenza binaria f_b per la sequenza $\{b'\}$.
- 4) Valutare P_e^b per la sequenza $\{b'\}$ e mostrare che il numero di errori per unità di tempo in $\{b'\}$ è lo stesso che in $\{c'\}$.
- 5) Mostrare che, adottando una codifica di canale a ripetizione 3 : 1, la probabilità di errore P_e^a per la sequenza $\{a'\}$ risulta pari a circa $P_e^a \simeq 3 (P_e^b)^2$.
- 6) Indicare la frequenza binaria f_a per le sequenze $\{a\}$ ed $\{a'\}$.

Supponiamo ora che $P_e^a = 0$, e si desideri un $SNR = \mathcal{P}_x / \mathcal{P}_{z-x} = 10000$. Nel caso in cui $x(t)$ sia un processo con densità di probabilità $p(x)$ uniforme, ed indicando con W la banda di $x(t)$;

- 7) Determinare il minimo numero di bit/campione M .
- 8) Determinare la massima banda W .
- 9) Se la banda è ridotta a $W' = \frac{1}{2}W$, determinare il nuovo valore di SNR ottenibile.

Svolgimento

- 1) La banda B occupata dal segnale y vale $B = \frac{f_s}{2} (1 + \gamma)$, e quindi deve risultare $f_s = \frac{2B}{1+\gamma} = \frac{2 \cdot 31.5 \cdot 10^3}{1.5} = 42 \cdot 10^3 = 42.000$ baud (baud = simboli/secondo).
- 2) Osserviamo che in questo caso la (5.15) non può essere applicata direttamente, in quanto non essendo ancora nota la f_b , non è possibile calcolare il valore di $E_b = \frac{\mathcal{P}_{y'}}{f_b}$. Notiamo però che essendo $f_b = f_s \cdot \log_2 L$, indicando con y l'argomento dell' $erfc\{\cdot\}$, questo può essere riscritto come

$$y = \sqrt{\frac{E_b}{N_0} \frac{3 \log_2 L}{(L^2 - 1) (1 + \gamma) (1 - \frac{\gamma}{4})}} = \sqrt{\frac{\mathcal{P}_{y'}}{f_s \cdot \log_2 L} \frac{1}{N_0} \frac{3 \cdot \log_2 L}{(L^2 - 1) 1.31}}$$

$$= \sqrt{\frac{\mathcal{P}_{y'}}{f_s \cdot N_0} \cdot \frac{2.29}{L^2 - 1}}$$

avendo tenuto conto che se $\gamma = 0.5$, allora $(1 + \gamma) (1 - \frac{\gamma}{4}) \simeq 1.31$. Inoltre, se $L \gg 1$ (come verificheremo), la (5.15) può essere approssimata come $P_e \simeq erfc\{y\}$, e dunque per $P_e = 10^{-4}$ la figura di pag. 155 ci permette di individuare il valore di $y \simeq 2.7$, e pertanto

$$\frac{\mathcal{P}_{y'}}{f_s \cdot N_0} \cdot \frac{2.29}{L^2 - 1} = y^2 = (2.7)^2 = 7.29$$

e, conoscendo i valori di f_s , $\mathcal{P}_{y'}$ e N_0 , scriviamo

$$\begin{aligned} L^2 &= 1 + \frac{\mathcal{P}_{y'}}{f_s \cdot N_0} \cdot \frac{2.29}{7.29} = 1 + \frac{10^{-2}}{42 \cdot 10^{-3} \cdot 4.61 \cdot 10^{-4}} \cdot 0.31 \\ &= 1 + 5.16 \cdot 10^6 \cdot 0.31 \simeq 1.6 \cdot 10^6 \end{aligned}$$

e quindi $L = \sqrt{1.6 \cdot 10^6} = 1265$ che, essendo un valore massimo, limitiamo a $L = 1024$ livelli

- 3) Dato che ad ogni simbolo di $\{c\}$ ad L livelli, con frequenza di emissione pari a f_s , corrisponde ad un gruppo di $N_b = \log_2 L = 10$ bit della sequenza $\{b\}$, la frequenza f_b è di 10 volte f_s , e quindi $f_b = 10 \cdot f_s = 10 \cdot 42 \cdot 10^3 = 420$ Kbps.
- 4) Grazie all'adozione del codice di Gray, in caso di errore tra livelli contigui per i simboli di $\{c'\}$, nella sequenza $\{b'\}$ solo uno (tra N_b) dei bit associati ad un simbolo è errato; il bit errato è uno qualsiasi del gruppo di N_b , e pertanto la probabilità che un bit specifico sia errato (quando è errato il simbolo di $\{c'\}$) è $\frac{1}{N_b}$. Pertanto $P_e^b = P_e^{b/c} P_e^c = \frac{1}{N_b} P_e^c$, in cui $P_e^{b/c}$ è la probabilità condizionata che un generico bit di $\{b'\}$ sia sbagliato quando è sbagliato il simbolo di $\{c'\}$ da cui ha origine.
 - Il numero di bit (della sequenza $\{b'\}$) errati per unità di tempo è dato da $P_e^b \cdot f_b$; sostituendo: $P_e^b \cdot f_b = \frac{P_e^c}{N_b} \cdot f_b = P_e^c \cdot \frac{f_b}{N_b} = P_e^c \cdot f_s$, ovvero è numericamente pari ai simboli errati (nella sequenza $\{c'\}$) per unità di tempo;
 - risulta dunque infine:
 - $P_e^b = \frac{P_e^c}{N_b} = \frac{10^{-4}}{10} = 10^{-5}$;
 - $P_e^b \cdot f_b = P_e^c \cdot f_s = 10^{-5} \cdot 420 \cdot 10^3 = 10^{-4} \cdot 42 \cdot 10^3 = 4.2 \frac{\text{errori}}{\text{secondo}}$
- 5) Ogni bit di $\{a'\}$ è sbagliato solo se sono sbagliati 2 o più bit in un gruppo di 3; come mostrato al capitolo seguente, la probabilità di 2 bit errati su 3 è calcolabile dalla distribuzione di Bernoulli, e vale $\binom{3}{2} p_e^2 (1 - p_e) = 3p_e^2 (1 - p_e)$, a cui va sommata la probabilità di 3 bit errati, pari a p_e^3 . Pertanto $p_e^a = 3p_e^2 (1 - p_e) + p_e^3 = 3p_e^2 - 3p_e^3 + p_e^3 \simeq 3p_e^2$ in cui ovviamente $p_e = P_e^b$, e l'approssimazione è legittima in quanto se $p_e = 10^{-5}$ allora $p_e^2 = 10^{-10}$ e $p_e^3 = 10^{-15}$, trascurabili rispetto a p_e . Lo stesso risultato si ottiene osservando che 2 bit errati su 3 hanno probabilità $p_e^2 (1 - p_e)$, e questi possono essere scelti in tre modi diversi (1° e 2° , 1° e 3° , 2° e 3°). In definitiva, risulta $P_e^a \simeq 3 (P_e^b)^2 = 3 \cdot 10^{-10}$.
- 6) Dato che ad ogni 3 bit di $\{b'\}$ corrisponde un solo bit di $\{a'\}$, si ottiene $f_a = \frac{f_b}{3} = \frac{420 \cdot 10^3}{3} = 140$ Kbps, a cui corrisponde $P_e^a \cdot f_a = 3 \cdot 10^{-10} \cdot 140 \cdot 10^3 = 4.2 \cdot 10^{-5} \frac{\text{errori}}{\text{secondo}}$.
- 7) Sappiamo che per un processo uniforme l' SNR_q di quantizzazione risulta approssimativamente $SNR_q = (L - 1)^2$, in cui L è il numero di livelli del quantizzatore, a cui corrisponde l'utilizzo di $M = \log_2 L$ bit/campione. Risulta pertanto $L = 1 + \sqrt{SNR_q} = 1 + \sqrt{10^4} = 101$ livelli. Per ottenere un numero intero di bit/campione ed un SNR_q migliore od uguale a quello desiderato, determiniamo l'intero superiore: $M = \lceil \log_2 L \rceil = 7$ bit/campione (equivalente a 128 livelli).
- 8) Come sappiamo, la frequenza di campionamento $f_c = \frac{1}{T_c}$ non può essere inferiore a $2W$; inoltre, la frequenza binaria f_a risulta pari al prodotto dei bit/campione per i campioni a secondo: $f_a = f_c \cdot M$; pertanto $f_c = \frac{f_a}{M} = \frac{140 \cdot 10^3}{7} = 20$ KHz e dunque la W massima risulta $W_{Max} = \frac{f_c}{2} = 10$ KHz.
- 9) Nel caso in cui $W' = \frac{1}{2}W$, allora si può dimezzare anche la frequenza di campionamento $f'_c = \frac{f_c}{2} = 10$ KHz, e pertanto utilizzare un $M' = 2M$ per ottenere la stessa f_a . Pertanto il nuovo SNR'_q risulta $SNR'_q = (L' - 1)^2 = (2^M - 1)^2 = (2^7 - 1)^2 \simeq 2.68 \cdot 10^8$, ovvero SNR'_q (dB) = 84.3 dB.

5.6 Sincronizzazione dati

Come mostrato al § 5.1.1, il segnale dati deve essere campionato al ricevitore, con cadenza pari alla frequenza di simbolo f_s , possibilmente al centro dell'intervallo di simbolo, in modo da contrastare gli effetti della limitazione di banda (vedi fig. 5.2); per questo motivo, occorre che il temporizzatore mostrato in fig. 5.1 determini gli istanti di campionamento più idonei, effettuando la *sincronizzazione di simbolo*⁶³; le diverse scelte per l'onda elementare $g(t)$ discusse al § 5.2.1, determinano differenti gradi di "difficoltà" nel conseguimento della sincronizzazione di simbolo.

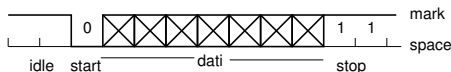
Un differente aspetto della sincronizzazione, riguarda il problema di ricostruire la struttura *sintattica*⁶⁴ del segnale binario. Infatti, la sequenza di bit ottenuta al ricevitore, è spesso il risultato della serializzazione (al lato del trasmettitore) di informazioni *a carattere* (o *parola*, o WORD), come nel caso di un *file* di testo, oppure dei campioni di un segnale⁶⁵. Pertanto, il ricevitore deve essere in grado di delimitare, nell'ambito del flusso di bit ricevuti, le unità di *parola*, e di raggruppare le parole in *trame*.

Nel seguito, analizziamo le esigenze e le soluzioni di sincronizzazione che emergono nell'ambito di due diverse tecniche di trasmissione, indicate come *asincrona* e *sincrona*, che si differenziano per il fatto che mentre nella prima le parole sono separate tra loro, nella seconda fluiscono senza interruzione.

5.6.1 Trasmissione asincrona

Viene adottata, ad esempio, nel caso di un terminale *stupido*⁶⁶ collegato ad un computer centrale: la trasmissione in questo caso avviene in modo discontinuo, ossia quando l'operatore *digita* sui tasti del terminale, e per questo la modalità di trasmissione è indicata come *asincrona*. In tal caso, la linea di comunicazione permane abitualmente in uno stato di libero (IDLE), contraddistinta da uno stato di tensione positiva, indicato anche come stato *mark*⁶⁷.

Quando è pronto un carattere da trasmettere, il segnale viene portato nello stato *zero* (detto SPACE) per la durata di 1 simbolo, che prende il nome di *bit di start*: la transizione *in discesa* viene rilevata dal ricevitore, che si predispone a contare un numero fisso di simboli (7 in figura) basandosi su di un suo orologio indipendente.



Segue poi un bit di *parità* (se prevista, vedi § 5.4.2.1 per la sua definizione) ed uno o due di *stop* (realizzato con MARK), presenti per assicurare una durata minima dello stato di IDLE, prima della trasmissione del carattere successivo.

⁶³In alternativa al recupero del sincronismo da parte del ricevitore, l'informazione di temporizzazione può essere trasmessa su di una diversa linea, come avviene nel caso di dispositivi ospitati su di uno stesso circuito stampato.

⁶⁴Una sintassi definisce un linguaggio, prescrivendo le regole con cui possono essere costruite sequenze di simboli noti (l'alfabeto), e l'analisi delle sequenze eseguita nei termini degli elementi definiti dalla sintassi, ne permette una interpretazione semantica. Il parallelismo linguistico porta spontaneamente ad indicare i simboli trasmessi come *alfabeto*, gruppi di simboli come *parole*, e gruppi di parole come *frasi*, od in alternativa, *trame* (FRAME, ovvero *telaio*).

⁶⁵In appendice 5.9 è riportata la codifica in termini di sequenze binarie dei caratteri stampabili, definita dallo standard ASCII; al § 6.3.1 si mostra la struttura della *trama* PCM, che trasporta i campioni di più sorgenti analogiche campionate.

⁶⁶Un DUMB TERMINAL non ha capacità di calcolo, e provvede solo alla visualizzazione di informazioni testuali. Fino agli anni '70, è stato l'unico meccanismo di interazione (comunque migliore delle schede perforate !!!) con un computer.

⁶⁷In tal caso la linea ".. IS MARKING TIME" (sta marcando il tempo).

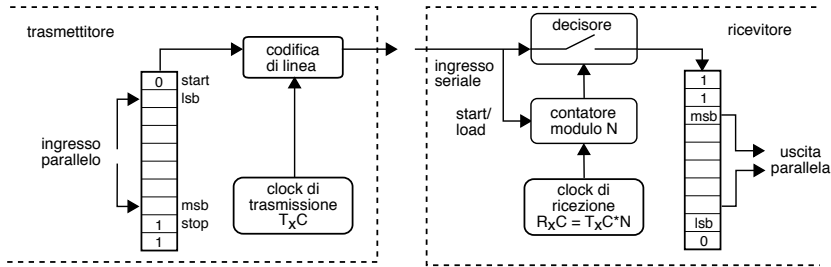


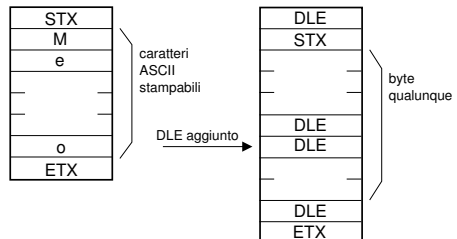
Figura 5.12: Trasmissione asincrona

Il vantaggio di una simile modalità operativa è che il ricevitore non ha bisogno di generare con estrema esattezza la temporizzazione del segnale entrante; si basa infatti su di un proprio orologio locale, di precisione non elevata⁶⁸, che viene *risvegliato* in corrispondenza del bit di start. Tale semplicità operativa produce una inefficienza, in quanto oltre ai dati ed al bit di parità, si introduce anche lo start e lo stop, utili solo ai fini della sincronizzazione ma privi di contenuto informativo.

In fig. 5.12 è mostrato uno schema di funzionamento per trasmettitore e ricevitore, che mostra come le parole entrino in modo parallelo nel trasmettitore, che produce un segnale dati in forma seriale a velocità f_s pari al clock $T_x\text{Clock}$ di trasmissione, con un bit di start, la parola dal *less* al *most significant bit*, e quindi due bit di stop.

Sincronizzazione di bit e di parola Il ricevitore dispone di un orologio interno operante ad un ritmo ($R_x\text{Clock}$) multiplo di quello di trasmissione T_xC , che incrementa un contatore modulo N , il quale quindi *divide* l'orologio per N , e che quando si azzerza determina il processo di decisione per un bit, il suo caricamento nel registro di ricezione, e lo scorrimento in basso dei contenuti dello stesso. Alla ricezione del bit di start, il contatore è inizializzato con il valore $N/2$, in modo che il suo azzeramento avvenga a metà del periodo di bit. A seguito dell'azzeramento, il contatore torna a dividere R_xC per N , ed i suoi successivi azzeramenti avvengono sempre a metà del periodo di bit, fino alla ricezione di una intera parola. Al suo termine (segnalato dall'arrivo in prima posizione del bit di start a zero), il registro di ricezione è letto in modo parallelo, recuperando una intera parola.

Sincronizzazione di trama I caratteri trasmessi possono far parte di messaggi più estesi, come ad esempio i paragrafi di un *file* di testo. Per questo motivo, può essere necessario inserire dei caratteri speciali tra quelli trasmessi, con lo scopo di delimitare i gruppi di caratteri che appartengono ad uno stesso messaggio. Se le diverse parole da trasmet-



⁶⁸Ovviamente, occorre stabilire un accordo a priori a riguardo la velocità, sia pure approssimata, della trasmissione.

tere non sono tutte quelle possibili in base alla lunghezza di parola adottata⁶⁹, la delimitazione può essere attuata mediante l'uso di caratteri speciali (di controllo) che non compaiono nel messaggio, come ad esempio i caratteri STX (*Start of Text*) e ETX (*End of Text*) dell'insieme ASCII (pag. 118), come mostrato nella figura precedente.

Se invece le parole trasmissibili sono qualsiasi, come nel caso della trasmissione di campioni di segnale, allora il carattere speciale ETX potrebbe essere *simulato* dai dati trasmessi⁷⁰, causando un *troncamento* prematuro del messaggio. In tal caso, sia STX che ETX vengono fatti precedere da un terzo carattere speciale, il DLE (*Data Link Escape*). Il trasmettitore, dopo aver inserito la coppia DLE-STX iniziale, ispeziona ogni carattere da inviare, e se questo *simula* un DLE, *inserisce* un secondo DLE, attuando una procedura detta CHARACTER (o BYTE) STUFFING. Il ricevitore a sua volta, per ogni DLE ricevuto, controlla se la parola successiva è un ETX, nel qual caso considera terminata la trasmissione; altrimenti, controlla se è un secondo DLE, che è stato inserito dal trasmettitore, e lo rimuove. Altri casi non sono possibili, e se si verificano, rivelano la presenza di un errore di trasmissione.

5.6.2 Trasmissione sincrona

La trasmissione dei bit di start e di stop necessaria per effettuare una trasmissione asincrona è fonte di inefficienza, e per questo a velocità più elevate si preferisce non inframmezzare i dati da trasmettere con delimitatori aggiuntivi. Ciò comporta l'esigenza di adottare in ricezione soluzioni apposite per individuare gli istanti di decisione corretti, e quindi conseguire il sincronismo di simbolo. Il sincronismo di parola si basa in generale sull'uso di parole di lunghezza costante, mentre quello di trama prevede due possibili soluzioni, l'una orientata al carattere, e l'altra al bit.

Sincronizzazione di simbolo La figura 5.13-a mostra uno schema idoneo ad estrarre la temporizzazione RxC dal segnale ricevuto, basata sull'uso di un circuito DPLL (*Digital Phase Lock Loop*), il cui funzionamento richiede la presenza di transizioni nel segnale ricevuto. Analogamente allo schema già analizzato nel caso di trasmissione asincrona, un orologio locale opera ad una frequenza N volte più elevata di quella nominale, e il DPLL (fig. 5.13-b) ne divide l'orologio per N , fornendo il segnale RxC necessario al decisore per individuare gli istanti posti al centro di un intervallo di simbolo. La divisione per N è realizzata all'interno del DPLL mediante un contatore all'indietro, che al suo azzeramento produce il segnale RxC di sincronismo, ed è ri-caricato con la costante N . Nel caso in cui si verifichi uno *slittamento di fase* tra il segnale ricevuto e l'orologio locale, questo può essere rilevato osservando che la transizione (quando presente) nel segnale non ricorre nella posizione presunta, ossia a metà del conteggio, ma in anticipo od in ritardo (fig. 5.13-c). In tal caso, il contatore che realizza la divisione viene inizializzato con un numero rispettivamente minore o maggiore di N , in modo che il successivo impulso di sincronismo RxC risulti spostato verso il centro del periodo di simbolo⁷¹.

⁶⁹Una parola di M bit descrive uno spazio di 2^M diversi elementi. Se le parole trasmissibili non sono tutte le 2^M possibili, alcune di queste (che non compariranno mai all'interno del messaggio) possono essere usate per la sua delimitazione.

⁷⁰Cioè, i dati trasmessi, che ora riempiono tutto lo spazio delle configurazioni possibili, contengono al loro interno la configurazione che è propria del carattere ETX.

⁷¹In termini generali, questo circuito è assimilabile ad un circuito di controllo, in quanto il suo principio di funzionamento si basa sul tentativo di azzerare una grandezza di errore. Infatti, la sincronizzazione dell'orologio del decisore di ricezione con il periodo di simbolo del segnale ricevuto,

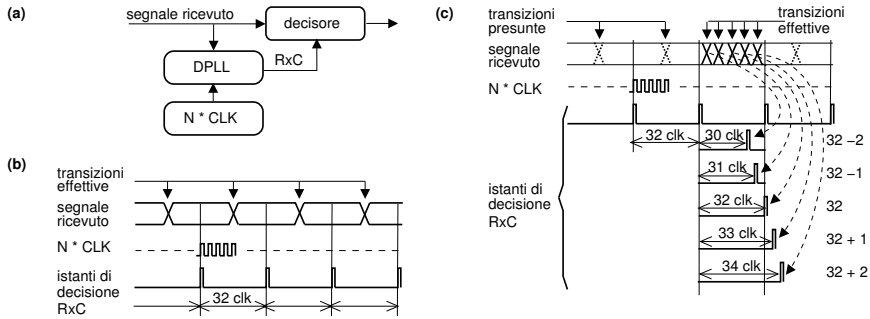
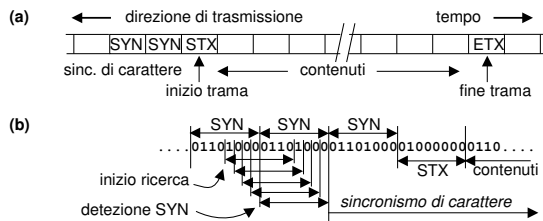


Figura 5.13: Funzionamento del DPLL: (a) schema circuitale; (b) condizioni di sincronismo; (c) correzione di fase

Nel caso di una differenza di velocità tra l'orologio di ricezione ed il ritmo di segnalazione, le correzioni avvengono di rado, e sono di entità ridotta. Al contrario, all'inizio di una trasmissione la differenza di fase può essere qualsiasi: per questo motivo, prima dei dati veri e propri, viene trasmessa una sequenza *trailer*, allo scopo di permettere appunto l'acquisizione del sincronismo di simbolo. La durata del trailer dipende dalla velocità di convergenza della procedura, per accelerare la quale sono imposte correzioni di maggiori entità in corrispondenza di errori di fase più elevati.

Trasmissione orientata al carattere La trasmissione orientata al carattere è usata principalmente nel caso di contenuti testuali, come per i file ASCII. In assenza dei bit di start e di stop, la sincronizzazione di carattere è ottenuta per mezzo della trasmissione, prima dei dati veri e propri, di una sequenza di caratteri SYN (*Synchronous Idle*), che permettono sia di conseguire (o mantenere) il sincronismo di bit, che di consentire l'individuazione dei confini di carattere, e quindi il sincronismo di carattere.

La figura a fianco mostra (a) che la sincronizzazione di trama è ottenuta come per il caso asincrono, racchiudendo il blocco da trasmettere entro una coppia di caratteri STX-ETX. Una volta che il ricevitore ha conseguito il sincronismo di bit, passa in una modalità di ricerca, verificando (b) se l'allineamento di 8 bit consecutivi corrisponde al carattere SYN, ed in caso negativo, ripete il tentativo bit a bit.



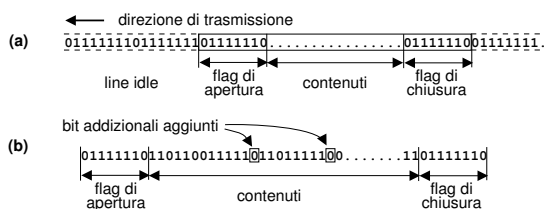
Una volta individuato il SYN, il ricevitore ha conseguito l'allineamento sul carattere, ed inizia ad aspettare il carattere STX, che indica l'inizio della trama, la quale è terminata da un ETX. Nel caso in cui la trasmissione contenga caratteri qualunque, e dunque L'ETX possa essere simulato dai dati, si ricorre alla stessa soluzione del caso

avviene effettuando un confronto tra la *velocità* dell'orologio locale ed un *ritmo* presente nel segnale in arrivo: questo segnale di errore alimenta quindi un circuito di controeazione, che mantiene il clock locale *al passo* con quello dei dati in arrivo. Un caso particolare di questo stesso principio è analizzato ai §11.2.1.2 e 11.3.1.1 a proposito del PLL.

asincrono, e cioè sia l'STX che l'ETX vengono fatti precedere da un DLE, ed all'interno dei dati si esegue la *byte stuffing*, sostituendo gli eventuali DLE simulati con una coppia di DLE.

Trasmissione orientata al bit Questa tecnica viene preferita sia nel caso in cui i dati da trasmettere non siano organizzati in caratteri, sia per ridurre l'inefficienza legata all'uso di caratteri di controllo aggiuntivi, nonché per evitare la dipendenza da quei particolari caratteri. Nella trasmissione orientata al bit, la sincronizzazione di bit e di trama non impiega i caratteri SYN e STX, bensì degli *idle bytes* 01111111 nei periodi di inattività, e dei *flag byte* 01111110 per indicare sia l'inizio che la fine di una trama.

La figura che segue mostra in (a) un esempio di trama, ed in (b) la soluzione della *bit stuffing*, necessaria ad evitare che il *flag byte* possa essere simulato dal contenuto della trasmissione. Ora



i dati trasmessi non devono essere necessariamente in numero multiplo della lunghezza di carattere, ed ogni qualvolta sono presenti 5 bit pari ad uno consecutivi, il trasmettitore inserisce forzatamente un bit pari a zero.

In tal modo, quando il ricevitore osserva un bit pari a 0 preceduto da 5 bit pari ad uno consecutivi, lo rimuove, conseguendo così la *trasparenza dai dati*, e permettendo il corretto rilevamento del flag byte di fine trama. Ovviamente, la procedura di bit stuffing/destuffing è applicata solamente al *contenuto* della trama.

5.7 Ricevitore ottimo

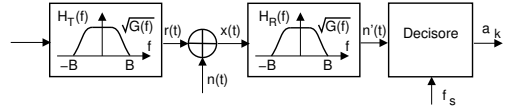
In questa sezione rimettiamo in discussione i risultati ottenuti ai § 5.2.2.3 e 5.3.3.1. Infatti, come sarà illustrato al § 9.5 in relazione al *filtro adattato*, in presenza di rumore bianco il valore di SNR presente nel punto di decisione è *massimo* se si usa un filtro di ricezione $h_R(t)$ *adattato* alla forma dell'impulso trasmesso $g(t) = h_T(t)$, ovvero (a meno di traslazioni temporali) per il quale risulti $H_R(f) = G^*(f)$. Invece, nello schema adottato per la figura a pag. 86, il filtro di ricezione possiede il solo scopo di limitare la banda del rumore, ed è sempre un passa-basso ideale, indipendente da $g(t)$. Allora, se si adotta una $G(f)$ di Nyquist *non* a banda minima, i campioni di rumore sovrapposti a quelli di segnale, danno luogo a v.a. $x(kT_s)$ gaussiane ma non più indipendenti⁷², e la P_e che si ottiene non è la minima possibile⁷³.

⁷²Infatti, il segnale $n(t)$ uscente da $H_R(f) = \text{rect}_{2B}(f)$ ha autocorrelazione $\mathcal{R}_N(\tau) = \mathcal{F}^{-1}\{|H(f)|^2\} = 2B\text{sinc}(2B\tau)$ (vedi § 9.2.3), che passa da zero per $\tau = \frac{1}{2B}$. Se si utilizza una $G(f)$ a coseno rialzato con $\gamma > 0$, occorre estendere la banda di ricezione a $B = \frac{f_s}{2}(1 + \gamma)$, a cui corrispondono campioni di rumore incorrelati se prelevati a distanza multipla di $\tau = \frac{1}{2B} = \frac{1}{f_s(1 + \gamma)}$. Invece, il rumore è campionato con frequenza pari a quella di simbolo f_s , e dunque con campioni a distanza $\tau = T_s = \frac{1}{f_s}$. Pertanto, i campioni di rumore sono correlati, con autocorrelazione pari a $\mathcal{R}_N(T_s) = 2B\text{sinc}(1 + \gamma)$.

⁷³Infatti, la dipendenza statistica tra i campioni di rumore, permetterebbe di realizzare un dispositivo *predittore lineare* che, in base alla conoscenza dei precedenti valori di rumore, calcoli una stima del valore corrente la quale, sottratta al valore effettivamente ricevuto, consente di ridurre la varianza della grandezza di osservazione, permettendo una riduzione della probabilità di errore.

Per rendere incorrelati i campioni di rumore, e ridurre la P_e al minimo, realizzando al contempo le condizioni di Nyquist in ricezione, tentiamo di verificare le condizioni $H(f) = G^*(f)$ di filtro adattato, decomponendo la caratteristica di Nyquist $G(f)$ in parti uguali tra trasmettitore e ricevitore, e dando quindi luogo allo schema di figura, in cui

$$H_T(f) = H_R(f) = \sqrt{G(f)}$$



In tal modo, al decisore giunge esattamente lo stesso segnale di prima⁷⁴, mentre la densità di potenza del rumore non è più costante, ma ora vale

$$\mathcal{P}_N(f) = \frac{N_0}{2} |H_R(f)|^2 = \frac{N_0}{2} G(f)$$

Pertanto, i campioni di rumore presi a distanza T_s sono statisticamente indipendenti (e quindi incorrelati perché gaussiani, vedi § 9.10.2), in quanto $\mathcal{R}_N(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_N(f)\}$ è ora un impulso di Nyquist, che passa da zero per $\tau = kT_s$. Notiamo che, essendo $G(f)$ reale pari, la fattorizzazione di $G(f)$ realizza effettivamente la condizione $H_R(f) = H_T^*(f)$ che definisce un filtro adattato.

Per determinare le nuove prestazioni nel caso in cui $G(f)$ sia a coseno rialzato, notiamo che mentre la banda passante di $H_R(f)$ (e dunque del rumore) si è mantenuta pari a $B = \frac{f_s}{2}(1 + \gamma)$, la potenza del rumore ora vale⁷⁵

$$\mathcal{P}_N = \int_{-\infty}^{\infty} \frac{N_0}{2} G(f) df = \frac{N_0}{2} f_s = \frac{N_0}{2T_b \log_2 L}$$

riducendosi di un fattore $(1 + \gamma)$ se confrontata con (5.11), e causando un aumento equivalente per l' SNR ; lo stesso fattore $(1 + \gamma)$ è quindi rimosso anche nella (5.17), portando a

$$P_e^{bit} = \frac{1}{\log_2 L} \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0} \frac{3 \log_2 L}{(L^2 - 1)(1 - \frac{\gamma}{4})}} \right\} \quad (5.19)$$

il valore della probabilità di errore sul bit adottando il ricevitore ottimo ed il codice di Gray. Dato che al massimo $1 + \gamma = 2$, questo corrisponde ad un *miglioramento massimo* di 3 dB nel valore di E_b/N_0 , permettendo di usare ancora le curve di fig. 5.8. D'altra parte, il fatto che la (5.19) coincida con la (5.17) per $\gamma = 0$ non è un risultato inatteso: infatti, se $\gamma = 0$ si attua una trasmissione a banda minima, e dunque un $H_R(f)$ rettangolare passabasso costituisce proprio il filtro adattato!

Pur non entrando nei dettagli dei metodi di predizione lineare, notiamo che la *correlazione* tra grandezze aleatorie ne determina la *dipendenza*, e che la conoscenza di valori passati consente di ridurre l'incertezza relativa ai nuovi valori. Il valore dei campioni di rumore precedenti, si calcola a partire da quello del simbolo *deciso* senza commettere errore, sottratto al valore del segnale ricevuto in quell'istante.

⁷⁴ Infatti, se $G(f)$ è tutta al trasmettitore, il segnale generato (e ricevuto) ha espressione (5.8) (vedi anche la (5.1)); indicando ora $g^\vee(t) = \mathcal{F}^{-1}\{\sqrt{G(f)}\}$, ed eseguendo un calcolo del tutto analogo a quello svolto in § 5.1.2.2, si ottiene che il segnale ricevuto nel caso di scomposizione di $G(f)$ ha espressione

$$r(t) = h_T(t) * h_R(t) * \sum_k a_k \cdot \delta(t - kT_s) = \sum_k a[k] \cdot g(t - kT_s)$$

in quanto $h_T(t) * h_R(t) = g^\vee(t) * g^\vee(t) = g(t)$ per la proprietà di prodotto in frequenza.

⁷⁵ Il risultato si può ottenere visivamente, a partire dalla $G(f)$ a coseno rialzato mostrata in fig. 5.4 a pag. 83, considerando la risposta di ampiezza nominale pari ad 1, e in base alle sue proprietà di simmetria attorno ad f_s : non è nient'altro che l'area di un rettangolo.

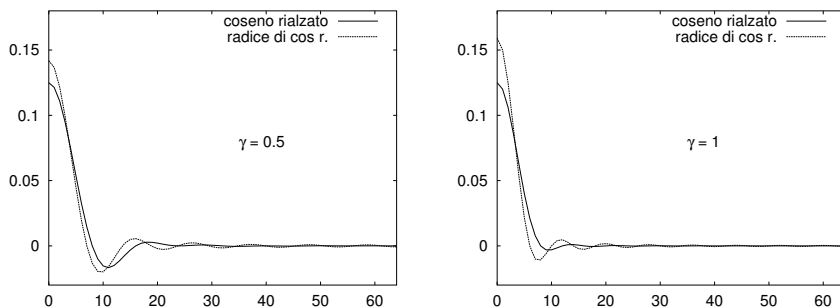


Figura 5.14: Confronto della risposta impulsiva del filtro ottimo e subottimo

Conseguenze L'adozione di un filtro di trasmissione $H_T(f) = \sqrt{G(f)}$ comporta che ora nel segnale trasmesso è presente ISI, che può essere rimossa solo mediante filtraggio dello stesso attraverso il filtro adattato $H_R(f) = \sqrt{G(f)}$.

In figura 5.14 si mostra l'andamento di $g^\vee(t) = \mathcal{F}^{-1}\{\sqrt{G(f)}\}$, posto a confronto con una $g(t)$ a coseno rialzato, per valori di roll-off pari a 0.5 ed 1, ottenuto mediante IFFT della corrispondente risposta in frequenza di modulo unitario nell'origine. Notiamo un aumento sia della durata che della ampiezza delle oscillazioni: questa circostanza determina una maggiore complessità realizzativa del filtro di trasmissione, che deve avere una risposta impulsiva più lunga⁷⁶.

Una seconda considerazione può essere svolta a riguardo del caso in cui sia presente un canale di trasmissione con risposta in frequenza $H(f)$ non ideale; in tal caso, occorre realizzare un filtro di trasmissione $H_T(f)$ tale che $H_T(f)H(f) = \sqrt{G(f)}$. Spesso l'equalizzazione è invece svolta al lato ricevente, in modo da ottenere $H_{eq}(f)H(f) = \sqrt{G(f)}$, ma in tal caso si ottiene una soluzione solamente *sub-ottima*, dato che si perde l'incorrelazione dei campioni di rumore, vedi § 9.2.3.

5.8 Equalizzazione numerica

Quando il segnale dati attraversa un canale che presenta una $H(f)$ non ideale (vedi pag. 349), la *distorsione lineare* introdotta causa la presenza di ISI nel segnale dati (vedi § 5.1.2.2), che appare quindi costituito da impulsi distorti $\tilde{g}(t) = g(t) * h(t)$ anziché $g(t)$. Tale fenomeno può essere *compensato* facendo transitare il segnale ricevuto attraverso un *filtro di equalizzazione* $H_{eq}(f)$ tale che

$$H(f)H_{eq}(f) = ae^{-j2\pi f\tau} \quad \text{ovvero} \quad H_{eq}(f) = \frac{ae^{-j2\pi f\tau}}{H(f)}$$

Se il lato ricevente conosce $H(f)$, può calcolare la $H_{eq}(f)$ teorica, e realizzare un filtro trasversale (vedi § 9.8) che la approssimi, calcolandone i coefficienti mediante uno sviluppo in serie. Il troncamento di tale sviluppo, necessario ad ottenere un $h_{eq}(t)$ di durata finita, comporta un errore residuo, che viene accettato purché piccolo.

⁷⁶Per una analisi degli effetti della limitazione temporale dell'impulso $g^\vee(t)$, vedere l'analisi svolta presso <https://engineering.purdue.edu/~ee538/SquareRootRaisedCosine.pdf>.

Ma spesso ciò che è direttamente disponibile al ricevitore *non è* l' $H(f)$ del canale, ma solamente *ciò che ne esce*; inoltre, nel caso delle trasmissioni numeriche siamo interessati ai soli valori di segnale *negli istanti* multipli del periodo di simbolo. Il metodo di equalizzazione presentato appresso tiene conto di queste due particolarità.

5.8.1 Equalizzazione zero forcing

Questa tecnica prevede che gli impulsi ricevuti distorti $\tilde{g}(t)$ attraversino un filtro trasversale, i cui coefficienti sono calcolati in modo da ripristinare in forma approssimata le condizioni di Nyquist (nel dominio del tempo) per l'impulso $g_{eq}(t)$ in uscita dal filtro. Mostriamo come.

A lato è mostrato un filtro trasversale (vedi § 9.8) con $2N + 1$ coefficienti ed un ritardo totale $2NT$. Il messaggio informativo

è preceduto dalla trasmissione di un impulso isolato *di apprendimento*, la cui forma d'onda distorta ricevuta $\tilde{g}(t)$ presenta un picco a $t = 0$ ed ISI su entrambi i lati (Fig. 5.15a). Ponendo $\tilde{g}(t)$ in ingresso al filtro $h_{eq}(t)$, all'uscita si ottiene l'impulso

$$g_{eq}(t) = \sum_{n=-N}^N c_n \tilde{g}(t - nT - NT) \quad (5.20)$$

che campionato agli istanti $t_k = kT + NT$ fornisce

$$g_{eq}(t_k) = \sum_{n=-N}^N c_n \tilde{g}(kT - nT) = \sum_{n=-N}^N c_n \tilde{g}_{k-n} \quad (5.21)$$

avendo adottato l'abbreviazione $\tilde{g}_{k-n} = \tilde{g}(kT - nT)$; il risultato prende quindi la forma di una *convoluzione discreta*. L'operazione di *equalizzazione Zero Forcing* consiste nel calcolare i coefficienti c_n in modo che $g_{eq}(t)$ soddisfi le condizioni di Nyquist almeno sui $2N + 1$ termini centrati sull'origine, ovvero

$$g_{eq}(t_k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, \dots, \pm N \end{cases}$$

assicurando così l'assenza di ISI *almeno* nei confronti degli N simboli precedenti e successivi. Il valore dei coefficienti c_n in grado di soddisfare queste condizioni si ottengono risolvendo il sistema di $2N + 1$ equazioni nelle $2N + 1$ incognite, impostato a partire dall'espressione (5.21), ed rappresentato in forma matriciale come

$$\begin{bmatrix} \tilde{g}_0 & \cdots & \tilde{g}_{-2N} \\ \vdots & & \vdots \\ \tilde{g}_{N-1} & \cdots & \tilde{g}_{-N-1} \\ \tilde{g}_N & \cdots & \tilde{g}_{-N} \\ \tilde{g}_{N+1} & \cdots & \tilde{g}_{-N+1} \\ \vdots & & \vdots \\ \tilde{g}_{2N} & \cdots & \tilde{g}_0 \end{bmatrix} \begin{bmatrix} c_{-N} \\ \vdots \\ c_{-1} \\ c_0 \\ c_1 \\ \vdots \\ c_N \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (5.22)$$

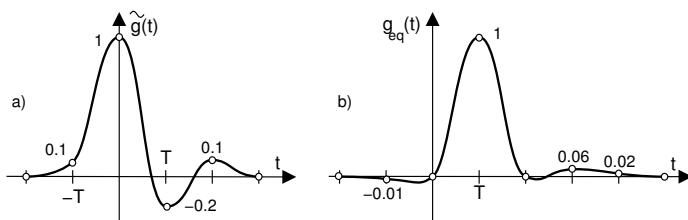


Figura 5.15: Impulso con ISI prima e dopo equalizzazione

in cui la matrice dei termini noti è costituita dai $4N + 1$ campioni di $\tilde{g}(t)$ prelevati agli istanti di simbolo $-2N, \dots, 2N$ posti in modo simmetrico rispetto allo zero. Anche se il metodo non garantisce nulla per istanti esterni a $(-NT, NT)$, è *ottimo* nel senso che minimizza l'ISI di picco, ed è semplice da realizzare. L'operazione di combinazione di più campioni ricevuti, eseguita dal filtro di equalizzazione, può avere l'effetto di aumentare la potenza di rumore, ma spesso questo peggioramento di qualità è più che compensato dal miglioramento dell'ISI.

Esempio Vogliamo realizzare un equalizzatore del terzo ordine, applicando il principio illustrato all'impulso con ISI mostrato alla fig. 5.15a. Inserendo quei valori di \tilde{g}_k nella matrice dei coefficienti (5.22), otteniamo

$$\begin{bmatrix} 1.0 & 0.1 & 0.0 \\ -0.2 & 1.0 & 0.1 \\ 0.1 & -0.2 & 1.0 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

da cui è possibile calcolare $c_{-1} = -0.096$ $c_0 = 0.96$ $c_1 = 0.2$. Inserendo ora ad uno ad uno i campioni di $\tilde{g}(t)$ in un filtro trasversale con questi coefficienti, si ottengono i valori di $g_{eq}(t)$ mostrati in fig. 5.15b assieme ad una curva interpolata. Notiamo che sebbene sia stato ottenuto il risultato desiderato di azzerare $g_{eq}(t)$ ai due istanti di simbolo ai lati del picco, risulta ancora essere $g_{eq} \neq 0$ ad istanti più remoti.

5.9 Codifica di carattere

Il codice ASCII (*American Standard Code for Information Inter-change*) è un codice a 7 bit, e molti codici ad 8 bit (come l'ISO 8859-1) si riducono ad ASCII nella loro metà bassa (con il bit più significativo a zero); i primi 32 codici corrispondono a caratteri *non stampabili*, detti codici di controllo, ottenibili su di una tastiera mediante la pressione del tasto CONTROL, e che hanno un significato speciale, come il *carriage return* (CR), il *line feed* (LF), *start of text* (STX), *backspace* (BS), *data link escape* (DLE). La tavola 5.1 mostra i 128 caratteri ASCII. La controparte internazionale dell'ASCII è nota come ISO 646; lo standard è stato pubblicato dallo *United States of America Standards Institute* (USASI) nel 1968.

5.9.1 Codifica UNICODE

Dal 2004 ISO/IEC non si occupa più della manutenzione delle codifiche di carattere ad 8 bit, supportando invece attivamente il consorzio UNICODE nella definizione dello *Universal Character Set*, che contiene centinaia di migliaia di caratteri di praticamente tutte le lingue del mondo, ognuno identificato in modo non ambiguo da un nome, e da un numero chiamato *Code Point*. Mentre per enumerare tutti i caratteri previsti

occorre una parola di ben 21 bit, sono state definite codifiche a lunghezza variabile, la più diffusa delle quali prende il nome di UTF-8, in base alla quale

- i primi 127 CodePoints, che corrispondono all'alfabeto ASCII, sono rappresentati da un singolo byte; pertanto un file ASCII è anche un file UTF-8 corretto
- i valori numerici associati ai caratteri dell'insieme ISO 8859-1 corrispondono ai CodePoints degli stessi caratteri
- i primi 1792 CodePoints, mediante i quali sono rappresentati i caratteri usati dalla totalità delle lingue occidentali, sono rappresentati (esclusi gli ASCII) mediante due byte
- i 65536 CodePoints del *Piano di Base* entro cui ricade la quasi totalità delle assegnazioni fatte finora, sono rappresentati (esclusi i casi precedenti) mediante tre byte
- i restanti CodePoints sono rappresentati mediante quattro byte

Tabella 5.1: codici ASCII

<i>dec</i>	<i>hex</i>	<i>char</i>	<i>dec</i>	<i>hex</i>	<i>char</i>	<i>dec</i>	<i>hex</i>	<i>char</i>	<i>dec</i>	<i>hex</i>	<i>char</i>
0	00	NUL	32	20		64	40	@	96	60	'
1	01	SOH	33	21	!	65	41	A	97	61	a
2	02	STX	34	22	"	66	42	B	98	62	b
3	03	ETX	35	23	#	67	43	C	99	63	c
4	04	EOT	36	24	\$	68	44	D	100	64	d
5	05	ENQ	37	25	%	69	45	E	101	65	e
6	06	ACK	38	26	&	70	46	F	102	66	f
7	07	BEL	39	27	'	71	47	G	103	67	g
8	08	BS	40	28	(72	48	H	104	68	h
9	09	HT	41	29)	73	49	I	105	69	i
10	0A	LF	42	2A	*	74	4A	J	106	6A	j
11	0B	VT	43	2B	+	75	4B	K	107	6B	k
12	0C	FF	44	2C	,	76	4C	L	108	6C	l
13	0D	CR	45	2D	-	77	4D	M	109	6D	m
14	0E	SO	46	2E	.	78	4E	N	110	6E	n
15	0F	SI	47	2F	/	79	4F	O	111	6F	o
16	10	DLE	48	30	0	80	50	P	112	70	p
17	11	DC1	49	31	1	81	51	Q	113	71	q
18	12	CD2	50	32	2	82	52	R	114	72	r
19	13	CD3	51	33	3	83	53	S	115	73	s
20	14	DC4	52	34	4	84	54	T	116	74	t
21	15	NAK	53	35	5	85	55	U	117	75	u
22	16	SYN	54	36	6	86	56	V	118	76	v
23	17	ETB	55	37	7	87	57	W	119	77	w
24	18	CAN	56	38	8	88	58	X	120	78	x
25	19	EM	57	39	9	89	59	Y	121	79	y
26	1A	SUB	58	3A	:	90	5A	Z	122	7A	z
27	1B	ESC	59	3B	;	91	5B	[123	7B	{
28	1C	FS	60	3C	<	92	5C	\	124	7C	
29	1D	GS	61	3D	=	93	5D]	125	7D	}
30	1E	RS	62	3E	>	94	5E	^	126	7E	~
31	1F	US	63	3F	?	95	5F	_	127	7F	DEL

Capitolo 6

Reti di trasmissione a circuito

In questo capitolo si espongono i principi di funzionamento della rete pubblica commutata, dalle origini della telefonia analogica, alle tecniche di segnalazione, di multiplexione plesiocrona e sincrona della telefonia numerica, fino alla commutazione, ed ai sistemi di trasporto.

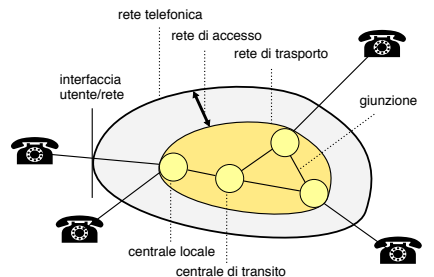
6.1 Introduzione

I casi di trasmissione finora analizzati fanno riferimento a collegamenti *punto-punto*, in cui una unica sorgente di informazione intende comunicare con un unico destinatario. Nella realtà, è assai più frequente il caso in cui i soggetti coinvolti nella comunicazione affidino la stessa ad una *rete* di collegamento, consegnando il messaggio al *nodo* di commutazione a cui hanno accesso. Il messaggio quindi, una volta determinato un percorso di *attraversamento* che coinvolga i nodi della rete più opportuni, giunge al destinatario, grazie anche alla presenza di informazioni addizionali, dette *di segnalazione*.

Le nozioni che seguono fanno esplicito riferimento alle *reti di telefonia*, dette anche a *commutazione di circuito*, allo scopo di fornire una panoramica dei principali aspetti delle stesse. Altri aspetti legati alle *reti di trasmissione dati* saranno illustrati a seguito della trattazione relativa alla teoria del traffico, assieme alle reti a *commutazione di pacchetto*.

6.1.1 Elementi della rete telefonica

Con riferimento alla figura, si anticipa che discuteremo prima dei metodi di *multiplexazione* che permettono alle diverse comunicazioni che terminano presso le *interfacce utente/rete* di essere aggregate da parte delle *centrali locali* per utilizzare un medesimo *collegamento di giunzione* interno alla *rete di trasporto*. Quindi, nel § 6.8 si esaminano i metodi di *commutazione* ed *instradamento* con cui viene individuato il percorso che una comunicazione deve intraprendere tra l'ingresso e l'uscita dalla rete di trasporto.



6.1.2 La rete di accesso

E' la parte più rilevante e *pesante* della rete, e consiste nel *doppino* in rame (pag. 384) che raggiunge la presa telefonica casalinga. All'interfaccia utente/rete sono così resi disponibili i *servizi* noti nel loro insieme come

- POTS (*Plain Old Telephone Service*) - vedi § 6.9.1;
- ISDN (*Integrated Service and Data Network*) - vedi § 6.9.2;
- ADSL (*Asymmetric Digital Subscriber Line*) - che in realtà usa POTS solo come tramite per raggiungere una *rete* IP vedi § 6.9.4.

Oltre a questi, sono ormai contemplati nella rete di accesso anche forme di collegamento diversi dal cavo, come

- accesso ottico - come nel caso FTTH (*Fiber To The Home*)¹, e che permette
 - di interconnettere un insieme più numeroso di collegamenti POTS già multiplati assieme, come nel caso di un grosso centralino aziendale, sovrapprendendosi allo scopo di un accesso ISDN-PRI;
 - di interconnettersi ad una rete IP ad una velocità maggiore di quella consentita dalla tecnologia ADSL;
- accesso radio
 - GSM² - noto anche come sistema cellulare di seconda generazione³, usa una rete diversa da PSTN, ma vi si interconnette in modo naturale. Il GSM nasce come standard aperto, favorendone la diffusione mondiale e l'interoperabilità tra gestori (*roaming*), e si sviluppa in forma completamente numerica, sia per la codifica vocale, che per il meccanismo di accesso multiplo al mezzo trasmissivo⁴, che adotta una organizzazione in trame; inoltre, ha introdotto la comunicazione dei messaggi SMS⁵;
 - GPRS⁶ e UMTS⁷ - mentre il primo (detto di *generazione 2.5*) usa la rete GSM per trasmettere dati a pacchetto, con velocità dell'ordine di 30-70 kbps, il secondo (detto anche di *terza generazione* o 3G) supporta in modo integrato sia le comunicazioni vocali, che i dati a pacchetto, con velocità dell'ordine dei 300 kbps, che salgono (teoricamente) a 3 e 14 Mbps con le estensioni UMTS 2+ e HDSPA rispettivamente;
 - WiFi⁸ e WiMax⁹ - mentre il primo distribuisce l'accesso ADSL su di un'area di estensione casalinga, il secondo ha una copertura di qualche chilometro, e permette collegamenti in mobilità. Entrambe permettono l'interconnessione ad un *Internet Service Provider* o ISP.

Altri tipi di offerte invece *non possono* essere considerate di accesso alla rete, pur se realizzate sfruttando sia la rete di accesso che quella di trasporto, come nel caso di

¹http://en.wikipedia.org/wiki/Fiber_to_the_x

²http://it.wikipedia.org/wiki/Global_System_for_Mobile_Communications

³La prima generazione si riferisce al sistema analogico TACS <http://it.wikipedia.org/wiki/TACS>

⁴http://en.wikipedia.org/wiki/Time_division_multiple_access

⁵<http://it.wikipedia.org/wiki/SMS>

⁶http://it.wikipedia.org/wiki/General_Packet_Radio_Service

⁷http://it.wikipedia.org/wiki/Universal_Mobile_Telecommunications_System

⁸<http://it.wikipedia.org/wiki/Wi-Fi>

⁹<http://it.wikipedia.org/wiki/WiMAX>

- CDN (*Circuito Diretto Numerico*) - offre la connettività diretta e continuativa con un'altra (ben specifica) interfaccia utente/rete, e pertanto viene a mancare la componente di commutazione;
- VPN (*Virtual Private Network*) - come sopra, con la differenza che in questo caso la connettività è basata su di una comunicazione a pacchetto anziché a circuito.

6.2 Moltiplicazione

In generale, raggruppare assieme più comunicazioni dirette alla medesima destinazione, in modo che condividano uno stesso mezzo trasmissivo, permette di

- tentare di occupare tutta la banda messa a disposizione dal mezzo trasmissivo
- massimizzare la percentuale di utilizzo del mezzo trasmissivo, nel caso di sorgenti non continuamente attive (vedi § 8.3.4)
- semplificare la gestione e la manutenzione dei collegamenti a lunga distanza, essendo questi minori in numero

Le tecniche di moltiplicazione, nella evoluzione storica delle telecomunicazioni, possono operare secondo le diverse modalità di

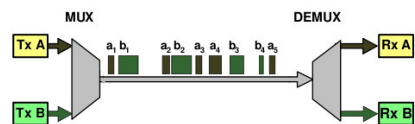
- *divisione di frequenza* - ogni comunicazione usa una banda di frequenze diversa, come descritto al § 10.1.1, nel contesto dello studio dei segnali modulati;
- *divisione di tempo* - le comunicazioni avvengono in intervalli di tempo disgiunti, mediante segnali numerici, e sono affrontate nel resto di questo capitolo;
- *divisione di codice* - tutte le comunicazioni usano la stessa banda allo stesso tempo, ma ogni diverso destinatario è ancora in grado di distinguere il proprio messaggio, mediante una operazione di correlazione tra segnali ortogonali (vedi § 9.5.2). Dato che il segnale trasmesso in questo caso si trova ad occupare una banda più estesa di quella originaria, la tecnica è detta *ad espansione di spettro* ed esposta al § 13.8.

6.2.1 Moltiplicazione a divisione di tempo

Per quanto riguarda la moltiplicazione a divisione di tempo, nella pratica questa è attuata solo a partire da segnali numerici, come sono sempre stati i segnali dati, ed in quel contesto si è sviluppato un approccio basato sull'uso di un *pacchetto dati* (vedi § 8.5.1), attuando uno schema detto

Moltiplicazione statistica e commutazione di pacchetto In questo caso il mezzo trasmissivo non è impegnato in modo esclusivo, ma la trasmissione può avvenire in modo sporadico, ed i dati inviati ad intervalli irregolari. Questo motivo, assieme alla dimensione variabile delle singole comunicazioni, porta a suddividere la comunicazione in unità autonome indicate come pacchetto dati.

La moltiplicazione dei pacchetti avviene quindi in modo *statistico*, senza riservare con esattezza risorse a questo o quel tributario: il moltiplicatore si limita ad inserire i pacchetti ricevuti in apposite code, da cui li preleva per poterli trasmettere in sequenza, attuando una modalità di trasferimento *orientata al ritardo* (vedi § 8.4). La presenza di code comporta



- il determinarsi di un ritardo variabile ed imprevedibile
- la possibilità che la coda sia piena, ed il pacchetto in ingresso venga scartato

D'altra parte, ogni pacchetto reca con sé le informazioni necessarie al suo recapito, facilitando il compito dell'instradamento (vedi § 6.7). A seconda dell'adozione di un principio di commutazione di tipo *a circuito virtuale* oppure a *datagramma* (vedi § 8.5.2.2), può essere presente o meno una *fase di setup* precedente l'inizio della comunicazione.

Multiplicazione deterministica e commutazione di circuito La modalità usata nella rete telefonica è invece basata su di uno schema di moltiplicazione con organizzazione di trama (vedi § 8.5.2.1) che determina un paradigma noto come *commutazione di circuito*, per il motivo che ora illustriamo.

Alle origini storiche della telefonia, nell'epoca dei telefoni *a manovella*, con la cornetta appesa al muro, la commutazione si basava sull'operato di un centralinista umano, che creava un vero e proprio *circuito elettrico* che collegando fisicamente tra loro le terminazioni dei diversi utenti. Nel caso in cui intervengano più centralinisti in cascata, la chiamata risulta instradata attraverso più centralini. Da allora, il termine commutazione di circuito individua il caso in cui



- è necessaria una fase di *setup* precedente alla comunicazione vera e propria, in cui vengono riservate le risorse;
- nella fase di setup si determina anche l'*instradamento* della chiamata nell'ambito della rete, che rimane lo stesso per tutta la durata della medesima;
- le risorse trasmissive restano impegnate in *modo esclusivo* per l'intera durata della conversazione.

Le cose non sono cambiate di molto (da un punto di vista concettuale) con l'avvento della telefonia numerica: in tal caso, più segnali vocali sono campionati e quantizzati in modo sincrono, ed il risultato (numerico) è multiplato in una *trama PCM* (§ 6.3.1), in cui viene riservato un intervallo temporale per ognuno dei flussi tributari.

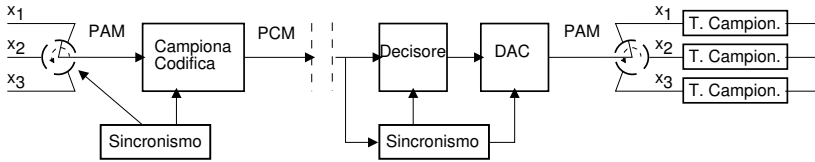
Ad ogni buon conto, si noti che un risultato della teoria del traffico (pag. 184) mostra come l'adozione di una strategia *orientata al ritardo* migliora notevolmente l'efficienza di utilizzo del mezzo stesso.

6.3 Rete plesiocrona

Questo termine si riferisce alla modalità di funzionamento quasi-sincrona adottata dalle centrali telefoniche, almeno finché la rete di trasporto non è divenuta capace di realizzare una modalità di moltiplicazione sincrona. In entrambi i casi, i segnali vocali sono trasportati in forma numerica, multiplandone i campioni a divisione di tempo in modo deterministico, in accordo ad una organizzazione di trama realizzata presso la centrale di accesso, come descritto di seguito.

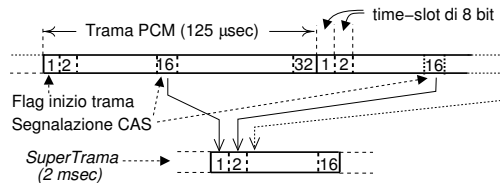
6.3.1 Trama PCM

Nella figura seguente sono rappresentati tre segnali *tributari*, campionati a turno alla stessa frequenza di 8 KHz, quantizzati ad 8 bit per campione con quantizzazione logaritmica (vedi § 4.2.2), e trasmessi (8 bit alla volta¹⁰) a turno su di un unico collegamento, producendo un segnale binario che prende il nome di PCM (*Pulse Code Modulation*¹¹). In figura è evidenziato inoltre un blocco di sincronismo necessario a ricostruire la corretta sequenza ricevuta, in modo da redistribuire correttamente i campioni ai filtri di restituzione.



La struttura temporale ripetitiva che ospita i campioni dei singoli tributari prende il nome di *trama* (FRAME¹²), ed è composta 32 intervalli detti *time-slot*. Trenta di questi ospitano a turno i bit di un campione proveniente da un numero massimo di 30 tributari¹³, mentre i rimanenti due intervalli convogliano le informazioni di segnalazione¹⁴, che indicano lo stato dei singoli collegamenti (il 16° intervallo) e forniscono il sincronismo relativo all'inizio della trama stessa (il primo). La velocità binaria complessiva risulta quindi di 32 intervalli * 8 bit/intervallo * 8000 campioni/secondo = 2048000 bit per secondo; per questo motivo, all'insieme ci si riferisce come alla *trama PCM a 2 Mbit*. La durata della trama deve essere invece la stessa del periodo di campionamento, ossia $1/8000 = 125 \mu\text{sec}$.

Il primo time-slot della trama contiene una configurazione di bit sempre uguale, chiamata FLAG (*bandiera*), che ha lo scopo di indicare ai circuiti di sincronismo l'inizio della trama stessa. I dati di segnalazione contenuti nel 16° intervallo devono essere *dihuiti* su più trame, per poter rappresentare tutti i 30 tribu-



¹⁰La tecnica di moltiplicare un blocco di bit (in questo caso 8) alla volta prende il nome di *word interleaving*, distinto dal *bit interleaving*, in cui l'alternanza è a livello di bit.

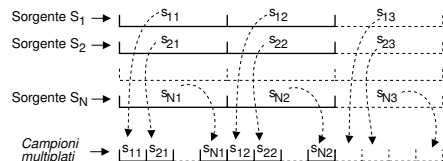
¹¹Il segnale PCM ispira il suo nome dal PAM (vedi § 6.9.5) in quanto ora, anziché trasmettere le ampiezze degli impulsi, si inviano i codici binari dei livelli di quantizzazione.

¹²frame SIGNIFICA PIÙ PROPRIAMENTE TELAIO, E IN QUESTO CASO HA IL SENSO DI INDIVIDUARE UNA STRUTTURA, DA "RIEMPIRE" CON IL MESSAGGIO INFORMATIVO.

¹³In figura è mostrato un esempio, in cui i campioni s_{ij} di N sorgenti S_i si alternano a formare una trama.

Durante l'intervallo temporale tra due campioni, devono essere collocati nella trama tutti gli M bit/campione delle N sorgenti, e quindi la frequenza binaria (in bit/secondo) complessiva sarà pari a $f_b = f_c$ (campioni/secondo/sorgente) $\cdot N$ (sorgenti) $\cdot M$ (bit/campione).

¹⁴Vedi anche la sezione 6.3.2.



tari¹⁵. Si è stabilito che occorra prelevare il 16° intervallo di 16 trame successive, per ricostruire una struttura detta *supertrama* (di $16 \cdot 8 = 128$ bit) che rappresenta le informazioni di tutti i tributari (disponendo così di 4 bit/tributario/supertrama), e che si ripete ogni $16 \cdot 125 = 2000 \mu\text{sec} = 2 \text{ msec}$.

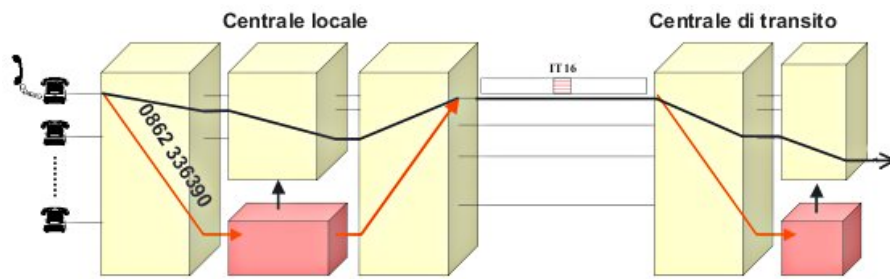
In effetti nel 16° time-slot della trama sono presenti a turno, oltre ai bit di segnalazione relativi allo stato dei tributari, anche bit necessari alla sincronizzazione della supertrama (ossia un *flag*), mentre le informazioni di segnalazione sono ripetute più volte nella stessa supertrama, per proteggersi da eventuali errori di ricezione, che danneggiando l'informazione sullo stato dei canali, potrebbero causare la "caduta della linea".

6.3.2 Messaggi di segnalazione

Come illustrato al § 6.9.1, la rete di accesso è sede di uno scambio di informazioni tra terminale e centrale locale, detta *segnalazione di utente*, e che ha lo scopo di indicare la disponibilità della rete, il numero chiamato, l'attivazione della suoneria, ed i messaggi a ritroso di libero/occupato. Queste informazioni, quando devono essere propagate verso il lato-rete della centrale di accesso, possono essere gestite secondo due diversi approcci.

Segnalazione associata al canale In questo caso la centrale di accesso inserisce le informazioni di segnalazione relative ad un tributario all'interno della *supertrama* di segnalazione ottenuta collezionando i valori presenti nel 16° time-slot. Questa modalità viene indicata come CAS (*Channel Associated Signaling*), ed ha origine dalla conversione dei precedenti collegamenti analogici, in cui la segnalazione relativa ad ogni terminale viaggiava in modo indissolubilmente associato al segnale vocale, condividendo con questo il mezzo trasmissivo a commutazione di circuito. Con la numerizzazione, si è inizialmente scelto di mantenere la segnalazione *associata* al segnale vocale, con la contropartita che quando, nell'attraversare una centrale di transito, una comunicazione è commutata su di una diversa linea di uscita, deve essere commutata anche la segnalazione associata.

La figura seguente mostra come la numerazione venga recepita da un organo di *controllo centrale*, che provvede a impostare il dispositivo di commutazione (§ 6.8), in modo che la comunicazione sia instradata verso la linea di uscita in direzione della destinazione. Quindi, l'informazione di segnalazione viene ri-associata nell'intervallo 16.



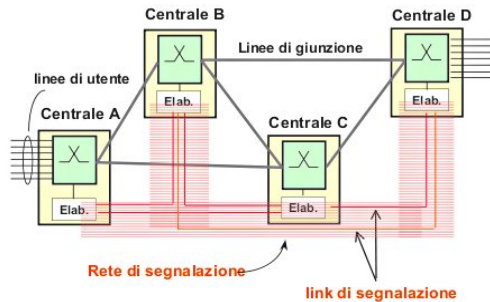
¹⁵Gli 8 bit del 16° intervallo sono infatti insufficienti a codificare lo stato dei 30 tributari che contribuiscono al segnale TDM.

Segnalazione a canale comune Il primo passo evolutivo è stato quindi quello di provvedere ad un *canale comune* di segnalazione direttamente collegato agli organi di controllo, su cui poter convogliare la segnalazione relativa a tutte le comunicazioni in transito tra le due centrali.



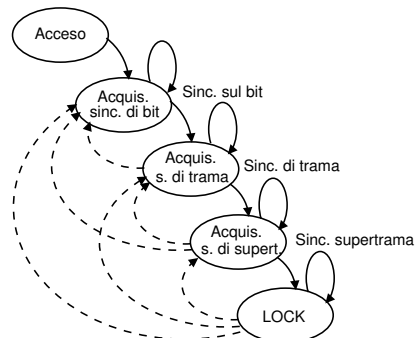
I messaggi di segnalazione, per loro natura, devono essere trasmessi solo quando si verificano degli eventi significativi, e per questo motivo sono ora inviati mediante dei *pacchetti dati*. Il passo successivo è quindi stato quello di realizzare una intera rete a *commutazione di pacchetto*, parallela a quella di transito su cui viaggiano (in modalità a circuito) le conversazioni vocali.

In tal modo, gli organi di controllo delle centrali sono in comunicazione diretta tra loro, secondo la modalità cosiddetta CCS (*Common Channel Signaling*), mediante una rete a pacchetto dedicata alla segnalazione, sulla quale viaggiano i messaggi definiti da un apposito *sistema di segnalazione* (vedi § 6.9.3), e che permette di centralizzare il controllo e la configurazione di tutte le centrali coinvolte nell'instradamento di una stessa comunicazione, rendendo così possibile la disponibilità di servizi come il trasferimento di chiamata, la conversazione a tre, l'avviso di chiamata...



6.3.3 Sincronizzazione di centrale

Nella figura a lato sono mostrati i diversi stati attraverso cui deve evolvere il dispositivo di sincronizzazione che opera sui flussi PCM CAS, prima di entrare nello stato di LOCK (*aggancio*) ed iniziare a poter leggere e smistare i contenuti dei diversi time-slot. Occorre infatti acquisire innanzitutto il sincronismo sul bit, sfruttando le caratteristiche del codice di linea utilizzato; quindi si sfrutta la conoscenza della configurazione scelta per il flag di inizio trama, per individuare da dove iniziare a conteggiare gli



intervalli temporali. Infine, viene individuato l'inizio della supertrama, grazie ad un'ulteriore configurazione prefissata, posta all'inizio della stessa. In ogni stato poi, esiste la possibilità (fortunatamente remota) di perdere il sincronismo ed *indietreggiare* (linee tratteggiate) nel diagramma di stato, perdendo le comunicazioni in corso.

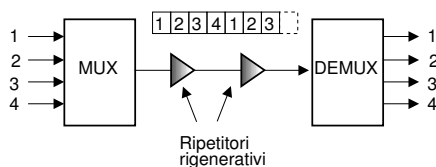
6.3.4 Multiplazione asincrona e PDH

L'argomento di questo paragrafo non va confuso con la *trasmissione* asincrona (quella START-STOP mostrata al § 5.6.1), e che descrive una modalità di *inviare* informazioni numeriche; qui invece si tratta di *multiplare*, ossia come *mettere assieme* più comunicazioni.

Via via che la rete di trasporto è interconnessa mediante centrali di livello gerarchico superiore, associate ad aree di influenza geografica più estesa (vedi § 6.5), i collegamenti di giunzione trasportano un numero di tributari sempre più elevato, ottenuti raggruppando assieme tutte le conversazioni contemporaneamente dirette verso la stessa destinazione. Considerando allo stesso tempo le problematiche legate al dover svolgere nelle centrali la funzione di commutazione, ci si pone il problema di individuare dei metodi efficienti di raggruppare assieme più tributari, anche a velocità diverse, facendo in modo che l'operazione di inserimento/rimozione di un singolo tributario sia relativamente agevole. Rimandiamo al § 6.4 l'analisi di come avvenga il processo di multiplazione nel caso in cui esista una perfetta sincronizzazione tra gli elementi della rete, e trattiamo nel seguito il caso della rete *plesiocrona*.

Nella trama PCM (§ 6.3.1), tutti i 30 canali sono campionati congiuntamente, e più flussi a 2 Mbit possono a loro volta essere "messi assieme" in modalità *bit interleaved* (prendendo un bit alla volta da ogni tributario) da appositi dispositivi *multiplatori* (o MULTIPLEXER, o MUX). Il collegamento può prevedere più dispositivi detti *ripetitori rigenerativi*, che oltre ad amplificare il segnale, lo "puliscono" dal rumore accumulato, decodificando i dati in ingresso per poi generare ex-novo il segnale numerico.

Il problema con questo modo di procedere è che i singoli tributari possono ragionevolmente avere origine da centrali differenti, ognuno con un proprio orologio indipendente, e quindi le velocità possono essere lievemente differenti l'una dall'altra¹⁶, pur essendo molto simili. In questo caso si dice che la rete opera in modo *plesiocrono*, ossia *quasi* isocrono (ma non del tutto).



N. Canali	sigla	Vel. (kbps)	Teorica
30	E1	2.048	
120	E2	8.448	8.192
480	E3	34.368	32.768
1920	E4	139.264	131.072
7680	E5	565.148	524.288

In tabella riportiamo la gerarchia CCITT¹⁷, nota come *Plesiochronous Digital Hierarchy* (PDH), secondo la quale ad esempio 4 flussi da 2 Mb/s sono multiplati in uno da 8 Mb/sec: notiamo che sebbene siano teoricamente sufficienti 8192 Mb/sec, in realtà il Multiplexer ne produce di più (8448). Questo avviene proprio per permettere

¹⁶Un oscillatore con precisione di una parte su milione, produce un ciclo in più o in meno ogni 10^6 ; ad una velocità di 2 Mb/s, ciò equivale a un paio di bit in più od in meno ogni secondo.

¹⁷Comité Consultatif International pour la Telephonie et Telegraphie. Questo organismo non esiste più, ed ora l'ente di standardizzazione ha nome ITU-T.

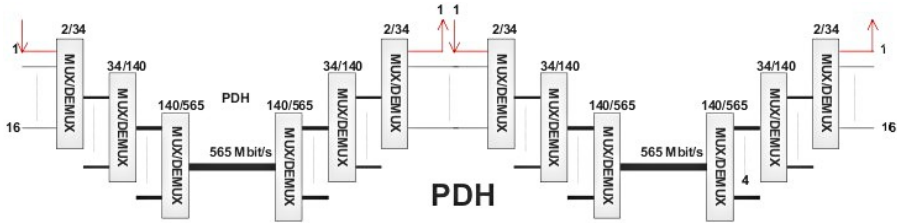
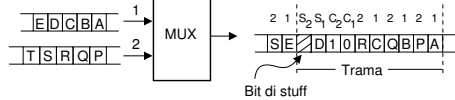


Figura 6.1: Gerarchia di multiplexazione PDH e complessità di un ADM

la trasmissione di segnali non necessariamente sincroni, mediante la tecnica del *Bit Stuffing*¹⁸.

6.3.4.1 Bit stuffing

Consideriamo 2 tributari i cui bit vengono inseriti alternativamente in una trama da 4 bit/canale; il secondo risulta lievemente più lento. I primi 3 + 3 bit vengono trasmessi comunque, mentre il 4° può essere trasmesso o meno, a seconda se i tributari lo abbiano pronto. Per ottenere questo risultato, i bit C_1 e C_2 (di controllo) valgono 0 oppure 1 a seconda se l'intervallo seguente (S_1 e S_2) contenga un dato valido oppure sia solo un *bit di stuffing*, cioè vuoto, in quanto il tributario corrispondente è più lento rispetto alla velocità nominale. Ecco perché le velocità delle gerarchie superiori sono *abbondanti*: per ospitare i bit di controllo, necessari a gestire tributari non sincronizzati.



Il metodo illustrato permette in ricezione di effettuare la *destuffing*, e riottenere i flussi originari. Nella realtà le informazioni di controllo sono molto ridondate, perché se scambiassimo un bit di stuffing per uno buono (o viceversa), distruggeremmo anche la struttura di trama del tributario che ha subito l'errore.

6.3.4.2 Add and Drop Multiplexer - ADM

La modalità *bit interleaved* con cui è realizzata la gerarchia PDH è particolarmente problematica qualora di desideri estrarre e/o introdurre un singolo tributario da/in un segnale multiplexato di ordine elevato, ovvero realizzare una funzione detta *Add and Drop*. In questo caso è infatti necessario eseguire un'operazione inversa a quella di multiplexazione, ovvero (vedi fig. 6.1) demultiplexare l'intero flusso, compresi tutti gli altri tributari, e successivamente ri-multiplexare di nuovo il tutto.

Questa caratteristica limita notevolmente la flessibilità nelle configurazioni di rete che si possono ottenere con questa tecnologia, e per i tributari passanti comporta l'aggiunta di un tempo di ritardo aggiuntivo dovuto alle operazioni di demultiplexazione e multiplexazione. Nella pratica vengono usati solo flussi di tipo E1, E3 ed E4, che sono quelli più adatti per essere trasportati nella gerarchia sincrona SDH, multiplexando direttamente sedici tributari a 2 Mbit/s all'interno di un unico flusso a 34 Mbit/s.

¹⁸Da: TO STUFF = riempire.

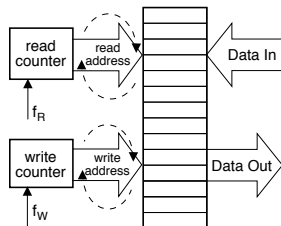
6.3.5 Sincronizzazione di rete

Se tutti i nodi della rete operassero alla stessa velocità, non sussisterebbero problemi nella moltiplicazione di più tributari. Nel caso in cui la sincronizzazione tra nodi sia completamente affidata ad un orologio di centrale di elevata precisione, si verifica il caso di funzionamento *plesiocrono*, che è quello prescritto per le centrali che interconnettono le reti di due diverse nazioni, o di due diversi operatori di telecomunicazioni. Ma questa non è l'unica soluzione.

Una alternativa è la sincronizzazione *mutua* tra centrali, in cui ognuna di queste emette dati in uscita ad una frequenza pari alla media delle frequenze dei dati in ingresso. A parte fenomeni transitori durante i quali la rete è soggetta ad oscillazioni di velocità, relativi all'inserimento od alla disattivazione di centrali "topologicamente importanti", il metodo funziona ragionevolmente bene. Una seconda soluzione è una sincronizzazione di rete di tipo *gerarchico* in cui le centrali ricevono informazioni di sincronismo da soggetti "più importanti", come per configurazioni *Master-Slave* in cui il Master è una centrale ad elevata precisione, od un riferimento in comune come ad esempio un satellite in orbita terrestre.

6.3.5.1 Elastic store

Si tratta di un accorgimento¹⁹ idoneo ad *assorbire* le fluttuazioni della velocità di trasmissione, come ad esempio nel caso della sincronizzazione mutua. Mentre il *bit stuffing* (§ 6.3.4.1) è adottato nella moltiplicazione di più tributari in un livello gerarchico più elevato, l'*elastic store* è usato per compensare le diverse velocità tra tributari di eguale livello gerarchico in ingresso ad un elemento di commutazione (§ 6.8).



È realizzato mediante un banco di memoria (di dimensione pari ad una trama), riempito (ciclicamente) con le parole (word) del flusso binario in ingresso, alla velocità f_W di quest'ultimo, alla posizione individuata dal contatore WRITE che si incrementa²⁰ appunto a velocità f_W , e che torna a puntare all'inizio della memoria una volta raggiunto l'indirizzo più elevato. Un secondo puntatore READ viene utilizzato per leggere la memoria, alla velocità f_R richiesta, e prelevare i dati da inviare in uscita: se f_R e f_W sono differenti, READ e WRITE prima o poi si sovrappongono, causando la perdita o la ripetizione di una intera trama, e nulla più²¹.

6.4 Gerarchia digitale sincrona

Definizione dei livelli della gerarchia Come anticipato, la *Synchronous Digital Hierarchy* (SDH²²) è una metodologia di moltiplicazione che presuppone un funzionamento perfettamente sincrono degli elementi di rete, ed ha solo una variante (nel Nord

¹⁹Letteralmente: *magazzino elastico*.

²⁰Il contatore WRITE, come anche READ, conta in binario, e si incrementa con frequenza f_W (f_R). Le parole binarie rappresentate da READ e WRITE forniscono l'indirizzo (all'interno del banco di memoria) in cui leggere i dati in uscita e scrivere quelli in ingresso rispettivamente.

²¹Infatti il sincronismo di trama viene preservato; inoltre l'evento di sovrapposizione dei puntatori può essere rilevato, e segnalato ai dispositivi di demoltiplicazione, in modo che tengano conto dell'errore che si è verificato.

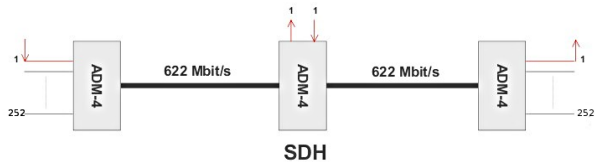
²²http://it.wikipedia.org/wiki/Synchronous_Digital_Hierarchy

SONET	SDH	payload (kbps)	v. trasm. (kbps)
STS-1	-	48.960	51.840
STS-3	STM-1	150.336	155.520
STS-12	STM-4	601.344	622.080
STS-24	STM-8	1.202.688	1.244.160
STS-48	STM-16	2.405.376	2.488.320
STS-96	STM-32	4.810.752	4.976.640
STS-192	STM-64	9.621.504	9.953.280
STS-768	STM-256	38.486.016	39.813.120
STS-1536	STM-512	76.972.032	79.626.120

Tabella 6.1: Nomenclatura della gerarchia ottica e relative velocità

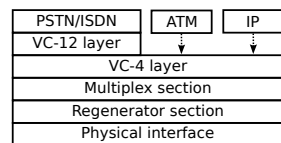
America), denominata SONET (*Synchronous Optical Network*), i cui livelli sono siglati STS oppure OC nel caso in cui ci si riferisca al segnale ottico corrispondente, e che interopera abbastanza bene con SDH. La tabella 6.1 elenca le velocità del *payload*²³ e di trasmissione associate ai diversi livelli della gerarchia di multiplexazione SDH/SONET.

Multiplexer Add and Drop La differenza strutturale rispetto a PDH, è che in SDH i tributari usano tutti lo stesso clock, da cui deriva la possibilità di aggiungere e togliere un singolo tributario senza alterare il flusso in cui è immerso, come esemplificato in figura, in cui 252 flussi PDH E1 concorrono a formare un multiplex STM-4.



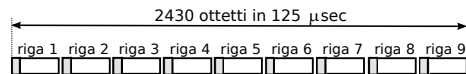
Eterogeneità del trasporto

SDH nasce allo scopo di consentire il trasporto di dati di diversa origine (PCM telefonico, ISDN, pacchetti Ethernet ed IP, celle ATM), come illustrato nella figura a fianco, che rappresenta impilate le diverse elaborazioni che i tributari devono subire per essere immessi nel flusso SDH.



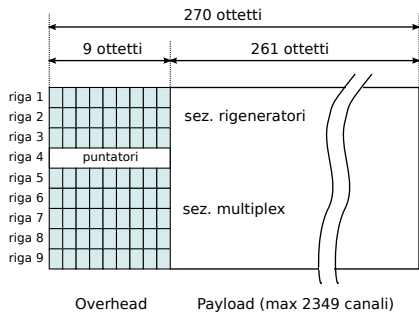
Struttura di trama

SDH si basa su di una struttura di trama di durata di 125 μ sec, durante i quali sono trasmessi in modalità *byte interleaved* una sequenza di ottetti provenienti da diversi tributari a 64 kbps che condividono la medesima sorgente di temporizzazione, cosicché ogni tributario può essere inserito o prelevato semplicemente scrivendo o leggendo sempre nello stesso punto (con la stessa fase) un ottetto ogni trama.



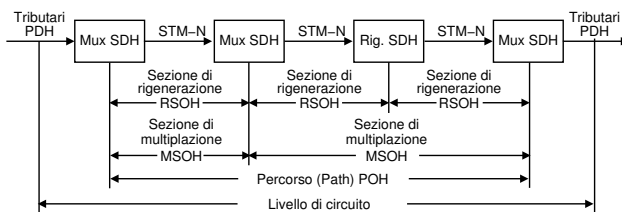
²³ Con il termine *payload* si indica il *carico pagante*, ossia i dati che vengono trasportati

Synchronous Transport Module STM-1 Il livello più basso della gerarchia è indicato come STM-1 (*Synchronous Transport Module*), che opera ad una velocità di 155.52 Mbit/s, può trasportare 63 flussi PDH E1 (ovvero 63 flussi * 32 time-slot/flusso = 2016 canali PCM), ed è caratterizzato da una trama composta da 2430 ottetti, di cui 81 di segnalazione e 2349 di dati²⁴, ovvero usando un ottetto di segnalazione ogni 30 totali, quasi come avviene per il flusso PDH E1 (in cui c'è un intervallo di segnalazione, il 16°, ogni 31 canali voce). Gli ottetti di segnalazione sono però ora raggruppati a gruppi di nove, seguiti da $29 \cdot 9 = 261$ ottetti di dati, ed il risultato è tradizionalmente rappresentato incolonnando le 9 sotto-sequenze di 270 ottetti come in figura, rappresentando così una trama come una matrice di 9 righe per 270 colonne.

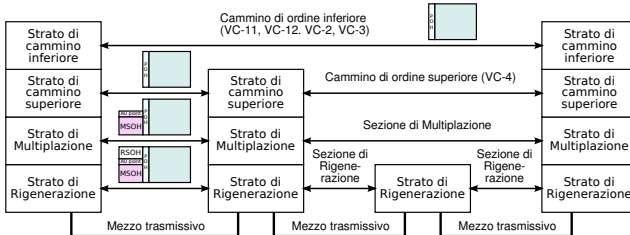


Le componenti dell'Overhead Le prime 9 colonne prendono il nome di *overhead* della trama, mentre la parte dati è indicata come *Payload* (o carico pagante). L'overhead contiene informazioni di segnalazione strettamente inerenti al processo di multiploazione, ossia finalizzate all'espletamento di funzioni OAM (*Operation, Administration, Maintenance*), che sono ora associate ad un annidamento di sezioni di trasmissione: *Path, Multiploazione e Rigenerazione*. Infatti, il percorso (*Path*) compiuto da un singolo tributario, si snoda tra un unico moltiploatore di ingresso ed un unico demoltiploatore di uscita, ma ad ogni moltiploatore *Add and Drop* (o commutatore) incontrato, viene definita una nuova *sezione di multiploazione*.

Allo stesso modo, per ogni ripetitore rigenerativo incontrato, è definita una nuova *sezione di rigenerazione*. Per ognuna di queste sezioni, è definito un *overhead* (OH) specifico per le operazioni OAM associate.



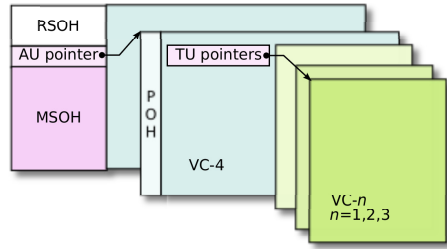
Dato che un ADM è anche rigeneratore, e che i dispositivi di ingresso - uscita del tributario sono anche ADM, si determina la *stratificazione funzionale* per la segnalazione raffigurata a fianco, in cui è evidenziato come l'overhead associato alle sezioni più esterne venga *impilato* su quello delle sezioni interne. Ma a differenza dell'incapsulamento (pag. 8.5.2.3) proprio



²⁴Notiamo che la differenza tra i 2349 ottetti di payload ed i 2016 canali voce fornisce 2349 - 2016 = 333 ottetti, che suddivisi per le nove righe, danno luogo a 37 ottetti per riga *in più*.

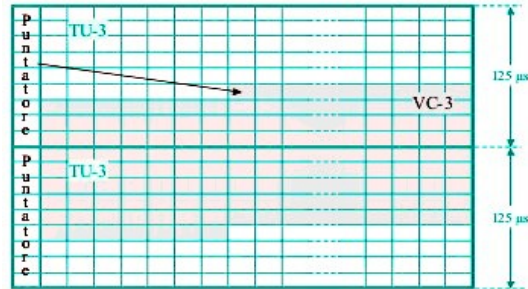
dei formati di trasmissione a pacchetto, in questo caso i tre tipi di overhead (*Path POH*, *Multiplex Section MSOH*, e *Regenerator Section RSOH*) sono inseriti nella trama STM-1 in punti diversi, come mostrato dalla figura che segue.

Il puntatore all'unità amministrativa Nelle prime tre righe dell'overhead della trama STM-1 trova posto l'RSOH, che viene scritto dai dispositivi di rigenerazione, e quindi letto e ri-scritto ad ogni rigeneratore successivo; in particolare, alla prima riga sono presenti i flag che consentono di acquisire il sincronismo di trama. Nelle ultime cinque righe dell'OH, troviamo il MSOH, scritto, letto e ri-creato dai dispositivi di moltiplicazione. Il POH trova posto all'interno del payload, e su questo torniamo tra brevissimo. Alla quarta riga dell'OH di trama, troviamo un puntatore (*AU Pointer*), che specifica la posizione di inizio del payload (chiamato ora AU, o *Administrative Unit*) nell'ambito della struttura di trama.



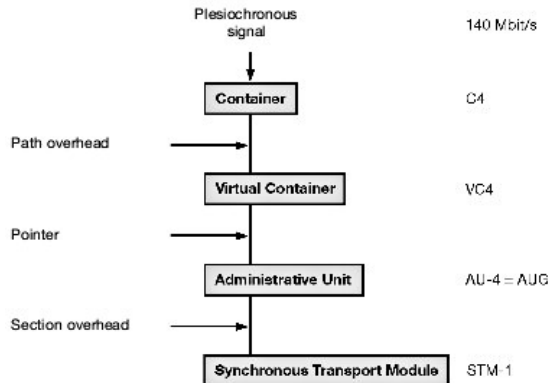
La presenza di questo puntatore deriva dalla volontà di ridurre al minimo l'uso di buffer e l'introduzione di ritardi di consegna; pertanto i dati da trasmettere *non* vengono inseriti nella struttura di trama all'inizio della stessa, bensì

al primo ottetto possibile al momento della disponibilità dei dati stessi. Quindi, è più che normale il caso in cui la AU inizi a metà di una trama, e termini a metà della trama successiva, come illustrato nella figura che a lato. La coppia AU ed AU Pointer prende quindi il nome di *Administrative Unit Group* (AUG).



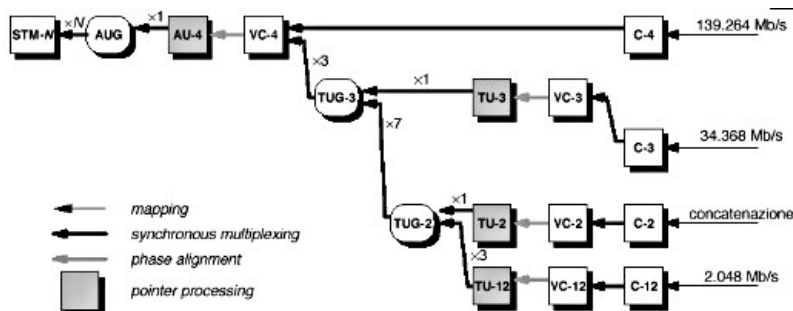
Virtual Container e Tributary Unit

Il riempimento della AU con i dati da trasmettere, avviene (vedi figura a lato) seguendo una serie di passi successivi, che vedono prima la creazione di una struttura dati detta *Container*, a cui si aggiunge il POH per ottenere un *Virtual Container*, da cui dopo l'aggiunta del puntatore deriva la AU. Notiamo ora che non necessariamente la AU deve essere riempita da un unico tributario; al



contrario, la multiplazione serve appunto ad ospitarne diversi!! A questo scopo, più VC a bassa velocità possono essere a loro volta multiplati in modalità *byte interleaved*, per produrre una struttura dati intermedia indicata TU (*Tributary Unit*), che a sua volta può essere inserita assieme ad altre TU, all'interno del VC di ordine superiore.

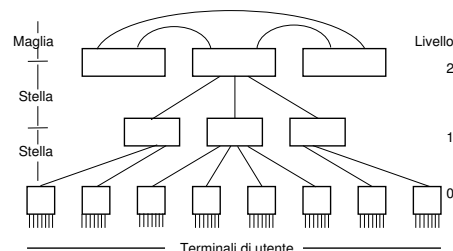
Non approfondiamo oltre questo argomento, che richiede una buona dose di pazienza per essere analizzato a fondo, e ci limitiamo ad inserire un diagramma che mostra le possibilità di combinazione di tributari differenti, in accordo alle specifiche di ETSI.



Esercizio Quanto canali voce entrano in un multiplex STM-1? *Risposta:* ci entra un AUG composto fino da 3 TUG-3, ognuno dei quali può contenere 7 TUG-2, che a loro volta contengono 3 flussi PDH a 2 Mbit/s, ognuno con 30 canali PCM. Pertanto: $30 \cdot 3 \cdot 7 \cdot 3 = 1890$ canali voce.

6.5 Topologia di rete

Nella figura seguente è riportata una possibile topologia di rete, a scopo puramente esemplificativo, dato che è un elemento su cui gli operatori godono di massima libertà.



Il nostro esempio si riferisce ad una rete a 2 livelli, in cui i terminali di utente che fanno capo ad una stessa centrale locale di livello zero accedono alla rete per mezzo di un collegamento che multipla le comunicazioni provenienti dalla stessa centrale.

Come vedremo al capitolo 8, il numero di collegamenti contemporaneamente possibili è inferiore al numero di terminali di utente, ed è dimensionato applicando i metodi della teoria del traffico al caso in questione. Le centrali locali sono connesse a stella a quelle di livello 1, che quindi realizzano instradamenti relativi alle comunicazioni tra utenti geograficamente vicini, ma connessi a centrali locali diverse. Se, al contrario, due utenti sono connessi alla stessa centrale di livello zero, l'instradamento non esce dalla centrale locale.

Le centrali di livello 1 sono a loro volta connesse a stella a quelle di livello 2, che gestiscono il traffico a livello nazionale; alcune di queste poi, consentono di instradare anche i collegamenti verso altre reti (ad es. di altre nazioni od operatori). Le centrali di 2° livello sono connesse tra loro a maglia completa (MESH in inglese) in modo da consentire instradamenti alternativi anche nel caso in cui un collegamento tra centrali vada fuori servizio.

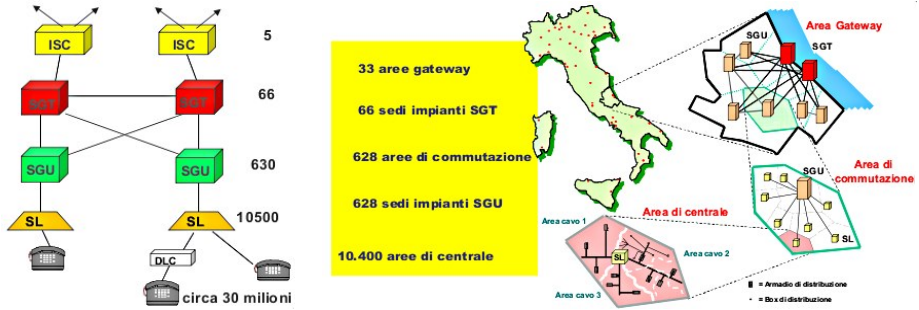


Figura 6.2: Topologia della rete telefonica Italiana

Nello specifico caso italiano, la denominazione attribuita alle centrali dei diversi livelli segue lo schema mostrato nella parte sinistra della figura 6.2: i terminali di utente sono attestati presso gli *Stadi di Linea* (SL) tramite la rete di accesso, mentre gli SL sono collegati agli *Stadi di Gruppo Urbano* (SGU) tramite la rete di trasporto; infine, gli SGU sono collegati agli *Stadi di Gruppo di Transito* (SGT) tramite rete di trasporto in fibra ottica. La parte destra della figura mostra inoltre come questi elementi siano dislocati geograficamente per la regione Abruzzo, individuando la ripartizione del territorio, e mostrando come ad un livello inferiore agli stadi di linea, la rete di accesso si dirami ulteriormente attraverso gli armadi ed i box di distribuzione.

6.6 Rete in fibra ottica

Nel periodo iniziale le fibre ottiche sono state usate prevalentemente nella rete di trasporto tra centrali di grado gerarchicamente elevato, mentre ora trovano impiego anche nella sezione di accesso. Per ciò che riguarda le modalità di trasmissione ottica, si rimanda al § 15.4; nel seguito illustriamo i dispositivi utilizzati, la topologia risultante, ed i sistemi di protezione.

6.6.1 Dispositivi SDH

Come anticipato, la trasmissione SDH si avvale di elementi (vedi Fig. 6.3) che possono essere descritti in termini funzionali secondo la seguente classificazione:

Rigeneratori Sono gli elementi di base, che consentono di suddividere su più tratte i collegamenti più lunghi, e che eliminano dal segnale in transito gli effetti del rumore e della dispersione temporale.

Multiplatori Combinano tributari PDH ed SDH, in modo da inserirli in flussi a velocità più elevate.

Multiplatori Add and Drop Permettono l'inserimento e l'estrazione di tributari a bassa velocità in/da un flusso in transito, e consentono la creazione di strutture ad anello.

Digital Cross Connect A differenza di un ADM, un DXC è interconnesso a più di un flusso SDH, e quindi può inserire un tributario (od un container) prelevato

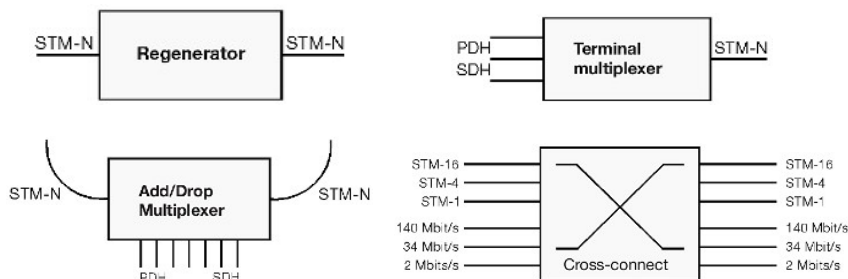
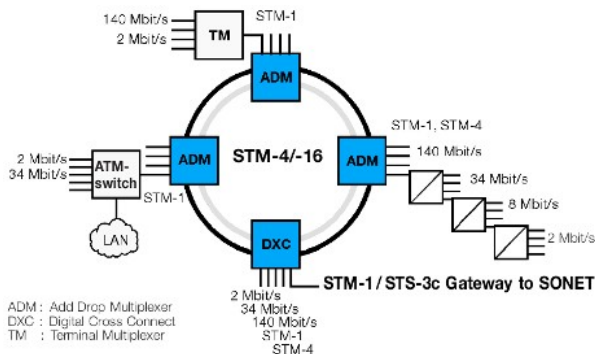


Figura 6.3: Dispositivi SDH

da un flusso entrante, all'interno di un diverso flusso uscente, realizzando così la funzione di commutazione.

6.6.2 Topologia ad anello

Le reti in fibra ottica sono quasi sempre realizzate mediante degli *anelli* che congiungono tra loro i nodi di commutazione in forma ciclica. I dispositivi DXC (*Digital Cross Connect*) sono infatti interconnessi a più di un anello, e svolgono la funzione di commutazione delle comunicazioni che devono essere inoltrate verso gli altri anelli.



6.6.2.1 Rete di trasporto

Al 2002, l'interconnessione dei collegamenti SDH nazionali risultava permessa dalla struttura su tre livelli riportata in Fig. 6.4.

6.6.2.2 Rete di accesso in fibra

La capacità del trasporto SDH di accettare tributari di tipo Ethernet o IP, facilita la realizzazione di una rete completamente ottica, anche nella sezione di accesso. La fig. 6.5 mostra alcuni casi pratici di accesso in fibra ottica. Iniziando da destra, sono mostrate delle reti Gigabit Ethernet (pag. 207) residenziali, interconnesse mediante *switch di livello 2* ad un POP (*Point of Presence*), il cui router (12000) si interconnette ad un anello SDH a 655 Mbps, sul quale sono instradati i pacchetti IP diretti verso Internet, per il tramite del POP primario. In basso a sinistra, sono mostrati accessi a due Megabit, contenenti sia traffico voce che dati, che vengono inseriti in anelli SDH da 155 Mbps: quello al centro inoltra i canali voce verso la PSTN, mentre quello di sinistra si interconnette nel *backbone* IP da 2.5 Gbps.

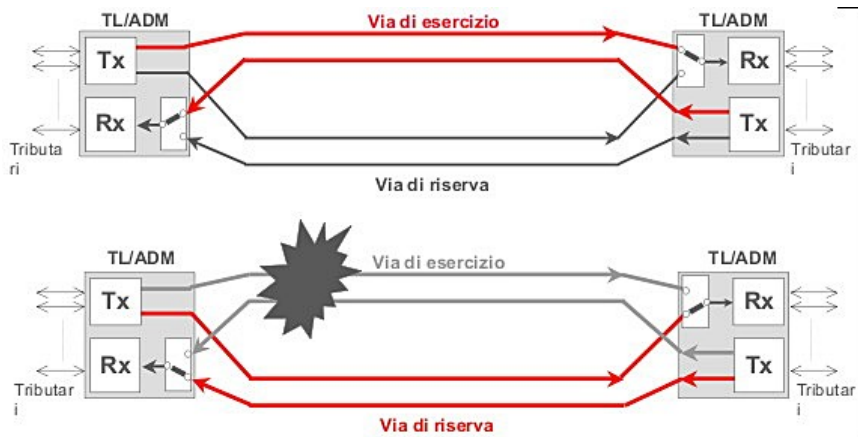


Figura 6.6: Sistema di protezione 1+1

automatico tra la linea andata fuori servizio, ed una riserva presente, come indicato nei seguenti schemi.

Protezione 1+1 In questo caso, ogni collegamento (vedi fig. 6.6) è provvisto di un collegamento di riserva. Qualora la via di esercizio vada fuori servizio, i terminali di linea che sono posti agli estremi se ne avvedono pressoché immediatamente, e provvedono a commutare la comunicazione sulla via di riserva.

Collegamento ad anello Come prima, per ognuna delle due direzioni di trasmissione è impegnata una diversa fibra ottica, ma in questo caso la via di ritorno (vedi fig. 6.7) si sviluppa investendo l'altra metà della circonferenza, percorsa nello stesso senso di rotazione. Aggiungendo un secondo anello di riserva (quello interno nella figura), anch'esso unidirezionale ma diretto in senso opposto al primo, la comunicazione può continuare anche nel caso in cui entrambi i collegamenti (generalmente co-locati) tra due nodi vadano fuori servizio.

6.7 Instradamento

Per questo argomento sono fornite solo alcune definizioni estremamente sommarie di tre possibili strategie, adottate nel corso della evoluzione delle reti telefoniche:

end to end o *Right-through* (da estremo ad estremo): la scelta del percorso è effettuata dalla centrale di origine, ad esempio in base al prefisso od all'inizio del numero, utilizzando delle *tabelle di routing* statiche. E' la modalità dell'inizio della telefonia, in cui i commutatori erano elettromeccanici, ed i collegamenti verso altre centrali erano *cablati*. Ha l'enorme svantaggio che i cambiamenti alla topologia della rete si devono riflettere in cambiamenti di tutte le tabelle - *o dei morsetti!*

Link-by-link o *Own-Exchange* (tratta per tratta): ogni centrale decide in autonomia dove instradare (in uscita) le connessioni entranti, in base a sue tabelle dinami-

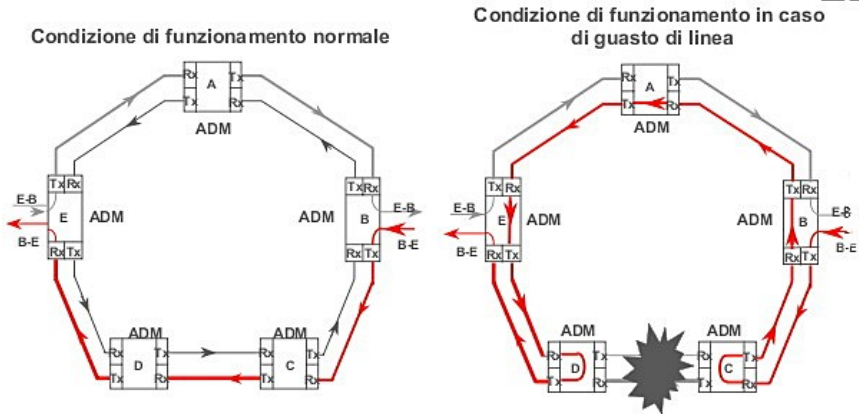


Figura 6.7: Configurazione ridondante ad anello

che, ovvero informazioni che giungono dalla rete stessa. Si adatta alle modifiche della topologia ma non è affidabile al 100 %, potendo ad esempio produrre dei *loop* (o circoli viziosi);

Tramite ccs (*Common Channel Signaling*, segnalazione a canale comune): le decisioni sull'instradamento sono demandate ad una rete di segnalazione parallela ed indipendente da quella del traffico smaltito, e che collega tutte le centrali ad un unico organo di controllo (il *canale comune*), il quale determina l'instradamento in base alla sua conoscenza dello stato del traffico nella rete, e comunica contemporaneamente a tutte le centrali coinvolte nell'instradamento, come configurare i propri organi di commutazione per realizzare il collegamento richiesto.

6.8 Commutazione

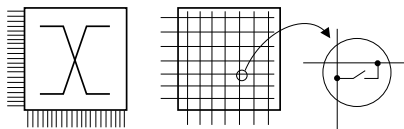
Illustriamo ora l'architettura di dispositivi che consentono la cosiddetta *commutazione di circuito*, ovvero la creazione di un collegamento *stabile* tra due porte del commutatore, con un *impegno permanente* di *risorse fisiche* per tutta la durata del collegamento. Un'altra modalità di commutazione, quella di *pacchetto*, sarà illustrata al Capitolo 8.

6.8.1 Reti a divisione di spazio

Sono chiamati così gli organi di commutazione che realizzano un collegamento fisico (elettrico) tra uno degli N ingressi ed una delle M uscite. Nel caso in cui $N > M$, la rete è un *concentratore*²⁵, mentre se $N < M$ la rete è un *espansore*; se $N = M$ la rete è quadrata e *non bloccante*.

Il commutatore è rappresentato da un blocco con una "X" (in inglese *cross*, od incrocio), e può essere pensata come una matrice binaria in cui ogni elemento (1 o

²⁵come ad esempio un centralino (PBX, PRIVATE BRANCH EXCHANGE) con 8 derivati (interni) e 2 linee esterne: se due interni parlano con l'esterno, un terzo interno che vuole anche lui uscire trova occupato. Si dice allora che si è verificata una condizione di *blocco*.

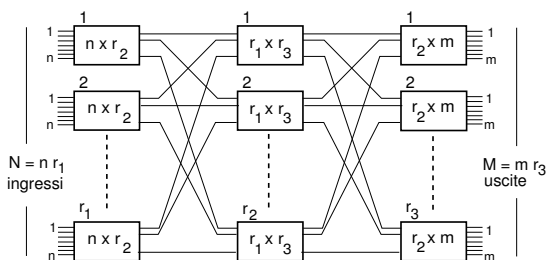


0) rappresenta lo stato (chiuso od aperto) di un interruttore (realizzato ad esempio mediante un transistor) che collega una linea di ingresso ad una di uscita.

Realizzare in questo modo una rete non bloccante prevede l'uso di un numero di interruttori pari ad $N \cdot M$, dei quali solo $\min(N, M)$ sono utilizzati, anche nelle condizioni di massimo carico. Inoltre, nessun interruttore può essere "eliminato" senza precludere irrimediabilmente la possibilità di collegare qualunque ingresso a qualunque uscita. Allo scopo di utilizzare un numero ridotto di interruttori, sia per costruire reti non bloccanti oppure bloccanti con bassa probabilità di blocco, si ricorre alle...

6.8.2 Reti multistadio

...di cui in figura è riportato un esempio a 3 stadi, in cui gli N ingressi sono ripartiti su r_1 reti più piccole con n ingressi, e le M uscite su r_3 reti con m uscite. Nel mezzo ci sono r_2 reti con r_1 ingressi ed r_3 uscite. Si può dimostrare che la rete complessiva è non bloccante se



il numero di matrici dello stadio intermedio è almeno $r_2 \geq n + m - 1$ (condizione di CLOS²⁶). Una connessione da sinistra a destra ha ora la possibilità di scegliere attraverso quale matrice intermedia passare.

Nel caso di reti quadrate ($N = M$), ponendo $n = m = \sqrt{\frac{N}{2}}$, si ottiene un numero complessivo di incroci pari a $4 \left(\sqrt{2} N^{\frac{3}{2}} - N \right)$, che risulta inferiore ad N^2 (e dunque vantaggioso rispetto ad un commutatore monostadio) a partire da $N \geq 24$. Ovviamente, la problematica relativa alle matrici di commutazione è molto articolata, coinvolgendo topologie più complesse, filosofie di instradamento, e tecniche per la stima delle probabilità di blocco. Tralasciamo ulteriori approfondimenti, per illustrare invece come realizzare dispositivi di commutazione per trasmissioni numeriche a *divisione di tempo*.

6.8.3 Commutazione numerica a divisione di tempo

Consideriamo il caso in cui si debbano commutare le comunicazioni associate ai singoli *time-slot* presenti in diversi flussi²⁷ numerici organizzati in trame. Avendo a disposi-

²⁶E' una condizione *sufficiente* a scongiurare il blocco anche nella condizione *peggiore*. Tale circostanza si verifica quando:

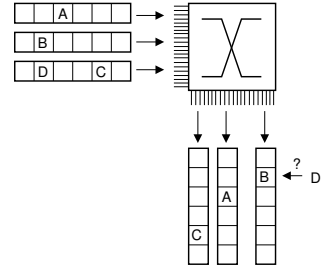
- ▷ una matrice del primo stadio (i) ha $n - 1$ terminazioni occupate
- ▷ una matrice del terzo stadio (j) ha $m - 1$ terminazioni occupate e
- ▷ tali terminazioni non sono connesse tra loro, anzi le connessioni associate impegnano ognuna una diversa matrice intermedia e

- ▷ si richiede la connessione tra le ultime due terminazioni libere di i e j

⇒ in totale si impegnano allora $m - 1 + n - 1 + 1 = m + n - 1$ matrici intermedie.

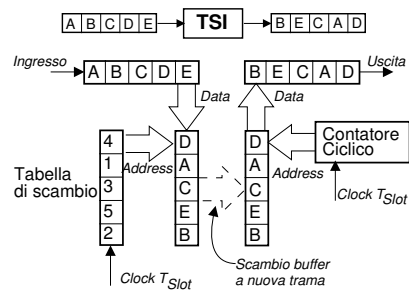
²⁷Le comunicazioni presenti in uno stesso flusso, ovvero appartenenti alla stessa trama, condividono la stessa origine/destinazione.

zione solamente una matrice di commutazione spaziale, quest'ultima può essere riprogrammata alla stessa frequenza dei time-slot, consentendo alle comunicazioni entranti di dirigersi verso i flussi uscenti in direzione delle rispettive destinazioni finali. La matrice spaziale, però, non può alterare l'ordine temporale dei dati in ingresso! Come rappresentato nella figura a lato, non può (ad esempio) inviare le conversazioni *B* e *D* sulla stessa linea uscente, in quanto si verifica un conflitto temporale. E' quindi evidente la necessità di introdurre uno stadio di *commutazione temporale*.



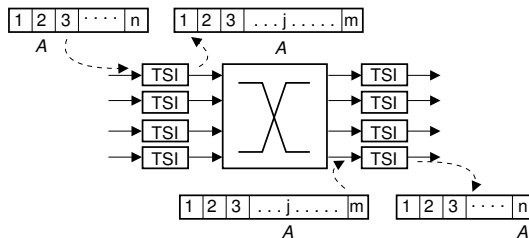
6.8.3.1 Time Slot Interchanger

Questo dispositivo è indicato come TSI (*Time Slot Interchanger*) ed ha la funzione di produrre in uscita una sequenza di dati identica a quella in ingresso, tranne per averne cambiato l'ordine temporale. In figura è mostrato un possibile schema di funzionamento: una trama entrante viene scritta, agli indirizzi ottenuti leggendo sequenzialmente la tabella di scambio, in un buffer di memoria (es.: *entra E e lo scrivo al 4° posto, poi entra D e va al 1° posto, etc.*). Prima dell'inizio di una nuova trama, il primo buffer è copiato in un secondo²⁸, che a sua volta viene letto con ordine sequenziale (partendo dall'alto), per creare la nuova trama in uscita. Ovviamente, è possibile anche la realizzazione opposta, con scrittura sequenziale e lettura secondo il nuovo ordinamento.



6.8.3.2 Commutazione bidimensionale

Così come un commutatore spaziale non è sufficiente, anche un TSI "da solo" è di scarsa utilità, non potendo instradare le comunicazioni su vie diverse. Combinando assieme le due funzioni, si giunge a realizzare commutatori sia di tempo che di spazio, come la struttura a 3 stadi in figura, chiamata "TST" perché alterna uno stadio temporale, uno spaziale ed uno temporale.



²⁸La tecnica prende il nome di *Double Buffering*.

Notiamo subito che, in questo schema, il numero di intervalli temporali in uscita dai TSI di ingresso è $m > n$ ⁽²⁹⁾: ciò determina, per lo stadio spaziale, una frequenza di commutazione più elevata della frequenza dei time-slot in ingresso. Una generica conversazione “A” che occupa il 2° slot del primo flusso può raggiungere (ad esempio) l’ultimo slot dell’ultimo flusso, occupando uno qualsiasi (j) degli m slot utilizzati dal commutatore spaziale. Aumentando il valore di m , si riduce la probabilità di blocco; in particolare, questa è nulla se $m = 2n - 1$ ⁽³⁰⁾.

Analizziamo i vantaggi conseguiti dalla commutazione numerica con un semplice esempio. Poniamo di voler commutare con lo schema illustrato 4 flussi PCM (con $n = 30$): i $4 * 30 = 120$ canali presenti sono commutati utilizzando solo $4 * 4 = 16$ interruttori, contro i $120 * 120 = 14.400$ interruttori necessari ad una matrice spaziale monostadio che svolga la commutazione dei 120 canali analogici !

6.9 Appendici

6.9.1 Plain old telephone service (POTS)

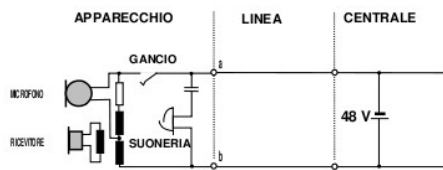
Il *buon vecchio servizio telefonico* consiste nel collegamento audio, nella banda del canale telefonico, attuato mediante un *terminale di utente* (telefono), e nella *segnalazione* (sempre *di utente*) necessaria ad instaurare il collegamento. L’insieme degli apparati che permette di interconnettere tra loro i telefoni di rete fissa è spesso indicato con l’acronimo PSTN (*Public Switched Telephone Network*), da cui si sono evoluti tutti gli sviluppi successivi delle telecomunicazioni.

Quando la centrale locale deve far squillare il telefono, invia sul doppino una tensione alternata che ne attiva la suoneria. Quando la cornetta dell’apparecchio telefonico viene sollevata ³¹, nel telefono si chiude un interruttore che determina lo scorrimento di una corrente continua nel *subscriber loop*, indicando la risposta da parte del chiamato. Se viceversa siamo dal lato chiamante, sollevando la cornetta *albertiamo* la centrale di accesso, la quale dopo aver riservato le risorse necessarie (ivi compreso un time-slot in uno dei flussi PCM uscenti) ci manifesta la sua disponibilità ad acquisire il numero che intendiamo comporre, mediante l’invio di un *tono di centrale*.

All’interno del telefono troviamo un particolare trasformatore a quattro porte, detto *ibrido* ³², in grado di separare il segnale in ingresso da quello in uscita, in modo da inviare il primo all’altoparlante, e di inviare al secondo quello del microfono.

Per comporre il numero, fino agli anni 80 erano in uso i *dischi combinatori*, che aprendo e chiudendo l’interruttore, determinavano una forma d’onda impulsiva, in cui il numero degli impulsi corrispondeva alla cifra immessa. Questo meccanismo è in diretta relazione alla presenza, nelle centrali telefoniche di prima generazione, dei motori passo-passo che determinavano l’azionamento dei commutatori di centrale.

Il disco combinatorio è stato poi soppiantato dalla attuale tastiera numerica DTMF (*Dual Tone Multi Frequency*), in cui ad ogni tasto (vedi lato sinistro della fig. 6.8) sono



²⁹Ovviamente, $m - n$ intervalli sono lasciati vuoti, in ordine *sperso* tra gli m .

³⁰Si confronti questo risultato con la condizione di Clos, fornita al § 6.8.2.

³¹In inglese si dice andare OFF-HOOK, con riferimento storico al gancio su cui riporre la cornetta, presente nei primi modelli di telefono.

³²http://en.wikipedia.org/wiki/Hybrid_coil

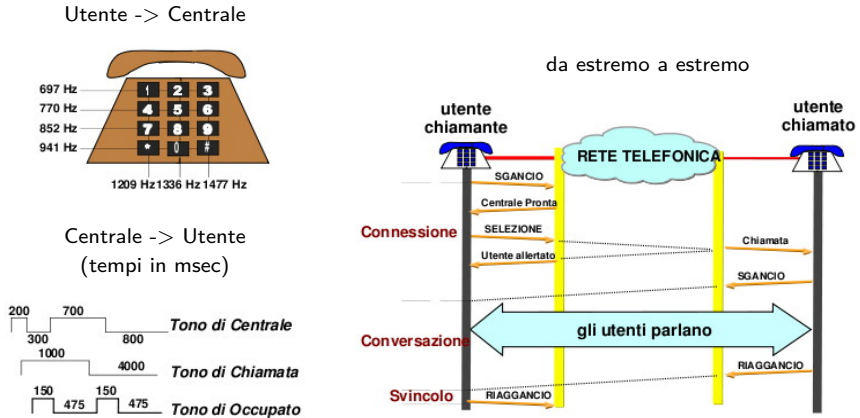


Figura 6.8: Segnalazione di utente

associate *due frequenze* che individuano la cifra (od il simbolo * e #) premuta, come descritto dalla figura. Viceversa, la segnalazione di utente nella direzione centrale -> utente avviene per mezzo di un codice basato su di un tono intermittente a 440 Hz³³, le cui durate sono descritte in basso a sinistra nella figura.

A seguito della ricezione del numero, la centrale *di origine* coinvolge il resto della rete, impegnando risorse della stessa, ed individuando quali nodi attraversare per giungere a destinazione (fase di *istadamento*, in inglese ROUTING). Una volta contattata la centrale di destinazione, questa provvede a far squillare il telefono chiamato, ed inviare indietro un segnale di *RingBack* che produce presso il chiamante un *tono di libero*, oppure un segnale di occupato (*Busy*), nel caso in cui il chiamato sia già impegnato in altra conversazione.

Il risultato dei messaggi di segnalazione di utente è esemplificato nel lato destro di fig. 6.8, in cui è evidenziato come ogni conversazione è in realtà composta da tre fasi imprescindibili:

- formazione della connessione (*call setup*), in cui sono svolte le funzioni di indirizzamento, e vengono riservate da parte della rete le risorse necessarie alla comunicazione
- mantenimento (*hold*), durante la quale le risorse impegnate sono utilizzate in modo esclusivo dalle parti in conversazione
- svincolo (*release*) in cui le risorse impegnate sono liberate

Il passaggio dalla telefonia analogia a quella numerica, in cui il segnale vocale è campionato e quantizzato come PCM, non ha di fatto alterato la presenza di queste tre fasi.

³³corrispondente al *la* centrale del pianoforte. Ho provato a verificare, e... a me arriva un la bemolle!

6.9.2 ISDN

La *Integrated Service Data Network*³⁴ è una modalità di accesso *numerico* alla rete telefonica, definito da una serie di standard reperibili presso l'ITU³⁵. In ISDN la conversione A/D avviene all'interno del terminale di utente, il quale può collegare allo stesso bus ISDN (interfaccia S³⁶ a quattro fili, utilizzando un codice di linea AMI), diversi dispositivi numerici, oppure anche analogici, interponendo per questi ultimi un dispositivo detto *Terminal Adapter* (TA).

L'accesso alla rete da parte del dispositivo NT (*Network Termination*) connesso al doppino, corrisponde alla *Interfaccia U*³⁷, su cui è trasmesso un segnale a quattro livelli noto come 2B1Q³⁸, per il quale sono standardizzate due diverse velocità di trasmissione. Nella modalità cosiddetta di *Accesso Base* (BRI, *Basic Rate Interface*), si ha a disposizione un collegamento numerico di banda base a 144 kbps, in cui trova posto una

struttura di trama³⁹ che ospita due canali voce (B1 e B2, da *Bearer*, ossia *portatore*, con dati PCM) a 64 kbps, in cui la trasmissione avviene in modo ininterrotto, e un canale dati (D) a 16 kbps, in cui la trasmissione avviene in modalità a pacchetto, ed in cui trovano posto le informazioni di segnalazione⁴⁰, come il protocollo Q.931⁴¹.

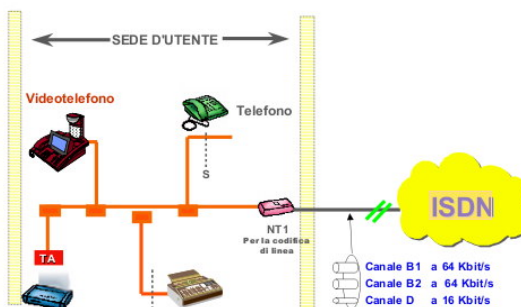
Nella modalità di *Accesso Primario* (PRI, *Primary Rate Interface*), adatta al collegamento di centralini, si hanno a disposizione 30 canali B (voce) a 64 kbps, ed un canale D (dati) di segnalazione a 64 kbps. Pertanto, PRI viene direttamente interconnesso al primo livello (E1) della gerarchia PDH descritta al § 6.3.4.

Dato che l'accesso ISDN preserva il flusso binario inviato sui canali B da estremo a estremo della rete, su quegli stessi canali posso essere inviate anche informazioni niente affatto vocali, ma bensì nativamente numeriche, purché il ricevente condivida le stesse modalità di interpretazione dei bit in arrivo. Sfruttando tale possibilità, sono ad esempio stati definiti i primi standard di videotelegrafia H.320⁴².

6.9.3 Sistema di segnalazione numero 7

Il *Signaling System #7* (SS7⁴³) è un insieme di protocolli di segnalazione telefonica a canale comune, usato per controllare la maggior parte delle chiamate telefoniche della

Accesso Base ISDN



³⁴<http://it.wikipedia.org/wiki/ISDN>

³⁵<http://www.itu.int/rec/T-REC-I/e>

³⁶<http://hea-www.harvard.edu/~fine/ISDN/n-isdn.html>

³⁷<http://www.ralpb.net/ISDN/ifaces.html>

³⁸<http://it.wikipedia.org/wiki/2B1Q>

³⁹http://telemat.die.unifi.it/book/corso_telematica/lez_100/grp_4.html

⁴⁰<http://www.rhyshaden.com/isdn.htm>

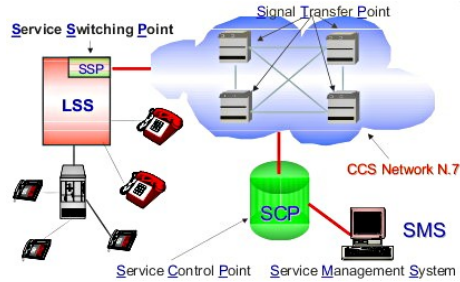
⁴¹<http://www.javvin.com/protocolQ931.html>

⁴²<http://it.wikipedia.org/wiki/H.320>

⁴³http://en.wikipedia.org/wiki/Signaling_System_7

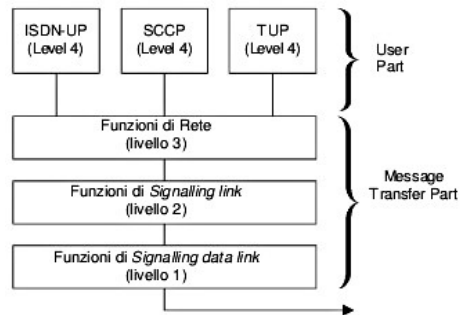
PSTN mondiale, che in questo caso prende il nome di *Intelligent Network* (IN⁴⁴). Oltre ad gestire la fasi di instaurazione e abbattimento della chiamata, permette altri servizi come reindirizzamento, carte prepagate, SMS, numero verde, conferenza, richiamata su occupato...

L'SS7 è descritto dalla serie di raccomandazioni ITU-T Q.700⁴⁵, a cui aderiscono anche le varianti regionali descritte da altri enti normativi. I messaggi SS7 sono trasferiti mediante connessioni numeriche tra entità di segnalazione, ospitate nelle centrali telefoniche, indicate con i termini di



- *Service switching point* (SSP⁴⁶), che termina la segnalazione di utente, ed invia una query all'SCP per determinare come gestire la richiesta di servizio;
- *Signal Transfer Point* (STP⁴⁷), che instrada i messaggi SS7 tra le diverse entità della IN;
- *Service Control Point* (SCP⁴⁸), che interroga un *Service Data Point* (SDP⁴⁹), il quale a sua volta detiene un database che (ad es.) identifica il numero geografico a cui deve essere inoltrata una chiamata diretta ad un numero verde. Alternativamente, l'SCP può determinare la riproduzione di messaggi preregistrati, o richiedere ulteriore input da parte del chiamante, in base all'*Intelligent Network Application Protocol* (INAP⁵⁰) che opera sopra il *Transaction Capabilities Application Part* (TCAP) della pila protocollare SS7.

Oltre alle entità che prendono parte alla architettura, SS7 è definito anche nei termini della gerarchia protocollare che descrive la stratificazione delle funzioni necessarie allo svolgimento dei servizi richiesti. Il semplice scambio dei messaggi tra le entità è basato su di una rete a commutazione di pacchetto, ed avviene in base alle procedure collettivamente indicate come *Message Transfer Part* (MTP⁵¹), responsabile della consegna affidabile dei messaggi SS7 tra le parti in comunicazione. Le funzioni di MTP sono stratificate su tre livelli, che dal basso in alto, si occupano degli aspetti di trasmissione tra le entità, della gestione degli errori in modo da garantire una comunicazione affidabile, e dell'instradamento dei messaggi tra le entità.



⁴⁴http://en.wikipedia.org/wiki/Intelligent_network

⁴⁵<http://www.itu.int/rec/T-REC-Q.700/en>

⁴⁶http://en.wikipedia.org/wiki/Service_switching_point

⁴⁷http://en.wikipedia.org/wiki/Signal_transfer_point

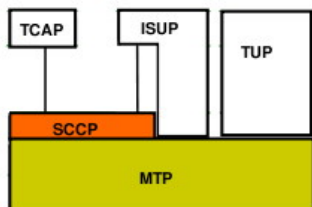
⁴⁸http://en.wikipedia.org/wiki/Service_Control_Point

⁴⁹http://en.wikipedia.org/wiki/Service_Data_Point

⁵⁰<http://en.wikipedia.org/wiki/INAP>

⁵¹http://en.wikipedia.org/wiki/Message_Transfer_Part

Al di sopra della MTP possono operare diversi protocolli indicati come *User Part*, come ad esempio il *Signalling Connection Control Part* (SCCP⁵²), che arricchisce le funzionalità di rete, offrendo ulteriori capacità di indirizzamento, ed un servizio orientato alla connessione anziché a pacchetto; attraverso SCCP possono operare processi applicativi basati sul *Transaction Capabilities Application Part* (TCAP⁵³).

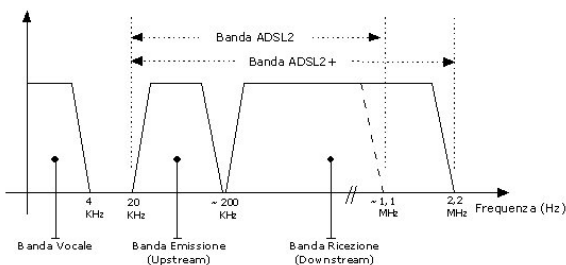


Altri esempi di *User Part* sono la *Telephone User Part* (TUP⁵⁴) e la *ISDN User Part* (ISUP⁵⁵). TUP è stata la prima UP ad essere definita, e fornisce il supporto all'offerta di servizi PSTN mediante la rete SS7. Attualmente è quasi ovunque rimpiazzato da ISUP, che offre altri servizi, come ad esempio l'identificazione del chiamante, e che può dialogare con l'MTP anche per il tramite di SCCP.

Qualora la rete di interconnessione tra le entità della IN sia una rete IP, allora sono da considerare gli ulteriori protocolli indicati come SIGTRAN⁵⁶ o *Signalling Transport*.

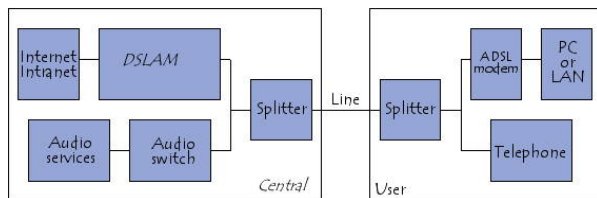
6.9.4 ADSL

L'*Asymmetric Digital Subscriber Line*⁵⁷ è l'insieme di tecnologie trasmissive e di rete per mezzo delle quali viene fornito l'accesso ad Internet *a banda larga* per il tramite del doppio telefonico (*subscriber loop*) in rame, già utilizzato per il normale servizio telefonico PO-



TS. L'uso condiviso del mezzo è reso possibile realizzando la trasmissione numerica ADSL su di una banda di frequenze *più elevate* di quelle usate da POTS, come mostrato in figura, dove sono rappresentati gli intervalli di frequenza riservati alla telefonia PSTN, ai dati in uscita (*upstream*) ed in ingresso (*downstream*).

La massima velocità di trasmissione è stata inizialmente posta rispettivamente pari ad 1 ed 8 Mbps per i due versi trasmissivi, anche in funzione della lunghezza del collega-



mento utente - centrale⁵⁸; successivamente, la massima velocità di ricezione è stata elevata rispettivamente a 12 e 20 Mbit/sec per gli standard ADSL2 e ADSL2+.

⁵²http://en.wikipedia.org/wiki/Signaling_Connection_and_Control_Part

⁵³http://en.wikipedia.org/wiki/Transaction_Capabilities_Application_Part

⁵⁴[http://en.wikipedia.org/wiki/Telephone_User_Part_\(TUP\)](http://en.wikipedia.org/wiki/Telephone_User_Part_(TUP))

⁵⁵http://en.wikipedia.org/wiki/ISDN_User_Part

⁵⁶<http://en.wikipedia.org/wiki/SIGTRAN>

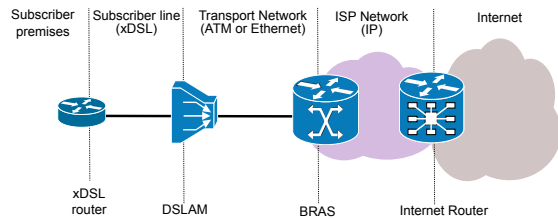
⁵⁷http://en.wikipedia.org/wiki/Asymmetric_Digital_Subscriber_Line

⁵⁸All'aumentare della lunghezza del collegamento, oltre a ridursi la potenza ricevuta e quindi peggiorare l'SNR, aumenta l'entità della diafonia tra utenti differenti, determinando un ulteriore peggioramento di SNR, che la tecnica di modulazione affronta riducendo la velocità trasmissiva.

I due segnali (vocale e numerico) sono poi separati su linee differenti⁵⁹ inserendo, a valle della presa telefonica casalinga, un doppio filtro passa-alto e passa-basso detto *splitter*. Un filtro del tutto simile esiste anche dal lato centrale, in modo da inoltrare la componente in banda audio alla centrale POTS, e la componente dati verso un dispositivo DSLAM.

DSLAM e oltre Il *Digital Subscriber Line Access Multiplexer* risiede nella centrale dell'operatore che offre il servizio POTS, provvede ad effettuare la demodulazione del segnale ADSL di ogni singolo utente, e si occupa di aggregare il traffico relativo a più utenti ed inviarlo verso gli ISP (*Internet Service Provider*) con cui gli utenti hanno un contratto di connessione ad Internet.

A questo fine può essere necessario attraversare prima una rete di trasporto⁶⁰ basata su ATM o ETHERNET, che termina il traffico sul *Broadband Remote Access Server* (BRAS)⁶¹ dell'ISP, utilizzato da quest'ultimo anche per terminare il protocollo di strato di collegamento PPP⁶², svolgere le funzioni di autenticazione dell'accesso, ed applicare eventuali *policy* a livello IP. Quindi, l'ISP provvede ad interconnettere il traffico del cliente con la rete Internet. Alternativamente, l'ISP può disporre di un *Point of Presence* (POP) nella stessa centrale in cui sono ospitati i DSLAM dei propri clienti⁶³, che in questo caso producono direttamente traffico IP, inoltrato verso la *core network* dell'ISP usando la sua stessa connettività.



DMT Il modem ADSL utilizza a tecnica di modulazione numerica multi-portante detta *Discrete Multi Tone*, in cui il flusso binario viene ripartito su più canali di frequenza contigui, ed il segnale analogico sintetizzato direttamente nel dominio della frequenza mediante il calcolo di una FFT, come previsto dalla tecnica di trasmissione OFDM (vedi § 13.7). In questo modo, oltre a semplificare le operazioni di equalizzazione, è possibile variare la velocità di trasmissione in modo indipendente per le diverse portanti, e mantenere buone prestazioni anche nel caso in cui l'SNR vari con la frequenza.

6.9.5 TDM mediante modulazione di ampiezza degli impulsi

Al tempo in cui la realizzazione del componente di quantizzazione (vedi § 4.2.1.1) presentava discrete difficoltà circuitali, si pensò⁶⁴ di sfruttare il teorema del campionamento (vedi § 4.1) per inviare su di un unico collegamento più comunicazioni multiplata a divisione di tempo (TDM = *Time Division Multiplex*).

⁵⁹ Ciò permette di non ascoltare nel telefono il fruscio della trasmissione ADSL, e di ridurre il rischio che le comunicazioni vocali determinino la caduta della connessione ADSL.

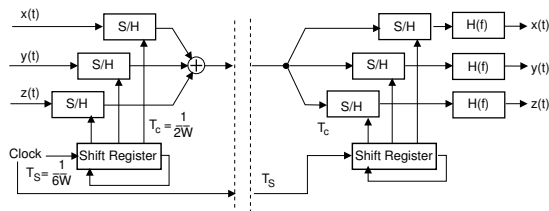
⁶⁰ Non lasciarsi fuorviare dal ruolo di *trasporto* della rete, che in effetti assolve unicamente un ruolo di *livello due* (o di collegamento), in quanto il punto di uscita non è qualsiasi, ma l'ISP fornitore dell'utente.

⁶¹ http://en.wikipedia.org/wiki/Broadband_Remote_Access_Server

⁶² http://en.wikipedia.org/wiki/Point-to-Point_Protocol

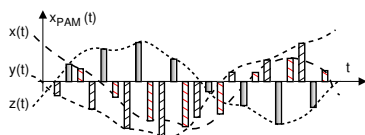
⁶³ Come ad es, nel caso dell'Unbundling: http://it.wikipedia.org/wiki/Unbundling_local_loop

⁶⁴ La *pensata* non ebbe molte applicazioni, se non in ambito della commutazione interna ad esempio ad un centralino, a causa della sensibilità del metodo agli errori di temporizzazione, ed alle caratteristiche del mezzo trasmissivo su cui inviare il segnale PAM.



E' sufficiente infatti sommare alla funzione $x^\circ(t)$ introdotta al § 4.1.4 altri segnali simili, ad esempio $y^\circ(t)$, $z^\circ(t)$ come mostrato in Figura, ognuno campionato a frequenza f_c , ma sfasato rispetto agli altri.

Da questa modalit  di multiplicit  analogica deriva il termine *onda PAM*, che sta per *Pulse Amplitude Modulation*, ovvero modulazione ad ampiezza di impulsi; gli impulsi sono separati da un intervallo $T_s = \frac{1}{Nf_c}$, con N pari al numero di segnali multipliciti. Il pedice s indica che si tratta di un *periodo di simbolo*.



Il segnale $x_{PAM}(t)$ composto dalle 3 sorgenti dell'esempio della figura in alto   mostrato a lato, e pu  essere nuovamente campionato estraendo $x(nT_c)$, $y(nT_c)$, $z(nT_c)$, mentre i segnali $x(t)$, $y(t)$ e $z(t)$ sono riprodotti facendone passare gli impulsi campionati a frequenza f_c in un filtro di ricostruzione con banda $W \leq f_c/2$.

6.10 Riferimenti

Per questo capitolo un po particolare, si elencano in modo distinto alcune fonti on-line a cui ci si   ispirati, e dalle quali sono state tratte alcune illustrazioni.

- La Rete di Telecomunicazioni http://net.infocom.uniroma1.it/corsi/impianti/lezioni_new/lez_1.pdf di *Stefano Paggi* <http://net.infocom.uniroma1.it/corsi/impianti/impianti.htm>
- ISDN <http://www-tlc.deis.unibo.it/Didattica/CorsiB0/RetiLB/lucidi/ISDN.pdf> di *Giorgio Corazza* <http://www-tlc.deis.unibo.it/Didattica/CorsiB0/RetiLB/>
- Sistema di Segnalazione SS No 7 <http://www.cedi.unipr.it/links/Corsi/telematica/Materiale/dispense/Telefonia/Ss7.pdf> di *A. Lazzari* <http://www.cedi.unipr.it/links/Corsi/telematica/Materiale/dispense.html>
- La segnalazione a canale comune http://primo.ing.uniroma1.it/materiale/Commutazione/2007_2008/IV.ppt di *Aldo Roveri*
- La Rete Telefonica <http://www.cedi.unipr.it/links/Corsi/telematica/Materiale/dispense/Telefonia/Telefonica.pdf> di *A. Lazzari*
- Understanding SONET/ SDH <http://www.electrosofts.com/sonet/index.html>
- Reti Ottiche <http://net.infocom.uniroma1.it/corsi/ro/ro.htm> di *Andrea Baiocchi*

Capitolo 7

Probabilità, processi, decisione statistica

In questo capitolo sono sviluppati argomenti attinenti la teoria delle probabilità, e che nel testo sono utilizzati quasi ovunque. Dopo aver brevemente riassunto assiomi, teoremi base e concetti di teoria delle probabilità, vengono introdotte le variabili aleatorie, caratterizzate mediante distribuzioni e momenti, e sono definite le v.a. uniforme e gaussiana. Si passa quindi a descrivere i processi aleatori, le diverse modalità di definirne delle medie, e le proprietà di stazionarietà ed ergodicità. Infine, un'ampia appendice è dedicata ad approfondire alcuni aspetti come funzione caratteristica, trasformazioni di v.a., detezione di sinusoidi nel rumore, elementi di statistica, decisione, e stima di parametro.

7.1 Teoria delle probabilità

Tratta delle caratteristiche *regolari* di fenomeni *irregolari* o *casuali*. Una prima definizione di probabilità è quella fornita dalla teoria frequentistica, la quale asserisce che se, ripetendo N volte un esperimento, si verifica la circostanza A per n_A volte, per essa si osserva una frequenza relativa n_A/N , da cui si deriva la probabilità di A come

$$Pr_A = \lim_{N \rightarrow \infty} \frac{n_A}{N}$$

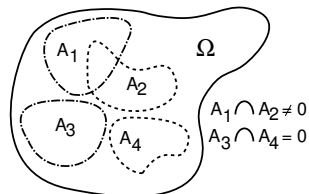
In termini più astratti, l'insieme di tutte le circostanze possibili può essere pensato come un insieme algebrico, i cui elementi (o punti) sono appunto le diverse circostanze. I punti possono essere raggruppati in sottoinsiemi (eventualmente vuoti o di un solo punto) per i quali valgono le proprietà di unione, intersezione, complemento, inclusione...

I fenomeni fisici sono posti in relazione con i punti degli insiemi suddetti mediante il concetto di *spazio campione* Ω , che è l'unione di tutti i possibili risultati di un *fenomeno aleatorio*. Sottoinsiemi dello spazio campione sono detti *eventi*. L'intero spazio è l'*evento certo*, mentre l'insieme vuoto corrisponde all'evento impossibile ϕ (od evento *nullo*). Una *unione* \cup di eventi, corrisponde all'evento che si verifica ogni qualvolta se ne verifichi *un suo componente*, mentre l'*intersezione* \cap è verificata se *tutti* i componenti lo sono. Esempio: il lancio di un dado genera uno spazio con 6 punti (eventi) disgiunti. Uno spazio campione può avere un numero di punti finito, infinito numerabile, o infinito.

7.1.1 Assiomi delle probabilità

Costituiscono le basi su cui sono costruiti i teoremi seguenti, ed affermano che:

- $0 \leq Pr(A) \leq 1$: la probabilità di un evento è compresa tra 0 ed 1;
- $Pr(\Omega) = 1$: la probabilità dell'evento certo è 1;
- Se $Pr(A_i \cap A_j) = \phi$ allora $Pr(\bigcup A_i) = \sum Pr(A_i)$: la probabilità dell'unione di eventi *disgiunti* è la somma delle singole probabilità.



7.1.2 Teoremi di base

- $Pr(\phi) = 0$: la probabilità dell'evento impossibile è nulla;
- $Pr(A \cap B) + Pr(A \cap \bar{B}) = Pr(A)$, e $Pr(B) + Pr(\bar{B}) = 1$: un evento ed il suo complemento riempiono lo spazio (detto anche teorema delle *probabilità totali*¹);
- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$: la probabilità dell'evento intersezione si conta una volta sola. Esempio alla nota²;
- Se $B \subseteq A$ allora $Pr(B) \leq Pr(A)$: quando l'evento B è contenuto in A il verificarsi del primo implica il secondo.

7.1.3 Probabilità condizionali

Può avvenire che il verificarsi di un evento *influenzi* il verificarsi o meno di un altro: Si dice allora che lo condiziona, ovvero che l'evento influenzato è *condizionato*. La probabilità che avvenga A , noto che B si sia verificato, si scrive $Pr(A/B)$, e si legge probabilità (condizionata) di A dato B , che è definita³ come

$$Pr(A/B) = \frac{Pr(A, B)}{Pr(B)}$$

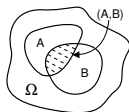
in cui $Pr(A, B) = Pr(A \cap B)$ è la probabilità *congiunta* che A e B si verifichino entrambi, ed a patto che $Pr(B) \neq 0$ (altrimenti anche $Pr(A/B)$ è zero!).

Esercizio: Valutare la probabilità condizionata $Pr(A/B)$ che lanciando un dado si ottenga un numero pari (evento $A = \{pari\}$), condizionatamente all'evento B che il numero sia > 2 .
 Soluzione alla nota⁴.

¹Utile per scrivere la probabilità di un evento come "1 meno" quella dell'evento complementare.

²Lanciando un dado, la probabilità $Pr(pari \cup > 2)$ di ottenere un numero pari, *oppure* più grande di due, è la *somma* delle probabilità dei singoli eventi $Pr(pari) = \frac{3}{6}$ e $Pr(> 2) = \frac{4}{6}$, meno quella che si verifichino assieme $Pr(pari \cap > 2) = \frac{2}{6}$. Pertanto: $Pr(pari \cup > 2) = \frac{3}{6} + \frac{4}{6} - \frac{2}{6} = \frac{5}{6}$.

³La relazione può essere verificata ricorrendo al diagramma in figura, ed interpretando $Pr(A/B)$ come il rapporto tra la misura di probabilità dell'evento congiunto, rispetto a quella dell'evento



condizionante.

⁴Il risultato è pari alla probabilità $Pr(A, B) = Pr(pari, > 2)$ che i due eventi si verifichino contemporaneamente, divisa per la probabilità $Pr(B) = Pr(> 2)$ che il numero sia > 2

Si rifletta sulla circostanza che la probabilità del pari $Pr(A) = \frac{1}{2}$, quella $Pr(B) = \frac{4}{6}$, o quella congiunta di entrambi $Pr(A, B) = \frac{2}{6}$, sono tutte riferite ad un qualunque lancio di dado, men-

A partire dalla precedente definizione, si ottiene quella della probabilità *congiunta*: $Pr(A, B) = Pr(A/B)Pr(B)$; inoltre, gli eventi condizionante e condizionato possono invertire i rispettivi ruoli, permettendo di scrivere anche: $Pr(A, B) = Pr(B/A)Pr(A)$. Eguagliando le due espressioni, si ottiene:

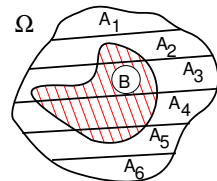
$$Pr(A/B) = \frac{Pr(B/A)Pr(A)}{Pr(B)} \quad \text{ed anche} \quad Pr(B/A) = \frac{Pr(A/B)Pr(B)}{Pr(A)}$$

Come ultima definizione, rammentiamo che le probabilità $Pr(A)$ e $Pr(B)$ sono indicate come *marginali*.

7.1.4 Teorema di Bayes

A volte, non tutti i possibili eventi sono direttamente osservabili: in tal caso la probabilità marginale $Pr(A)$ è indicata come probabilità *a priori*. Qualora l'evento A sia in qualche modo legato ad un secondo evento B , che invece possiamo osservare, la probabilità condizionata $Pr(A/B)$ prende il nome di probabilità *a posteriori* perché, a differenza di quella *a priori*, rappresenta un valore di probabilità valutata *dopo* la conoscenza di B .

In generale, però, si conosce solamente $Pr(A)$ e $Pr(B/A)$ (queste ultime sono dette probabilità condizionate *in avanti*), e per calcolare $Pr(A/B)$ occorre conoscere anche $Pr(B)$. Quest'ultima quantità si determina *saturo* la probabilità congiunta $Pr(A, B)$ rispetto a tutti gli eventi marginali A_i possibili:



$$Pr(B) = \sum_i Pr(B, A_i) = \sum_i Pr(B/A_i)Pr(A_i)$$

a patto che risultino $Pr(A_i, A_j) = 0$ e $\bigcup A_i = \Omega$, e cioè che l'insieme degli $\{A_i\}$ costituisca una partizione dello spazio degli eventi Ω . Tale circostanza è mostrata in figura.

L'ultima relazione ci permette di enunciare il *teorema di Bayes*, che mostra come ottenere le probabilità *a posteriori* a partire da quelle *a priori* e da quelle condizionate *in avanti*:

$$Pr(A_i/B_j) = \frac{Pr(B_j/A_i)Pr(A_i)}{\sum_k Pr(B_j/A_k)Pr(A_k)}$$

Al §17.2.1 è mostrata l'applicazione del risultato ad un problema di decisione statistica tipico delle telecomunicazioni, relativo alla ricezione binaria. Di seguito, invece, è illustrato un esempio più diretto.

tre $Pr(\text{pari} / > 2)$ è relativa ad un numero ridotto di lanci, ossia solo quelli che determinano un risultato > 2 . Pertanto, essendo $Pr(B) \leq 1$, si ottiene $Pr(A/B) \geq Pr(A, B)$; infatti per l'esempio del dado si ottiene $Pr(\text{pari} / > 2) = Pr(\text{pari}, > 2) / Pr(> 2) = \frac{2}{6} / \frac{4}{6} = \frac{1}{2}$, che è maggiore di $Pr(\text{pari}, > 2) = \frac{1}{3}$ (i valori di probabilità sono ottenuti come rapporto tra il numero di casi favorevoli e quello dei casi possibili).

Si ottiene invece $Pr(A/B) = Pr(A, B)$ solo se $Pr(B) = 1$, ossia se B corrisponde all'unione di tutti gli eventi possibili.

7.1.5 Indipendenza statistica

Si verifica quando

$$Pr(A/B) = Pr(A)$$

in quanto il verificarsi di B non influenza A . Come conseguenza, per due eventi statisticamente indipendenti avviene che

$$Pr(A, B) = Pr(A) Pr(B)$$

Esempi

- Un sistema di comunicazione radio è affetto da attenuazioni supplementari causate da pioggia. Indicando con FS l'evento che il sistema vada fuori servizio, e conoscendo le probabilità condizionate $Pr(FS/piove) = .5$, $Pr(FS/non\ piove) = .05$ e la probabilità marginale $Pr(piove) = .03$, determinare:

1. La probabilità di fuori servizio $Pr(FS)$, indipendentemente dal verificarsi o meno dell'evento piovoso;
2. La probabilità che stia piovendo, sapendo che il sistema è fuori servizio.

Risultato ⁽⁵⁾.

- Quale è la probabilità che, lanciando 3 volte un dado, esca 3 volte 1 ? Risultato ⁽⁶⁾.
- Un'urna contiene 2 biglie bianche e 3 nere. Qual è la probabilità che su 2 estrazioni consecutive, escano le 2 biglie bianche ? Risultato ⁽⁷⁾.
- Qual è la probabilità che 2 carte, estratte a caso da un mazzo da bridge da 52, siano K e Q ? Risultato ⁽⁸⁾.

7.2 Variabili aleatorie

Finora si è parlato di *eventi* in modo astratto, mentre spesso ci si trova ad associare ad ogni punto dello spazio campione un valore numerico: lo spazio campione Ω diventa allora l'*insieme dei numeri* e prende il nome di *variabile aleatoria*. La realizzazione di un evento corrisponde ora all'assegnazione di un valore (tra i possibili) alla variabile aleatoria; tale valore "prescelto" prende dunque il nome di *realizzazione* della v.a. Distinguiamo inoltre tra variabili aleatorie *discrete* e *continue*, a seconda se la grandezza

⁵La probabilità marginale di fuori servizio si calcola applicando il teorema delle probabilità totali:

$$\begin{aligned} Pr(FS) &= Pr(FS/piove) \cdot Pr(piove) + Pr(FS/non\ piove) \cdot Pr(non\ piove) = \\ &= .5 \cdot .03 + .05 \cdot .97 = .0635 = 6.35\% \end{aligned}$$

in quanto $Pr(non\ piove) = 1 - Pr(piove) = .97$. Applicando il teorema di Bayes si trova quindi:

$$Pr(piove/FS) = \frac{Pr(FS/piove) \cdot Pr(piove)}{Pr(FS)} = \frac{.5 \cdot .03}{.0635} = .236 = 23.6\%$$

Si noti come la probabilità *a priori* che piova (3 %) venga rimpiazzata dal suo valore *a posteriori* (23,6 %) grazie alla nuova informazione di cui disponiamo (collegamento fuori servizio). Per una definizione precisa delle probabilità a priori ed a posteriori si veda l'appendice 17.2.1.

⁶E' pari al prodotto delle probabilità marginali, essendo i lanci statisticamente indipendenti, visto che il dado è "senza memoria". Pertanto il risultato è $(\frac{1}{6})^3 = \frac{1}{216} \simeq 4.6296 \cdot 10^{-3}$.

⁷Anche qui l'urna è senza memoria; però dopo la prima estrazione le biglie restano in 4! Pertanto ora il prodotto delle probabilità marginali risulta $\frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$.

⁸ $Pr(K, Q) = Pr(K\ prima, Q\ seconda) + Pr(Q\ prima, K\ seconda) = Pr(K\ prima) \cdot Pr(Q\ seconda/K\ prima) + Pr(Q\ prima) \cdot Pr(K\ seconda/Q\ prima) = 2 \left(\frac{4}{52} \frac{4}{51} \right) = \frac{8}{663} \simeq 1.2 \cdot 10^{-2}$.

che descrivono abbia valori numerabili o continui⁹. La caratterizzazione della variabile aleatoria, in termini probabilistici, si ottiene indicando come la “massa di probabilità” si distribuisce sull’insieme di valori che la variabile aleatoria può assumere, per mezzo delle 2 funzioni di variabile aleatoria seguenti.

7.2.1 Densità di probabilità e funzione di distribuzione

Come come la massa di un oggetto *non omogeneo* è distribuita in modo più o meno denso in regioni differenti del suo volume complessivo, così la *densità di probabilità* (o d.d.p.) indica su quali valori della variabile aleatoria si concentra la probabilità. Ad esempio, la densità della v.a. discreta associata al lancio di un dado può essere scritta:

$$p_D(x) = \sum_{n=1}^6 \frac{1}{6} \delta(x - n)$$

il cui significato discutiamo subito, con l’aiuto dei due grafici seguenti, in cui D e x indicano rispettivamente la v.a. (il numero che uscirà) ed una sua realizzazione (una delle 6 facce). I 6 impulsi centrati in $x = n$ rappresentano una concentrazione di probabilità nei sei possibili valori; l’area di tali impulsi è proprio pari alla probabilità di ognuno dei sei risultati. E’ facile verificare che

$$\int_{-\infty}^{\infty} p_D(x) dx = 1$$

e che risulta

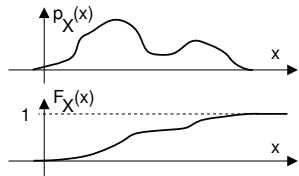
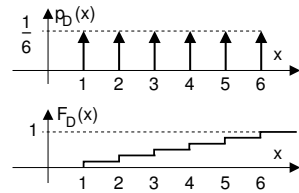
$$\int_a^b p_D(x) dx = Pr \{a < D \leq b\}$$

ovvero pari alla probabilità che la v.a. D assuma un valore tra a e b . In particolare, non potendosi verificare una probabilità negativa, si ha $p_D(x) \geq 0$ con $\forall x$.

Una funzione di v.a. strettamente collegata alla densità è la funzione *distribuzione di probabilità*¹⁰, definita come

$$F_X(x) = \int_{-\infty}^x p_X(\xi) d\xi = Pr \{X \leq x\}$$

che risulta una funzione non decrescente di x , limitata ad un massimo valore di 1, ed il cui andamento mostriamo sotto alla $p_D(x)$, nel caso dell’esempio del lancio del dado¹¹.



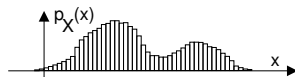
⁹Un esempio classico di v.a. discreta è quello del lancio di un dado, un altro sono i numeri del lotto. Una v.a. continua può essere ad esempio un valore di pressione atmosferica in un luogo, oppure l’attenuazione di una trasmissione radio dovuta a fenomeni atmosferici.

¹⁰In realtà, l’ordine storico è quello di definire prima $F_X(x)$ come la probabilità che X sia non superiore ad un valore x , ovvero $F_X(x) = Pr \{X \leq x\}$, e quindi $p_X(x) = \frac{dF_X(x)}{dx}$. Il motivo di tale “priorità” risiede nel fatto che $F_X(x)$ presenta minori “difficoltà analitiche” di definizione (ad esempio presenta solo discontinuità di prima specie, anche con v.a. discrete).

¹¹Si ricorda che la derivata di un gradino è un impulso di area pari al dislivello. Infatti, sapevamo già che l’integrale di un impulso è una costante - ammesso che l’impulso cada dentro gli estremi di integrazione!

Le definizioni date mantengono validità nel caso di v.a. continua, originando le curve mostrate nei due grafici a lato. Ora è ancora più evidente la circostanza che $p_X(x)$ è una *densità*, e diviene una probabilità solo quando moltiplicata per un intervallo di x ⁽¹²⁾.

Istogramma Evidenziamo ora la stretta relazione che intercorre tra la densità di probabilità e l'*istogramma*. Quest'ultimo può essere realizzato se si dispone di una serie di realizzazioni della v.a., e si ottiene suddividendo il campo di variabilità della grandezza X in sotto-intervalli, e disegnando rettangoli verticali, ognuno di altezza pari al numero di volte che (nell'ambito del campione statistico a disposizione) X assume un valore in quell'intervallo. Dividendo l'altezza di ogni rettangolo per il numero di osservazioni N , si ottiene una approssimazione di $p_X(x)$, via via più precisa con $N \rightarrow \infty$, e con una conseguente riduzione dell'estensione degli intervalli.



7.2.2 Valore atteso, momento e momento centrato

Indichiamo con $g(x)$ una funzione di variabile aleatoria¹³.

Valore atteso Si definisce *valore atteso* (o *media di insieme*¹⁴) di $g(x)$ rispetto alla variabile aleatoria X la quantità:

$$E_X \{g(x)\} = \int_{-\infty}^{\infty} g(x) p_X(x) dx$$

che corrisponde al valor medio di $g(x)$, i cui però i valori assunti ad ogni x sono *pesati*¹⁵ mediante il valore di probabilità $p_X(x) dx$; tale operazione di media integrale è indicata con la notazione $E_X \{ \cdot \}$ ¹⁶, mediante la quale si indica a pedice la v.a. (X) rispetto a cui eseguire la pesatura.

Momento Nel caso in cui $g(x) = x^n$, il valore atteso prende il nome di *momento di ordine n* , che corrisponde quindi al valore atteso della n -esima potenza della v.a., e che si indica come

$$m_X^{(n)} = E \{x^n\} = \int_{-\infty}^{\infty} x^n p_X(x) dx$$

Nel caso di variabili aleatorie discrete, i momenti sono definiti come $m_X^{(n)} = \sum_i x_i^n p_i$, in cui $p_i = Pr \{x = x_i\}$, pesando quindi le possibili realizzazioni x_i con le rispettive probabilità.

¹²Infatti la probabilità che X cada tra x_0 e $x_0 + \Delta x$ vale $\int_{x_0}^{x_0 + \Delta x} p_X(x) dx \simeq p_X(x_0) \Delta x$.

¹³Un esempio di funzione di v.a. potrebbe essere il valore della vincita associata ai 13 in schedina, che dipende dalla v.a. rappresentata dai risultati delle partite, una volta noto il montepremi e le giocate. Infatti, per ogni possibile vettore di risultati, si determina un diverso numero di giocate vincenti, e quindi un diverso modo di suddividere il montepremi. Essendo i risultati improbabili giocati da un ridotto numero di schedine, a queste compete un valore maggiore in caso di vincita, ben superiore al suo *valore atteso*, indicativo invece della vincita media.

¹⁴Per *insieme* ci si riferisce a quello dei possibili valori assunti dalla v.a. X .

¹⁵Notiamo che se al posto delle probabilità $p_X(x) dx$ utilizziamo i valori di un istogramma $Pr(x_i) = \frac{N(x_i < x \leq x_i + \Delta x)}{N} = \frac{N_i}{N}$, l'integrale si trasforma in una sommatoria, il cui sviluppo evidenzia l'equivalenza con una media pesata: $\sum x_i Pr(x_i) = \frac{x_1 N_1 + x_2 N_2 + \dots + x_n N_n}{N}$.

¹⁶In effetti, la E simboleggia la parola *Expectation*, che è il termine inglese usato per indicare il valore atteso.

Media Verifichiamo subito che $m_X^{(0)} = 1$, mentre il momento di *primo* ordine

$$m_X = m_X^{(1)} = \int_{-\infty}^{\infty} x p_X(x) dx$$

prende il nome di *media*¹⁷ della v.a. X (a volte denominata *centroide*), mentre con $n = 2$ si ha la *media quadratica* $m_X^{(2)} = \int_{-\infty}^{\infty} x^2 p_X(x) dx$.

Momento centrato Nel caso in cui $g(x) = (x - m_X)^n$, il relativo valore atteso è chiamato *momento centrato* di ordine n , ed indicato come

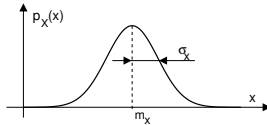
$$\mu_X^{(n)} = E\{(x - m_X)^n\} = \int_{-\infty}^{\infty} (x - m_X)^n p_X(x) dx$$

E' immediato constatare che $\mu_X^{(0)} = 1$ e che $\mu_X^{(1)} = 0$.

Varianza E' il nome dato al momento centrato del 2° ordine, ed è indicata come

$$\sigma_X^2 = \mu_X^{(2)} = E\{(x - m_X)^2\} = \int_{-\infty}^{\infty} (x - m_X)^2 p_X(x) dx$$

La radice quadrata della varianza, σ_X , prende il nome di *deviazione standard*. Mentre la media m_X indica dove si colloca il "centro statistico" della densità di probabilità, σ_X indica quanto le singole determinazioni della v.a. siano disperse attorno ad m_X .

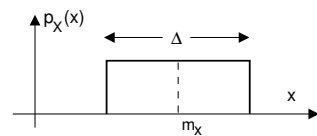


Una relazione notevole che lega i primi due momenti (centrati e non) è ⁽¹⁸⁾:

$$\sigma_X^2 = m_X^{(2)} - (m_X)^2 \quad (7.1)$$

7.2.3 Variabile aleatoria a distribuzione uniforme

Applichiamo la definizione di momento ad un caso pratico: la variabile aleatoria *uniforme* è caratterizzata da uno stesso valore di probabilità per tutta la gamma di realizzazioni possibili, limitate queste ultime ad un unico intervallo non disgiunto; pertanto, la densità di probabilità è esprimibile mediante una funzione rettangolare:



$$p_X(x) = \frac{1}{\Delta} \text{rect}_{\Delta}(x - m_X)$$

in cui Δ rappresenta l'estensione dell'intervallo di esistenza della variabile aleatoria.

¹⁷Supponiamo che X rappresenti l'altezza degli individui; l'altezza *media* sarà allora calcolabile proprio come momento del primo ordine.

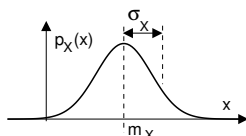
¹⁸Infatti risulta

$$\begin{aligned} \sigma_X^2 &= E\{(x - m_X)^2\} = E\{x^2 + (m_X)^2 - 2xm_X\} = E\{x^2\} + (m_X)^2 - 2m_X E\{x\} = \\ &= m_X^{(2)} + (m_X)^2 - 2(m_X)^2 = m_X^{(2)} - (m_X)^2 \end{aligned}$$

Si è preferito usare la notazione $E\{x\}$, più compatta rispetto all'indicazione degli integrali coinvolti; i passaggi svolti si giustificano ricordando la proprietà distributiva degli integrali (appunto), ed osservando che il valore atteso di una costante è la costante stessa.

E' facile verificare che il parametro m_X , che indica l'ascissa a cui è centrato il rettangolo, corrisponde esattamente al momento di primo ordine di X . Il calcolo della varianza¹⁹ invece fornisce: $\sigma_X^2 = \frac{\Delta^2}{12}$.

7.2.4 Variabile aleatoria gaussiana e funzione $erfc\{\cdot\}$

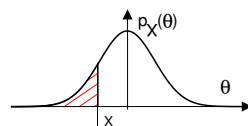


Una variabile aleatoria gaussiana x è descritta da una densità di probabilità di espressione

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}} \quad (7.2)$$

il cui andamento è mostrato in figura, dove è posto in evidenza come m_x e σ_x (media e deviazione standard) siano in relazione la prima con la centratura orizzontale, e la seconda con la dispersione della curva attorno alla media. Oltre che da un punto di vista grafico, i primi due momenti della v.a. descrivono completamente la densità anche dal punto di vista analitico; pertanto, la stima di questi (ad esempio a partire da un buon numero di realizzazioni²⁰) è sufficiente per descrivere completamente il fenomeno aleatorio. La v.a. gaussiana descrive bene una moltitudine di fenomeni naturali, ed è dimostrabile analiticamente che la sua densità è caratteristica²¹ di grandezze ottenute dalla somma di un numero molto elevato di cause aleatorie, tutte statisticamente indipendenti e con la medesima d.d.p.²² (*teorema centrale del limite*²³). La funzione di distribuzione di questa v.a. non è calcolabile in forma chiusa; fortunatamente però, stante la necessità di conoscere la probabilità di eventi gaussiani, sono disponibili tabelle e grafici, che riportano il valore numerico dell'integrale che definisce tali valori di probabilità.

Per renderci conto della situazione, proviamo a calcolare la probabilità che X non superi un certo valore x , pari per definizione alla funzione di distribuzione, e rappresentata dall'area tratteggiata in figura (per semplicità ci riferiamo ad una gaussiana a media nulla):



$$F_X(x) = Pr\{X \leq x\} = \int_{-\infty}^x p_X(\theta) d\theta = 1 - \int_x^{\infty} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{\theta^2}{2\sigma_x^2}} d\theta$$

Effettuiamo ora un cambio di variabile, ponendo $\frac{\theta^2}{2\sigma_x^2} = \eta^2$, per cui in corrispondenza di $\theta = x$ si ha $\eta = \frac{x}{\sqrt{2}\sigma_x}$, e risulta $d\theta = \sqrt{2}\sigma_x d\eta$. Con queste posizioni, possiamo

¹⁹ Anziché calcolare σ_X^2 per la $p_X(x)$ data, calcoliamo $m_X^{(2)}$ per una v.a. uniforme con $m_X = 0$: in tal caso infatti $m_X^{(2)} = \sigma_X^2$. Si ha: $\sigma_X^2 = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} x^2 \frac{1}{\Delta} dx = \frac{x^3}{3\Delta} \Big|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} = \frac{1}{3\Delta} \left(\frac{\Delta^3}{8} + \frac{\Delta^3}{8} \right) = \frac{1}{3\Delta} 2 \frac{\Delta^3}{8} = \frac{\Delta^2}{12}$.

²⁰ Disponendo di un insieme $\{x_n\}$ di N realizzazioni di una variabile aleatoria X , possiamo effettuare le seguenti stime: $\hat{m}_x = \frac{1}{N} \sum_{n=1}^N x_n$ e $\hat{m}_x^{(2)} = \frac{1}{N} \sum_{n=1}^N x_n^2$, il cui valore tende *asintoticamente* a quello delle rispettive medie di insieme, come N (la dimensione del campione statistico) tende a ∞ .

²¹ Il suo scopritore, K.F. Gauss, denominò la v.a. e la sua ddp come *Normale*, indicando con questo il fatto che il suo uso potesse essere "quotidiano".

²² Questa condizione è anche detta di v.a. *indipendenti e identicamente distribuite*, ovvero *i.i.d.*

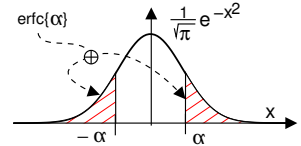
²³ http://it.wikipedia.org/wiki/Teoremi_centrali_del_limite, ovvero sperimentare presso http://local.disia.unifi.it/VL/VL_IT/applets/DiceExperiment.html

riscrivere

$$\begin{aligned}
 F_X(x) &= 1 - \int_{\frac{x}{\sqrt{2}\sigma_x}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\eta^2} \sqrt{2}\sigma_x d\eta \\
 &= 1 - \int_{\frac{x}{\sqrt{2}\sigma_x}}^{\infty} \frac{1}{\sqrt{\pi}} e^{-\eta^2} d\eta = 1 - \frac{1}{2} \operatorname{erfc} \left\{ \frac{x}{\sqrt{2}\sigma_x} \right\}
 \end{aligned}$$

Cosa è successo? Semplicemente, abbiamo espresso l'integrale (irrisolvibile in forma chiusa) nei termini della "funzione" $\operatorname{erfc}\{\cdot\}$, che rappresenta la probabilità che il valore assoluto di una v.a. gaussiana a media nulla e varianza $\frac{1}{2}$ superi il valore dato come argomento, come mostrato in figura, e pari a

$$\operatorname{erfc}\{\alpha\} = 2 \frac{1}{\sqrt{\pi}} \int_{\alpha}^{\infty} e^{-x^2} dx$$



I valori di erfc in funzione del suo argomento sono reperibili sia in forma di tabelle numeriche, sia in forma di diagrammi quotati²⁴.

In linea generale quindi, volendo calcolare la probabilità che una v.a. gaussiana X , con media m_x e varianza σ_x^2 , superi in ampiezza un determinato valore \bar{x} , l'unica strada percorribile è quella di utilizzare la funzione erfc , avendo cura di porre come argomento il valore di \bar{x} debitamente scalato, per ricondursi ad una gaussiana a media nulla e varianza $\frac{1}{2}$:

$$\Pr\{X > \bar{x}\} = \frac{1}{2} \operatorname{erfc} \left\{ \frac{\bar{x} - m_x}{\sqrt{2}\sigma_x} \right\} \quad (7.3)$$

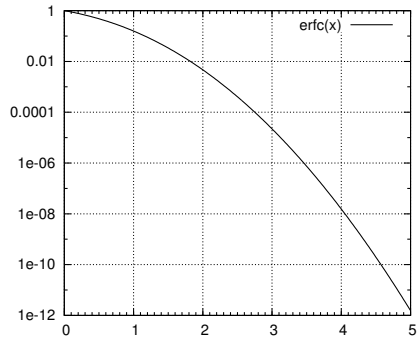


Figura 7.1: Valore di $\operatorname{erfc}\{\alpha\}$ per una gaussiana normalizzata

7.2.5 Altre variabili aleatorie

Altri tipi di variabile aleatoria prendono spesso il nome da quello di uno scienziato, ed in questo testo sono definite congiuntamente ai loro casi di utilizzo, come nel caso di esponenziale § 8.2.1, Bernoulli § 8.1, Poisson § 8.2, Rayleigh e Rice pag. 163.

7.3 Processi stazionari ed ergodici

Dopo aver descritto come caratterizzare statisticamente singoli valori (denominati variabili aleatorie), occupiamoci del caso in cui si voglia descrivere da un punto di vista probabilistico *un intero segnale*, la cui reale identità non sia nota a priori²⁵.

²⁴Il termine erfc sta per *funzione di errore complementare*, e trae origine dai risultati della misura di grandezze fisiche, in cui l'errore di misura, dipendente da cause molteplici, si assume appunto gaussiano.

²⁵Chiaramente, la maggioranza dei segnali trasmessi da apparati di TLC sono di questo tipo.

Un segnale siffatto viene detto *membro* (o realizzazione) di un *processo aleatorio*, e può essere indicato come $x(t, \theta)$, che corrisponde ad una descrizione formale che prevede una coppia di insiemi: il primo di questi è l'insieme \mathcal{T} degli istanti temporali (tipicamente un intervallo) su cui sono definiti i membri del processo; il secondo è relativo ad una variabile aleatoria Θ , i cui valori θ identificano una particolare realizzazione del processo. Pertanto, una singola realizzazione $\theta = \theta_i$, per così dire, *indica* il processo, le cui istanze effettive $x(t, \theta_i)$, con $t \in \mathcal{T}$, sono note solo dopo la conoscenza di $\theta_i \in \Theta$ ⁽²⁶⁾. Il processo aleatorio è quindi definito come l'insieme dei segnali $\{x(t, \theta)\}$, con $t \in \mathcal{T}$ e $\theta \in \Theta$.

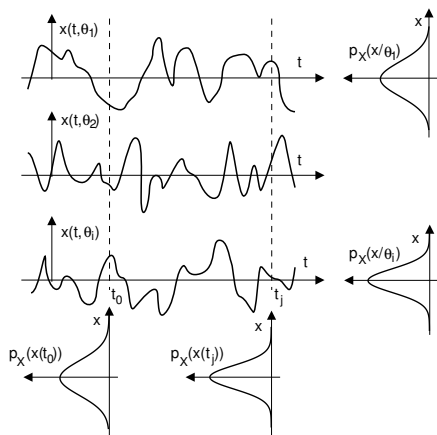


Figura 7.2: Un processo *non ergodico*

Se viceversa fissiamo un particolare istante temporale t_j , il valore $x(t_j, \theta)$ è una variabile aleatoria, la cui realizzazione dipende da quella di $\theta \in \Theta$; pertanto, è definita la densità $p_X(x(t_j))$ (indipendente da θ), che possiamo disegnare in corrispondenza dell'istante t_j in cui è prelevato il campione²⁷; a tale riguardo, si faccia riferimento alla figura precedente, che mostra le densità di probabilità definite a partire dai membri di un processo.

7.3.1 Media di insieme

E' definita come il *valore atteso* di una *potenza n-esima* dei valori del segnale, ossia un suo momento, eseguito rispetto alla variabilità dovuta a Θ , ed è pertanto calcolata come

$$m_X^{(n)}(t_j) = E_{\Theta} \{x^n(t_j, \theta)\} = \int_{-\infty}^{\infty} x^n(t_j, \theta) p_{\Theta}(\theta) d\theta = \int_{-\infty}^{\infty} x^n p_X(x(t_j)) dx$$

in cui l'ultima eguaglianza evidenzia come una media di insieme dipenda dalla d.d.p. $p_X(x(t_j))$ di $x(t_j, \theta)$ al variare di $\theta \in \Theta$, mostrata in basso in fig. 7.2. Notiamo che in linea di principio, la media di insieme dipende dall'istante t_j in cui è prelevato un valore²⁸.

7.3.2 Media temporale

In alternativa, possiamo fissare una particolare realizzazione θ_i di Θ , e quindi identificare un singolo membro $x(t, \theta_i)$, che è ora un segnale certo²⁹; per lo stesso, possono

²⁶Per fissare le idee, conduciamo parallelamente al testo un esempio "reale" in cui il processo aleatorio è costituito da... la selezione musicale svolta da un dj. L'insieme \mathcal{T} sarà allora costituito dall'orario di apertura delle discoteche (dalle 22 all'alba ?), mentre in θ faremo ricadere tutte le caratteristiche di variabilità (umore del dj, i dischi che ha in valigia, la discoteca in cui ci troviamo, il giorno della settimana...).

²⁷Nell'esempio, $x(t_0, \theta)$ è il valore di pressione sonora rilevabile ad un determinato istante (es. le 23.30) al variare di θ (qualunque dj, discoteca, giorno...).

²⁸Ad esempio, se in tutte le serate il volume aumenta progressivamente nel tempo, la $p_X(x(t_j))$ si *allargherà* per t_j crescenti.

²⁹ $x(t, \theta_i)$ rappresenta, nel nostro esempio, l'intera selezione musicale (detta *serata*) proposta da un ben preciso dj, in un preciso locale, un giorno ben preciso.

quindi essere calcolate le *medie temporali*, indicate con una linea sopra alla quantità di cui si calcola la media $\overline{(\cdot)}$:

$$m_X^{(n)}(\theta_i) = \overline{x^n(t, \theta_i)} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^n(t, \theta_i) dt$$

In particolare, ritroviamo il *valore medio*

$$m_X(\theta_i) = \overline{x(t, \theta_i)} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t, \theta_i) dt$$

e la *potenza*³⁰ (o *media quadratica*)

$$m_X^{(2)}(\theta_i) = \overline{x^2(t, \theta_i)} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t, \theta_i) dt$$

Notiamo che una generica media temporale:

- non dipende dal tempo;
- è una variabile aleatoria (dipende infatti dalla realizzazione di Θ).

7.3.3 Media temporale calcolata come media di insieme

L'estrazione da $x(t, \theta_i)$ di un valore ad un istante casuale $t \in \mathcal{T}$, definisce una ulteriore variabile aleatoria, descritta dalla densità di probabilità (condizionata) $p_X(x/\theta_i)$, che disegniamo a fianco dei singoli membri mostrati in fig. 7.2. Qualora la $p_X(x/\theta_i)$ sia nota, le medie temporali di ordine n possono essere calcolate (per quel membro) come i momenti:

$$m_X^{(n)}(\theta_i) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^n(t, \theta_i) dt = \int_{-\infty}^{\infty} x^n p_X(x/\theta_i) dx = E_{X/\Theta=\theta_i} \{x^n\}$$

Ciò equivale infatti ad effettuare una media ponderata, in cui ogni possibile valore di x è pesato per la sua probabilità $p_X(x/\theta_i) dx$.

7.3.4 Processo stazionario

Qualora $p_X(x(t_j))$ non dipenda da t_j , ma risulti $p_X(x(t_j)) = p_X^T(x)$ per qualsiasi $t_j \in \mathcal{T}$, il processo $\{x(t, \theta)\}$ è detto stazionario³¹ *in senso stretto*. In tal caso tutte le medie di insieme non dipendono più dal tempo, ossia $m_X^{(n)}(t) = m_X^{(n)}$ per $\forall t \in \mathcal{T}$, e le $p_X(x(t_j))$ in basso in fig. 7.2 sono tutte uguali.

Se invece solamente le prime due medie di insieme $m_X(t)$ e $m_X^{(2)}(t)$ non dipendono da t , il processo $\{x(t, \theta)\}$ è detto stazionario *in media* ed *in media quadratica*, od anche stazionario *in senso lato*³².

³⁰ $m_X^{(2)}(\theta_i)$ in questo caso rappresenta la potenza media con cui è suonata la musica nella particolare serata θ_i .

³¹ La "serata in discoteca" stazionaria si verifica pertanto se non mutano nel tempo il genere di musica, il volume dell'amplificazione... o meglio se eventuali variazioni in alcune particolari discoteche-realizzazioni sono compensate da variazioni opposte in altrettanti membri del processo.

³² In questo caso la $p_X(x(t))$ non è nota, oppure non è stazionaria, ma le maggiori applicazioni della proprietà di stazionarietà dipendono solo da $m_X(t)$ e $m_X^{(2)}(t)$, che possono essere misurati (o per meglio dire *stimati*), e risultare stazionari anche se $p_X(x(t))$ non lo è.

Supponiamo ora di suddividere il membro $x(t, \theta_i)$ in più intervalli temporali, e di calcolare per ciascuno di essi le medie temporali, limitatamente al relativo intervallo. Nel caso in cui queste risultino uguali tra loro, e di conseguenza uguali alla media temporale $m_X^{(n)}(\theta_i)$, il membro è (individualmente) stazionario³³. Ovviamente, se tutti i membri sono individualmente stazionari, lo è anche il processo a cui appartengono.

7.3.5 Processo stazionario ed ergodico

Questa importante sottoclasse di processi stazionari identifica la circostanza che *ogni membro del processo è statisticamente rappresentativo di tutti gli altri*. Ciò si verifica quando la densità di probabilità (a destra in fig. 7.2) dei valori estratti da un singolo membro $p_X(x/\theta_i)$ è sempre la stessa, indipendentemente dal particolare θ_i , ottenendo in definitiva $p_X(x/\theta_i) = p_X^\Theta(x)$ indipendentemente dalla realizzazione e, per la stazionarietà, anche $p_X(x/\theta_i) = p_X^\mathcal{T}(x)$, e dunque $p_X^\Theta(x) = p_X^\mathcal{T}(x) = p_X(x)$. In questo caso le medie temporali $m_X^{(n)}(\theta_i)$, calcolabili come momenti sulla singola realizzazione come illustrato al §7.3.3, sono identiche per tutti i membri³⁴ θ_i , ed identiche anche alle medie di insieme $m_X^{(n)}(t_j)$ calcolate per un qualunque istante. Enunciamo pertanto la definizione:

Un processo stazionario è ergodico se la media temporale calcolata su di una qualunque realizzazione del processo, coincide con la media di insieme relativa ad una variabile aleatoria estratta ad un istante qualsiasi (per la stazionarietà) da una realizzazione qualsiasi (per l'ergodicità).

Esempio: la potenza di segnale Mostriamo come il calcolo della potenza di un membro di un processo ergodico sia equivalente a quello del momento di 2° ordine del processo:

$$\begin{aligned} \mathcal{P}_X(\theta) &= \overline{x^2(\theta)} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(\tau, \theta) d\tau = \int_{-\infty}^{\infty} x^2 p_X(x/\theta) dx = \\ &= \int_{-\infty}^{\infty} x^2 p_X(x) dx = m_X^{(2)} = E\{x^2\} = \mathcal{P}_X \end{aligned}$$

Questo risultato mostra come sia possibile calcolare la potenza di una realizzazione di un processo, senza conoscerne la forma d'onda.

Potenza, varianza, media quadratica e valore efficace In particolare osserviamo che in base alla (7.1) possiamo scrivere

$$\mathcal{P}_X = m_X^{(2)} = \sigma_x^2 + (m_x)^2$$

e per i segnali a media nulla ($m_x = 0$) si ottiene $\mathcal{P}_X = \sigma_x^2$; in tal caso il valore efficace $\sqrt{\mathcal{P}_X}$ coincide con la deviazione standard σ_x . La radice della potenza è inoltre spesso indicata come *valore RMS* (ROOT MEAN SQUARE), definito come $x_{RMS} = \sqrt{\overline{x^2(t)}}$, ovvero la radice della *media quadratica* (nel tempo). Se il segnale è a media nulla, x_{RMS} coincide quindi con il valore efficace; se $x(t)$ è membro di un processo ergodico a media nulla, x_{RMS} coincide con la deviazione standard.

³³Questo accade se la selezione musicale di una particolare serata si mantiene costante (es. solo raggamuffin) oppure variata ma in modo omogeneo (es. senza tre "lenti" di fila).

³⁴Volendo pertanto giungere alla definizione di una serata *ergodica* in discoteca, dovremmo eliminare quei casi che, anche se individualmente stazionari, sono decisamente "fuori standard" (tutto metal, solo liscio...).

7.3.6 Riassumendo

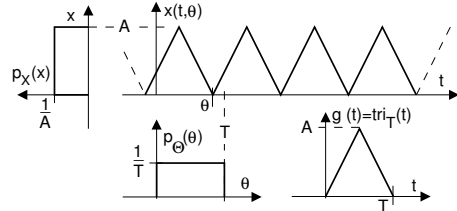
- Se un processo è ergodico, è anche stazionario, ma non il viceversa. Esempio: se $x(t, \theta) = C_\theta$ pari ad una costante (aleatoria), allora è senz'altro stazionario, ma $p_X(x/\theta) = \delta(x - C_\theta)$, e quindi non ergodico.
- Se un processo è ergodico è possibile:
 - calcolare le medie di insieme in forma di medie temporali a partire da una singola realizzazione *oppure*
 - ottenere le medie temporali di una qualunque realizzazione a partire dalle medie di insieme, disponendo della statistica $p_X(x)$.
- Se l'eguaglianza tra medie di insieme e temporali sussiste solo fino ad un determinato ordine e non oltre, il processo *non* è ergodico *in senso stretto*. Per ciò che concerne le Telecomunicazioni, è spesso sufficiente la proprietà di ergodicità *in senso lato*, ovvero limitata al 2° ordine, che garantisce $\overline{x(t)} = E\{x\} = m_x$; $\overline{x^2(t)} = E\{x^2\} = m_x^{(2)}$.

7.3.7 Processo ad aleatorietà parametrica

Questo è il nome dato a processi $\{x(t, \theta)\}$ per quali il parametro θ compare in modo esplicito nella espressione analitica dei segnali membri. Ad esempio, il segnale periodico

$$x(t, \theta) = \sum_{n=-\infty}^{\infty} A \cdot g_T(t - \theta - nT)$$

rappresentato in figura, ha come parametro un ritardo θ , che è una variabile aleatoria che ne rende imprecisata la fase iniziale. Se θ è (come in figura) una v.a. uniformemente distribuita tra 0 e T (ovvero $p_\Theta(\theta) = \frac{1}{T} \text{rect}_T(\theta - \frac{T}{2})$), allora il processo è stazionario ed ergodico, e la sua densità di probabilità per $g(t) = \text{tri}_T(t)$ risulta pari a



$$p_X(x) = \frac{1}{A} \text{rect}_A(x - \frac{A}{2})$$

Il valor medio $m_X = E\{x\}$ è pari alla media temporale $\frac{A}{2}$, la varianza è quella della d.d.p.³⁵ uniforme $\sigma_X^2 = \frac{A^2}{12}$ e la potenza vale

$$P_X = \sigma_X^2 + m_X^2 = \frac{A^2}{12} + \frac{A^2}{4} = \frac{4A^2}{12} = \frac{A^2}{3}$$

Se la $p_\Theta(\theta)$ fosse stata diversa, il processo avrebbe potuto perdere ergodicità. Se ad esempio $p_\Theta(\theta) = \frac{4}{T} \text{rect}_{\frac{T}{4}}(\theta)$, si sarebbe persa la stazionarietà: infatti prendendo ad esempio $-\frac{T}{8} < t < \frac{T}{8}$, tutte le realizzazioni avrebbero valori minori del valor medio $\frac{A}{2}$.

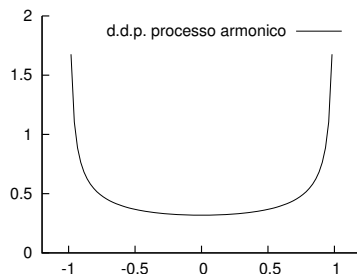
³⁵d.d.p. è l'abbreviazione comunemente usata per Densità di Probabilità.

Processo armonico Si tratta di un processo ad aleatorietà parametrica, i cui membri hanno espressione

$$x(t, \theta) = A \cos(2\pi f_0 t + \theta)$$

dove θ è una v.a. uniforme con d.d.p. $p_{\Theta}(\theta) = \frac{1}{2\pi} \text{rect}_{2\pi}(\theta)$. In tal caso il processo è stazionario ed ergodico, e si ottiene che un valore estratto a caso da un membro qualsiasi è una v.a. con d.d.p.

$$p_X(x) = \frac{1}{\pi\sqrt{A^2 - x^2}} \tag{7.4}$$



mostrata in figura per $A = 1$.

Segnale dati Riprendiamo qui l'espressione (5.1) a cui aggiungiamo un elemento di indeterminazione per quanto riguarda la relazione temporale tra l'origine dei tempi e gli istanti caratteristici, scrivendo

$$x(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT + \theta)$$

con θ v.a. a distribuzione uniforme tra $\pm \frac{T}{2}$, in modo da rendere il processo ergodico. Mentre il calcolo della sua densità di potenza sarà affrontato al § 9.2.4, qui ci limitiamo ad osservare che, considerando i valori a_n come determinazioni di v.a. indipendenti ed identicamente distribuite, la densità di probabilità di $x(t)$ può euristicamente essere desunta dalla analisi del corrispettivo diagramma ad occhio. Ad esempio, nel caso di segnale a banda infinita e a_n a due livelli equiprobabili (vedi fig. 5.2 a pag. 76) la $p_X(x)$ sarà costituita da due impulsi di area $1/2$, mentre nei casi di limitazione in banda e/o adozione di un impulso con caratteristica a coseno rialzato, la stessa assumerà un andamento continuo³⁶.

7.4 Trasformazione di v.a. e cambio di variabili

Quando più v.a. si combinano con leggi diverse dalla somma, il risultato del § 7.6.1 non è più sufficiente a fornire una espressione per la d.d.p. risultante. Illustriamo allora il procedimento analitico generale, necessario ad ottenere una espressione per la d.d.p. di generiche funzioni di v.a.

7.4.1 Caso unidimensionale

Consideriamo una prima v.a. X , ed una seconda Y da essa derivata per mezzo della relazione $y = f(x)$, che si applica alle determinazioni di x di X . La caratterizzazione statistica di Y in termini della sua d.d.p. $p_Y(y)$ può ottenersi a partire da quella di X , nei termini della funzione di distribuzione di Y , come:

$$F_Y(y) = Pr\{Y \leq y\} = Pr\{X \leq f^{-1}(x)\} \tag{7.5}$$

³⁶In una prossima edizione, potrei calcolare le ddp corrispondenti ai diagrammi ad occhio di fig. 5.5

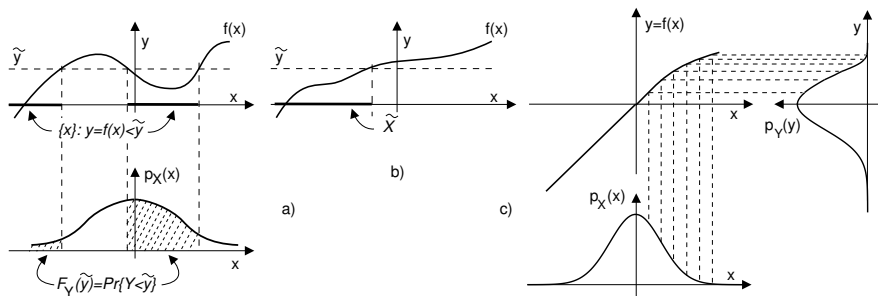


Figura 7.3: Trasformazioni tra variabili aleatorie

e calcolando poi $p_Y(y) = \frac{dF_Y(y)}{dy}$.

L'espressione (7.5) può dare luogo a risultati più o meno usabili a seconda della natura della trasformazione $f(x)$: in fig. 7.3a) troviamo un esempio in cui i valori $y \leq \tilde{y}$ hanno origine da due diversi intervalli di X ; in corrispondenza di questi, l'area sottesa dalla $p_X(x)$ individua la probabilità cercata.

Nel caso in cui $f(x)$ sia monotona crescente come in fig. 7.3 b), per ogni valore di \tilde{y} esiste un solo intervallo di $\tilde{X} \subset X$ tale che $y = f(x)|_{x \in \tilde{X}} \leq \tilde{y}$, e la (7.5) può essere riscritta come

$$F_Y(y) = Pr \{X \leq f^{-1}(y)\} = F_X(x = f^{-1}(y))$$

che, derivata, permette di giungere alla espressione di calcolo della $p_Y(y)$:

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(x)}{dx} \Big|_{x=f^{-1}(y)} \frac{df^{-1}(y)}{dy} = p_X(f^{-1}(y)) \frac{df^{-1}(y)}{dy} \quad (7.6)$$

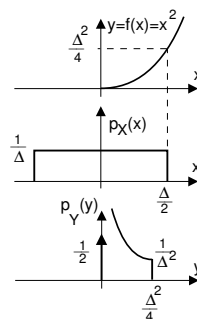
La (7.6) può essere interpretata a parole, osservando che la nuova v.a. $y = f(x)$ possiede una d.d.p. pari a quella di x , calcolata con argomento pari alla funzione inversa $x = f^{-1}(y)$, moltiplicata per la derivata di $f^{-1}(y)$. La d.d.p. risultante da una trasformazione di v.a. si presta inoltre ad un processo di costruzione grafica, come esemplificato nella fig. 7.3 c).

Esempio Determinare $p_Y(y)$, qualora risulti $y = f(x) = \begin{cases} 0 & \text{con } x \leq 0 \\ x^2 & \text{con } x > 0 \end{cases}$, nel caso in cui $p_X(x) = \frac{1}{\Delta} \text{rect}_{\Delta}(x)$.

Osserviamo innanzitutto che tutte le determinazioni $x \leq 0$ danno luogo ad un unico valore $y = 0$; pertanto si ottiene $p_Y(0) = \frac{1}{2} \delta(y)$. Per $0 < y \leq \frac{\Delta^2}{4}$ (corrispondente ad $0 < x \leq \frac{\Delta}{2}$) si applica la teoria svolta, ottenendo $F_Y(y) = Pr \{x \leq \sqrt{y}\} = F_X(\sqrt{y})$, e dunque

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(x)}{dx} \Big|_{x=\sqrt{y}} \frac{d(x = \sqrt{y})}{dy} = \frac{1}{\Delta} \frac{1}{2\sqrt{y}}$$

in cui l'ultima eguaglianza tiene conto che $\frac{dF_X(x)}{dx} = p_X(x)$, che vale $\frac{1}{\Delta}$ per tutti gli x nell'intervallo in considerazione. L'ultima curva mostra la d.d.p. risultante per questo esempio.



7.4.2 Caso multidimensionale

Descriviamo questo caso per mezzo del vettore di v.a. $\mathbf{X} = (x_1, x_2, \dots, x_n)$, a cui è associata una d.d.p. congiunta $p_{\mathbf{X}}(x_1, x_2, \dots, x_n)$, e di un secondo vettore aleatorio \mathbf{Y} dipendente dal primo mediante la trasformazione $\mathbf{Y} = \mathbf{G}(\mathbf{X})$, ovvero

$$\begin{cases} y_1 = g_1(x_1, x_2, \dots, x_n) \\ y_2 = g_2(x_1, x_2, \dots, x_n) \\ \vdots \\ y_n = g_n(x_1, x_2, \dots, x_n) \end{cases}$$

Se esiste la relazione inversa $\mathbf{X} = \mathbf{G}^{-1}(\mathbf{Y})$, composta dall'insieme di funzioni $x_i = g_i^{-1}(y_1, y_2, \dots, y_n)$ per $i = 1, 2, \dots, n$, allora per la d.d.p. di \mathbf{Y} sussiste³⁷ un risultato formalmente molto simile a quello valido nel caso monodimensionale, e cioè

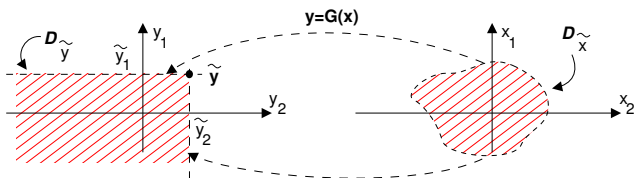
$$p_{\mathbf{Y}}(y_1, y_2, \dots, y_n) = p_{\mathbf{X}}(\mathbf{X} = \mathbf{G}^{-1}(\mathbf{Y})) \cdot |\mathbf{J}(\mathbf{X}/\mathbf{Y})|$$

in cui $p_{\mathbf{X}}(\mathbf{x} = \mathbf{G}^{-1}(\mathbf{y}))$ è la d.d.p. di \mathbf{X} calcolata con argomento dipendente da \mathbf{Y} , e $|\mathbf{J}(\mathbf{X}/\mathbf{Y})|$ è lo *jacobiano* della trasformazione inversa \mathbf{G}^{-1} , ossia il determinante della matrice costituita da tutte le sue derivate parziali:

$$\mathbf{J}(\mathbf{X}/\mathbf{Y}) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \dots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \dots & \dots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

³⁷La dimostrazione segue le medesime linee guida del caso precedente, ed è impostata sulla base della considerazione che la funzione di distribuzione di \mathbf{Y} , calcolata in un generico punto $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n)$, rappresenta la probabilità che \mathbf{Y} appartenga alla regione (dominio) delimitata dal punto $\tilde{\mathbf{y}}$, indicata con $\mathcal{D}_{\tilde{\mathbf{y}}}$:

$$F_{\mathbf{Y}}(\tilde{\mathbf{y}}) = Pr\{\mathbf{Y} \leq \tilde{\mathbf{y}}\} = Pr\{\mathbf{Y} \in \mathcal{D}_{\tilde{\mathbf{y}}}\}$$



Alla stessa regione $\mathcal{D}_{\tilde{\mathbf{y}}}$, ne corrisponde una diversa $\mathcal{D}_{\tilde{\mathbf{x}}}$ nello spazio \mathbf{X} , tale che per ogni valore $\mathbf{x}^\circ \in \mathcal{D}_{\tilde{\mathbf{x}}}$ risulti $\mathbf{y}^\circ = \mathbf{G}(\mathbf{x}^\circ) \in \mathcal{D}_{\tilde{\mathbf{y}}}$. Con queste posizioni, la $F_{\mathbf{Y}}(\tilde{\mathbf{y}}) = Pr\{\mathbf{Y} \in \mathcal{D}_{\tilde{\mathbf{y}}}\}$ si calcola a partire dalla d.d.p. $p_{\mathbf{X}}(\mathbf{x})$, integrata sul dominio $\mathcal{D}_{\tilde{\mathbf{x}}}$:

$$F_{\mathbf{Y}}(\tilde{\mathbf{y}}) = Pr\{\mathbf{X} \in \mathcal{D}_{\tilde{\mathbf{x}}}\} = \int_{\mathcal{D}_{\tilde{\mathbf{x}}}} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Infine, osservando che

$$p_{\mathbf{Y}}(y_1, y_2, \dots, y_n) = \frac{\partial^n F_{\mathbf{Y}}(y_1, y_2, \dots, y_n)}{\partial y_1 \partial y_2 \dots \partial y_n}$$

si ottiene il risultato mostrato.

Variabile aleatoria di Rayleigh Come applicazione della teoria sopra illustrata, consideriamo la relazione che lega il modulo ρ e la fase φ di un numero complesso $z = x + jy$, alle sue parti reale ed immaginaria, assieme alle corrispondenti funzioni inverse:

$$\begin{cases} \rho = \sqrt{x^2 + y^2} \\ \varphi = \arctan \frac{y}{x} \end{cases} \quad \begin{cases} x = \rho \cos \varphi \\ y = \rho \sin \varphi \end{cases} \quad (7.7)$$

Nel caso in cui x ed y siano due v.a. gaussiane indipendenti, a media nulla e uguale varianza σ^2 , la d.d.p. congiunta di (x, y) si ottiene come prodotto delle d.d.p. marginali, e vale

$$p_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (7.8)$$

La $p_{P,\Phi}(\rho, \varphi)$ si ottiene come descritto, valutando³⁸ i due termini $p_{X,Y}(x(\rho, \varphi), y(\rho, \varphi))$ e $|\mathbf{J}(x, y/\rho, \varphi)|$, ed ottenendo quindi

$$p_{P,\Phi}(\rho, \varphi) = \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \quad \text{con} \quad \begin{cases} 0 < \rho < \infty \\ -\pi < \varphi < \pi \end{cases}$$

Le d.d.p. marginali $p_P(\rho)$ e $p_\Phi(\varphi)$ si ottengono saturando³⁹ la d.d.p. congiunta rispetto all'altra variabile, ricavando

$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \quad \text{con} \quad \rho \geq 0; \quad p_\Phi(\varphi) = \frac{1}{2\pi} \quad \text{con} \quad -\pi < \varphi \leq \pi \quad (7.9)$$

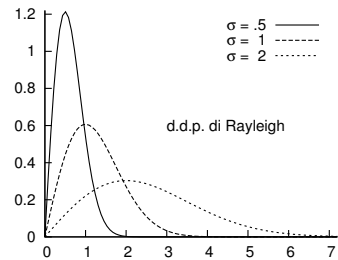
L'espressione di $p_P(\rho)$ in (7.9) prende nome di d.d.p. di RAYLEIGH, graficata appresso, mentre il valor medio e la varianza della v.a. ρ valgono rispettivamente

$$m_P = \sigma \sqrt{\frac{\pi}{2}} \quad \text{e} \quad \sigma_P^2 = \sigma^2 \left(2 - \frac{\pi}{2}\right)$$

E' inoltre possibile mostrare che per essa vale la proprietà

$$Pr\{\rho > \gamma\} = \int_\gamma^\infty p_P(\rho) d\rho = \exp\left(-\frac{\gamma^2}{2\sigma^2}\right) \quad (7.10)$$

Quest'ultimo valore può rappresentare la probabilità di *mancare un bersaglio* per una distanza superiore a γ , nell'ipotesi che gli errori di puntamento orizzontale e verticale siano entrambi gaussiani, indipendenti, a media nulla ed uguale varianza.



³⁸Il calcolo dei due termini si esegue come

$$p_{X,Y}(x(\rho, \varphi), y(\rho, \varphi)) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\rho^2(\cos^2 \varphi + \sin^2 \varphi)}{2\sigma^2}\right) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

$$|\mathbf{J}(x, y/\rho, \varphi)| = \left| \begin{bmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \varphi} \end{bmatrix} \right| = \left| \begin{bmatrix} \cos \varphi & -\rho \sin \varphi \\ \sin \varphi & \rho \cos \varphi \end{bmatrix} \right| = \rho (\cos^2 \varphi + \sin^2 \varphi) = \rho$$

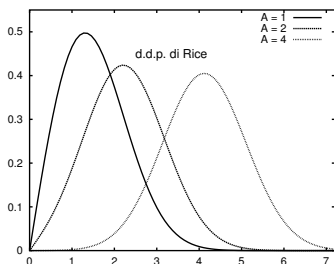
³⁹Svolgiamo il calcolo solo per la prima relazione:

$$p_P(\rho) = \int_{-\pi}^{\pi} p_{P,\Phi}(\rho, \varphi) d\varphi = \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \cdot \int_{-\pi}^{\pi} d\varphi = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

Variabile aleatoria di Rice Consideriamo nuovamente la trasformazione (7.7), in cui si considera, al posto di x , la v.a. x' , somma di una v.a. gaussiana x e di una costante A . In tal caso, la d.d.p. $p_P(\rho)$ è detta di RICE, ed ha⁴⁰ espressione

$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{\rho A}{\sigma^2}\right) \text{ per } \rho \geq 0 \quad (7.11)$$

dove $I_0(z) = \frac{1}{2\pi} \int_0^{2\pi} e^{z \cos \varphi} d\varphi$ è la funzione *modificata* di Bessel del primo tipo ed ordine zero, la cui espressione non ne permette il calcolo in forma chiusa, ma che può



essere approssimata come $I_0(z) \sim e^{\frac{z^2}{4}}$ per $z \ll 1$, e come $I_0(z) \sim \frac{e^z}{\sqrt{2\pi z}}$ per $z \gg 1$.

A lato è mostrato l'andamento di $p_P(\rho)$ con $\sigma = 1$ e tre diversi valori di A , che possiamo rapportare a quelle del secondo grafico per la d.d.p. di Rayleigh, ottenuto per lo stesso valore di σ . Notiamo infine che per $A = 0$ si torna al caso di Rayleigh, mentre per valori crescenti di A , l'andamento della d.d.p. di Rice approssima sempre più quello di una gaussiana.

7.5 Elementi di statistica

La statistica si occupa dell'utilizzo numerico di osservazioni e misure, il cui esito viene pensato come il risultato di un fenomeno aleatorio, ad es. al fine di *inferire*⁴¹ dai dati la plausibilità di ipotesi... Ma le cose *si rendono interessanti* per il fatto che

- potendo ripetere la misura e/o l'osservazione più volte, è possibile ridurre il margine di errore, e migliorare l'affidabilità della decisione;
- non è detto che la d.d.p. del fenomeno in esame sia completamente definita.

Ad esempio, la d.d.p. che produce i valori di osservazione può essere nota, a meno del valore dei suoi parametri media e varianza: compito della statistica è allora anche quello di *stimare* grandezze altrimenti ignote. Senza alcuna ambizione di affrontare compiutamente l'argomento, limitiamoci qui ad esporre brevemente questi particolari campi applicativi.

⁴⁰Osserviamo innanzitutto che d.d.p. congiunta di partenza si scrive ora come

$$p_{X',Y}(x',y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x'-A)^2 + y^2}{2\sigma^2}\right)$$

in quanto x' è una v.a. gaussiana con media A e varianza σ^2 . Sostituendo quindi nell'esponente $x' = \rho \cos \varphi$ e $y = \rho \sin \varphi$, si ottiene

$$(x'-A)^2 + y^2 = \rho^2 \cos^2 \varphi + A^2 - 2\rho A \cos \varphi + \rho^2 \sin^2 \varphi = \rho^2 + A^2 - 2\rho A \cos \varphi$$

Osservando ora che il giacobiano della trasformazione ha un valore pari a ρ anche in questo caso, otteniamo

$$\begin{aligned} p_{P,\Phi}(\rho, \varphi) &= p_{X',Y}(x'(\rho, \varphi), y(\rho, \varphi)) |\mathbf{J}(x', y/\rho, \varphi)| \\ &= \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right) \exp\left(\frac{2\rho A \cos \varphi}{2\sigma^2}\right) \end{aligned}$$

La saturazione di questa d.d.p. congiunta, operata eseguendo $p_P(\rho) = \int_{-\pi}^{\pi} p_{P,\Phi}(\rho, \varphi) d\varphi$, determina il risultato mostrato.

⁴¹Vedi ad es. http://it.wikipedia.org/wiki/Inferenza_statistica

7.5.1 Verifica di ipotesi

Si tratta del tema illustrato al § 7.6.2, ossia di affermare se una certa ipotesi H_1 sia da ritenere valida, e con quali margini probabilistici. La terminologia generale prevede che il valore di probabilità prima indicato come P_{fa} (decidere per H_1 quando è vera H_0) sia detto errore di primo tipo o *livello di significatività* del test, e indicato con α , mentre il valore P_d (decidere correttamente per H_1) è detto *potenza* del test, ed indicato come $1 - \beta$, dove β evidentemente è la probabilità di perdita P_p o *errore di secondo tipo*. Conoscendo la d.d.p. $p_X(x/H_0)$ della grandezza di osservazione X quando è vera H_0 , si calcola una soglia γ tale che $Pr(x > \gamma/H_0) = \int_{\gamma}^{\infty} p_X(x/H_0) = \alpha_{min}$, tale cioè da rendere $P_{fa} = \alpha_{min}$, in genere posto pari a 0.05 (ovvero il 5%).

Qualora l'osservazione di $\tilde{x} > \gamma$ faccia decidere per H_1 , spesso al risultato si aggiunge anche la probabilità $\int_{\tilde{x}}^{\infty} p_X(x/H_0)$ detta *p-value*⁴², che descrive quanto sia *poco probabile* osservare un campione x che si discosti ancor più da H_0 di quanto non sia riuscito il “nostro” campione \tilde{x} , e che perciò indica meglio di α la significatività del test, ovvero la *discrepanza* tra i dati osservati e l'ipotesi nulla H_0 . Naturalmente, per tutti i possibili campioni $x > \gamma$ si osserva un *p-value* $< \alpha_{min}$.

7.5.2 Funzione di verosimiglianza

Viene definita a partire da una probabilità condizionata, qualora nella stessa espressione siano invertiti i ruoli di evento condizionato e condizionante: rimanendo nel campo della verifica di ipotesi, le d.d.p. condizionate $p_X(x/H_0)$ e $p_X(x/H_1)$, qualora valutate entrambe per lo stesso $x = \tilde{x}$ (la effettiva osservazione), possono essere interpretate come il valore di verosimiglianza per le ipotesi H_0 ed H_1 , che quindi ribaltano il ruolo da evento condizionante a variabile indipendente. Ma attenzione, una verosimiglianza *non* è una probabilità, in quanto $p_X(\tilde{x}/H_0) + p_X(\tilde{x}/H_1) \neq 1$! La funzione di verosimiglianza (o *likelihood*) è spesso indicata con la lettera \mathcal{L} , in modo da poter scrivere $\mathcal{L}(H_{0,1}/\tilde{x}) = p_X(\tilde{x}/H_{0,1})$.

7.5.3 Decisione di massima verosimiglianza

Riprendendo il tema della verifica di ipotesi § 7.5.1, osserviamo che qualora la soglia di decisione γ *non* sia assegnata in base al livello di significatività $P_{fa} = \alpha_{min}$ desiderato, ma in corrispondenza della intersezione tra $p_X(x/H_0)$ e $p_X(x/H_1)$, come avviene in fig. 7.4, la decisione stessa è detta di *massima verosimiglianza*, dato che sceglie l'ipotesi H_i per la quale la funzione di verosimiglianza $\mathcal{L}(H_i/\tilde{x})$ è massima. Ma anche in questo caso, alla decisione ottenuta per \tilde{x} è possibile associare il relativo *p-value*. A volte, anziché confrontare tra di loro i valori di \mathcal{L} , se ne esegue il *rapporto*, decidendo quindi in base all'esito maggiore o minore di uno (vedi ad es. pag. 448).

7.5.4 Stima di parametro

Nel caso in cui si conosca il tipo di v.a. che caratterizza un fenomeno aleatorio in esame, ma non i suoi parametri (es. media e varianza), gli stessi valori osservati $\mathbf{x} = (x_1, x_2, \dots, x_N)$ possono essere usati per addivenire ad una loro stima. Indicando con θ il vero (incognito) valore di un parametro da stimare, il suo stimatore $\hat{\theta}(\mathbf{x})$ dipende

⁴²Vedi ad es. <http://en.wikipedia.org/wiki/P-value>

dalla variabilità di \mathbf{x} e quindi è esso stesso una v.a., che può essere caratterizzata nei termini della sua d.d.p. $p_{\hat{\theta}}(\hat{\theta})$, media $m_{\hat{\theta}}$ e varianza $\sigma_{\hat{\theta}}^2$.

Mentre per uno specifico vettore di osservazioni \mathbf{x} lo stimatore commette un errore $\varepsilon(\mathbf{x}) = \hat{\theta}(\mathbf{x}) - \theta$, desidereremmo che il valore atteso di tale errore fosse nullo, ovvero che $E\{\hat{\theta}(\mathbf{x})\} = m_{\hat{\theta}} = \theta$, nei cui caso lo stimatore è detto *non polarizzato*. Per quanto riguarda la varianza $\sigma_{\hat{\theta}}^2 = E\left\{\left(\hat{\theta}(\mathbf{x}) - m_{\hat{\theta}}\right)^2\right\}$, vorremmo che fosse la più piccola possibile. Uno stimatore non polarizzato e che consegue la minima varianza tra i possibili stimatori è detto *efficiente* e soddisfa la disuguaglianza di Rao-Cramér⁴³; se all'aumentare della dimensione N del campione \mathbf{x} , tende ad uno la probabilità che $\hat{\theta}(\mathbf{x}) = \theta$, allora lo stimatore è detto *consistente*. Infine, uno stimatore consistente è anche *asintoticamente normale* se all'aumentare di N la sua d.d.p. tende ad una gaussiana⁴⁴, con valor medio θ e varianza che si riduce come $1/N$.

Al § seguente esaminiamo una tra le diverse tecniche⁴⁵ che consentono di arrivare ad una formula di stima.

7.5.5 Stima di massima verosimiglianza

In questo caso il valore stimato è quello che massimizza la funzione di verosimiglianza (§ 7.5.2) calcolata in corrispondenza ai dati campionari: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\{\mathcal{L}(\theta/\bar{x})\}$. Ad esempio, nel caso in cui i valori osservati $\bar{\mathbf{x}} = (x_1, x_2, \dots, x_N)$ siano ottenuti da un fenomeno aleatorio con d.d.p. gaussiana con media m_x e varianza σ_x^2 , e gli esperimenti siano statisticamente indipendenti, la funzione di verosimiglianza risulta⁴⁶

$$\mathcal{L}(m_x, \sigma_x^2/\bar{\mathbf{x}}) = p_X(\bar{\mathbf{x}}/m_x, \sigma_x^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - m_x}{\sigma_x}\right)^2\right\}$$

ma a questa espressione si preferisce il suo logaritmo naturale $L = \ln \mathcal{L}$, che essendo *monotono* non altera la massimizzazione cercata:

$$L(m_x, \sigma_x^2/\bar{\mathbf{x}}) = \ln \mathcal{L}(m_x, \sigma_x^2/\bar{\mathbf{x}}) = -\frac{n}{2} \ln(2\pi\sigma_x^2) - \frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - m_x}{\sigma_x}\right)^2$$

Per trovare le stime \hat{m}_x e $\hat{\sigma}_x^2$ di massima (log)verosimiglianza, si eguagliano a zero le derivate parziali di L :

$$\begin{aligned} \frac{\partial L}{\partial m_x} &= \frac{1}{\sigma_x^2} \sum_{i=1}^N (x_i - m_x) = 0 \\ \frac{\partial L}{\partial \sigma_x^2} &= -\frac{n}{2} \frac{1}{\sigma_x^2} + \frac{1}{2\sigma_x^4} \sum_{i=1}^N (x_i - m_x)^2 = 0 \end{aligned}$$

⁴³Vedi ad es. http://it.wikipedia.org/wiki/Disuguaglianza_di_Cramér-Rao

⁴⁴Ad esempio, il teorema centrale del limite (§ 7.2.4) fa sì che la media campionaria $\hat{m}_x = \frac{1}{N} \sum_{i=1}^N x_i$, in quanto somma di v.a. indipendenti e identicamente distribuite, tenda ad una gaussiana per $N \rightarrow \infty$.

⁴⁵Vedi ad es. http://en.wikipedia.org/wiki/Point_estimation

⁴⁶Vedi § 9.10.2 per l'espressione di una gaussiana multidimensionale.

ottenendo gli stimatori (o *statistiche*)⁴⁷

$$\begin{cases} \hat{m}_x = \frac{1}{N} \sum_{i=1}^N x_i & \text{media campionaria} \\ \hat{\sigma}_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{m}_x)^2 & \text{varianza campionaria} \end{cases}$$

Verifichiamo ora le loro proprietà. Per quanto riguarda \hat{m}_x , la media campionaria risulta non polarizzata, ovvero⁴⁸ $E\{\hat{m}_x\} = m_x$, mentre la sua varianza risulta⁴⁹ $\sigma_{\hat{m}_x}^2 = \sigma_x^2/N$: pertanto, la media campionaria è uno stimatore consistente e asintoticamente normale. In particolare, la proprietà $\sigma_{\hat{m}_x}^2 = \sigma_x^2/N$ è quella che ci permette di ridurre a piacimento la variabilità della osservazione, ripetendola un adeguato numero di volte.

Per quanto riguarda la stima di varianza $\hat{\sigma}_x^2$, i calcoli per determinare il suo valore atteso $m_{\hat{\sigma}_x^2} = E\left\{\frac{1}{N} \sum_{i=1}^N (x_i - \hat{m}_x)^2\right\}$ passano prima per la dimostrazione⁵⁰ che $\sum_i (x_i - \hat{m}_x)^2 = \sum_i (x_i - m_x)^2 - N(\hat{m}_x - m_x)^2$, in modo da ottenere

$$\begin{aligned} m_{\hat{\sigma}_x^2} &= \frac{1}{N} \sum_{i=1}^N E\{(x_i - m_x)^2\} - \frac{N}{N} E\{(\hat{m}_x - m_x)^2\} = \\ &= \frac{1}{N} N \sigma_x^2 - \frac{\sigma_x^2}{N} = \sigma_x^2 \frac{N-1}{N} \end{aligned}$$

dato che il primo termine corrisponde all'espressione della varianza di X , ed il secondo a quella della varianza di \hat{m}_x , calcolata prima. Il risultato ottenuto indica come la varianza campionaria si riveli uno stimatore *polarizzato*. Per questo, la formula dello stimatore (di massima verosimiglianza) della varianza viene ordinariamente sostituita dalla espressione

$$\hat{\sigma}_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{m}_x)^2$$

⁴⁷Per quanto riguarda \hat{m}_x , imponendo $\sum_i (x_i - \hat{m}_x) = 0$ si perviene facilmente al risultato, mentre per $\hat{\sigma}_x^2$ l'eguaglianza a zero produce $\frac{1}{2\hat{\sigma}_x^4} \sum_i (x_i - \hat{m}_x)^2 = \frac{n}{2} \frac{1}{\hat{\sigma}_x^2} \rightarrow \frac{1}{\hat{\sigma}_x^2} \sum_i (x_i - \hat{m}_x)^2 = N$ e quindi il risultato.

⁴⁸Infatti $E\{\hat{m}_x\} = E\left\{\frac{1}{N} \sum_{i=1}^N x_i\right\} = \frac{1}{N} \sum_{i=1}^N E\{x_i\} = \frac{1}{N} N m_x = m_x$

⁴⁹In questo caso riscriviamo \hat{m}_x come $\hat{m}_x = \sum_{i=1}^N \frac{x_i}{N}$, consideriamo che la varianza di una somma di v.a. i.i.d. è la somma delle varianze (vedi § 9.7.2), e che $\sigma_{aX}^2 = a^2 \sigma_X^2$: pertanto si ottiene $\sigma_{\hat{m}_x}^2 = \sum_{i=1}^N \frac{\sigma_x^2}{N^2} = \frac{\sigma_x^2}{N}$.

⁵⁰Occorre innanzitutto riscrivere $x_i - \hat{m}_x$ come $x_i - m_x + m_x - \hat{m}_x = (x_i - m_x) - (\hat{m}_x - m_x)$, in modo da ottenere $(x_i - \hat{m}_x)^2 = (x_i - m_x)^2 - 2(x_i - m_x)(\hat{m}_x - m_x) + (\hat{m}_x - m_x)^2$. Eseguendo ora la sommatoria su i si ottiene

$$\begin{aligned} \sum_i (x_i - \hat{m}_x)^2 &= \sum_i (x_i - m_x)^2 - 2(\hat{m}_x - m_x) \sum_i (x_i - m_x) + \sum_i (\hat{m}_x - m_x)^2 \\ &= \sum_i (x_i - m_x)^2 - 2N(\hat{m}_x - m_x)^2 + N(\hat{m}_x - m_x)^2 = \sum_i (x_i - m_x)^2 - N(\hat{m}_x - m_x)^2 \end{aligned}$$

in quanto $\sum_i (x_i - m_x) = \sum_i x_i - \sum_i m_x = N\hat{m}_x - Nm_x = N(\hat{m}_x - m_x)$.

7.6 Appendici

7.6.1 Funzione caratteristica

E' definita come l'antitrasformata di Fourier di una densità di probabilità, ovvero (equivalentemente) come il valore atteso di e^{jwz} :

$$\Phi_z(w) = \mathcal{F}^{-1} \{p_Z(z)\} = E_Z \left\{ e^{jwz} \right\} = \int p_Z(z) e^{jwz} dz$$

Osserviamo che, se applicata alla somma di v.a. *indipendenti*, si ha:

$$\begin{aligned} \Phi_z(w) &= E_Z \left\{ e^{jw(x+y)} \right\} = E_Z \left\{ e^{jwx} e^{jwy} \right\} = E_X \left\{ e^{jwx} \right\} E_Y \left\{ e^{jwy} \right\} \\ &= \Phi_x(w) \Phi_y(w) \end{aligned}$$

in quanto se x e y sono statisticamente indipendenti (§ 7.1.5) la loro d.d.p. congiunta si fattorizza, ed il valore atteso si scompone nel prodotto di due integrali. Pertanto, la funzione caratteristica di una somma di v.a. indipendenti, è pari al prodotto delle funzioni caratteristiche.

Effettuando ora l'operazione inversa (trasformata di Fourier della funzione caratteristica della somma) e ricordando che un prodotto in frequenza è una convoluzione nel tempo (e viceversa) si ottiene il risultato $p_Z(z) = \mathcal{F} \{ \Phi_z(w) \} = \mathcal{F} \{ \Phi_x(w) \Phi_y(w) \} = p_X(x) * p_Y(y)$ che ci permette di enunciare:

La densità di probabilità della somma di v.a. indipendenti è pari alla convoluzione tra le rispettive densità di probabilità marginali.

La funzione caratteristica ha altri usi... ma non approfondiamo oltre.

7.6.2 Detezione di sinusoidi nel rumore

Applichiamo i principi della decisione statistica ad un caso "classico", mostrando come i risultati ottenuti al 7.4.2 si applichino allo schema di demodulazione incoerente mostrato al § 11.2.2. Qualora in ingresso al demodulatore sia presente un rumore di potenza σ^2 , pari a quella delle singole componenti analogiche, ed a questo sia sovrapposta una sinusoide di ampiezza A , per il modulo dell'involuppo complesso $\rho = |\underline{x}(t)|$ valgono esattamente le considerazioni svolte per la v.a. di Rice.

La ricezione del rumore dà infatti luogo (vedi § 12.1.2) a due processi di rumore $x_c(t)$ ed $x_s(t)$ gaussiani ed ergodici, che in assenza di sinusoidi hanno media nulla e varianza σ^2 , e quindi la d.d.p. di ρ assume un andamento di Rayleigh, come mostrato alla figura seguente, per il caso indicato con H_0 , e relativo a $\sigma = 1$. La curva H_1 mostra invece la d.d.p. di ρ nel caso di presenza di segnale, con andamento di Rice, per $A = 4$ e $\sigma = 1$.

Il problema nasce qualora il tono sinusoidale non sia presente con continuità, e si desideri operare una decisione relativa all'ipotesi di una sua presenza (indicata con H_1) od assenza (H_0). Nel caso in cui A e σ siano note, il problema è ben posto, e si riduce a determinare una soglia γ con cui confrontare ρ , e decidere per H_0 od H_1 nei casi in cui $\rho < \gamma$ e $\rho > \gamma$ rispettivamente. In tal caso, vengono definiti tre possibili eventi, assieme alle rispettive probabilità, che dipendono dal valore assegnato a γ :

- Probabilità di detezione $P_d = \int_{\gamma}^{\infty} H_1(\rho) d\rho$

- Probabilità di falso allarme $P_{fa} = \int_{\gamma}^{\infty} H_0(\rho) d\rho$
- Probabilità di perdita $P_p = \int_0^{\gamma} H_1(\rho) d\rho$

in cui gli ultimi due valori sono riferiti ad eventi di errore.

La nomenclatura adottata è quella tipica dei radar; in tale contesto, può aver senso tentare di spostare γ in modo da favorire l'uno o l'altro evento in base a considerazioni strategiche⁵¹. Nel caso in cui i due errori siano equivalenti, e se le probabilità a priori di H_0 ed H_1 sono uguali, la probabilità di errore

$$P_e = Pr(H_0) Pr(e/H_0) + Pr(H_1) Pr(e/H_1) = \frac{1}{2}P_{fa} + \frac{1}{2}P_p \quad (7.12)$$

risulta minimizzata qualora si adotti una decisione di *massima verosimiglianza* (vedi § 7.5.3), ponendo γ nel punto in cui le due curve si intersecano (come in figura 7.4), in modo da preferire l'ipotesi H_i per la quale la probabilità condizionata $Pr(\rho/H_i)$ è più grande.

Una valutazione asintotica delle prestazioni può essere svolta notando che all'aumentare di $\frac{A}{\sigma}$, il valore di γ si avvicina (da destra) ad $\frac{A}{2}$; ponendo quindi $\gamma = \frac{A}{2}$ e sostituendo le espressioni (7.9) e (7.11) delle d.d.p. a posteriori in quella (7.12) della P_e , otteniamo

$$P_e = \frac{1}{2} \int_{\frac{A}{2}}^{\infty} \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) d\rho + \frac{1}{2} \int_0^{\frac{A}{2}} \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{\rho A}{\sigma^2}\right) d\rho \quad (7.13)$$

Per ciò che riguarda il primo termine, applicando il risultato (7.10) si trova il valore $\int_{\frac{A}{2}}^{\infty} \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) d\rho = \exp\left(-\frac{A^2}{8\sigma^2}\right)$. Per il secondo termine, osserviamo che il suo valore è ben più piccolo del primo (si veda la figura tracciata per $A = 4$, o le considerazioni riportate in nota⁵²), e quindi può essere trascurato, fornendo in definitiva

$$P_e = \frac{1}{2} \exp\left(-\frac{A^2}{8\sigma^2}\right)$$

⁵¹In ambito militare, può aver senso ridurre la probabilità di perdita, a spese di un aumento di quella di falso allarme, tranne nel caso in cui quest'ultimo non provochi conseguenze del tutto irreversibili, e "sbagliate" in caso di errore. Ragionamenti opposti possono essere svolti in campo medico, in cui si dovrebbe preferire un falso allarme, piuttosto che trascurare l'importanza di un sintomo.

⁵²Come già osservato, la funzione modificata di Bessel può essere approssimata come $I_0\left(\frac{\rho A}{\sigma^2}\right) \sim \frac{\exp\left(\frac{\rho A}{\sigma^2}\right)}{\sqrt{2\pi \frac{\rho A}{\sigma^2}}}$ per $\frac{\rho A}{\sigma^2} \gg 1$, e quindi il secondo termine di (7.13) diviene

$$\begin{aligned} & \frac{1}{2} \int_0^{\frac{A}{2}} \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \exp\left(-\frac{A^2}{2\sigma^2}\right) \exp\left(\frac{\rho A}{\sigma^2}\right) \frac{\sigma}{\sqrt{2\pi \rho A}} d\rho = \\ & = \frac{1}{2} \int_0^{\frac{A}{2}} \sqrt{\frac{\rho}{2\pi \sigma^2 A}} \exp\left(\frac{(\rho - A)^2}{2\sigma^2}\right) d\rho \end{aligned}$$

Notiamo ora che per $\frac{A}{\sigma} \gg 1$ l'integrando è trascurabile tranne che per valori di ρ vicini ad A , cosicché il limite inferiore può essere esteso a $-\infty$, e il termine $\sqrt{\frac{\rho}{2\pi \sigma^2 A}}$ può essere sostituito con $\frac{1}{\sqrt{2\pi}\sigma}$, ottenendo quindi la maggiorazione $\frac{1}{2} \int_{-\infty}^{\frac{A}{2}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{\rho^2 - A^2}{2\sigma^2}\right) d\rho$, che è l'espressione dell'integrale di una gaussiana con media A^2 , e che può essere espresso come $\frac{1}{2} \cdot \frac{1}{2} \operatorname{erfc}\left(\frac{A}{2\sqrt{2}\sigma}\right)$.

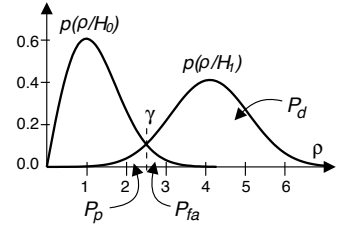


Figura 7.4: Soglia e probabilità di decisione statistica

per $\frac{A}{\sigma} \gg 1$. Notiamo infine che $\frac{A^2}{2}$ rappresenta la potenza della sinusoide, e che σ^2 è la potenza del rumore. Pertanto, il risultato trovato ha una immediata interpretazione in termini di $SNR = \frac{A^2/2}{\sigma^2}$:

$$P_e = \frac{1}{2} \exp\left(-\frac{SNR}{4}\right)$$

Se $z \gg 1$, $\frac{1}{2}erfc(z)$ può essere approssimata come $\frac{1}{2z\sqrt{\pi}} \exp(-z^2)$, ottenendo

$$\frac{1}{2} \int_{-\infty}^{\frac{A}{2}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\rho^2 - A^2}{2\sigma^2}\right) d\rho \sim \frac{\sigma}{A\sqrt{2\pi}} \exp\left(-\frac{A^2}{8\sigma^2}\right)$$

che, essendo per ipotesi $\frac{A}{\sigma} \gg 1$, risulta trascurabile rispetto a $\frac{1}{2} \exp\left(-\frac{A^2}{8\sigma^2}\right)$.

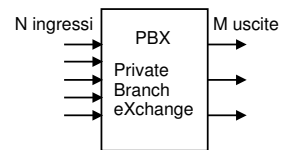
Capitolo 8

Traffico, code e reti a pacchetto

Qui trovano spazio argomenti di *ingegneria delle reti*, a carattere sia teorico che estremamente pratico. Gli aspetti teorici si fondano sui concetti di probabilità affrontati al capitolo 7, e dopo aver caratterizzato in tal senso il traffico informativo, forniscono le metodologie di dimensionamento per collegamenti condivisi da più flussi ai fini del conseguimento di prestazioni individuate come la probabilità di blocco nei sistemi di servizio orientati alla perdita, nei casi di popolazione finita ed infinita, ed il tempo medio di servizio per i sistemi orientati al ritardo, come nel caso di coda infinita e servente unico. Dal punto di vista pratico, sono quindi fornite definizioni e concetti relativi alle reti per trasmissione dati, con particolare riguardo alle reti a pacchetto, nei termini dei diversi modi di funzionamento e delle gerarchie di protocolli adottate, come IP, Ethernet e ATM.

8.1 Distribuzione binomiale per popolazione finita

Iniziamo con il chiederci quante linee uscenti M siano necessarie ad un centralino con N interni, in modo che la probabilità di trovare tutte le linee occupate sia inferiore ad un valore massimo, chiamato *grado di servizio*¹. Per trovare il risultato, calcoliamo prima la probabilità che tutte le linee uscenti siano occupate, assumendo noti N ed M .



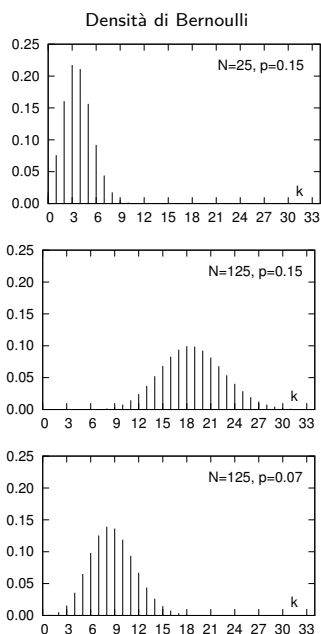
Affrontiamo il problema in termini ancor più generali, chiedendoci quale sia la probabilità $p_B(k)$ che un numero k di persone (su N) sia contemporaneamente al telefono. Assumiamo che ognuno degli N interni abbia una probabilità p di telefonare, ossia passi il $p \cdot 100\%$ del suo tempo al telefono, e che le telefonate siano statisticamente indipendenti. Allora, ci saranno in media Np telefoni occupati, e la probabilità che un ben preciso gruppo di k individui telefoni (e $N - k$ no), è pari a $p^k q^{N-k}$ (in cui $q = 1 - p$).

Dato che il numero di differenti modi di scegliere k oggetti tra N è pari a $\binom{N}{k} = \frac{N!}{k!(N-k)!} = \frac{N(N-1)\cdots(N-k+1)}{k!}$, allora la probabilità di avere k (qualsiasi) persone al

¹Il termine grado di servizio esprime un concetto di *qualità*, ed è usato in contesti diversi per indicare differenti grandezze associate appunto alla qualità dei servizi di telecomunicazione. Nel caso presente, una buona qualità corrisponde a una bassa probabilità di occupato.

telefono è pari a

$$p_B(k) = \binom{N}{k} p^k q^{N-k}$$



Dato che $\sum_{k=0}^N p_B(k) = 1$, la funzione $p_B(k)$ è una densità di probabilità di v.a. discreta, nota con il nome di variabile aleatoria di *Bernoulli*².

Al variare di k , si ottengono tutte le probabilità cercate, rappresentate nella figura a lato nel caso in cui $p = 0.15$ e $N = 25$, oppure $N = 125$; nel secondo caso, si utilizza anche il valore $p = 0.07$, che causa una concentrazione di $p_B(k)$ attorno a valori k inferiori. Inoltre, osserviamo che non si possono avere più di N utenti al telefono.

Per conoscere il numero di linee occorrenti a garantire una probabilità di *congestione* (o di blocco) P_B inferiore ad un massimo, si sommano (partendo *da destra*) i valori di probabilità $p_B(k)$, finché non si supera la probabilità prefissata: allora M sarà pari all'ultimo indice k . Infatti in tal modo la probabilità che ci siano più di M interni a voler telefonare è pari a

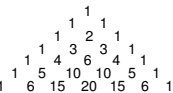
$$Pr(k > M) = \sum_{k=M+1}^N p_B(k) = \sum_{k=M+1}^N \binom{N}{k} p^k q^{N-k} < P_B$$

La distribuzione Binomiale è detta anche *delle prove ripetute* poiché può essere usata per calcolare la probabilità di un certo numero di eventi favorevoli, a seguito della ripetizione dello stesso fenomeno aleatorio³.

Il valore medio della distribuzione Binomiale è $m_B = Np$, e la varianza $\sigma_B^2 = Npq$. Tornando al caso del centralino, il numero medio di linee occupate è Np : tale quantità rappresenta il *traffico offerto medio*, che si misura in ERLANG: ad esempio, un traffico medio di 3 Erlang corrisponde ad osservare in media 3 linee occupate.

Il rapporto $\frac{\sigma_B^2}{m_B} = \frac{Npq}{Np} = q < 1$ è un indice di come la variabile aleatoria si distribuisce attorno alla media. Il caso di Bernoulli in cui $\frac{\sigma_B^2}{m_B} < 1$ è rappresentativo di un traffico *dolce*, che deriva dall'ipotesi di popolazione finita, e che si sostanzia nel

²La $p_B(k)$ è detta anche *Binomiale*, in quanto i fattori $\binom{N}{k}$ sono quelli della potenza di un binomio $(p+q)^N$, calcolabili anche facendo uso del triangolo di *Pascal* (ma definito prima da



Tartaglia, e prima ancora da *Hayyām*), mostrato per riferimento di seguito.

³Esempio: si voglia calcolare la probabilità di osservare 3 volte testa, su 10 lanci di una moneta. Questa risulta pari a $p_B(3) = \binom{10}{3} p^3 q^7 = 120 \cdot .5^3 \cdot .5^7 = 0.117$, ovvero una probabilità dell'11,7%. Come ulteriore esempio, citiamo l'uso della distribuzione Binomiale per calcolare la probabilità di errore complessiva in una trasmissione numerica realizzata mediante un collegamento costituito da N tratte, collegate da ripetitori rigenerativi. In una trasmissione binaria, si ha errore se un numero dispari di tratte causa un errore per lo stesso bit, e cioè $P_e = \sum_{k=1, k \text{ dispari}}^N \binom{N}{k} p^k q^{N-k}$, in cui p è la prob. di errore sul bit per una tratta; inoltre risulta che se $p \ll 1$ e $Np \ll 1$, allora $P_e \approx Np$.

fatto che all'aumentare delle linee occupate, diminuisce la probabilità di una nuova chiamata, in quanto diminuiscono le persone *non* al telefono.

Esercizio Una linea telefonica risulta occupata per l'80 % del tempo, e le telefonate non durano mai più di 5 minuti. Provandola a chiamare con una cadenza fissa di un tentativo ogni 10 minuti, determinare

1. la probabilità di trovare libero *entro* 3 tentativi
2. la probabilità di trovare libero *almeno* una volta in due ore
3. la probabilità di trovare libero *esattamente* tre volte in due ore

Indichiamo con $p = 0.2$ la probabilità di successo di un singolo tentativo, e con $q = 1 - p = 0.8$ quella di fallimento, identificando così il problema nel contesto delle *prove ripetute*.

1. Assumendo gli eventi indipendenti, la prob. di trovare libero entro tre tentativi è la somma delle prob. degli eventi favorevoli, ossia subito libero, oppure al secondo, od al terzo tentativo, ovvero $p + p \cdot q + p \cdot q \cdot q = .2 + .2 \cdot .8 + .2 \cdot .2 \cdot .8 = 0.488 = 48.8$ %.
2. In due ore si effettuano $\frac{120}{10} = 12$ tentativi. Conviene in questo caso valutare la probabilità dell'evento complementare p_0 , quello di fallire tutti i tentativi, pari a $p_B(k)_{k=0}$, ovvero $p_0 = \binom{12}{0} p^0 q^{12} = \frac{12!}{12!} \cdot 8^{12} = 0.0687195$, e quindi la prob. p_1 di libero almeno una volta vale $p_1 = 1 - p_0 = 93.12$ %.
3. Trovare libero esattamente tre volte infine ha probabilità $\binom{12}{3} p^3 q^9 = \frac{12 \cdot 11 \cdot 10}{3 \cdot 2} \cdot .2^3 \cdot .8^9 = 0.23$.

8.2 Distribuzione di Poisson

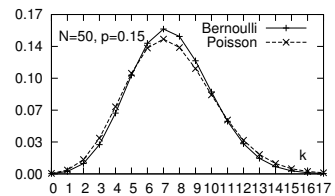
Al crescere del numero N di utenti, l'utilizzo della distribuzione Binomiale può risultare disagiata, per via dei fattoriali, e si preferisce trattare il numero di conversazioni attive k come una variabile aleatoria di POISSON, la cui densità di probabilità ha espressione

$$p_P(k) = e^{-\alpha} \frac{\alpha^k}{k!} \quad (8.1)$$

ed è caratterizzata da valor medio e varianza $m_P = \sigma_P^2 = \alpha$.

La *Poissoniana* costituisce una buona approssimazione della ddp di Bernoulli, adottando per la prima lo stesso valor medio della seconda $m_P = m_B$, ossia $\alpha = Np$, come mostrato in figura. Ma più in generale, questa densità è impiegata per descrivere la probabilità che si verifichino un numero di eventi *indipendenti e completamente casuali* di cui è noto solo il numero medio α ⁽⁴⁾.

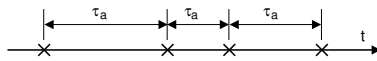
D'altra parte, al tendere di N ad ∞ , il modello Bernoulliano precedentemente adottato perde di validità. Infatti, nel caso di una popolazione infinita, il numero di nuove chiamate *non diminuisce* all'aumentare del numero dei collegamenti in corso. In questo caso, gli eventi corrispondenti all'inizio di una nuova chiamata sono invece considerati *indipendenti e completamente casuali*, e descritti unicamente in base ad



⁴Usando il modello Poissoniano pertanto, la probabilità che (ad esempio) si stiano svolgendo *meno* di 4 conversazioni contemporanee è pari a $p_P(0) + p_P(1) + p_P(2) + p_P(3) = e^{-\alpha} \left(1 + \alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{6}\right)$.

una *frequenza media di interarrivo* λ che rappresenta la velocità⁵ con cui si presentano le nuove chiamate⁶.

L'inverso di λ rappresenta un tempo, ed esattamente $\bar{\tau}_a = 1/\lambda$ è il *valor medio* della variabile aleatoria τ_a costituita dall'intervallo di tempo tra l'arrivo di due chiamate.



Tempi di Interarrivo

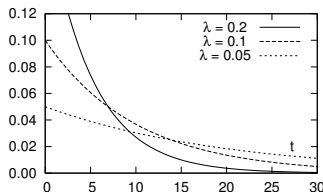
Con queste definizioni, è possibile riferire la v.a. di Poisson ad un intervallo temporale di osservazione T , durante il quale si presentano un numero medio α di chiamate⁷ pari a $\alpha = \lambda T$. Pertanto, possiamo scrivere la ddp della v.a. Poissoniana come

$$p_P(k)|_T = e^{-\lambda T} \frac{(\lambda T)^k}{k!}$$

che indica la probabilità che in un tempo T si verifichino k eventi (indipendenti e completamente casuali) la cui frequenza media è λ (⁸).

8.2.1 Variabile aleatoria esponenziale negativa

La descrizione statistica che la ddp di Poisson fornisce per il numero di eventi che si verificano in un (generico) tempo t , è strettamente legata al considerare questi come *indipendenti, identicamente distribuiti*, e per i quali l'intervallo di tempo tra l'occorrenza degli stessi è una determinazione di variabile aleatoria *completamente casuale*⁹, descritta da una densità di probabilità



Densità di v.a. Esponenziale

⁵ λ viene espresso in *richieste per unità di tempo*.

⁶La trattazione può facilmente applicarsi a svariate circostanze: dalla frequenza con cui si presentano richieste di collegamento ad una rete di comunicazioni, alla frequenza con cui transitano automobili sotto un cavalcavia, alla frequenza con cui particelle subatomiche transitano in un determinato volume, alla frequenza con cui gli studenti si presentano a lezione...

⁷Esempio: se da un cavalcavia osserviamo (mediamente) $\lambda = 3$ auto/minuto, nell'arco di $T = 2$ minuti, transiteranno (in media) $3 \cdot 2 = 6$ autovetture.

⁸Esempio: sapendo che l'autobus (completamente casuale!) che stiamo aspettando ha una frequenza di passaggio (media) di 8 minuti, calcolare: **A**) la probabilità di non vederne nessuno per 15 minuti e **B**) la probabilità che ne passino 2 in 10 minuti.

Soluzione: si ha $\lambda = 1/8$ passaggi/minuto e quindi: **A**) $p_P(0)|_{15} = e^{-\frac{15}{8}} = 0.15$ pari al 15%;

B) $p_P(2)|_{10} = e^{-\frac{10}{8}} \frac{(\frac{10}{8})^2}{2} = 0.224$ pari al 22.4%

⁹Da un punto di vista formale, per eventi *completamente casuali* si intende che gli eventi stessi non hanno memoria di quando siano accaduti l'ultima volta, permettendo quindi di scrivere

$$Pr(t > t_0 + \theta | t > t_0) = Pr(t > \theta)$$

ossia che la probabilità di attendere altri θ istanti, avendone già attesi t_0 , non dipende da t_0 . Per verificare che la ddp esponenziale consente di soddisfare questa condizione, svolgiamo i passaggi, applicando al terzultimo la (8.2):

$$\begin{aligned} Pr(t > t_0 + \theta | t > t_0) &= \frac{Pr(t > t_0 + \theta; t > t_0)}{Pr(t > t_0)} = \frac{Pr(t > t_0 + \theta)}{Pr(t > t_0)} \\ &= \frac{e^{-\lambda(t_0 + \theta)}}{e^{-\lambda t_0}} = e^{-\lambda \theta} = Pr(t > \theta) \end{aligned}$$

esponenziale¹⁰, espressa analiticamente come

$$p_E(t) = \lambda e^{-\lambda t}$$

valida per $t \geq 0$, e mostrata in figura; tale v.a. è caratterizzata dai momenti $m_E = \frac{1}{\lambda}$ e $\sigma_E^2 = \frac{1}{\lambda^2}$. La probabilità che il tempo di attesa di una v.a. esponenziale superi un determinato valore t_0 , è allora calcolabile come

$$Pr(t > t_0) = \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_{t_0}^{\infty} = e^{-\lambda t_0} \quad (8.2)$$

e questo risultato ci permette di verificare il legame con la Poissoniana¹¹.

Esempio Se la durata media di una telefonata è di 5 minuti, e la durata complessiva è completamente casuale, quale è la probabilità che la stessa duri più di 20 minuti?

Risposta: ci viene fornito un tempo di attesa medio τ_a , a cui corrisponde una frequenza di servizio $\lambda = \frac{1}{\tau_a}$, e quindi la soluzione risulta $Pr(t > 20) = \int_{20}^{\infty} \frac{1}{\tau_a} e^{-t/\tau_a} dt = e^{-20/5} = 0.0183 = 1.83\%$.

Un corollario¹² della (8.2) è che, se $t_0 \rightarrow 0$, allora la probabilità che si verifichi un evento entro un tempo t_0 , è *direttamente proporzionale* (a meno di un infinitesimo di ordine superiore di t_0) al valore di t_0 , ossia

$$Pr(t \leq t_0) \Big|_{t_0 \rightarrow 0} = \lambda t_0 + o(t_0) \quad (8.3)$$

8.3 Sistema di servizio orientato alla perdita

Un *sistema di servizio* è una entità in grado di accogliere delle *richieste di servizio*, ovvero eventi che definiscono il cosiddetto *processo di ingresso* al sistema, fino al raggiungimento della capacità limite, determinata dal numero M di *serventi* di cui il sistema dispone¹³.

Una volta occupati tutti i serventi, e finché non se ne libera qualcuno, le successive richieste possono essere poste in coda, individuando così un sistema *orientato al ritardo*

¹⁰La ddp esponenziale è spesso adottata come un modello approssimato ma di facile applicazione per rappresentare un tempo di attesa, ed applicato ad esempio alla durata di una conversazione telefonica, oppure all'intervallo tra due malfunzionamenti di un apparato.

¹¹Consideriamo un ospedale in cui nascono *in media* 6 bimbeti al giorno (o 0.25 nascite l'ora), e consideriamo l'intervallo tra questi eventi come una v.a. completamente casuale. Allora, se assumiamo che la probabilità di k nascite in un tempo T sia descritta da una v.a. di Poisson, ossia a cui compete una probabilità $p_P(k) = e^{-\lambda T} \frac{(\lambda T)^k}{k!}$, allora la probabilità che durante un tempo T non avvenga nessuna nascita, dovrebbe corrispondere a calcolare $p_P(0)$, ovvero $e^{-\lambda T} \frac{(\lambda T)^0}{0!} = e^{-\lambda T}$, che è esattamente il risultato che fornisce la v.a. esponenziale per la probabilità $Pr(t > T)$ che non vi siano nascite per un tempo T .

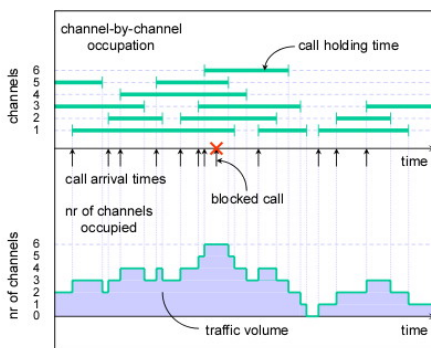
¹²La dimostrazione della (8.3) si basa sulla considerazione che $Pr(t \leq t_0) = 1 - Pr(t > t_0)$, e sulla espansione in serie di potenze $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots$ che si riduce a $e^x = 1 + x + o(t_0)$ se $x \rightarrow 0$. Pertanto, la (8.2) diviene $Pr(t > t_0) \Big|_{t_0 \rightarrow 0} = 1 - \lambda t_0 + o(t_0)$, e quindi $Pr(t \leq t_0) = 1 - 1 + \lambda t_0 + o(t_0) = \lambda t_0 + o(t_0)$.

¹³Gli esempi dalla vita reale sono molteplici, dal casello autostradale presso cui arrivano auto richiedenti il servizio del casellante (M =numero di caselli aperti), al distributore automatico di bevande (servente unico), all'aereo che per atterrare richiede l'uso della pista (servente unico)... nel contesto delle telecomunicazioni, il modello si applica ogni qualvolta vi siano un numero limitato di risorse a disposizione, come ad esempio (ma non solo!) il numero di linee telefoniche uscenti da un organo di commutazione, od il numero di *time-slot* presente in una trama PCM, od il numero di operatori di un *call-center*...

(che affrontiamo al § 8.4), oppure rifiutate (vedi la figura a fianco), come avviene per i sistemi *orientati alla perdita*. Scopo della presente sezione sarà pertanto quello di determinare il numero di serveri necessario a garantire una *probabilità di rifiuto* della richiesta di servizio pari ad un valore che descrive il *grado di servizio* che si intende fornire.

8.3.1 Frequenza di arrivo e di servizio

Mentre il processo di ingresso è descritto in termini della *frequenza media* di arrivo λ , il tempo medio di occupazione dei serveri (indicato come *processo di servizio*) è descritto nei termini del *tempo medio di servizio* τ_S , ovvero dal suo inverso $\mu = 1/\tau_S$, pari alla *frequenza media* di servizio. Nella trattazione seguente si fa l'ipotesi che entrambi i processi (di ingresso e di servizio) siano descrivibili in termini di v.a. a distribuzione esponenziale¹⁴, ovvero che le durate degli eventi “nuova richiesta” e “servente occupato” siano *completamente casuali*¹⁵.



Richieste di servizio e occupazione serveri

8.3.2 Intensità media di traffico

Il rapporto $A_o = \frac{\lambda}{\mu}$ è indicato come *intensità media* del traffico *offerto*¹⁶ e descrive quanti serveri (in media) *sarebbero* occupati ad espletare le richieste arrivate e non ancora servite, nel caso in cui M fosse infinito. L'aggettivo *offerto* indica la circostanza che, essendo invece M finito, alcune richieste non sono accolte, ed A_o risulta diverso dal traffico A_s che può essere effettivamente *smaltito*. L'unità di misura dell'intensità di traffico è l'ERLANG, il cui valore indica appunto il numero medio di serveri occupati.

Esempio Ad un centralino giungono una media di $\lambda = 3$ chiamate al minuto, e la durata media di una conversazione è $1/\mu = 3$ minuti. In tal caso l'intensità media di traffico risulta $A_o = 3 \cdot 3 = 9$ Erlang, corrispondenti al potenziale impegno di una *media* di 9 centralinisti (e nove linee telefoniche).

8.3.3 Probabilità di rifiuto

La teoria che porta a determinare la probabilità che una nuova richiesta di servizio non possa essere accolta a causa dell'esaurimento dei serveri, si basa sull'analisi di un cosiddetto *processo di nascita e morte*, che descrive da un punto di vista statistico

¹⁴L'ipotesi permette di valutare la probabilità che l'intervallo temporale tra due eventi di ingresso sia superiore a θ , in base alla (8.2), come $e^{-\lambda\theta}$ (ad esempio, la prob. che tra due richieste di connessione in ingresso ad una centrale telefonica passi un tempo almeno pari a θ); allo stesso modo, la probabilità che il servizio abbia una durata maggiore di θ è pari a $e^{-\mu\theta}$ (ad esempio, la prob. che una telefonata duri più di θ).

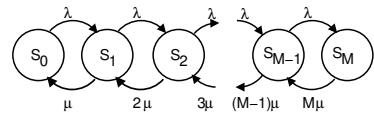
¹⁵Le ipotesi poste fanno sì che i risultati a cui giungeremo siano conservativi, ovvero il numero di serveri risulterà maggiore od uguale a quello realmente necessario; l'altro caso limite (di attese deterministiche) corrisponde a quello in cui il tempo di servizio non varia, ma è costante, come ad esempio il caso del tempo necessario alla trasmissione di una cella ATM di dimensioni fisse. In questi casi, la stessa intensità media di traffico $A_o = \frac{\lambda}{\mu}$ può essere gestita con un numero molto ridotto di serveri; nella realtà, ci si troverà in situazioni intermedie.

¹⁶Si noti che il pedice o è una “o” e non uno “0”, ed identifica appunto l'aggettivo *offerto*.

l'evoluzione di una popolazione, nei termini di una frequenza di nascita (nuova conversazione) e di morte (termine della conversazione). Istante per istante, il numero esatto di individui della popolazione può variare, ma in un istante a caso, possiamo pensare alla numerosità della popolazione come ad una variabile aleatoria discreta, descritta in base ai valori di probabilità p_k che la popolazione assumi esattamente a k individui. La determinazione di questi valori p_k dipende dalla caratterizzazione dei processi di ingresso e di servizio, e nel caso in cui questi siano descritti da v.a. esponenziali (o poissoniane, a seconda se ci riferiamo ai tempi medi di interarrivo/partenza, od al loro numero medio per unità di tempo) si può procedere nel modo che segue.

Descriviamo innanzitutto l'evoluzione dello stato del sistema, in cui il numero di *serventi occupati* evolve aumentando o diminuendo di una unità alla volta (come per i processi di nascita e morte), con l'ausilio della figura seguente, dove il generico stato S_k rappresenta la circostanza che k serventi siano occupati, circostanza a cui compete una probabilità $p_k = Pr(S_k)$.

Gli stati del grafo sono collegati da archi etichettati con la frequenza λ delle transizioni tra gli stati, ovvero dal ritmo con cui si passa da S_k a S_{k+1} a causa di una nuova richiesta, indipendente (per ipotesi) dal numero di serventi già occupati, e dal ritmo $(k+1) \cdot \mu$ con cui si torna da S_{k+1} ad S_k , a causa del termine del servizio espletato da uno tra i $k+1$ serventi occupati, e proporzionale quindi a questo numero¹⁷. Se λ e μ non variano nel tempo, esaurito un transitorio iniziale, il sistema di servizio si troverà in *condizioni stazionarie*, permettendoci di scrivere le *equazioni di equilibrio statistico*



$$\lambda p_k = \mu (k+1) p_{k+1} \quad \text{con } k = 0, 1, 2, \dots, M-1 \quad (8.4)$$

che eguagliano la frequenza media con cui il sistema evolve dallo stato k verso $k+1$, alla frequenza media con cui avviene la transizione inversa¹⁸. La (8.4) può essere riscritta come $p_{k+1} = \frac{\lambda}{\mu(k+1)} p_k = \frac{A_0}{(k+1)} p_k$, che applicata ricorsivamente, porta a scrivere

$$p_k = \frac{A_0^k}{k!} p_0 \quad (8.5)$$

Non resta ora che trovare il modo per dare un valore a p_0 , e questo è oltremodo semplice, ricordando che deve risultare¹⁹ $1 = \sum_{m=0}^M p_m = p_0 \sum_{m=0}^M \frac{A_0^m}{m!}$, e quindi

$$p_0 = \left(\sum_{m=0}^M \frac{A_0^m}{m!} \right)^{-1} \quad (8.6)$$

¹⁷Pensiamo ad un ufficio postale visto dall'esterno: la frequenza media λ con cui entrano nuove persone non dipende da quanti siano già all'interno, mentre invece la frequenza con la quale escono dipende sia dal tempo medio $1/\mu$ di permanenza allo sportello, che dal numero di sportelli (serventi) M in funzione. La differenza con il caso che stiamo trattando, scaturisce dal fatto che l'ufficio postale è un sistema a coda, e dato che la coda c'è *praticamente sempre* (ossia i serventi sono generalmente tutti occupati) possiamo dire che la frequenza media di uscita è proprio $M\mu$.

¹⁸E' un po' come se il numero medio di nuove richieste per unità di tempo λ si distribuisse, in accordo alle probabilità p_k , tra tutti gli stati possibili del sistema: come dire che del totale di λ , una parte λp_0 trovano il sistema vuoto, una parte λp_1 con un solo occupante, eccetera. Per quanto riguarda le richieste servite per unità di tempo, la frequenza di uscita dal sistema è quella che si otterrebbe con un unico servente, moltiplicata per il numero di serventi occupati. Dato che questa ultima quantità è una grandezza probabilistica, la reale frequenza di uscita μ_r può essere valutata come valore atteso, ossia $\mu_r = \sum_{k=1}^M \mu \cdot k \cdot p_k$

¹⁹Usiamo il pedice m anziché k per non creare confusione nella (8.7)

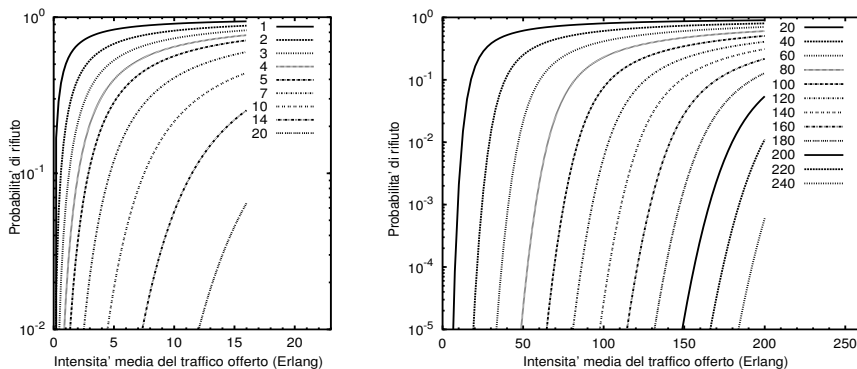


Figura 8.1: Andamento della probabilità di blocco P_B in un sistema orientato alla perdita, al variare di A_o , per il numero di server indicato sulle curve

Nei due casi distinti in cui i server siano in numero finito (e pari ad M) od infinito ($M = \infty$) otteniamo rispettivamente il caso cercato, ed un caso limite. Se poniamo $M = \infty$, tenendo conto dell'espansione in serie $\sum_{m=0}^{\infty} \frac{A_o^m}{m!} = e^{A_o}$, si ottiene che la (8.6) fornisce appunto $p_0 = e^{-A_o}$, e la (8.5) diviene $p_k = e^{-A_o} \frac{A_o^k}{k!}$, che come riconosciamo immediatamente è proprio la ddp di Poisson (8.1) con valore medio A_o ²⁰. Se invece poniamo M finito, la sommatoria che compare in (8.6) non corrisponde ad una serie nota, e dunque rimane come è, fornendo il risultato

$$p_k = Pr(S_k) = \frac{\frac{A_o^k}{k!}}{\sum_{m=0}^M \frac{A_o^m}{m!}}$$

Notiamo ora che p_M è la probabilità che tutti i server siano occupati, pari dunque alla probabilità che una nuova richiesta di servizio sia rifiutata. Chiamiamo allora questo valore *Probabilità di Blocco, di Rifiuto o di Perdita*, la cui espressione prende il nome di FORMULA B DI ERLANG, del primo tipo, di ordine M ed argomento A_o :

$$P_B = Pr(S_M) = p_M = \frac{\frac{A_o^M}{M!}}{\sum_{m=0}^M \frac{A_o^m}{m!}} = E_{1,M}(A_o) \quad (8.7)$$

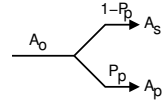
L'andamento di P_B in funzione di M e di A_o è graficato in Fig. 8.1, e mostra come (ad esempio) per una intensità di traffico offerto pari a 40 Erlang, siano necessari più di 50 server per mantenere una P_B minore dell'1%, che salgono a più di 60 per una $P_B = 10^{-3}$.

8.3.4 Efficienza di giunzione

In presenza di una intensità media di traffico offerto A_o , ed una probabilità di perdita $P_p = P_B$, solamente il $(1 - P_p) \cdot 100$ % delle richieste è smaltito, e quindi

²⁰Questo risultato è in perfetto accordo con le la (8.1), quando abbiamo sostituito alla ddp di Bernoulli quella di Poisson, mantenendo inalterato il numero medio di server occupati, che ora indichiamo con A_o , come definito al § 8.3.2.

A_o si ripartisce tra l'intensità media di *traffico smaltito* $A_s = A_o(1 - P_p)$, e l'intensità media di *traffico perso* $A_p = A_oP_p$. Possiamo definire un coefficiente di utilizzazione, o efficienza



$$\rho = \frac{A_s}{M} = \frac{A_o}{M} (1 - P_p)$$

che rappresenta la percentuale di impegno dei server, e di cui la figura 8.2 mostra l'andamento al variare di A_o , per una P_B assegnata e pari a $2 \cdot 10^{-3}$, assieme al numero di server necessario a garantire tale probabilità di blocco.

Come si può osservare, una volta fissato il grado di servizio, all'aumentare del numero di server il traffico smaltito cresce più in fretta di quanto non crescano i server²¹, cosicché (a parità di P_p) l'efficienza aumenta con l'intensità di traffico offerto, e per questo i collegamenti (*giunzioni*) in grado di smaltire un numero più elevato di connessioni, garantiscono anche una maggiore economicità di esercizio.

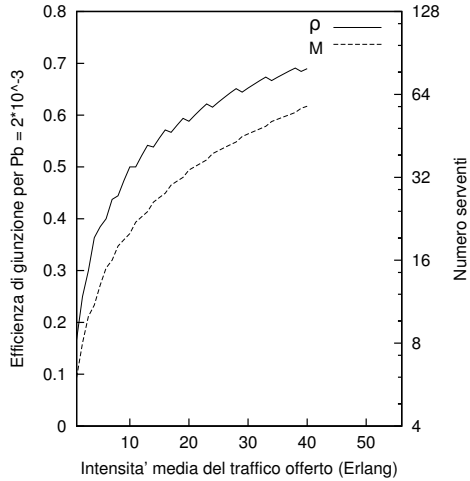


Figura 8.2: Efficienza di giunzione

8.3.5 Validità del modello

Le considerazioni esposte si riferiscono ad una ipotesi di traffico completamente casuale con tempi di interarrivo e di servizio esponenziali²², ossia con un processo di traffico incidente di Poisson. In queste ipotesi, il rapporto $\frac{\sigma_P^2}{m_P} = 1$ tra la varianza e la media delle distribuzioni di Poisson, è rappresentativo appunto di un traffico *completamente casuale*.

Del tutto diversa può risultare l'analisi, nel caso di una giunzione usata solo nel caso di trabocco del traffico da una giunzione piena. In questo caso λ non è più costante, anzi aumenta con l'aumentare delle connessioni già avvenute, tipico di *traffico a valanga*²³.

Esempio Un numero molto elevato di sorgenti analogiche condivide uno stesso mezzo trasmissivo, caratterizzato da una capacità complessiva netta di 25.6 Mbps. Le sorgenti sono campionate a frequenza $f_c = 21.33$ KHz e con una risoluzione di 12 bit/campione; ogni sorgente trasmette ad istanti casuali per un tempo casuale, quindi gli intervalli di interarrivo e di servizio sono entrambi v.a. a distribuzione esponenziale negativa, di valor medio rispettivamente $\lambda = 20$ richieste/minuto e $\frac{1}{\mu} = 4.25$ minuti.

1. Determinare la f_b di una sorgente nelle fasi di attività;

²¹ovvero, all'aumentare del traffico offerto, M aumenta più lentamente di A_o . Ad esempio, dalla figura si può verificare che se per $A_o = 10$ occorrono circa 21 server, per una intensità doppia $A_o = 20$ il numero di server necessario a mantenere la stessa P_B risulta poco più di 32.

²²In effetti, è stato dimostrato che i risultati ottenuti per i sistemi di servizio orientati alla perdita possono essere considerati validi anche nel caso di tempi di servizio a distribuzione qualsiasi, non necessariamente esponenziale.

²³Un esempio di tale tipo di traffico potrebbe essere... l'uscita da uno stadio (o da un cinema, una metropolitana,...) in cui il flusso di individui non è casuale, ma aumenta fino a saturare le vie di uscita.

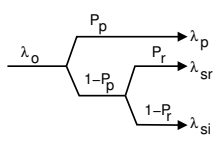
- determinare il numero massimo di sorgenti contemporaneamente attive;
- determinare il grado di servizio (Probabilità di rifiuto) ottenibile con il mezzo trasmissivo indicato;
- indicare la capacità da aggiungere al collegamento per garantire un grado di servizio cento volte migliore.

Risposte

- $f_b = \frac{\text{bit}}{\text{campione}} \cdot \frac{\text{campioni}}{\text{secondo}} = 12 \cdot 21.33 \cdot 10^3 = 256 \text{ Kbps}$;
- Il numero massimo di sorgenti contemporaneamente attive coincide con il numero di serventi M del collegamento, e quindi $M = \frac{25.6 \cdot 10^6}{256 \cdot 10^3} = 100$ serventi;
- L'intensità media di traffico offerto risulta pari a $A_o = \frac{\lambda}{\mu} = \frac{20}{1/4.25} = 85$ Erlang, e pertanto dalle curve di Fig. 8.1 si trova una probabilità di rifiuto pari a circa 10^{-2} ;
- Si richiede quindi una probabilità di rifiuto 100 volte inferiore, e cioè pari a 10^{-4} : si ottiene che la banda deve essere aumentata del 20%. Infatti, dalle curve di Fig. 8.1 si osserva che ciò richiede (a parità di A_o) almeno 120 (circa) serventi, 20 in più, pari ad una capacità aggiuntiva di $20 \cdot 256 \cdot 10^3 = 5.12 \text{ Mbps}$.

8.4 Sistemi di servizio orientati al ritardo

Mentre i sistemi orientati alla perdita rappresentano il modo di operare delle reti di telecomunicazione a *commutazione di circuito*, in cui ogni connessione impegna in modo esclusivo alcune risorse di rete, che una volta esaurite producono un *rifiuto* della richiesta di connessione, i sistemi *orientati al ritardo* sono rappresentativi di reti a *commutazione di pacchetto*, in cui i messaggi sono suddivisi in unità elementari (detti pacchetti, appunto) la cui ricezione non deve più avvenire in tempo reale, e che condividono le stesse risorse fisiche (degli organi di commutazione e di trasmissione) con i pacchetti di altre comunicazioni. Pertanto, l'invio di un pacchetto può essere *ritardato* se il sistema di servizio è in grado di gestire delle *code di attesa*, in cui accumulare le richieste che eccedono il numero di serventi a disposizione, e da cui prelevare (con ritardo) i pacchetti stessi non appena si rendano disponibili le risorse trasmissive necessarie.



In questo caso, il grafico che mostra la ripartizione dei flussi di richieste si modifica come in figura, dove è evidenziato come la frequenza di richieste λ_o si suddivide tra la frequenza delle richieste perse λ_p , quelle servite con ritardo λ_{sr} , e quelle servite immediatamente λ_{si} , in funzione della probabilità di perdita P_p

e di ritardo P_r . Nei termini di queste quantità, valgono le relazioni:

$$\lambda_p = P_p \lambda_o; \quad \lambda_{sr} = P_r (1 - P_p) \lambda_o; \quad \lambda_{si} = (1 - P_r) (1 - P_p) \lambda_o$$

Indicando con $\tau_S = \frac{1}{\mu}$ il tempo medio di servizio di ogni richiesta, (che non comprende quindi il tempo di accodamento), si definisce, come già noto, una intensità di traffico offerto $A_o = \frac{\lambda_o}{\mu} = \lambda_o \tau_S$, che deve risultare

$$A_o = A_p + A_{sr} + A_{si} \quad \text{e quindi} \quad A_{sr} = \frac{\lambda_{sr}}{\mu}, \quad A_{si} = \frac{\lambda_{si}}{\mu}$$

Considerando il caso in cui la coda abbia una lunghezza finita e pari ad L , osserviamo che, a prima vista, anche le L richieste successive all'impegno di tutti gli M serventi sono accolte (e poste in coda), come se i serventi fossero divenuti $M + L$. In realtà

l'analisi fornisce risultati differenti, in quanto le richieste accodate devono essere ancora servite, e quindi il calcolo della P_p non è una diretta estensione dei risultati ottenuti per i sistemi orientati alla perdita. E' comunque abbastanza semplice verificare²⁴ che ora la P_p risulta inferiore alla P_B del caso senza coda, e pertanto l'intensità di traffico smaltito $A_s = A_{sr} + A_{si} = (1 - P_p) A_o$ aumenta, a parità di offerta.

8.4.1 Risultato di Little

Si tratta di un risultato molto generale, valido per qualsiasi distribuzione dei tempi di interarrivo e di servizio, la cui applicazione può tornare utile nell'analisi, e che recita:

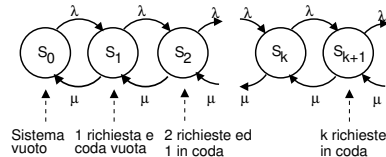
Il numero medio \bar{N} di utenti contemporaneamente presenti in un sistema di servizio è pari al prodotto tra frequenza media di smaltimento delle richieste λ_s ed il tempo medio di permanenza τ_p dell'utente nel sistema

e quindi in definitiva $\bar{N} = \lambda_s \cdot \tau_p$. Nell'applicazione al caso di servizi orientati alla perdita, si ha $\tau_p = \tau_S$, mentre nei servizi a coda risulta $\tau_p = \tau_c + \tau_S$ in cui τ_c rappresenta il tempo medio di coda.

8.4.2 Sistemi a coda infinita ed a servente unico

Prima di fornire risultati più generali, svolgiamo l'analisi per questo caso particolare, in cui la frequenza di richieste perse λ_p è nulla, dato che una coda di lunghezza infinita le accoglie comunque tutte.

Il sistema è descritto da un punto di vista statistico mediante il diagramma di nascita e morte riportato a fianco, in cui ogni stato S_k rappresenta k richieste nel sistema, di cui una sta ricevendo servizio e $k - 1$ sono accodate.



Per procedere nell'analisi, si applica lo stesso principio di equilibrio statistico già adottato a pag. 177, il quale asserisce che, esaurito un periodo transitorio iniziale, la frequenza media delle transizioni tra S_k e S_{k+1} deve eguagliare quella da S_{k+1} ad S_k . Indicando con $p_k = Pr(S_k)$ la probabilità che il sistema contenga k richieste, l'equilibrio statistico si traduce nell'insieme di equazioni

$$\lambda_o p_k = \mu p_{k+1} \quad \text{con } k = 0, 1, 2, \dots, \infty \tag{8.8}$$

Infatti, in base alle stesse considerazioni svolte nella prima parte della nota 18 di pag. 177, $\lambda_o p_k$ è pari alla frequenza media (frazione di λ_o) con cui il numero di richieste accolte passa da k a $k + 1$; essendo il servente unico, la frequenza di servizio è sempre $\mu = \frac{1}{\tau_S}$, indipendentemente dal numero di richieste accodate, e dunque μp_{k+1} è proprio la frequenza media con cui il sistema passa da $k + 1$ a k richieste accolte.

La relazione (8.8) è di natura ricorsiva, e può esprimersi come

$$p_k = \left(\frac{\lambda_o}{\mu} \right)^k p_0 = A_o^k p_0$$

²⁴Se P_B è la probabilità di blocco derivante dalla disponibilità di M serventi, una frequenza di richieste pari a $P_B \cdot \lambda_o$ non può essere servita immediatamente; adottando una coda, la frequenza delle richieste non servite immediatamente $P_B \cdot \lambda_o$ è uguale a $\lambda_o (P_p + P_r (1 - P_p))$, ed eguagliando le due espressioni si ottiene $P_p = \frac{P_B - P_r}{1 - P_r}$, che è sempre minore di P_B .

Per determinare il valore $p_0 = Pr(S_0)$, uguale alla probabilità che il sistema sia vuoto, ricordiamo²⁵ che deve risultare

$$1 = \sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} p_0 A_o^k = p_0 \frac{1}{1 - A_o}$$

da cui otteniamo $p_0 = 1 - A_o$ e dunque

$$p_k = (1 - A_o) A_o^k$$

che corrisponde ad una densità di probabilità esponenziale discreta.

Siamo ora in grado di determinare alcune grandezze di interesse:

Probabilità di ritardo P_r : risulta pari alla probabilità che il sistema non sia vuoto, e cioè che ci sia già almeno una richiesta accolta, ed è pari a²⁶

$$P_r = 1 - p_0 = 1 - (1 - A_o) = A_o$$

Ricordiamo di aver già definito l'efficienza come il rapporto $\rho = \frac{A_s}{M}$ tra il traffico smaltito ed il numero dei serventi; nel nostro caso $M = 1$ e $A_s = A_o$: dunque $\rho = A_o$. Pertanto, il risultato $P_r = A_o = \rho$ indica come, al tendere ad 1 dell'efficienza, la probabilità di ritardo tenda anch'essa ad 1.

Lunghezza media di coda indicata con \bar{L} : risulta essere semplicemente il valore atteso del numero di richieste presenti nel sistema, ovvero²⁷

$$\bar{L} = E\{k\} = \sum_{k=0}^{\infty} k p_k = (1 - A_o) \sum_{k=0}^{\infty} k A_o^k = \frac{A_o}{1 - A_o}$$

da cui risulta che per $A_o \rightarrow 1$ la coda tende ad una lunghezza infinita.

Tempo medio di permanenza indicato con τ_p , e scomponibile nella somma $\tau_p = \tau_s + \tau_c$ tra il tempo medio di servizio ed il tempo medio di coda. Possiamo applicare qui il risultato di Little $\bar{N} = \lambda_s \cdot \tau_p$, che esprime la relazione tra numero medio \bar{N} di richieste presenti, frequenza di smaltimento (qui pari a quella di offerta²⁸), e tempo medio di permanenza; infatti accade che $\bar{N} = \bar{L}$, ed utilizzando il risultato $\bar{L} = \frac{A_o}{1 - A_o}$ si ottiene

$$\tau_p = \frac{\bar{N}}{\lambda_s} = \frac{\bar{L}}{\lambda_o} = \frac{A_o}{1 - A_o} \frac{1}{\lambda_o} = \frac{\lambda_o}{\mu} \frac{1}{\lambda_o} \frac{1}{1 - \lambda_o/\mu} = \frac{1}{\mu - \lambda_o}$$

da cui si osserva che, se la frequenza di offerta tende al valore della frequenza di servizio, il tempo medio di permanenza tende ad ∞ .

²⁵Nella derivazione del risultato si fa uso della relazione $\sum_{k=0}^{\infty} \alpha^k = \frac{1}{1 - \alpha}$, nota con il nome di *serie geometrica*, e valida se $\alpha < 1$, come infatti risulta nel nostro caso, in quanto necessariamente deve risultare $A_o = \frac{\lambda_o}{\mu} < 1$; se il servente è unico infatti, una frequenza di arrivo maggiore di quella di servizio preclude ogni speranza di funzionamento, dato che evidentemente il sistema non ha modo di smaltire in tempo le richieste che si presentano.

²⁶Ricordiamo che p_0 è la probabilità che il sistema sia vuoto, e dunque $1 - p_0$ quella che *non* sia vuoto.

²⁷si fa uso della relazione $\sum_{k=0}^{\infty} k \alpha^k = \alpha \sum_{k=0}^{\infty} k \alpha^{k-1} = \alpha \frac{\partial}{\partial \alpha} \sum_{k=0}^{\infty} \alpha^k = \alpha \frac{\partial}{\partial \alpha} \frac{1}{1 - \alpha} = \frac{\alpha}{(1 - \alpha)^2}$

²⁸Non può essere $\lambda_s > \lambda_o$, perché si servirebbero più richieste di quante se ne presentano. Se fosse invece $\lambda_s < \lambda_o$, la coda crescerebbe inesorabilmente e sarebbe quindi inutile.

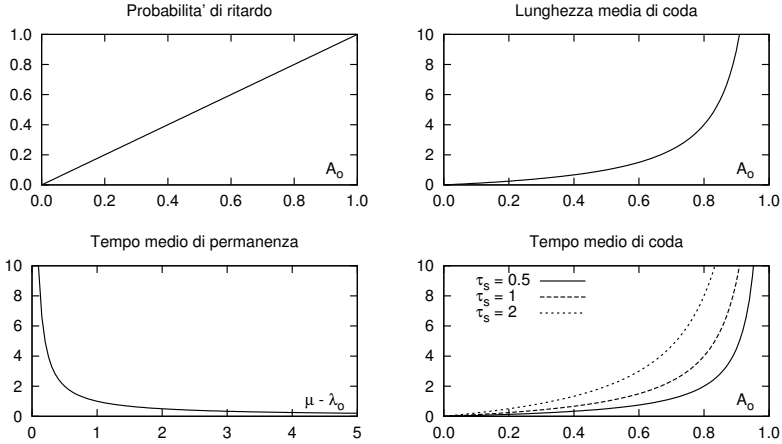


Figura 8.3: Grandezze di interesse per il sistema a coda infinita ed unico servente

Tempo medio di coda si calcola come

$$\tau_c = \tau_p - \tau_S = \frac{1}{\mu - \lambda_o} - \frac{1}{\mu} = \frac{\mu - \mu + \lambda_o}{\mu(\mu - \lambda_o)} = \frac{A_o}{\mu(1 - A_o)} = \frac{1}{\mu} \frac{\rho}{1 - \rho} = \tau_S \frac{\rho}{1 - \rho}$$

Questo risultato mostra che il tempo medio di coda è legato al tempo medio di servizio e all'efficienza di giunzione, confermando ancora i risultati per $\rho \rightarrow \infty$.

La fig. 8.3 mostra l'andamento delle grandezze appena calcolate.

8.4.3 Sistemi a coda finita e con più serventi

Riportiamo solo i risultati, validi se entrambi i processi di ingresso e di servizio sono esponenziali con frequenza media λ_o e μ , la coda è lunga L , i serventi sono M e le sorgenti infinite.

Probabilità di k richieste nel sistema

$$p_k(A_o) = \begin{cases} \frac{A_o^k}{k! \alpha(A_o)} & 0 \leq k \leq M \\ \frac{A_o^k}{M^{k-M} M! \alpha(A_o)} & M \leq k \leq M + L \end{cases}$$

in cui $\alpha(A_o) = \frac{1}{p_0(A_o)} = \sum_{k=0}^{M+L} \frac{A_o^k}{k!}$ e $A_o = \frac{\lambda_o}{\mu}$. Si noti come per $0 \leq k \leq M$ ed $L = 0$ si ottenga lo stesso risultato già esposto per i sistemi orientati alla perdita, mentre per $M = 1$ ed $L = \infty$ ci si riconduca al caso precedentemente analizzato.

Probabilità di ritardo

$$P_r = \sum_{k=M}^{M+L} p_k(A_o) = p_M(A_o) \frac{1 - \rho^{L+1}}{1 - \rho} \quad \text{in cui} \quad \rho = \frac{A_o}{M}$$

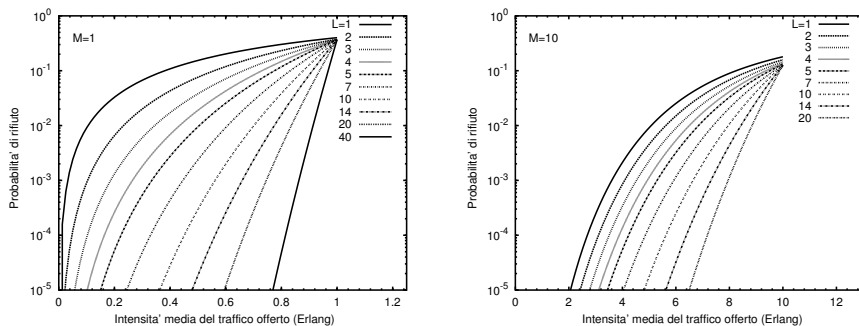


Figura 8.4: Probabilità di perdita per un sistema a coda finita con uno o dieci serventi

Probabilità di perdita

$$P_p = p_{M+L}(A_o) = \frac{A_o^{M+L}}{M^L M! \cdot \alpha(A_o)}$$

Tempo medio di coda

$$\tau_c = \tau_S \frac{P_r - L \cdot P_{M+L}(A_o)}{M - A_o}$$

La Figura 8.4 descrive la probabilità di perdita per un sistema a servente singolo (a sinistra) e con 10 serventi (a destra), in funzione dell'intensità di traffico offerto e della lunghezza di coda, così come risulta dalla applicazione delle formule riportate. Nel caso di trasmissione di pacchetti di lunghezza fissa, per i quali il tempo di servizio è fisso e *non* a distribuzione esponenziale²⁹, i risultati ottenuti costituiscono una *stima conservativa* delle prestazioni del sistema (che potranno cioè essere migliori). L'analisi delle curve permette di valutare con esattezza il vantaggio dell'uso di una coda (a spese del tempo di ritardo). Infatti, aumentando il numero di posizioni di coda si mantiene una probabilità di blocco accettabile anche per traffico intenso.

Ad esempio, per $P_b = 1\%$ ed $M = 1$, osserviamo che una coda con $L = 20$ posizioni gestisce un traffico di $A_o = 0.83$ Erlang, contro gli $A_o = 0.11$ Erlang del caso senza coda. Ciò corrisponde ad un aumento dell'efficienza di $\frac{0.83}{0.11} = 7.54$ volte. D'altra parte ora il tempo medio di coda (calcolato in modo conservativo applicando la relazione per coda infinita) è $\tau_c = \tau_S \frac{\rho}{1-\rho} = \frac{0.83}{1-0.83} \tau_S = 4.9 \tau_S$, ed è quindi aumentato (rispetto a τ_S) di quasi 5 volte.

Esercizio Un nodo di una rete per dati effettua la moltiplicazione di pacchetti di dimensione media di 8 Byte³⁰ su collegamenti con velocità binaria $f_b = 100$ Mbps³¹

- 1) Determinare il tempo medio di servizio di ogni singolo pacchetto;

²⁹In una trasmissione a pacchetto, operata a frequenza binaria f_b e con pacchetti di lunghezza media \bar{L}_p bit, il tempo *medio* di servizio per un singolo pacchetto è pari a quello medio necessario alla sua trasmissione, e cioè $\tau_S = \bar{L}_p / f_b$.

³⁰1 byte = 8 bit, 1 K = $2^{10} = 1024$. Il "K" in questione è "un K informatico". Nel caso invece in cui ci si riferisca ad una velocità di trasmissione, il prefisso K torna a valere $10^3 = 1000$.

³¹In virtù di quanto esposto alla nota precedente, in questo caso $1M = 10^6 = 1000000$.

- 2) determinare il tempo medio di interarrivo τ_a tra pacchetti corrispondente ad un traffico di ingresso di 1200 pacchetti/secondo, e l'associata intensità A_o ;
- 3) assumendo che la dimensione dei pacchetti sia una v.a. con densità esponenziale negativa, così come il tempo di interarrivo tra pacchetti, e che la memoria del multiplatore sia così grande da approssimare le condizioni di coda infinita, determinare il ritardo medio di un pacchetto, ossia il tempo medio trascorso tra quando un pacchetto si presenta in ingresso al nodo e quando ne esce;
- 4) calcolare la quantità di memoria necessaria ad ospitare i dati che si accumulano in un intervallo temporale pari al ritardo medio, considerando pacchetti di lunghezza fissa e pari alla media.

Risposte

- 1) Il tempo medio di servizio di un pacchetto è pari al tempo occorrente per trasmetterlo:

$$\tau_S = \frac{1}{\mu} = \text{durata di un bit} \cdot \frac{\text{bit}}{\text{pacchetto}} = \frac{1}{10^8} \left[\frac{\text{secondi}}{\text{bit}} \right] \cdot 1024 \left[\frac{\text{byte}}{\text{pacchetto}} \right] \cdot 8 \left[\frac{\text{bit}}{\text{byte}} \right] \simeq 655 \mu\text{sec};$$
- 2) $\tau_a = \frac{1}{\lambda} = \frac{1}{1200} = 833 \mu\text{sec}$; $A_o = \frac{\lambda}{\mu} = 1200 \cdot 655 \cdot 10^{-6} = 0.786$ Erlang;
- 3) Le condizioni poste corrispondono a quelle di traffico poissoniano e sistema a singolo servente e coda infinita, per il quale la teoria fornisce per il tempo di permanenza il risultato $\tau_p = \frac{1}{\mu - \lambda_o} = \frac{1}{\frac{10^6}{655} - \frac{10^6}{833}} = \frac{1}{326} \simeq 3 \text{ msec}$;
- 4) La memoria necessaria è pari al prodotto tra il tempo medio di permanenza ed il numero di bit che si accumulano in quel periodo, ovvero $3 \cdot 10^{-3} [\text{sec}] \cdot 1200 \left[\frac{\text{pacch}}{\text{sec}} \right] \cdot 1024 \left[\frac{\text{byte}}{\text{pacch}} \right] \simeq 3.7 \text{ Kbyte}$.

8.5 Reti per trasmissione dati

In questa sezione illustriamo le particolarità legate alle *trasmissioni dati*, e come queste possano essere vantaggiosamente sfruttate per conseguire la *maggiore efficienza* che i sistemi di servizio a coda presentano rispetto a quelli orientati alla perdita. Le particolari *modalità e funzioni* legate alle trasmissioni dati saranno classificate secondo uno schema che ne consente il confronto in termini di prestazioni e vincoli sulla realizzazione della rete. Infine, verranno formalizzate le esigenze legate alla soluzione dei problemi di trasmissione dati, introducendo i concetti legati alle *architetture protocollari*, assieme ad alcuni esempi reali.

Le trasmissioni dati si prestano bene a comunicazioni in cui siano possibili ritardi temporali variabili, attuando una filosofia di tipo *ad immagazzinamento e rilancio* (STORE AND FORWARD) basata sul suddividere il messaggio in unità informative elementari denominate *pacchetti*, che possono essere inoltrati sulla rete di comunicazione, assieme a quelli prodotti da altre trasmissioni. L'applicazione della stessa metodologia a trasmissioni (ad esempio) vocali non è per nulla semplice, in quanto la presenza di un ritardo variabile per la trasmissione dei pacchetti comporta problemi non trascurabili, a meno di attuare speciali meccanismi di priorità e prenotazione della banda, tuttora oggetto di ricerca.

8.5.1 Il pacchetto dati

Discutiamo brevemente, in termini generali, i possibili contenuti di un pacchetto dati; il suo formato effettivo dipenderà dal particolare protocollo di trasmissione adottato.

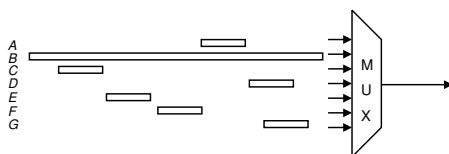
La prima osservazione da fare è che la suddivisione del messaggio in pacchetti comporta un aumento delle informazioni da trasmettere, in quanto ognuno di questi dovrà contenere informazioni aggiuntive per consentire un suo corretto recapito e la sua ricombinazione con gli altri pacchetti dello stesso messaggio. Occorre inoltre affrontare gli ulteriori problemi tipici di una comunicazione dati, ovvero come contrastare gli errori di trasmissione, e come gestire le risorse di rete.

In termini generali, un pacchetto è composto da una *intestazione* (HEADER), dalla parte di messaggio che trasporta (*dati*), e da un campo *codice di parità* (CRC) necessario a rivelare l'occorrenza di errori di trasmissione³². L'*header* a sua volta può essere suddiviso in campi, in cui trovano posto (tra le altre cose) gli *indirizzi* del destinatario e



della sorgente, un *codice di controllo* che causa in chi lo riceve l'esecuzione di una procedura specifica, un *numero di sequenza* che identifica il pacchetto all'interno del messaggio originale, ed un campo che indica la *lunghezza* del pacchetto. Nonostante la presenza delle informazioni aggiuntive³³, la trasmissione a pacchetto consegue una efficienza maggiore di quella a circuito, in quanto è attuata mediante sistemi a coda.

Può sembrare vantaggioso mantenere la dimensione dei pacchetti elevata, riducendo così la rilevanza delle informazioni aggiuntive, ma si verificano controindicazioni. Infatti, suddividere messaggi lunghi in pacchetti più piccoli garantisce l'inoltro



di (altre) comunicazioni più brevi durante la trasmissione di messaggi lunghi, che altrimenti *bloccherebbero* i sistemi di coda se realizzate con un unico "pacchettone": in figura è mostrato un esempio in cui *B*, presentandosi in ingresso al multiplexer con lieve anticipo rispetto agli altri pacchetti più piccoli, ne impedisce l'inoltro, monopolizzando la linea di uscita per tutta la durata della sua trasmissione.

Infine, all'aumentare della lunghezza di un pacchetto aumenta proporzionalmente la probabilità di uno (o più) bit errati (vedi anche la formula (5.18) a pag. 94 e la discussione al § 5.5.2.3), e dunque l'uso di dimensioni contenute riduce le necessità di ritrasmissione.

8.5.2 Modo di trasferimento delle informazioni

È definito in base alla specificazione di 3 caratteristiche che lo contraddistinguono: *lo schema di multiplazione*, *il principio di commutazione* e *l'architettura protocollare*.

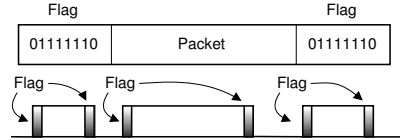
³²La sigla CRC significa *Cyclic Redundancy Check* (controllo ciclico di ridondanza) ed indica una parola binaria i cui bit sono calcolati in base ad operazioni algebriche (vedi § 5.4.2.3) attuate sui bit di cui il resto del messaggio è composto. Dal lato ricevente sono eseguite le stesse operazioni, ed il risultato confrontato con quello presente nel CRC, in modo da controllare la presenza di errori di trasmissione.

³³L'entità delle informazioni aggiuntive rispetto a quelle del messaggio può variare molto per i diversi protocolli, da pochi bit a pacchetto fino ad un 10-20% dell'intero pacchetto (per lunghezze ridotte di quest'ultimo).

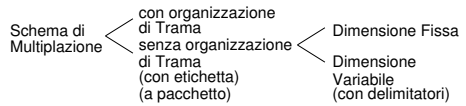
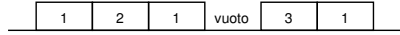
8.5.2.1 Schema di multiplazione

È stato già illustrato uno schema a divisione tempo che prevede l'uso di una trama in cui trovano posto diverse comunicazione vocali³⁴, e che necessita di un funzionamento sincronizzato dei nodi di rete. La trasmissione a pacchetto invece non prevede l'uso esclusivo di risorse da parte delle singole comunicazioni, e non fa uso di una struttura di trama e pertanto occorrono soluzioni particolari per permettere la delimitazione dei pacchetti.

Ad esempio, i protocolli HDLC ed X.25 presentano pacchetti di dimensione variabile, e fanno uso di un byte di flag (vedi pag. 112) costituito da sequenza 01111110 in testa ed in coda, per separare tra loro i pacchetti di comunicazioni differenti. Per evitare che i dati "propri" del pacchetto possano simulare un flag, in trasmissione viene inserito un bit 0 dopo 5 uni di fila, che (se presente) viene rimosso al ricevitore. Se dopo 5 uni c'è ancora un 1 (e poi uno zero), allora è un flag.



Nel caso in cui il pacchetto invece abbia una dimensione fissa³⁵, ci si trova ad operare in una situazione simile a quella in presenza di trama, tranne che... la trama non c'è, e dunque l'ordine dei pacchetti è qualsiasi, ma viene meno l'esigenza dei flag di delimitazione.



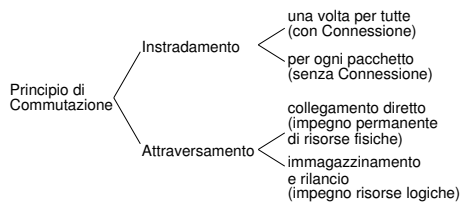
In entrambi i casi (lunghezza di pacchetto fissa o variabile) i nodi della rete non necessitano di operare in sincronismo tra loro; lo schema di multiplazione è quindi detto a divisione di tempo senza organizzazione di trama, asincrono, con etichetta. Il termine etichetta (o label) indica che ogni pacchetto deve recare con sé le informazioni idonee a ricombinarlo assieme agli altri dello stesso messaggio.

8.5.2.2 Principio di commutazione

È definito in base a come sono realizzate le due funzioni di instradamento, ovvero come individuare un percorso nella rete, e di attraversamento, ossia come permettere l'inoltro del messaggio tra le porte di ingresso e di uscita del commutatore.

Se l'instradamento (ROUTING) viene determinato una volta per tutte all'inizio del collegamento, il modo di trasferimento viene detto con connessione. Se al contrario l'instradamento avviene in modo indipendente per ogni pacchetto, il collegamento è detto senza connessione ed ogni pacchetto di uno stesso messaggio può seguire percorsi differenti.

L'attraversamento di un nodo di rete consiste invece nel demoltiplicare le informazioni in ingresso e multiplarle di nuovo su uscite diverse: ciò può avvenire mediante un collegamento diretto o per immagazzinamento e rilancio.



³⁴Come nel PCM telefonico, vedi § 6.3.1

³⁵Un modo di trasferimento con pacchetti di dimensione fissa è l'ATM (Asynchronous Transfer Mode che viene descritto in appendice 8.6.2.

Virtuale per similitudine con il caso di commutazione di circuito, con la differenza che ora il percorso individuato è definito solo in termini di tabelle e di etichette, e non di risorse fisiche (tranne che per la memoria della tabella).

Al termine della comunicazione, sul circuito virtuale viene inviato un apposito pacchetto di controllo, che provoca la rimozione del routing dalle tabelle.

Congestione e controllo di flusso Durante la fase di instradamento, il percorso nella rete è determinato in base alle condizioni di traffico del momento, ed eventualmente la connessione può essere rifiutata nel caso in cui la memoria di coda nei nodi coinvolti sia quasi esaurita, evento indicato con il termine di *congestione*.

D'altra parte, se alcune sorgenti origine dei Canali Virtuali già assegnati e che si incrociano in uno stesso nodo intermedio, iniziano ad emettere pacchetti a frequenza più elevata del previsto, il nodo intermedio si congestiona (ossia esaurisce la memoria di transito) ed inizia a *perdere pacchetti*, penalizzando tutti i Canali Virtuali che attraversano il nodo.

Per questo motivo, sono indispensabili strategie di *controllo di flusso* che permettano ai nodi di regolare l'emissione delle sorgenti. Il controllo di flusso è attuato anch'esso mediante pacchetti (di controllo), privi del campo di dati, ma contenenti un codice identificativo del comando che rappresentano. Ad esempio, un nodo non invia nuovi pacchetti di un circuito virtuale finché non riceve un *pacchetto di riscontro* relativo ai pacchetti precedenti. D'altra parte, nel caso di una rete congestionata, la perdita di pacchetti causa il mancato invio dei riscontri relativi, e dunque i nodi a monte cessano l'invio di nuovi pacchetti⁴⁰. Dopo un certo periodo di tempo (TIMEOUT) il collegamento è giudicato interrotto e viene generato un pacchetto di *Reset* da inviare sul canale virtuale, e che causa, nei nodi attraversati, il rilascio delle risorse logiche (tabelle) relative al Canale Virtuale.

Discutiamo ora invece di un ulteriore possibile principio di commutazione:

Commutazione di pacchetto a datagramma Anche in questo caso, *l'attraversamento* dei nodi avviene per *immagazzinamento e rilancio*, mentre la funzione di *instradamento* è svolta in modo distribuito tra i nodi di rete *per ogni pacchetto*, i quali (chiamati ora *datagrammi*) devono necessariamente contenere l'indirizzo completo della destinazione. Infatti, in questo caso manca del tutto la fase iniziale del collegamento, in cui prenotare l'impegno delle risorse (fisiche o logiche) che saranno utilizzate⁴¹. Semplicemente, non è previsto alcun impegno a priori, ed ogni pacchetto costituisce un collegamento individuale che impegna i nodi di rete solo per la durata del proprio passaggio. L'instradamento avviene mediante tabelle presenti nei nodi, sia di tipo statico che dinamico (nel qual caso tengono conto delle condizioni di carico e di coda dei nodi limitrofi) che indicano le possibili porte di uscita per raggiungere la destinazione scritta sul pacchetto. Quest'ultimo quindi viene fatto uscire *senza nessuna alterazione* dalla porta di uscita.

Uno dei maggiori vantaggi dei datagrammi rispetto ai circuiti virtuali è una migliore resistenza ai guasti e malfunzionamenti: in questo caso infatti, a parte una eventuale

⁴⁰In realtà vengono prima fatti dei tentativi di inviare nuovamente i pacchetti "vecchi". Questi ultimi infatti sono conservati da chi li invia (che può anche essere un nodo intermedio), finché non sono riscontrati dal ricevente. Quest'ultimo fatto può causare ulteriore congestione, in quanto restano impegnate risorse di memoria "a monte" della congestione che così *si propaga*.

⁴¹Per questo motivo, il collegamento è detto *senza connessione*.

necessità di ritrasmettere i pacchetti persi, il collegamento prosegue attraverso percorsi alternativi; inoltre l'elevato numero di percorsi alternativi, può permettere (in condizioni di carico leggero) di soddisfare brevi richieste di trasmissione a velocità elevate. Allo stesso tempo, in presenza di messaggi molto brevi, l'invio di un singolo datagramma è più che sufficiente, mentre nel caso a circuito virtuale le fasi di instaurazione ed abbattimento sarebbero state un lavoro in più da svolgere (tanto che ad es. l'X.25, che è nato a c.v., prevede anche il funzionamento a datagramma).

Consegna ordinata e congestione Uno dei maggiori problemi legati all'uso di datagrammi è che l'ordine di arrivo dei pacchetti può essere diverso da quello di partenza, potendo questi seguire percorsi differenti. Per questo motivo, nei datagrammi è presente un *numero di sequenza* che si incrementa ad ogni pacchetto trasmesso, ed alla destinazione sono predisposti dei *buffer*⁴² di memoria nei quali ricostruire l'ordine esatto dei pacchetti.

Nel caso di un pacchetto mancante, il ricevente non sa se questo è semplicemente ritardato oppure è andato perso, rendendo problematico il controllo di flusso. In questo caso si produce un impegno anomalo dei buffer di ingresso, che non possono essere rilasciati perché incompleti, e ciò può causare il rifiuto dell'accettazione di nuovi pacchetti, provocando un impegno anomalo anche per i buffer di uscita di altri nodi, causando congestione⁴³.

Prima di effettuare un trasferimento a datagramma, è opportuno (a parte il caso di messaggi composti da un singolo datagramma) verificare la disponibilità del destinatario finale, e preavvisarlo di riservare una adeguata quantità di memoria. Ad esempio, in Internet avviene proprio questo (vedi pag. 198).

Proseguiamo la descrizione delle reti per dati con l'ultima caratteristica di un modo di trasferimento:

8.5.2.3 Architettura protocollare

Definisce la stratificazione delle funzioni di comunicazione, sia per gli apparati terminali che per i nodi di transito, e di come queste interagiscono reciprocamente sia tra nodi diversi, che nell'ambito di uno stesso nodo. Alcune di queste sono già state introdotte, e le citiamo per prime, seguite da quelle più rilevanti illustrate di seguito:

- il *controllo di flusso*, che impedisce la saturazione dei buffer;
- la *consegna ordinata*, per riassemblare messaggi frammentati su più datagrammi;
- la *segmentazione e riassemblaggio*, che definisce le regole per frammentare un messaggio in pacchetti e ricomporli, ad esempio in corrispondenza dei "confini" tra sottoreti con differente lunghezza di pacchetto;
- il *controllo di connessione*, che provvede ad instaurare la connessione, eseguire l'instradamento, impegnare le risorse, supervisionare il controllo di flusso, abbattere la connessione al suo termine;
- il *controllo di errore*, che provvede a riscontrare le unità informative, a rilevare gli errori di trasmissione, a gestire le richieste di trasmissione;

⁴²Il termine *buffer* ha traduzione letterale "respingente, paracolpi, cuscinetto" ed è a volte espresso in italiano dalla locuzione memoria tampone.

⁴³La soluzione a questa "spirale negativa" si basa ancora sull'uso di un allarme a tempo (timeout), scaduto il quale si giudica interrotto il collegamento, e sono liberati i buffer.

- *l'incapsulamento*, che aggiunge ai pacchetti di dati da trasmettere le informazioni di protocollo come l'header, gli indirizzi, il controllo di parità...

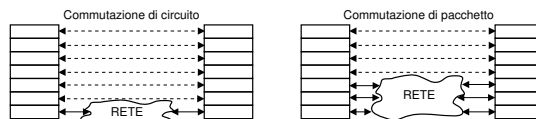
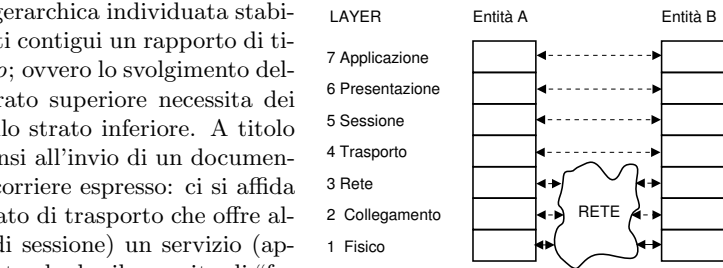
Stratificazione ISO-OSI Per aiutare nella schematizzazione delle interazioni tra le funzioni illustrate, l'*International Standard Organization* (ISO) ha formalizzato un modello concettuale per sistemi di comunicazione denominato *Open System Interconnection* (OSI)⁴⁴, che individua una relazione gerarchica tra i protocolli. In particolare sono definiti sette *strati* o *livelli* (LAYERS) ognuno dei quali raggruppa un insieme di funzioni affini. Gli strati più elevati (4-7) sono indicati anche come *strati di utente*, in quanto legati a funzioni relative ai soli apparati terminali; gli *strati di transito* invece (1-3) riguardano funzioni che devono essere presenti anche nei nodi intermedi.

La relazione gerarchica individuata stabilisce tra due strati contigui un rapporto di tipo *utente-servizio*; ovvero lo svolgimento delle funzioni di strato superiore necessita dei servizi offerti dallo strato inferiore. A titolo di esempio, si pensi all'invio di un documento mediante un corriere espresso: ci si affida allora ad uno strato di trasporto che offre all'utente (strato di sessione) un servizio (appuntamento) di trasporto che ha il compito di "far apparire" il documento presso il destinatario. La sede locale del corriere si affida quindi alla propria divisione interna che gestisce la rete dei corrispondenti, la quale si affida a sua volta ai corrispondenti stessi, che hanno il compito di assistere alla consegna ed all'arrivo (collegamento) del documento. Il trasferimento fisico dello stesso può quindi avvenire mediante un ultimo strato funzionale (treno, nave, aereo, auto...) che provvede al recapito in base alle informazioni ricevute dallo strato di collegamento.

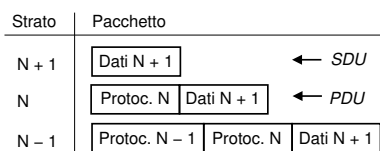
Per terminare l'esempio, facciamo notare come in ogni livello avvengano *due* tipi di colloqui (regolati da altrettanti protocolli): uno è *orizzontale*, detto anche *tra pari* (PEER-TO-PEER), come è ad esempio il contenuto del documento che spediamo, od i rapporti tra corrispondenti locali (che nel caso di un sistema di comunicazione corrisponde allo strato di collegamento, relativo ai protocolli tra singole coppie di nodi di rete); il secondo tipo di colloquio avviene invece in forma *verticale*, o *tra utente e servizio*, in quanto per realizzare le funzioni di uno strato *utente* ci si affida ad un *servizio* di comunicazione offerto dallo strato inferiore (che a sua volta può avvalersi dei servizi degli altri strati ancora inferiori)⁴⁵.

⁴⁴In virtù dell'intreccio di sigle, il modello di riferimento prende il nome (palindromo) di modello *ISO-OSI*.

⁴⁵Il modo di trasferimento è completamente definito dopo che sia stato specificato in quale strato siano svolte le funzioni di commutazione e moltiplicazione. In una rete a commutazione di circuito, queste sono realizzate dallo strato fisico che, esaurita la fase di instradamento ed impegno di risorse fisiche, collega in modo trasparente sorgente e destinazione. Nella commutazione di pacchetto, invece, le funzioni di moltiplicazione e commutazione coinvolgono (per tutti i pacchetti del messaggio) tutti i nodi di rete interessati; si dice pertanto che i protocolli di collegamento e di rete devono *essere terminati* (nel senso di gestiti) da tutti i nodi di rete.



Incapsulamento La modalità con cui un protocollo tra pari di strato N affida i suoi dati ad un servizio di strato $N - 1$, si avvale (nella commutazione di pacchetto) della funzione di *incapsulamento*, di cui viene data una interpretazione grafica alla figura seguente.



I dati che lo strato $N + 1$ vuol trasmettere al suo pari, indicati anche come *Service Data Unit* (SDU), sono prefissi dalle informazioni di protocollo necessarie alla gestione del collegamento tra entità allo strato N . Questa nuova unità informativa

prende il nome di *Protocol Data Unit* (PDU) per lo strato N , e viene passata in forma di SDU al servizio di collegamento offerto dallo strato $N - 1$, che ripete l'operazione di incapsulamento con le proprie informazioni di protocollo, generando una nuova PDU (di strato $N - 1$). Pertanto, lo strato fisico provvederà a trasmettere pacchetti contenenti tutte le informazioni di protocollo degli strati superiori.

Indipendenza dei servizi tra pari dal servizio di collegamento Quando uno strato affida il collegamento con un suo pari allo strato inferiore, quest'ultimo può mascherare al superiore la modalità con cui viene realizzato il trasferimento.

In particolare, se ci riferiamo all'interfaccia tra gli strati di trasporto e di rete, lo strato di rete può realizzare con il suo pari collegamenti con o senza connessione, mentre quello di trasporto offre allo stesso tempo (ma in modo indipendente) agli strati superiori un servizio con o senza connessione, dando luogo alle seguenti 4 possibilità:

Servizio di rete	SERVIZIO DI TRASPORTO	
	CIRCUITO VIRTUALE	DATAGRAMMA
Circuito Virtuale	SNA, X.25	Insolito
Datagramma	Arpanet, TCP/IP	Decnet

SNA (SYSTEM NETWORK ARCHITECTURE) è una architettura proprietaria IBM, in cui il trasferimento avviene in modo ordinato, richiedendo al livello di trasporto un circuito virtuale, che è realizzato da una serie di canali virtuali tra i nodi di rete. La stessa architettura è adottata anche dall'X.25, che costituisce l'insieme di protocolli che descrivono il funzionamento di reti pubbliche a commutazione di pacchetto, presenti in tutto il mondo: quella italiana prende il nome di ITAPAC.

Arpanet è l'architettura di Internet, in cui sebbene lo strato di rete operi con un principio di commutazione a datagramma, mediante il protocollo *IP* (INTERNET PROTOCOL), lo strato di trasporto (*TCP*, TRANSFER CONTROL PROGRAM) offre a quelli superiori un servizio con connessione, attuato mediante circuiti virtuali, in modo da garantire il corretto sequenziamento delle unità informative, ed offrire canali di comunicazione formalmente simili ai files presenti localmente su disco. Il mascheramento del servizio di rete interna a datagramma in un servizio con connessione avviene a carico dello strato *TCP* di trasporto presente nei nodi terminali, che appunto affronta il riassetto ordinato dei datagrammi ricevuti dallo strato di rete.

Decnet è (o meglio era) l'architettura Digital, in cui il controllo di errore, la sequenzializzazione, ed il controllo di flusso sono realizzati dal livello di trasporto.

Soluzione insolita non è praticata perché equivale a fornire alla rete pacchetti disordinati, farli consegnare nello stesso identico disordine a destinazione, dove poi sono riassemblati. Può avere un senso se la comunicazione è sporadica, ma sempre per la stessa destinazione, nel qual caso somiglia ad un circuito virtuale permanente.

8.6 Appendici

In questa sezione trova spazio la descrizione di aspetti delle reti per dati come Internet, e le reti IP e ATM, che costituiscono l'immediata applicazione degli argomenti esposti nel capitolo, ossia i sistemi a coda ed a trasmissione di pacchetto, e le architetture protocollari. Viceversa, la realizzazione di sistemi orientati alla perdita ed a commutazione di circuito è discussa al Cap. 6.

Iniziamo subito con il dire che il modello a strati ISO-OSI è una astrazione concettuale utile per individuare raggruppamenti di funzioni, e serve ottimamente come modello per stimolare l'interoperabilità di apparati di diversi costruttori. D'altra parte, alcune delle realizzazioni esistenti (come ad esempio INTERNET) si sono sviluppate precedentemente alla definizione del modello, mentre altre (come ATM) seguono filosofie che solo successivamente sono state incorporate nel modello di riferimento. Pertanto, utilizzeremo le classificazioni ISO-OSI come parametro di riferimento, mediante il quale analizzare le funzioni delle reti reali.

8.6.1 La rete Internet

8.6.1.1 Storia

Nel 1964 L. Kleinrock (UCLA) propone un modello di rete non gerarchica e con parti ridondanti, che realizza una modalità di trasferimento senza connessione e senza garanzie di qualità del servizio, rimandando queste ultime ai livelli superiori dell'architettura protocollare. Tale tipologia di servizio è oggi indicata con il termine *best effort*⁴⁶. Nel '69 sono operativi cinque nodi nelle università americane, e nel '72 avviene la prima dimostrazione pubblica di ARPANET, basata su NCP. Nel '73 Kahn e Cerf iniziano a definire TCP, da cui viene successivamente separato l'IP per la convenienza di non dover necessariamente aprire sempre una connessione. Fino all'80, il DoD⁴⁷ sovvenziona le università per implementare in ambiente UNIX i protocolli, che nel frattempo si vanno arricchendo di servizi, mentre la trasmissione Ethernet (del 1973) è adottata per realizzare LAN.

Nel 1983 il DoD decreta che tutti i calcolatori connessi a ARPANET adottino i protocolli TCP/IP, e separa la rete in due parti: una civile (ARPANET) ed una militare (MILNET). Negli anni seguenti i finanziamenti dalla *National Science Foundation* permettono lo sviluppo di una rete di trasporto a lunga distanza e di reti regionali, che interconnettono LAN di altre università e di enti di ricerca alla rete ARPANET, alla quale si collegano poi anche le comunità scientifiche non americane.

Nel 1990 ARPANET cessa le sue attività, e Barners-Lee (CERN) definisce il WWW, mentre nel '93 Andreessen (NCSA) sviluppa il primo *browser* WWW. Dal 1995 L'NSF non finanzia più la rete di interconnessione, ed il traffico inizia ad essere trasportato da operatori privati.

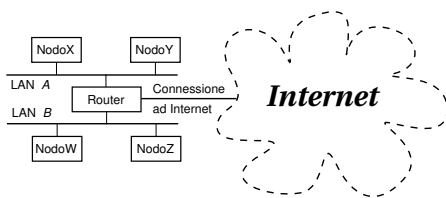
⁴⁶Migliore sforzo, ossia la rete dà il massimo, senza però garantire nulla.

⁴⁷Department of Defense.

8.6.1.2 Le caratteristiche

La parola *Internet* in realtà è composta da due parole, INTER e NET, in quanto le caratteristiche della rete Internet sono quelle di fondere in una unica architettura una infinità di singole reti locali, potenzialmente disomogenee, e permettere la comunicazione tra i computer delle diverse sottoreti.

Ogni nodo della rete è connesso ad una rete locale (LAN⁴⁸), la quale a sua volta è interconnessa ad Internet mediante dei nodi detti *router*⁴⁹ che sono collegati ad una o più LAN e ad Internet, e svolgono la funzione di instradare le comunicazioni verso l'esterno. L'instradamento ha luogo in base ad un *indirizzo IP*⁵⁰, che individua i singoli nodi in modo univoco su scala mondiale.



Come anticipato, lo strato di rete (o strato IP) realizza un modo di trasferimento a datagramma e non fornisce garanzie sulla qualità di servizio (QoS, QUALITY OF SERVICE) in termini di ritardi, errori e pacchetti persi. La situazione è mitigata dalla strato di trasporto (TCP, TRANSMISSION CONTROL

PROTOCOL) che offre ai processi applicativi un servizio a circuito virtuale.

I protocolli di Internet sono realizzati in software e sono pubblici; gli utenti stessi e molte sottoreti private contribuiscono significativamente al trasporto, all'indirizzamento, alla commutazione ed alla notifica delle informazioni. Queste sono alcune ragioni fondamentali per cui Internet *non è di nessuno* ed è un patrimonio dell'umanità.

8.6.1.3 Gli indirizzi

Strato	Indirizzo
Applicazione	<i>protocollo://nodo.dominio.tld</i>
Trasporto	<i>socket TCP o porta</i>
Rete	<i>indirizzo IP x.y.w.z</i>
Collegamento	<i>indirizzo Ethernet a:b:c:d:e:f</i>

Iniziamo l'argomento discutendo subito la stratificazione degli indirizzi coinvolti in una comunicazione via Internet. Ogni livello funzionale infatti utilizza le proprie convenzioni di indirizzamento, come illustrato nella tabella a fianco. Se a prima

vista questa abbondanza di indirizzi può apparire esagerata, è proprio in questo modo che si realizza l'interoperabilità tra ambienti di rete differenti.

IP ed Ethernet I computer connessi ad Internet (detti *nodi*) sono le sorgenti e le destinazioni dell'informazione, e sono individuati da *un indirizzo IP*, che consiste in

⁴⁸LOCAL AREA NETWORK, ossia *rete locale*. Con questo termine si indica un collegamento che non si estende oltre (approssimativamente) un edificio.

⁴⁹La funzione di conversione di protocollo tra reti disomogenee è detta di *gateway*, mentre l'interconnessione tra reti locali è svolta da dispositivi *bridge* oppure da *ripetitori* se le reti sono omogenee. Con il termine *router* si indica più propriamente il caso in cui il nodo svolge funzioni di instradamento, che tipicamente avviene nello *strato di rete*. Nel caso in cui invece si operi un instradamento a livello dello *strato di collegamento*, ossia nell'ambito di sezioni diverse (collegate da bridge o ripetitori) di una stessa LAN, il dispositivo viene detto *switch*. Infine, un *firewall* opera a livello di trasporto, e permette di impostare *regole di controllo* per restringere l'accesso alla rete interna in base all'indirizzo di *sorgente*, al tipo di *protocollo*, e/o a determinati *servizi*.

⁵⁰IP = Internet Protocol.

un gruppo di 4 byte⁵¹ e che si scrive $x.y.w.z$ con ognuna delle 4 variabili pari ad un numero tra 0 e 255.

I nodi sono connessi alla rete mediante una interfaccia a volte indicata come MAC (MEDIA ACCESS CONTROL). Prendendo come esempio⁵² i nodi connessi ad una LAN Ethernet, l'interfaccia di rete è individuata a sua volta da un *indirizzo Ethernet* composto da 6 byte. Quest'ultimo è unico in tutto il mondo, ed impresso dal costruttore nella scheda di interfaccia. L'indirizzo Ethernet viene però utilizzato solo nell'ambito della LAN di cui il nodo fa parte, ossia dopo che i pacchetti sono stati instradati dai router, per mezzo dell'indirizzo IP, verso la LAN.

Sottoreti Ogni nodo conosce, oltre al proprio indirizzo IP, anche una *maschera di sottorete* composta da una serie di uni seguita da zeri, in numero complessivo di 32 bit, tanti quanti ne sono presenti nell'indirizzo IP. Il termine *maschera* è dovuto all'operazione di AND binario (vedi tabella) operata tra la maschera e gli indirizzi IP, per determinare se questi appartengano alla propria stessa LAN oppure risiedano altrove.

Indirizzo IP	Maschera Sottorete	Indirizzo sottorete
151.100.8.33	255.255.255.0	151.100.8.0

Nel caso in cui la sottorete di un nodo Y verso cui il nodo X deve inviare un pacchetto è la stessa su cui è connesso X, allora questi può individuare l'indirizzo Ethernet del destinatario⁵³ ed inviargli il pacchetto direttamente. In caso contrario, X invierà il pacchetto al proprio *default gateway* verso Internet.

Intranet Alcuni gruppi di indirizzi IP (come quelli $192.168.w.z$ oppure $10.y.w.z$) non vengono instradati dai router, e possono essere riutilizzati nelle *reti private* di tutto il mondo per realizzare le cosiddette *reti intranet* operanti con gli stessi protocolli ed applicativi che funzionano via Internet.

Domain Name Server (dns) L'utente di una applicazione internet in realtà non è a conoscenza degli indirizzi IP dei diversi nodi, ma li identifica per mezzo di nomi simbolici del tipo *nodo.dominio.tld*, detti anche *indirizzi Internet*. Il processo di risoluzione che individua l'indirizzo IP associato al nome avviene interrogando un particolare nodo, il DOMAIN NAME SERVER (*servente dei nomi di dominio*). La struttura dei nomi, scandita dai punti, individua una gerarchia di autorità per i diversi campi. Il campo tld è chiamato *dominio di primo livello* (TOP LEVEL DOMAIN⁵⁴), mentre il campo dominio in genere è stato registrato da qualche organizzazione che lo giudica rappresentativo della propria offerta informativa. Il campo *nodo* rappresenta invece

⁵¹Con 4 byte si indirizzano (in linea di principio) $2^{32} = 4.29 \cdot 10^9$ diversi nodi (più di 4 miliardi). E' tuttora in sviluppo il cosiddetto IPv6, che estenderà l'indirizzo IP a 16 byte, portando la capacità teorica a $3.4 \cdot 10^{38}$ nodi. L'IPv6 prevede inoltre particolari soluzioni di suddivisione dell'indirizzo, allo scopo di coadiuvare le operazioni di *routing*.

⁵²Evidentemente esistono molte diverse possibilità di collegamento ad Internet, come via telefono (tramite provider), collegamento satellitare, Frame Relay, linea dedicata, ISDN, ADSL... ma si preferisce svolgere un unico esempio per non appesantire eccessivamente l'esposizione. La consapevolezza delle molteplici alternative consente ad ogni modo di comprendere la necessità di separare gli strati di trasporto e di rete dall'effettiva modalità di trasmissione.

⁵³Mostriamo in seguito che questo avviene mediante il protocollo ARP.

⁵⁴I top level domain possono essere pari ad un identificativo geografico (.it, .se, .au...) od una delle sigle .com, .org, .net, .mil, .edu, che sono quelle utilizzate quando internet era solo americana.

una ben determinata macchina, il cui indirizzo Internet completo è `nodo.dominio.tld`, e che non necessariamente è collegato alla stessa LAN a cui sono connessi gli altri nodi con indirizzi che termina per `dominio.tld`.

Quando un `nodoX` generico deve comunicare con `nodoY.dominio.tld`, interroga il proprio DNS⁵⁵ per conoscerne l'IP. Nella rete sono presenti molti DNS, alcuni dei quali detengono informazioni *autorevoli*⁵⁶ riguardo ai nodi di uno o più domini, altri (i DNS *radice*, o *ROOT*) detengono le informazioni relative a quali DNS siano autorevoli per i domini di primo livello, ed altri fanno da tramite tra i primi due ed i *client* che richiedono una risoluzione di indirizzo. Se il DNS di `nodoX` non è *autorevole per nodoY*, allora⁵⁷ provvede ad inoltrare la richiesta, interrogando prima un DNS radice per individuare chi è autorevole per `.tld`, quindi interroga questo per trovare chi è autorevole per `.dominio.tld`, e quindi usa la risposta ottenuta per dirigere la richiesta di risoluzione originaria. Se la cosa può sembrare troppo macchinosa per funzionare bene, è perché la stessa sequenza di operazioni *non deve* essere effettuata sempre: il DNS utilizzato da `nodoX` riceve infatti, assieme all'IP di `nodoY`, anche una informazione detta *TIME TO LIVE* (TTL o *tempo di vita*) che descrive la scadenza della coppia *nome-IP* ottenuta. Genericamente il TTL è di qualche giorno, e fino alla sua scadenza il DNS *ricorda*⁵⁸ la corrispondenza, in modo da fornire la propria copia in corrispondenza delle richieste future, e ridurre sensibilmente il traffico legato alla risoluzione degli indirizzi Internet. L'insieme delle risoluzioni apprese è denominata *cache* del DNS⁵⁹.

Indirizzi TCP Si è detto che ogni nodo è individuato in Internet mediante il proprio indirizzo IP, ma questo non è sufficiente ad indicare con quale particolare programma (che può essere uno specifico *server* come nel caso del DNS) si vuole entrare in comunicazione. I programmi che sono pronti a ricevere connessioni si pongono *in ascolto* su ben determinate *porte* (o *socket*⁶⁰), identificate da numeri⁶¹, e che sono referenziati in modo simbolico (es. `http://`, `ftp://`) dagli applicativi di utente che si rivolgono allo strato di trasporto (il TCP) per stabilire un collegamento con un server presente su di un nodo remoto.

Alcuni servizi rispondono ad indirizzi *ben noti*, fissi per tutti i nodi, in quanto il chiamante deve sapere a priori a quale porta connettersi. Il nodo contattato invece,

⁵⁵Il "proprio" DNS viene configurato per l'host in modo fisso, oppure in modo dinamico da alcuni Service Provider raggiungibili per via telefonica, e convenientemente corrisponde ad un nodo situato "vicino" al nodo che lo interroga.

⁵⁶Chi registra il dominio deve disporre necessariamente di un DNS in cui inserire le informazioni sulle corrispondenze tra i nomi dei nodi del proprio dominio ed i loro corrispondenti indirizzi IP. In tal caso quel DNS si dice *autorevole* per il dominio ed è responsabile di diffondere tali informazioni al resto della rete.

⁵⁷In realtà esiste anche una diversa modalità operativa, che consiste nel delegare la ricerca ad un diverso DNS (detto *forwarder*), il quale attua lui i passi descritti appresso, e provvede per proprio conto alla risoluzione, il cui esito è poi comunicato al primo DNS e da questi ad *hostX*. Il vantaggio di tale procedura risiede nella maggiore ricchezza della *cache* (descritta appresso) di un DNS utilizzato intensivamente.

⁵⁸Il DNS ricorda anche le altre corrispondenze ottenute, come il DNS autorevole per `.tld` e per `.dominio.tld`; nel caso infine in cui si sia utilizzato un forwarder, sarà quest'ultimo a mantenere memoria delle corrispondenze per i DNS intermedi.

⁵⁹CACHE è un termine generico, che letteralmente si traduce *nascondiglio dei viveri*, e che viene adottato ogni volta si debba indicare una memoria che contiene copie di riserva, o di scorta...

⁶⁰*Socket* è un termine che corrisponde alla... presa per l'energia elettrica casalinga, ed in questo contesto ha il significato di una *presa* a cui si "attacca" il processo che richiede la comunicazione. Per l'esattezza, un *Socket internet* è individuato dal numero di porta TCP e dall'indirizzo IP.

⁶¹Spesso gli indirizzi che identificano i punti di contatto di servizi specifici vengono indicati come SERVICE ACCESS POINT (SAP), anche per situazioni differenti dal caso specifico delle porte del TCP.

apre con il chiamante una connessione di ritorno, su di un numero di porta diverso, che è stato comunicato dal chiamante al momento della richiesta di connessione, e per il quale il chiamante non ha già aperto altre connessioni differenti.

8.6.1.4 TCP

Discutiamo ora del TCP⁶², che offre ai processi applicativi un servizio di trasporto a circuito virtuale, *attaccato* ad una porta⁶³ di un nodo remoto individuato dall'indirizzo IP. Il suo compito è quello di ricevere dai processi applicativi dei dati, suddividerli in pacchetti, ed inviarli al suo pari che svolge il processo inverso.

Il pacchetto TCP La struttura di un pacchetto TCP è mostrata in figura, e comprende una intestazione composta da 6 gruppi (o più) di 4 byte per un minimo di 192 bit, a cui segue un numero variabile di gruppi di 4 byte di Dati, provenienti dagli strati applicativi superiori.

Troviamo subito i numeri delle porte a cui si riferisce la connessione, mentre gli indirizzi IP sono aggiunti dallo strato di rete. I numeri di *Sequenza* e di *Riscontro* servono rispettivamente a numerare i bytes dei pacchetti uscenti, ed a notificare l'altro lato del collegamento del numero di sequenza del prossimo byte che si aspetta di ricevere⁶⁴, riscontrando implicitamente come correttamente arrivati i pacchetti con numero di sequenza più basso.

Offset (4 bit) codifica il numero di parole da 4 byte dell'intestazione, mentre nei 6 bit *Riservati* non è mai stato scritto nulla. I 6 bit del campo *Controllo* hanno ognuno un nome ed un significato preciso, qualora posti ad uno. Il primo (URG) indica che il campo urgent pointer contiene un valore significativo; ACK indica che si sta usando il Numero di Riscontro; PSH indica un pacchetto urgente che non può rispettare la coda in ricezione; RST segnala un malfunzionamento e impone il reset della connessione;

1		8		16		24	
Porta Sorgente				Porta Destinazione			
Numero di Sequenza (Tx)							
Numero di Riscontro (Rx)							
Offset	Riserva	Contr.	Finestra				
Checksum				Puntatore Urgente			
Opzioni				Riempimento			
Dati							
Dati							
...							

⁶²TCP = *Transport Control Protocol*.

⁶³Il numero di porta costituisce in pratica l'*identificativo di connessione* del circuito virtuale. Nel caso in cui un server debba comunicare con più client, dopo avere accettato la connessione giunta su di una *porta ben nota*, apre con i client diversi canali di ritorno, differenziati dall'uso di porte di risposta differenti.

La lista completa dei servizi standardizzati e degli indirizzi ben noti (*socket*) presso i quali i server sono in attesa di richieste di connessione, è presente in tutte le distribuzioni Linux nel file */etc/services*.

⁶⁴Il numero di sequenza si incrementa ad ogni pacchetto di una quantità pari alla sua dimensione in bytes, ed ha lo scopo di permettere le operazioni di controllo di flusso. Il valore iniziale del numero di sequenza e di riscontro è diverso per ogni connessione, e generato in modo pseudo-casuale da entrambe le parti in base ai propri orologi interni, allo scopo di minimizzare i problemi dovuti all'inaffidabilità dello strato di rete (l'IP) che può perdere o ritardare i datagrammi, nel qual caso il TCP trasmittente ri-invia i pacchetti precedenti dopo un time-out. Questo comportamento può determinare l'arrivo al lato ricevente di un pacchetto duplicato, e consegnato addirittura dopo che la connessione tra i due nodi è stata chiusa e riaperta. In tal caso però la nuova connessione adotta un diverso numero di sequenza iniziale, cosicché il pacchetto duplicato e ritardato risulta fuori sequenza, e non viene accettato.

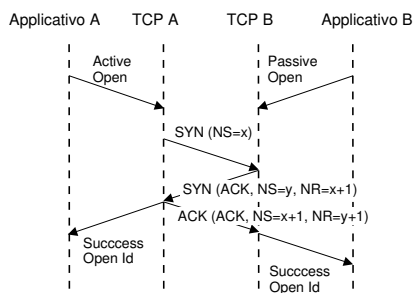
SYN è pari ad uno solo per il primo pacchetto inviato per richiedere di creare una connessione; FIN indica che la sorgente ha esaurito i dati da trasmettere.

I 16 bit di *Finestra* rappresentano il numero di byte che, a partire dal numero espresso dal *Numero di Riscontro*, chi invia il pacchetto è in grado di ricevere, ed il suo utilizzo sarà meglio illustrato tra breve nel contesto del controllo di flusso. Il *Checksum* serve al ricevente per verificare se si sia verificato un errore, il *Puntatore Urgente* contiene il numero di sequenza dell'ultimo byte di una sequenza di dati urgenti, e le *Opzioni* (di lunghezza variabile) sono presenti solo raramente, ed utilizzate a fini di controllo, ad esempio per variare la dimensione della finestra. Infine, il *Riempimento* conclude l'ultima parola da 32 bit.

Uno stesso pacchetto TCP può svolgere funzioni di sola segnalazione, o di sola trasmissione dati, od entrambe.

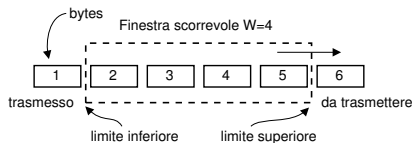
Apertura e chiusura della connessione Il TCP offre un servizio di di trasporto a circuito virtuale, e prima di inviare dati, deve effettuare un colloquio iniziale con il nodo remoto di destinazione. In particolare, il colloquio ha lo scopo di accertare la disponibilità del destinatario ad accettare la connessione, e permette alle due parti di scambiarsi i rispettivi numeri di sequenza descritti alla nota 64.

L'estremo che viene "chiamato" riveste il ruolo di *server*, e l'altro di *client*. Dato che anche quest'ultimo deve riscontrare il numero di sequenza fornito dal server, occorrono tre pacchetti per terminare il dialogo, che prende il nome di THREE WAY HANDSHAKE⁶⁵. Il diagramma a lato mostra l'evoluzione temporale del colloquio tra un processo applicativo client (A), ed un server (B) che si pone in ascolto, mostrando come al primo SYN che pone $NS_A = x$, ne segua un altro che pone $NS_B = y$, seguito a sua volta dall'ACK di chi ha iniziato⁶⁶. La chiusura può avvenire per diverse cause: o perché è terminato il messaggio, segnalato dal bit FIN, o per situazioni anomale, che il TCP indica con il bit RST.



Protocollo a finestra Allo scopo di realizzare un controllo di flusso, il TCP prevede l'uso dell' NR inviato dal ricevente per dosare il ritmo con cui trasmettere i propri pacchetti.

La lunghezza di *Finestra* comunicata con il SYN del ricevente, determina la quantità di memoria riservata per i buffer dedicati alla connessione, che viene gestita come una memoria a scorrimento o *finestra scorrevole* (SLIDING WINDOW). Questa memoria è presente per gestire i casi di pacchetti ritardati o fuori sequenza, e contiene i bytes già trasmessi. Il trasmittente (vedi figura) non fa avanzare il limite inferiore finché non



⁶⁵HANDSHAKE = stretta di mano.

⁶⁶Per ciò che riguarda i valori dei numeri di riscontro NR , questi sono incrementati di 1, perché la *finestra* (descritta nel seguito) inizia dai bytes del prossimo pacchetto, a cui competeranno appunto valori di NS incrementati di uno.

riceve un riscontro con NR maggiore di tale limite. In questo modo non occorre attendere il riscontro di tutti i bytes, o di tutti i pacchetti (che devono comunque essere di dimensione inferiore alla finestra), ma ci può avvantaggiare trasmettendo l'intero contenuto della finestra.

Una finestra del tutto analoga è utilizzata dal ricevente, allo scopo di ricomporre l'ordine originario dei pacchetti consegnati disordinatamente dallo strato IP di rete. Non appena il ricevente completa un segmento contiguo al limite inferiore, sposta quest'ultimo in avanti di tanti bytes quanti ne è riuscito a leggere in modo contiguo, ed invia un riscontro con NR pari al più basso numero di byte che ancora non è pervenuto⁶⁷.

Nel caso in cui sia settato il bit URG ⁶⁸, si stanno inviando dati urgenti fuori sequenza, e che non devono rispettare il protocollo a finestra, come ad esempio per recapitare un segnale di interrupt relativo ad una sessione Telnet per terminare una applicazione remota.

Controllo di errore Trascorso un certo tempo (detto *timeout*) nell'attesa di un riscontro, il trasmittente ritiene che alcuni pacchetti sono andati persi, e li re-invia⁶⁹. Il valore del *timeout* viene calcolato dinamicamente dal TCP in base alle sue misure di *round-trip delay*⁷⁰, ossia del tempo che intercorre in media tra invio di un pacchetto e ricezione del suo riscontro. In questo modo il TCP si adatta alle condizioni di carico della rete ed evita di ri-spedito pacchetti troppo presto o di effettuare attese inutili. In particolare, nel caso di rete congestionata aumenta la frequenza dei pacchetti persi, e valori di *timeout* troppo ridotti potrebbero peggiorare la situazione.

Controllo di flusso Il meccanismo a finestra scorrevole determina, istante per istante, il numero massimo di bytes che possono essere trasmessi verso il destinatario, e pertanto consente al nodo meno veloce di adeguare la velocità di trasmissione alle proprie capacità. La dimensione della finestra può essere variata (su iniziativa del ricevente) nel corso della connessione, in accordo al valore presente nel campo *Finestra* dell'intestazione TCP. Ad esempio, una connessione può iniziare con una dimensione di finestra ridotta, e poi aumentarla nel caso in cui non si verificano errori, la rete sopporti il traffico, ed i nodi abbiano memoria disponibile.

Controllo di congestione Il TCP può usare la sua misura di *round-trip delay* come un indicatore di congestione della rete, e lo scadere di un *timeout* come un segnale del peggioramento della congestione. In tal caso quindi, può essere ridotta la dimensione della finestra di trasmissione, caricando così di meno la rete.

UDP Lo *User Datagram Protocol* è ancora un protocollo di trasporto, che opera senza connessione, e sostituisce il TCP per inviare pacchetti isolati, o serie di pac-

⁶⁷Il riscontro può viaggiare su di un pacchetto già in "partenza" con un carico utile di dati e destinato al nodo a cui si deve inviare il riscontro. In tal caso quest'ultimo prende il nome di PIGGYBACK (*rimorchio*), o *riscontro rimorchiato*.

⁶⁸In tal caso, il campo *Puntatore Urgente* contiene il numero di sequenza del byte che delimita superiormente i dati che devono essere consegnati urgentemente.

⁶⁹Il mancato invio del riscontro può anche essere causato dal verificarsi di un *checksum* errato dal lato ricevente, nel qual caso quest'ultimo semplicemente evita di inviare il riscontro, confidando nella ritrasmissione per timeout.

⁷⁰Con licenza poetica: *il ritardo del girotondo*, che qui raffigura un percorso di andata e ritorno senza soste.

chetti la cui ritrasmissione (se perduti) sarebbe inutile. Ad esempio, è utilizzato nella trasmissione di dati in tempo reale, oppure per protocolli di interrogazione e controllo come il DNS.

8.6.1.5 IP

L'*Internet Protocol* costituisce l'ossatura della rete internet, realizzandone i servizi di rete ed interfacciando le diverse sottoreti a cui sono connessi i nodi. Le sue principali funzioni sono pertanto l'indirizzamento, l'instradamento e la variazione della dimensione⁷¹ dei pacchetti prodotti dal TCP o da altri protocolli degli strati superiori. Ogni pacchetto è inviato come un messaggio indipendente, in modalità datagramma; la consegna dei datagrammi non è garantita⁷², e questi possono essere persi, duplicati o consegnati fuori sequenza.

L'IP riceve dallo strato superiore (il TCP od un altro protocollo) un flusso di byte suddivisi in pacchetti, a cui si aggiunge l'indirizzo IP di destinazione; tale flusso è utilizzato per riempire un proprio buffer di dimensione opportuna, che quando pieno (od al termine del pacchetto ricevuto *dall'alto*) è *incapsulato* aggiungendo una intestazione (*l'header*) che codifica la segnalazione dello strato di rete realizzato dal protocollo IP.

L'**intestazione IP** contiene le informazioni mostrate alla figura successiva.

1	5	9	17	20	32
VER	HLEN	TOS	TLEN		
Identificazione			Flags	Frag. Offset	
TTL	Protocollo		Checksum		
IP Address Sorgente					
IP Address Destinazione					
Opzioni			Riempimento		

Il campo *VER* indica quale versione si sta utilizzando, e permette sperimentazioni e miglioramenti senza interrompere il servizio. *HLEN* e *TLEN* indicano rispettivamente la lunghezza dell'header e di tutto il pacchetto, mentre *TOS* codifica una *Type of Service* per differenziare ad esempio la QoS⁷³

richiesta. L'*identificazione* riporta lo stesso valore per tutti i frammenti di uno stesso datagramma, mentre l'*Offset di frammento* indica la posizione del frammento nel datagramma (con frammenti di dimensione multipla di 8 byte).

Solo 2 dei tre bit di *Flag* sono usati, *DF* (*Don't Fragment*) per richiedere alla rete di non frammentare il datagramma, e *MF* (*More Fragments*) per indicare che seguiranno altri frammenti. Il *TTL* (Time To Live) determina la massima permanenza del pacchetto nella rete⁷⁴, il *protocollo* indica a chi consegnare il datagramma all'arrivo (ad es. TCP o UDP), e *Checksum* serve per verificare l'assenza di errori nell'header⁷⁵.

⁷¹L'IP può trovarsi a dover inoltrare i pacchetti su sottoreti che operano con dimensioni di pacchetto inferiori. Per questo, deve essere in grado di frammentare il pacchetto in più datagrammi, e di ricomporli nell'unità informativa originaria all'altro estremo del collegamento.

⁷²Si suppone infatti che le sottoreti a cui sono connessi i nodi non garantiscano affidabilità. Ciò consente di poter usare sottoreti le più generiche (incluse quelle affidabili, ovviamente).

⁷³La Qualità del Servizio richiesta per il particolare datagramma può esprimere necessità particolari, come ad esempio il ritardo massimo di consegna. La possibilità di esprimere questa esigenza a livello IP fa parte dello standard, ma per lunghi anni non se ne è fatto uso. L'avvento delle comunicazioni multimediali ha risvegliato l'interesse per il campo TOS.

⁷⁴Lo scopo del TTL è di evitare che si verifichino fenomeni di loop infinito, nei quali un pacchetto "rimbalza" tra due nodi per problemi di configurazione. Per questo, TTL è inizializzato al massimo numero di nodi che il pacchetto può attraversare, e viene decrementato da ogni nodo che lo riceve (e ritrasmette). Quando TTL arriva a zero, il pacchetto è scartato.

⁷⁵In presenza di un frammento ricevuto con errori nell'header, viene scartato tutto il datagramma di cui il frammento fa parte, delegando allo strato superiore le procedure per l'eventuale recupero dell'errore.

Gli *Indirizzi IP* di sorgente e destinazione hanno l'evidente funzione di recapitare correttamente il messaggio, mentre il campo *Opzioni* ha una lunghezza variabile, può essere omesso, e consente ad esempio di richiedere il tracciamento della serie di router attraversati.

Indirizzamento e Routing A pagina 195 si è anticipata la relazione che lega la parte iniziale dell'indirizzo IP ad una determinata sottorete, in modo da partizionare i 2^{32} indirizzi su di una gerarchia a due livelli e delegare la consegna all'host finale ad uno o più router responsabili di servire la sottorete⁷⁶. In realtà la gerarchia presenta una ulteriore suddivisione, dettata sia da esigenze amministrative che funzionali.

I bit più significativi dell'indirizzo IP identificano 5 diversi gruppi (o *classi*) di indirizzi, descritti dalla seguente tabella:

Inizio IP addr	Classe	bit rete/nodo	N. reti	N. nodi per rete
0	A	7/24	128	16 777 216
10	B	14/16	16 384	65 536
110	C	21/8	2 097 152	256
1110	D	28 bit di indirizzo multicast per 268 435 456 canali		
11110	E	27 bit per usi futuri e ricerca		

Quando una organizzazione decide di essere presente in internet, richiede l'assegnazione di un lotto di indirizzi IP ad apposti organismi, i quali attribuiscono all'organizzazione un gruppo di indirizzi di classe A, B o C in base al numero di nodi che l'organizzazione prevede di mettere in rete. Una rete in classe B ad esempio è individuata da 14 bit (ossia, assieme ai bit di classe, dai primi due bytes dell'indirizzo IP), e quindi esistono $2^{14} = 16384$ diverse reti in classe B, ognuna con una capacità di $2^{16} = 65536$ diversi nodi. Chi è intestatario di un gruppo di indirizzi, provvede ad assegnarli ai singoli nodi della propria sottorete.

Subnetting e Supernetting Osserviamo ora che la maschera di sottorete presentata a pag. 195 *non* coincide con il gruppo di bit che identifica la classe e la rete: infatti, l'insieme di indirizzi 151.100.x.y corrisponde ad una rete in classe B, mentre la maschera di sottorete 255.255.255.0 individua una sottorete in classe C. Praticamente, la rete in classe B è stata ulteriormente suddivisa (*subnettata*) in 256 sottoreti di classe C, permettendo di realizzare un instradamento gerarchico su due livelli nell'ambito dell'organizzazione intestataria della rete in classe B⁷⁷. L'operazione inversa (detta *supernetting*), ossia quella di aggregare più reti di dimensione ridotta in una di dimensione maggiore, ha senso all'interno del router che instrada il traffico verso l'organizzazione intestataria delle sottoreti, in quanto permette di ridurre la dimensione delle tabelle di routing, che contengono così un solo elemento relativo alla super-rete, anziché un elemento per ogni singola sottorete.

⁷⁶Possiamo portare come analogia un indirizzo civico, a cui il postino consegna la corrispondenza, che viene poi smistata ai singoli condomini dal portiere dello stabile. Il servizio postale, così come la rete Internet, non ha interesse di sapere come sono suddivise le sottoreti delle diverse organizzazioni, ed i router instradano i pacchetti IP in base alla parte "rete" dell'indirizzo, delegando ai router della rete di destinazione il completamento dell'instradamento.

⁷⁷In questo caso, l'Università di Roma "La Sapienza" è intestataria della rete 151.100.

Classless Interdomain Routing - cidr Nella prima metà degli anni '90 apparve evidente che il partizionamento degli indirizzi nelle tre classi A, B e C non era rispondente alle richieste dell'utenza; accadeva infatti che le reti in classe C erano troppo "piccole", mentre quelle in classe B rischiavano di esaurirsi a breve, pur essendo sfruttate molto poco⁷⁸. Per questo motivo, è stata rimossa la suddivisione rigida nelle tre classi, e si è sistematicamente applicato il principio del supernetting. In pratica, si è ridefinita la maschera di sottorete, come una sequenza di *uni* allineata a sinistra, permettendo così di definire reti di dimensione pari a una potenza di due qualsiasi. Come risultato, ora una sottorete è identificata da una coppia indirizzo/maschera del tipo (ad es.) 172.192.0.0/12, che rappresenta tutti 2^{20} indirizzi che vanno da 172.192.0.0 a 172.207.255.255, che hanno i 12 bit più elevati uguali a 101011001100: questa sequenza prende il nome di *prefisso* della rete. In definitiva quindi, la maschera è espressa come il numero di bit più significativi in comune a tutti i nodi della sottorete.

Longest Match Un router decide su che porta instradare un pacchetto IP in base al confronto tra l'indirizzo di destinazione e tutti i prefissi presenti nella tabella di routing, associati ciascuno alla "migliore" porta di uscita verso la sottorete definita dal prefisso. Nel caso in cui si verifichi più di una uguaglianza, si sceglie l'instradamento caratterizzato dal *maggior numero* di bit coincidenti, ossia relativo al prefisso *più lungo*. Infatti, in tal modo viene preferita la direzione *più specifica* verso la destinazione finale. In assenza di uguaglianze invece, il pacchetto è inoltrato in base ad una *default route*, che tipicamente rimanda la decisione ad un router "gerarchicamente più elevato"⁷⁹.

Sistemi Autonomi e Border Gateway Vi sono router collegati direttamente con le LAN, e configurati per instradare correttamente i pacchetti diretti a destinazioni locali. Vi sono poi router collegati solo ad altri router, che *apprendono* gli instradamenti verso le reti locali mediante appositi *protocolli di routing* che consentono ai router di primo tipo di *pubblicizzare* (ADVERTISE) le reti raggiungibili direttamente, ed ai router del secondo tipo di fare altrettanto nei confronti dei loro pari.

L'insieme di sottoreti (e router, nodi e DNS) gestite da una stessa organizzazione prende il nome di *Autonomous System* (AS), e nel suo ambito sono attivi protocolli di routing detti *Interior Gateway Protocols* (IGP), che distribuiscono le informazioni di raggiungibilità interna. Alcuni router di uno stesso AS svolgono il ruolo di *Border Gateway* (BG), e comunicano con i BG di altri AS mediante appositi *Exterior Gateway Protocols* (EGP), pubblicizzando all'esterno le proprie sottoreti, apprendendo dagli altri BG la raggiungibilità delle sottoreti esterne, e distribuendo tali informazioni ai router interni. Un compito particolare dell'EGP, è quello di attuare qualche politica nei confronti del *traffico di transito* tra due AS diversi dall'AS di cui il BG fa parte: in tal caso, il protocollo prende il nome di *Border Gateway Protocol* (BGP).

L'applicazione del CDIR comporta, per ogni scambio di informazioni di routing, la necessità di aggregare o disaggregare i prefissi di sottorete, in modo da mantenere al minimo la dimensione delle tabelle di instradamento.

⁷⁸Ad esempio, organizzazioni con poco più di un migliaio di nodi erano costrette a richiedere una intera classe B con capacità di 65536 nodi.

⁷⁹Sebbene la topologia di Internet possa essere qualunque, nella pratica esistono dei *carrier* internazionali che svolgono la funzione di *backbone* (spina dorsale) della rete, interconnettendo tra loro i continenti e le nazioni.

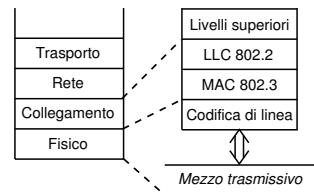
Multicast Tornando all'esame della tabella di pag. 201, in cui la classe E costituisce evidentemente una "riserva" di indirizzi per poter effettuare sperimentazioni, la classe D individua invece dei canali *multicast*⁸⁰. Quando un nodo decide di aderire ad un canale multicast, invia un messaggio⁸¹ in tal senso al proprio router più vicino, che a sua volta si occupa di informare gli altri router. Questi ultimi provvederanno quindi, qualora osservino transitare un pacchetto avente come destinazione un canale multicast, ad instradarlo verso l'host aderente. In presenza di più nodi nella stessa sottorete in ascolto dello stesso canale, solo una copia dei pacchetti attraverserà il router: il traffico multicast⁸² evita infatti di aprire una connessione dedicata per ogni destinatario, ma si suddivide via via nella rete solo quando i destinatari sono raggiungibili da vie diverse.

8.6.1.6 Ethernet

Ci occupiamo qui di un caso particolare di realizzazione dei primi due livelli del modello ISO-OSI. Come anticipato a pag. 195, molti nodi di Internet sono univocamente individuati da un indirizzo (Ethernet) di 6 byte che, sebbene sia unico al mondo, viene usato solamente nell'ambito della LAN a cui il nodo è connesso, in quanto la distribuzione mondiale degli indirizzi Ethernet è casuale⁸³: se infatti questi fossero usati come indirizzi a livello di rete, le tabelle di instradamento dovrebbero essere a conoscenza di *tutti* i nodi esistenti⁸⁴. Puntualizziamo inoltre che un nodo di Internet può essere connesso alla rete anche per via telefonica, o con svariati altri metodi; ci limitiamo qui a descrivere il caso delle LAN Ethernet, peraltro particolarmente diffuso.

Ethernet individua un particolare tipo di pacchetto dati, adottato inizialmente dalla Xerox, adatto ad incapsulare dati provenienti da protocolli diversi. Successivamente, il formato è stato standardizzato dall'IEEE, e per ciò che ci interessa le specifiche sono quelle identificate dalle sigle 802.2 (LOGICAL LINK CONTROL, LLC) e 802.3 (CARRIER SENSE MULTIPLE ACCESS - COLLISION DETECT, CSMA/CD).

La figura mostra il legame tra queste due sigle e gli strati del modello; lo strato MAC in cui si realizza il CSMA/CD individua il MEDIA ACCESS CONTROL. Il mezzo trasmissivo è un cavo, coassiale o coppia simmetrica, sul quale sono collegati tutti i nodi della LAN, che si *contendono* il mezzo trasmissivo, in quanto vi può trasmettere solo un nodo per volta. Inoltre, tutti i nodi sono in ascolto sullo stesso mezzo per ricevere i pacchetti a loro destinati, riconoscibili per la presenza del proprio indirizzo Ethernet nel campo destinazione. Un pacchetto Ethernet può inoltre riportare un indirizzo di destinazione particolare, detto di *Broadcast*, che obbliga *tutti* i nodi presenti alla ricezione del pacchetto.



⁸⁰Il termine *multicast* è ispirato alle trasmissioni *broadcast* effettuate dalle emittenti radio televisive.

⁸¹Mediante il protocollo IGMP (*Internet Group Management Protocol*) che opera sopra lo strato IP, ma (a differenza del TCP) fa uso di datagrammi non riscontrati, similmente all'UDP ed all'ICMP.

⁸²Data l'impossibilità a stabilire un controllo di flusso con tutti i destinatari, il traffico multicast viaggia all'interno di pacchetti UDP.

⁸³E rappresenta quindi ciò che viene detto uno *spazio di indirizzi piatto* (FLAT ADDRESS SPACE).

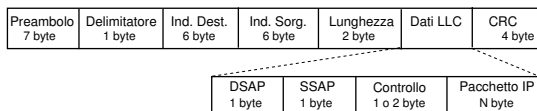
⁸⁴Al contrario, il partizionamento dell'indirizzo IP in rete+nodo permette di utilizzare tabelle di routing di dimensioni gestibili.

Address Resolution Protocol - arp Quando un pacchetto IP giunge ad un router, e l'indirizzo IP indica che il destinatario è connesso ad una delle LAN direttamente raggiungibili dal router⁸⁵, questo invia su quella LAN un pacchetto *broadcast*, su cui viaggia una richiesta ARP (ADDRESS RESOLUTION PROTOCOL), allo scopo di individuare l'indirizzo Ethernet del nodo a cui è assegnato l'indirizzo IP di destinazione del pacchetto arrivato al router. Se tale nodo è presente ed operativo, riconosce che la richiesta è diretta a lui, ed invia un pacchetto di risposta comunicando il proprio indirizzo Ethernet, che viene memorizzato dal router in una apposita tabella⁸⁶.

Operazioni simili sono svolte da ognuno dei nodi della LAN, ogni volta che debbano inviare un pacchetto ad un altro nodo direttamente connesso alla stessa rete locale. Se al contrario l'IP di destinazione non fa parte della stessa LAN, il pacchetto è inviato al *default gateway*.

Formato di pacchetto Il pacchetto Ethernet è generato dall'LLC e dal MAC, ognuno dei quali incapsula il pacchetto IP con le proprie informazioni di protocollo.

Nella figura seguente è mostrato il risultato finale delle operazioni. In testa troviamo 7 byte di *preambolo*, necessario a permettere la sincronizzazione dell'orologio del ricevente con quello in trasmissione; dato che la sincronizzazione richiede un tempo non noto a priori, un byte di *flag* segnala l'inizio del pacchetto. Troviamo quindi gli *indirizzi Ethernet* di sorgente e destinazione, due byte che indicano la *lunghezza* della restante parte del pacchetto, e quindi l'incapsulamento dei dati prodotti dall'LLC. In fondo, sono presenti 4 byte che realizzano il *controllo di errore*.



D'altra parte, per ovviare al numero limitato di possibili incapsulamenti esprimibili utilizzando solo gli 8 bit dei campi SAP, è stata introdotta una estensione all'LLC denominata SNAP (*Subnetwork Access Protocol*)⁸⁷ che pone i campi DSAP, SSAP e controllo pari a 0xAAAA03, a cui aggiunge altri 5 bytes, dei quali i primi tre sono denominati OUI (*Organizationally Unique Identifier*) che, se posti tutti a zero, stabiliscono che i due byte seguenti (indicati come *protocol ID*) debbano essere interpretati come un codice *Ethertype*⁸⁸, lo stesso usato nel formato *Ethernet II* discusso appresso, permettendo quindi di specificare finalmente il protocollo incapsulato.

⁸⁵ Ad ogni porta del router è associata una coppia sottorete/maschera (vedi pag. 195) che descrive l'insieme degli indirizzi direttamente connessi alla porta. La verifica di raggiungibilità (o *adiacenza*) è attuata mettendo in AND l'IP di destinazione con le maschere, e confrontando il risultato con quello dell'AND tra le maschere e gli indirizzi delle sottoreti collegate.

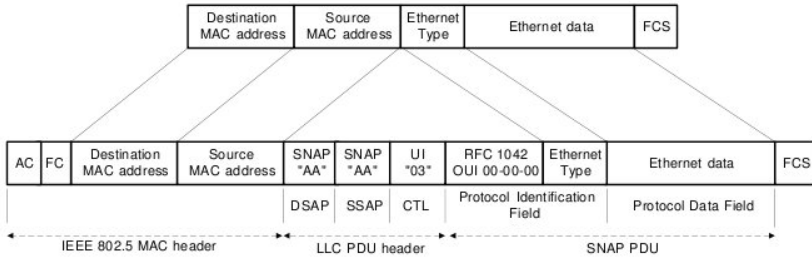
⁸⁶ Dato che i nodi possono essere spostati, possono cambiare scheda di rete e possono cambiare indirizzo IP assegnatogli, la corrispondenza IP-Ethernet è tutt'altro che duratura, ed ogni riga della tabella ARP indica anche quando si sia appresa la corrispondenza, in modo da poter stabilire una scadenza, ed effettuare nuovamente la richiesta per verificare se sono intervenuti cambiamenti topologici.

Se il nodo ha cambiato IP, ma non il nome, sarà il TTL del DNS (mantenuto aggiornato per il dominio del nodo) a provocare il rinnovo della richiesta dell'indirizzo.

⁸⁷ http://en.wikipedia.org/wiki/Subnetwork_Access_Protocol

⁸⁸ <http://en.wikipedia.org/wiki/EtherType>

Infine, viene molto frequentemente usato un formato di trama ancora diverso, detto *Ethernet II* o DIX⁸⁹, che corrisponde a quello definito inizialmente prima che l'IEEE emettesse gli standard della serie 802, e che usa i 16 bit del campo *lunghezza* per indicare direttamente l'*Ethertype* della SDU incapsulata, ed omette i campi DSAP, SSAP e di controllo. In tal caso, il campo *lunghezza* rappresenta un numero più grande di 0x0600, maggiore della massima lunghezza prevista, e ciò fa sì che venga interpretato come codice *Ethertype*, e che se sono incapsulati pacchetti IP, vale 0x0800. La figura seguente, tratta dal documento dell'IEEE, illustra la corrispondenza tra i campi del formato SNAP e DIX.



Collisione Come anticipato, il mezzo trasmissivo è in comune con tutti i nodi, e dunque si è studiata una particolare soluzione il cui nome CSMA/CD indica che l'*Accesso Multiplo* avviene in due fasi: prima di trasmettere, si ascolta se non vi sia già qualcuno che trasmette (CARRIER SENSE), e durante la trasmissione, si verifica che nessun altro stia trasmettendo contemporaneamente (COLLISION DETECT). Pertanto, ogni nodo che debba trasmettere si pone prima in ascolto, e se osserva che già vi sono trasmissioni in corso, attende un tempo casuale e riprova. Quando trova il mezzo "libero", inizia a trasmettere, ma contemporaneamente verifica che nessun altro inizi a sua volta la trasmissione: questo fatto può accadere, in virtù del tempo di propagazione⁹⁰ non nullo, e determina un periodo (detto di *contesa*, e che dipende dalla massima lunghezza del cavo) entro il quale un nodo può erroneamente credere che nessun altro stia trasmettendo.

Qualora sia rilevata una contesa, i due nodi smettono di trasmettere, e riprovano solo dopo una attesa di durata casuale.

Trasmissione Il segnale relativo al pacchetto Ethernet viene trasmesso adottando una codifica di linea di tipo Manchester differenziale. La configurazione con tutti i nodi collegati su di uno stesso cavo è detta *a bus*, e sono state coniate apposite sigle per identificare il tipo di connessione, come ad esempio 10BASE5 e 10BASE2, relative al

⁸⁹Vedi http://en.wikipedia.org/wiki/Ethernet_II_framing. La sigla DIX deriva dalle iniziali delle aziende che l'hanno definito, ossia DEC, Intel and Xerox

⁹⁰Su di un cavo coassiale *tick* da 50 Ω, la velocità di propagazione risulta di 231·10⁶ metri/secondo. Su di una lunghezza di 500 metri, occorrono 2.16 μsec perché un segnale si propaghi da un estremo all'altro. Dato che è permesso di congiungere fino a 5 segmenti di rete per mezzo di ripetitori, e che anch'essi introducono un ritardo, si è stabilito che la minima lunghezza di un pacchetto Ethernet debba essere di 64 byte, che alla velocità di trasmissione di 10 Mbit/sec corrisponde ad una durata di 54.4 μsec, garantendo così che se si è verificata una collisione, le due parti in causa possano accorgersene.

collegamento di banda base a 10 Mbps, su cavo *tick* e *thin*⁹¹, con estensione massima 500 e 200 metri⁹².

8.6.1.7 Fast e Gigabit Ethernet

Mentre si proponeva ATM come una soluzione idonea per quasi tutti gli ambiti, la tecnologia Ethernet ha incrementato la velocità trasmissiva di un fattore pari a mille, e si propone sempre più come soluzione generalizzata.

Fast Ethernet Nel 1995 è stato definito lo standard IEEE 802.3u detto *Fast Ethernet*, che eleva la velocità di trasmissione a 100 Mbps ed impiega due diversi cavi UTP⁹³ per le due direzioni di trasmissione, rendendo eventualmente la comunicazione *full-duplex*⁹⁴. In quest'ambito sono definiti i sistemi 10BASE-T e 100BASE-T, relativi all'uso del cavo UTP anziché di un coassiale, e prevedono una topologia *a stella* per la LAN, realizzata utilizzando una unità centrale (detta HUB=*mozzo di ruota*) da cui si dipartono tanti cavi, ognuno che collega un unico nodo. Nel caso di un HUB economico, questo svolge solo le funzioni di ripetitore (ritrasmette tutto su tutte le sue porte) e dunque le collisioni possono ancora verificarsi.

LAN Switch D'altra parte, i dispositivi detti BRIDGE o LAN SWITCH *apprendono* dai pacchetti in transito gli indirizzi ethernet dei nodi collegati alle porte, ed evitano di ritrasmettere i pacchetti sulle porte dove *non si trova* il destinatario. Dato che gran parte del traffico è inviato verso il *gateway* della LAN, lo SWITCH apprende in fretta su che porta questo si trovi, cosicché tutti i pacchetti destinati all'esterno non sono ritrasmessi sugli altri rami della LAN, ed il traffico tra i nodi connessi allo SWITCH non si propaga al resto della LAN.

La lunghezza massima dei collegamenti è ora ridotta a 100 metri, per il motivo che un pacchetto di dimensione minima di 64 byte trasmesso a 100 Mbps, impiega un tempo che è $\frac{1}{10}$ di quello relativo alla velocità di 10 Mbps, e quindi per consentire la detezione di collisione, si è dovuta ridurre di pari misura la massima distanza tra nodi trasmettenti.

Dominio di broadcast e VLAN Anche se i dispositivi BRIDGE e SWITCH evitano di trasmettere traffico verso le porte diverse da quella di destinazione, alcuni pacchetti devono comunque essere ritrasmessi in tutte le direzioni: si tratta del traffico *broadcast*, diretto verso un ben preciso insieme di indirizzi ethernet, ed usato per funzioni di coordinamento tra i nodi della LAN, come ad esempio *l'esplorazione delle risorse di rete*. Il traffico broadcast non esce dalla LAN, arrestandosi al router di livello IP; una eccessiva presenza di traffico broadcast può però pregiudicare l'efficienza sia della LAN che dei suoi nodi, oltre che produrre problemi di sicurezza; per questo si è sviluppata la possibilità di assegnare le porte di uno switch a diversi *domini di broadcast*, detti LAN *virtuali* (VLAN), che non scambiano traffico, realizzando di fatto molteplici LAN con

⁹¹TICK = *duro* (grosso), THIN = *sottile*. Ci si riferisce al diametro del cavo.

⁹²Le sigle indicano infatti la velocità, se in banda base o meno, e la lunghezza della tratta.

⁹³UNSHIELDED TWISTED PAIR (UTP), ossia la coppia ritorta non schermata.

⁹⁴La trasmissione *full-duplex* si instaura quando entrambe le interfacce agli estremi ne sono capaci. Una interfaccia *half-duplex* deve invece gestire situazioni *interne* di collisione, quando un pacchetto uscente da un nodo si scontra con uno entrante.

uno stesso cablaggio. Per interconnettere le LAN, occorre attraversare un dispositivo router.

Gigabit Ethernet Nel giugno 1998 viene standardizzato l'IEEE 802.3z, che porta ad 1 Gbps la velocità di trasmissione delle trame Ethernet, rimpiazzando lo strato di codifica di linea dell'802.3 con i due strati inferiori dell'ANSI X3T11 *Fiber Channel*. In questo modo, si mantiene la compatibilità con gli strati LLC e MAC di Ethernet, mentre la trasmissione avviene su fibra ottica o su cavo in accordo alla tabella seguente.

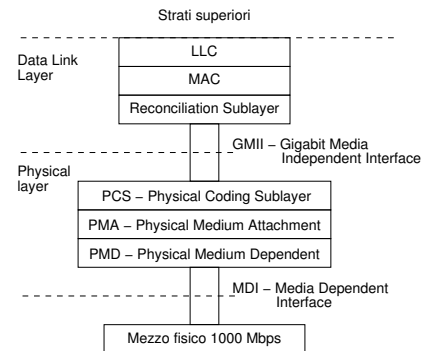
media	distanza	mezzo	sorgente
1000BASE-SX	300 m	f.o. multimodo ϕ 62.5 μ m	laser 850 nm
	550 m	f.o. multimodo ϕ 50 μ m	laser 850 nm
1000BASE-LX	550 m	f.o. multimodo ϕ 50 o 62.5 μ m	laser 1300 nm
	3000 m	f.o. monomodo ϕ 9 μ m	laser 1300 nm
1000BASE-CX	25 m	cavo STP (<i>shielded twisted pair</i>)	
1000BASE-T	25-100 m	4 coppie di cavo UTP	

Packet bursting Dato che ora la velocità di trasmissione è 10 volte quella del fast Ethernet, la compatibilità con il MAC CSMA/CD richiederebbe di ridurre la massima lunghezza del collegamento a 10 metri. Al contrario, è stata aumentata la durata minima di una trama portandola a 512 byte, in modo da aumentare la durata della trasmissione e garantire la detezione di collisione. In effetti, il MAC ethernet continua a produrre pacchetti di durata minima 64 byte, e questi sono riempiti (PADDED) fino a 512 byte con una *carrier extension* di simboli speciali. Questa operazione è particolarmente inefficiente se i pacchetti da 64 byte sono frequenti; in tal caso si attua allora il *packet bursting* che, esauriti i 512 byte minimi realizzati come indicato, accoda gli ulteriori pacchetti nello stesso burst trasmissivo, fino ad una lunghezza di 1500 byte.

Architettura La figura seguente mostra la pila protocollare per Gigabit Ethernet. La GMII permette di usare lo strato MAC con qualunque strato fisico, ed opera sia in full-duplex che in half-duplex, alle velocità di 10, 100 e 1000 Mbps, mediante due percorsi dati (Tx e Rx) da 8 bit, più due segnali di strato per indicare presenza di portante e detezione di collisione, che sono mappati dal RS nelle primitive riconosciute dallo strato MAC preesistente.

Lo strato fisico è suddiviso in tre sottolivelli. Il PCS fornisce una interfaccia uniforme al RS per tutti i media. Provvede alla conversione 8B/10B tipica del *Fiber Channel*, che rappresenta gruppi di 8 bit mediante *code group* da 10 bit, alcuni dei quali rappresentano i simboli, ed altri sono codici di controllo, come quelli usati per la *carrier extension*. Il PCS genera inoltre le indicazioni sulla portante e sulla collisione, e gestisce la auto-negoziazione sulla velocità di trasmissione e sulla bidirezionalità del media.

Il PMA provvede alla conversione parallelo-serie e viceversa, mentre il PMD definisce l'MDI, ossia la segnalazione di strato fisico necessaria ai diversi media, così come il tipo di connettore.



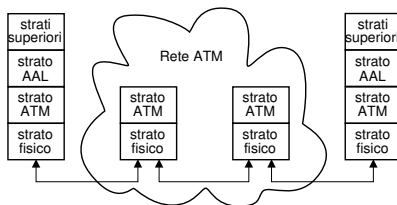
Ripetitore full-duplex e controllo di flusso Qualora tutte le porte di un ripetitore siano di tipo full-duplex, non può più verificarsi contesa di accesso al mezzo, e la contesa avviene all'interno del ripetitore, che (non essendo un SWITCH) copia tutte le trame in ingresso (debitamente bufferizzate in apposite code) in tutte le code associate alle porte di uscita. Pertanto, la lunghezza massima dei collegamenti non è più dettata dalla necessità di rilevare collisioni, ma dalle caratteristiche del mezzo trasmissivo. D'altra parte, possono verificarsi situazioni di *flooding* delle code di ingresso; il comitato IEEE 802.3x ha quindi definito un meccanismo di controllo di flusso, che mette in grado i ripetitori (e gli switch) di richiedere ai nodi connessi la sospensione temporanea della trasmissione.

10 Gigabit Ethernet Nel 2002 viene definito lo standard IEEE 802.3ae, che stabilisce le modalità operative di un collegamento Ethernet operante solo in full duplex su fibra ottica. Lo standard prevede di interoperare con la trasmissione SONET/SDH.

8.6.2 Rete ATM

La sigla ATM sta per *Asynchronous Transfer Mode*, ed identifica una particolare rete progettata per trasportare indifferentemente traffico di diversa natura, sia di tipo dati che real-time⁹⁵, che per questo motivo è indicata anche come B-ISDN⁹⁶. Il suo funzionamento si basa sul principio della *commutazione di cella* (CELL SWITCHING), dove per cella si intende un pacchetto di lunghezza fissa di 53 byte. I primi 5 byte delle celle contengono un identificativo di connessione, ed il loro instradamento avviene mediante dei circuiti virtuali. La commutazione delle celle tra i nodi di rete ha luogo in maniera particolarmente efficiente, e questa è una delle caratteristiche più rilevanti dell'ATM.

Architettura La rete ATM viene indicata anche come una *Overlay Network*, in quanto operativamente si *sovrappone* ai livelli inferiori di una rete esterna.



Dal canto suo, ATM è strutturata sui tre strati funzionali di adattamento (AAL), di commutazione ATM, e fisico. Mentre i nodi ai bordi della rete devono realizzare tutti e tre gli strati, i nodi interni svolgono solo le funzioni attuate da quelli inferiori. La tabella 8.1 riporta le principali funzioni svolte dai tre strati, e pone in evidenza come in uno stesso

strato siano identificabili diverse sotto-funzioni.

Strato fisico Il mezzo primario di trasmissione (con cui è in contatto il sotto-strato PM) per ATM è la fibra ottica, in accordo alla struttura di trama dell'SDH/SONET, per la quale sono state standardizzate le velocità di 1.5 e 2 Mbps (DS1/E1), 155 Mbps

⁹⁵ Per traffico real-time si intende sia quello telefonico, sia più in generale quello di natura multimediale.

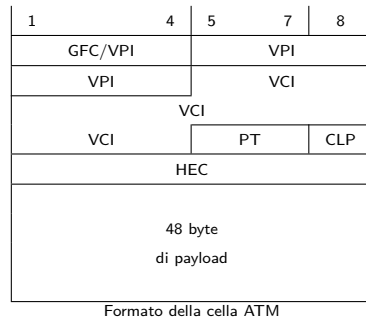
⁹⁶ Siamo alla fine degli anni '80, e la definizione *Integrated Service Data Network* (ISDN) si riferisce ad una rete in grado di permettere, oltre al normale trasporto dei dati, anche servizi di rete. La rete ISDN era però limitata ad una velocità massima (presso l'utente) di 2 Mbps, e per questo venne chiamata *narrow-band ISDN* (N-ISDN). A questa, avrebbe fatto seguito la *broad-band ISDN* (B-ISDN) che ha poi dato luogo alla definizione dell'ATM.

strato	sotto-strato	funzioni
ATM Adaptation Layer (AAL)	<ul style="list-style-type: none"> • Convergenza (CS) • Segmentazione e Riasssemblaggio (SAR) 	Definisce il servizio offerto agli strati superiori Suddivide i dati in modo compatibile con la dimensione di cella, e li ricostruisce in ricezione
ATM layer		Moltiplicazione e demoltiplicazione delle celle Traslazione delle etichette VPI/VCI Generazione/estrazione dell'HEADER della cella Gestione del controllo di flusso GFC
Physical Layer (PL)	<ul style="list-style-type: none"> • Convergenza di trasmissione (TC) • Mezzo Fisico (PM) 	Delimitazione delle celle Inserimento celle IDLE per adattamento velocità Generazione e verifica dell'HEC (controllo di errore) Generazione della trama di trasmissione Temporizzazione e sincronizzazione Gestione del mezzo

Tabella 8.1: Stratificazione delle funzioni in una rete ATM

(OC3) e 622 Mbps (OC12c). La velocità di 155 Mbps è disponibile anche su FIBRE CHANNEL, e su cavo ritorto, mentre la velocità di 100 Mbps è disponibile su FDDI. Infine, sono previste anche velocità di interconnessione di 139, 52, 45, 34 e 25 Mbps.

In funzione del mezzo trasmissivo, può variare la *struttura di trama*⁹⁷ (mostrata in figura) in cui vanno inserite le celle. Il quinto byte della intestazione di cella, contiene l'*Header Error Code* (HEC) calcolato sui 4 byte precedenti, che viene usato in ricezione per rivelare due errori e correggerne uno⁹⁸. Nel caso in cui la sorgente produca dati a velocità inferiore a quella del collegamento, sono inserite celle aggiuntive di tipo IDLE, rimosse al ricevitore⁹⁹. Infine, la *delimitazione delle celle* è attuata in ricezione in base alla correlazione tra i primi quattro byte dell'header, ed il campo HEC dello stesso.



Strato ATM Mentre lo strato fisico si occupa di trasmettere e ricevere celle, lo strato ATM si occupa di elaborarle. Nei nodi di frontiera, le celle sono multiplate e demultiplate, mentre dentro la rete, sono commutate tra gli ingressi e le uscite.

Nei primi quattro byte dell'header di cella trova posto l'*etichetta* necessaria a realizzare il trasferimento a circuito virtuale; questa etichetta è suddivisa in due campi, il *Virtual Path Identifier* (VPI) ed il *Virtual Channel Identifier* (VCI)¹⁰⁰.

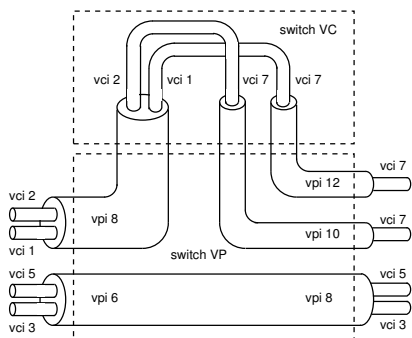
⁹⁷ Sono definite due tipi di *interfaccia utente-network* (UNI): quella SDH/SONET, in cui le celle sono inserite nel *payload* della trama SDH, e quella CELL-BASED, che prevede un flusso continuo di celle. Mentre nel primo caso il bit rate lordo comprende l'*overhead* di trama, nel secondo comprende la presenza di celle di tipo *Operation and Maintenance* (OAM).

⁹⁸ Nel primo caso la cella viene scartata, mentre nel secondo inoltrata correttamente. La presenza di più di due errori, provoca un errato inoltro della cella.

⁹⁹ Le celle IDLE sono riconoscibili in base ad una particolare configurazione dei primi 4 byte dell'header, così come avviene per le celle OAM, nonché per altri tipi particolari di cella, che trasportano la segnalazione degli strati superiori.

¹⁰⁰ Mentre per VCI sono riservati 16 bit, per VPI si usano 12 bit all'interno della rete, e 8 bit ai

Il motivo della suddivisione risiede nella possibilità di raggruppare logicamente diversi circuiti virtuali che condividono lo stesso percorso nella rete. Nei collegamenti di cui è composto il percorso comune, viene usato uno stesso vpi per tutte le celle, mentre le diverse connessioni su quel percorso sono identificate mediante diversi vci. L'instradamento congiunto delle celle con uguale vpi è effettuato nei nodi (vp switch), che si occupano solo¹⁰¹ di scambiare il VPI delle celle, e di porle sulla porta di uscita corretta, come indicato dalle tabelle di instradamento.



La creazione delle tabelle di instradamento può essere di tipo manuale, dando luogo ad una *Permanent Virtual Connection* (PVC), oppure può essere il risultato di una richiesta estemporanea, dando luogo ad una *Switched Virtual Connection* (SVC)¹⁰⁶; L'OGGETTO DELLA RICHIESTA PUÒ ESSERE UNA VCC od una VPC, ed in questo secondo caso la VPC verrà usata per tutte le VCC future tra i due nodi.

Classi di traffico e Qualità del Servizio (QoS) Nella fase di *setup*, sono attuate delle verifiche dette *Connession Admission Control* (CAC) per assicurarsi che la nuova connessione non degradi le prestazioni di quelle già in corso, nel qual caso la chiamata è rifiutata. La sua accettazione determina invece la stipula tra utente e rete di un *Traffic Contract* a cui la sorgente si deve attenere. Nel corso della trasmissione, i nodi ATM verificano che le caratteristiche del traffico in transito nelle VCC siano conformi al rispettivo contratto, svolgendo un *Usage Parameter Control* (UPC) detto anche *policing*¹⁰⁷. Prima di proseguire, forniamo però alcune definizioni.

suoi bordi, riservando 4 bit indicati come *Generic Flow Control* (GFC) per regolare il flusso delle sorgenti.

¹⁰¹Questa semplificazione del lavoro di instradamento, quando confrontata con quello relativo ad una rete IP, è all'origine della vocazione *fast switching* della rete ATM. Per di più, permette la realizzazione *hardware* dei circuiti di commutazione. D'altra parte, mentre per IP l'instradamento avviene al momento della trasmissione, in ATM avviene durante il *set-up* della connessione, quando le tabelle di instradamento sono inizializzate.

¹⁰²Nel caso in cui venga invece scambiato solo il VCI, si ottiene uno switch VC puro.

¹⁰³La rete ATM assicura la consegna delle celle di una stessa VCC nello stesso ordine con cui sono state trasmesse, mentre non assicura l'ordinamento per le celle di una stessa VPC.

¹⁰⁴Può accadere infatti di incontrare uno switch VC puro, in cui è scambiato solo il VCI, ed al quale fanno capo due diverse VCC.

¹⁰⁵I nodi di ingresso ed uscita sono indicati come *ingress* ed *egress* nella terminologia ATM.

¹⁰⁶Nella richiesta di una SVC, l'utente invia i messaggi di *setup* su di una particolare (*well known*) coppia VPI/VCI=0/5. In generale, le prime 32 VCI di ogni VPI sono riservate per propositi di controllo. In queste, sono contenuti dei messaggi di segnalazione che aderiscono alle specifiche Q.2931, che fanno parte di *User Network Interface* (UNI) 3.1, e che sono un adattamento di Q.931 per N-ISDN. Le specifiche UNI 4.0 prevedono la negoziazione della QoS, e la capacità di richiedere una SVC per una VPC.

¹⁰⁷Letteralmente: POLIZIOTTAMENTO. Il controllo può anche essere effettuato su di una intera VPC.

Come anticipato, ATM si è sviluppata per trasportare diversi tipi di traffico, classificabili come segue, nei termini dei parametri indicati di seguito:

- *Constant Bit Rate* (CBR) identifica il traffico real-time come la voce¹⁰⁸ ed il video non codificato;
- *Variable Bit Rate* (VBR) può essere di tipo real time (es. video MPEG) oppure no, ed allora può tollerare variazioni di ritardo (CDV) ma non l'eccessiva perdita di dati (CLR);
- *Available Bit Rate* (ABR) tenta di sfruttare al meglio la banda disponibile. Il contratto prevede la fornitura di un MCR da parte della rete, e le sorgenti sono in grado di rispondere ad una indicazione di congestione, riducendo di conseguenza l'attività;
- *Unspecified Bit Rate* (UBR) condivide la banda rimanente con ABR, ma non gli è riconosciuto un MBR, né è previsto nessun controllo di congestione. Le celle in eccesso sono scartate. Idonea per trasmissioni insensibili a ritardi elevati, e che dispongono di meccanismi di controllo di flusso indipendenti¹⁰⁹.

Le classi di traffico sono descrivibili mediante i parametri

- *Peak Cell Rate* (PCR) applicabile a tutte le classi, ma è l'unico parametro per CBR;
- *Sustainable Cell Rate* (SCR) assieme ai tre seguenti, descrive le caratteristiche di VBR: velocità comprese tra SCR e PCR sono non-conformi, se di durata maggiore di MBS;
- *Minimum Cell Rate* (MCR) caratterizza la garanzia di banda offerta alla classe ABR;
- *Maximum Burst Size* (MBS) descrive la durata dei picchi di traffico per sorgenti VBR.

Il contratto di traffico, mentre impegna la sorgente a rispettare i parametri di traffico dichiarati, vincola la rete alla realizzazione di una *Quality of Service* (QoS), rappresentata dalle grandezze (tra le altre)

- *Cell Transfer Delay* (CTD) assieme alla seguente, è molto importante per la classe CBR;
- *Cell Delay Variation* (CDV) rappresenta la variabilità nella consegna delle celle, dannosa per le applicazioni real-time. La presenza di una CDV elevata può inoltre provocare fenomeni di momentanea congestione all'interno della rete, e può essere ridotta adottando degli *shaper*¹¹⁰, che riducono la variabilità di ritardo a spese un aumento di CTD;
- *Cell Loss Ratio* (CLR) rappresenta il tasso di scarto di celle del collegamento.

Nel caso in cui il policing rilevi che una connessione viola le condizioni contrattuali¹¹¹, può intraprendere svariate azioni, e se può, non scarta immediatamente la cella, ma

¹⁰⁸La classe CBR si presta bene a trasportare traffico telefonico PCM. In questo caso, può trasportare solo gli intervalli temporali realmente occupati.

¹⁰⁹La classe UBR è particolarmente adatta al trasporto di traffico IP, in quanto questo è un protocollo senza connessione, e gli strati superiori (ad es. il TCP) sono in grado di gestire correttamente un servizio di collegamento con perdita di dati.

¹¹⁰Un *sagomatore* è composto in prima approssimazione da un buffer di memoria, il cui ritmo di svuotamento *non è mai* superiore ad un valore costante.

¹¹¹Ad esempio, una CBR supera il proprio PCR, od una VBR oltrepassa il PCR per più tempo di MBS, oppure il traffico generato da una UBR non può essere instradato per l'esaurimento della banda.

provvede comunque a segnalare l'anomalia, ponendo pari ad uno il bit *Cell Loss Priority* (CLP) dell'header. Ciò fa sì che la cella divenga *scartabile*¹¹² in caso di congestione in altri nodi. Un ulteriore campo dell'header, il *Payload Type* (PT), può infine ospitare una *segnalazione in avanti*, che manifesta il fatto che la cella in questione ha subito congestione.

Indirizzamento I nodi di una rete ATM sono identificati da un indirizzo di 20 byte, di diverso significato nei casi di reti private o pubbliche, come indicato dal primo byte

Rete Privata				
AFI	ICD/DCC	HO-DSP	ESI	SEL
Rete Pubblica				
AFI	E.164	HO-DSP	ESI	SEL

(AFI). Nel primo caso, detto *formato NSAP*¹¹³, il DCC o l'ICD sono assegnati dall'ISO, e l'indirizzo del nodo è disposto nei 10 byte indicati come *High-Order*

Domain Specific Part (HO-DSP). I sei byte dell'*End System Identifier* (ESI) sono forniti dal dispositivo connesso ai bordi della rete, e coincidono con il suo indirizzo *Ethernet*: in tal modo la rete comunica un prefisso che identifica il nodo di ingresso, ed il dispositivo lo associa al proprio ESI per forgiare il proprio indirizzo completo. Infine, il byte SEL può essere usato per moltiplicare più entità presso il terminale, ed è ignorato dalla rete.

Nel caso di rete pubblica, il campo HO-DSP è ristretto a 4 byte, e gli 8 byte di E.164 contengono un indirizzo appartenente alla numerazione telefonica mondiale.

Strato di adattamento Come mostrato in tab. 8.1, l'AAL è suddiviso in due componenti, *Segmenting and Reassembly* (SAR) e *Convergence Sublayer* (CS); le funzioni di quest'ultimo sono ulteriormente ripartite tra una *Common Part* (CPCS) ed una *Service Specific cs* (SSCS).

Il compito di AAL è quello di generare i 48 byte del payload per le celle ATM, a partire dalle SDU ricevute, e di ricomporre queste ultime in ricezione, a partire dal risultato della loro demoltiplicazione operata (in base alle etichette VPI/VCI) dallo strato ATM ricevente. Mentre il SAR si interfaccia con lo strato ATM, il CS interagisce con i protocolli superiori, e le esatte operazioni svolte dipendono dalla natura del traffico trasportato: la fig. 8.5 mostra quattro diverse situazioni.

La classe A è un classico caso CBR, ed in tal caso si adotta un AAL di tipo 1, in cui lo strato CS è assente, ed il SAR utilizza il primo dei 48 byte di cella per inserire informazioni di controllo sull'ordine di consegna, e di ausilio al recupero della temporizzazione di sorgente presso la destinazione.

La classe B (AAL 2) individua sorgenti multimediali a pacchetto, mentre per la C (AAL 3/4 o 5) siamo più tipicamente in presenza di una connessione dati a circuito virtuale. In questa categoria rientra il trasporto di collegamenti X.25 e *frame relay*,

¹¹²Alcune classi di traffico pongono CLP=1 già in partenza, sia per una capacità indipendente di risolvere situazioni di perdita di dati, sia per la diversa natura dei dati che possono inviare, come ad esempio una codifica di segnale in cui alcuni dati possono essere interpolati, mentre altri no. Al contrario, alcune sorgenti confidano molto nel rispetto del proprio CLP=0, come ad esempio nel caso in cui queste inviino pacchetti di dati ben più grandi delle celle ATM, e che sono di conseguenza frammentati in molte unità, ed in presenza di una sola cella mancante, devono ritrasmettere l'intero pacchetto. In quest'ultimo caso, sono state elaborate strategie di *scarto precoce* (EARLY DISCARD) di tutte le celle di un pacchetto, per il quale si è già verificato lo scarto di una cella componente.

¹¹³Il formato NSAP si ispira alla *Network Service Access Point* dell'OSI, e se ne differenzia per aver fuso i campi *Routing Domain* e *Area* in un solo campo HO-DSP, per il quale si è adottata una gerarchia di instradamento basata su di un prefisso mobile, in modo simile al CDIR dell'IP.

A	B	C	D
servizio isocrono		ritardo variabile consentito	
bit rate costante	bit rate variabile		
con connessione			senza connessione
AAL 1	AAL 2	AAL 3/4 o 5	AAL 3/4 o 5

Figura 8.5: Classi di servizio della rete ATM

sia di tipo ABR che UBR. Lo stesso tipo di AAL (3/4 o 5) è infine usato anche per la classe D, in cui rientra pienamente il trasporto di traffico IP su ATM.

Quando il CS di AAL 3/4 riceve una SDU (di dimensione massima $2^{16} - 1$) dagli strati superiori, la allinea ad un multiplo di 32 byte, e vi aggiunge 32 byte in testa ed in coda con informazioni di lunghezza e di controllo di integrità. La CS-PDU risultante è passata al SAR, che la suddivide in blocchi di 44 byte, a cui ne aggiunge 2 in testa e due in coda¹¹⁴, e completa così la serie di 48 byte da passare allo strato ATM. Al contrario, il SAR dell'AAL 5 suddivide la CS-PDU in blocchi da 48 byte e non aggiunge informazioni¹¹⁵, demandando il riconoscimento dell'ultima cella di una stessa CS-PDU ad un bit del campo PT presente nell'header di cella ATM. D'altra parte, la lunghezza della CS-PDU dell'AAL 5 è multipla di 48 byte, aggiungendone un numero appropriato, oltre ai 64 byte di intestazione (ora posta in coda), in cui ora sono presenti anche 8 bit di informazione da utente ad utente.

Ip su ATM classico Allo stesso tempo in cui si diffonde l'uso di ATM tra gli operatori di TLC, il TCP/IP emerge come *lo* standard comune per l'interconnessione tra elaboratori. Sebbene il TCP/IP si appoggi ad *Ethernet* in area locale, per i collegamenti a lunga distanza¹¹⁶ l'ATM presenta indubbi vantaggi come la disponibilità di banda su richiesta, la coesistenza con il traffico di tipo diverso, l'elevata efficienza della commutazione, e la possibilità di raggiungere diverse destinazioni. Una prima soluzione, subito scartata, fu quella nota come *peer model*, in cui i nodi ATM possiedono un indirizzo IP, ed usano i protocolli di routing IP. ATM risulta così *appaiata* alla rete IP, ma ciò complica la realizzazione dei nodi ATM, ed il metodo non si generalizza per protocolli diversi da IP.

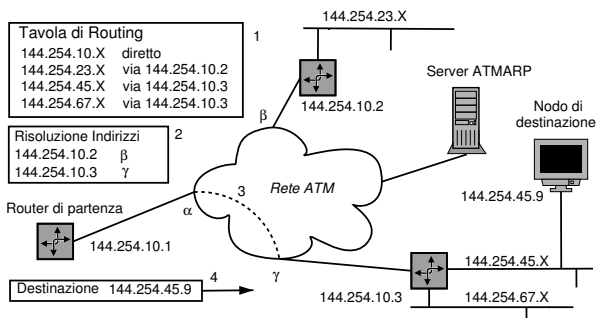
L'alternativa seguita, detta *overlay model*, vede ATM come uno stato di collegamento su cui opera l'IP, che si comporta come se si trovasse su di una LAN. In particolare, solo i nodi di frontiera tra IP ed ATM prendono un doppio indirizzo, ed individuano una *Logical Subnet* (LIS) definita da uno stesso prefisso IP ed una stessa maschera di sottorete. Con riferimento alla figura che segue, quando il router di partenza vuole contattare il nodo di destinazione, trova (1) prima l'IP del router di destinazione, e quindi invia una richiesta ARP al server ATMARP presente nella LIS¹¹⁷, che risponde comunicando

¹¹⁴Questi ultimi 4 byte contengono l'indicazione (2 bit) se si tratti della prima, ultima od intermedia cella di una stessa CS-PDU, la lunghezza dei dati validi se è l'ultima (6 bit), un numero di sequenza (4 bit), un controllo di errore (10 bit), ed una etichetta (10 bit) che rende possibile interallacciare temporalmente le celle di diverse CS-PDU.

¹¹⁵In questo modo si risparmiano 4 byte ogni 48. Ora però è indispensabile che le celle arrivino in sequenza, e non è più possibile alternare diverse CS-PDU.

¹¹⁶Quando la distanza tra i nodi oltrepassa dimensioni di un edificio, si parla di *Campus Network* o di *Wide Area Network* (WAN), ed a volte è usato il termine *Metropolitan Area Network* (MAN) per estensione cittadina. Per estensioni ancora maggiori si parla di *reti in area geografica*.

¹¹⁷Tutti i nodi della LIS hanno configurato manualmente l'indirizzo ATM del server ATMARP.



l'indirizzo γ , il quale è così risolto (2). A questo punto si può instaurare una VCC con B mediante la segnalazione ATM (3), ed effettuare la comunicazione (4). Questa soluzione è nota come *VC multiplexing*, ed i dati sono incapsulati direttamente nella CPCS-PDU di AAL5. In ricezione, l'etichetta VPI/VCI è usata per consegnare il pacchetto al protocollo di strato superiore che ha realizzato la connessione ATM. D'altra parte, questa elaborazione deve avvenire a *diretto contatto* con AAL5, e ciò preclude la possibilità di interlavoro con nodi esterni alla rete ATM.

Nel caso in cui sia antieconomico creare un gran numero di VC, o se si dispone unicamente di un PVC¹¹⁸, il pacchetto IP viene incapsulato in un header LLC IEEE 802.2 prima di essere consegnato all'AAL5. In tal modo, il router ricevente esamina l'header LLC del pacchetto ricevuto dal nodo ATM di *egress*, per consegnare il pacchetto al protocollo appropriato, realizzando così un *trasporto multiprotocollo* su ATM.

LANE, NHRP e MPOA Discutiamo qui brevemente ulteriori possibilità di utilizzo di ATM come trasporto IP, ma a cui verosimilmente sarà preferito l'MPLS.

Mentre l'approccio classico aggiunge un substrato tra IP ed AAL5, per così dire *esterno* alla rete ATM, l'approccio LANE (LAN *Emulation*) ne aggiunge uno *esterno* alla rete IP, che *crede* di avere a che fare con una LAN ethernet. In questo caso anziché una LIS, si definisce una *Emulated LAN* (ELAN), il cui esatto funzionamento prevede diversi passaggi¹¹⁹.

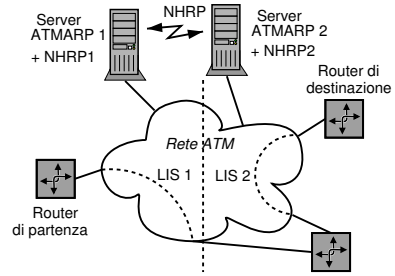
Sia nel caso classico che in quello LANE, se due router IP sono su due LAN (LIS o ELAN) differenti (con prefissi differenti) la comunicazione tra i due deve necessariamente attraversare un terzo router IP, anche se esiste un collegamento diretto tra i primi due, tutto interno alla rete ATM. La situazione è illustrata nella figura seguente, per il caso classico. Come possiamo notare, i router di partenza e di destinazione potrebbero dialogare direttamente tramite la rete ATM, diminuendo il carico di traffico della

¹¹⁸Un VC permanente collega solamente una coppia di nodi, ed in tal caso è possibile anche fare a meno del server ATMARP, in quanto un PVC è configurato manualmente. Nei fatti, questo è l'uso più diffuso del trasporto IP over ATM, ed è tipicamente utilizzato per collegare sedi distanti di uno stesso sistema autonomo, eliminando la necessità di sviluppare in proprio un impianto di TLC tra le sedi.

¹¹⁹La emulazione di una LAN da parte della rete ATM è possibile dopo aver definito per ogni ELAN un LAN *Emulation Server* (LES) a cui ogni LAN *Emulation Client* (LEC) si rivolge per conoscere l'indirizzo ATM di un altro LEC, a partire da suo indirizzo MAC (la traduzione da IP a MAC è già avvenuta tramite ARP a livello IP). In una ELAN deve inoltre essere presente un dispositivo *Broadcast and Unknown Server* (BUS) che diffonde a tutti i LEC i pacchetti broadcast Ethernet (come ad es. le richieste ARP), e che viene usato dai LEC che devono inviare un broadcast. Infine, occorre un LAN *Emulation Configuration Server* (LECS) che conosce, per ogni ELAN della rete ATM, l'elenco dei LEC, del LES e del BUS.

All'accensione di un LEC, questo contatta il LECS (conoscendone l'indirizzo ATM, oppure su di una VCC ben nota, o tramite segnalazione ATM) per apprendere gli indirizzi ATM del proprio LES e del BUS. Quindi, registra presso il LES la corrispondenza tra i propri indirizzi MAC ed ATM. Quando un LEC desidera inviare dati ad un altro LEC, dopo averne risolto l'indirizzo MAC interrogando il LES, incapsula le trame IP con un header LLC IEEE 802.2 proprio come nel caso classico.

stessa, e risparmiando al router intermedio il compito di riassembleare e disassemblare i pacchetti IP in transito, oltre a riclassificarli ai fini del routing. Se i server ATMARP delle due LIS possono scambiarsi le proprie informazioni, il router di partenza può arrivare a conoscere l'indirizzo ATM di quello di destinazione, e creare un collegamento diretto. Lo scambio delle corrispondenze $\langle ind. IP; ind. ATM \rangle$ avviene per mezzo del *Next Hop Resolution Protocol* (NHRP)



tra entità indicate come NHRP *Server* (NHS), che possono appartenere ognuno a più LIS, e che instaurano tra di loro un meccanismo di *passa-parola*¹²⁰, per rispondere alle interrogazioni che ricevono. L'applicazione di un meccanismo in parte simile, porta nel caso delle ELAN alla definizione del *Multi Protocol over ATM* (MPOA)¹²¹.

MPLS Il *Multi Protocol Label Switching* (MPLS) è una metodo di realizzare una trasmissione a circuito virtuale su reti IP, la cui architettura è descritta nella RFC 3031 dell'IETF, e che verrà esposto meglio in una prossima edizione. Qui illustriamo i legami che MPLS presenta con ATM.

Lo sviluppo di MPLS ha origine dalle iniziative industriali tese a realizzare router internet economici di prestazioni elevate, e capaci di gestire la banda in modo appropriato. Lo IETF ha ricevuto il compito di armonizzare in una architettura standardizzata i diversi approcci, basati sul principio di inoltrare i pacchetti in base ad una etichetta (LABEL) impostata dal primo router della rete, proprio come avviene in ATM. Dato che erano già disponibili i dispositivi hardware per realizzare i nodi di switching ATM, i primi prototipi hanno semplicemente utilizzato tali switch sotto il diretto controllo di un router IP, collegato ad altri simili tramite la rete ATM. L'MPLS è tuttavia più generale, sia verso l'alto (è *multiprotocollo* in quanto si applica oltre che ad IP, a qualunque altro strato di rete) che verso il basso (funziona indifferentemente dall'implementazione dello strato di collegamento, sia ATM, *ethernet* od altro).

La *label* apposta dal primo MPLS *Router* (LSR) dipende dalla destinazione IP del pacchetto; diverse destinazioni possono coincidere con una sola *Forwarding Equivalence Class* (FEC)¹²², identificata da una singola *label*. Tutti i pacchetti di una stessa FEC sono inoltrati verso il medesimo *next hop*, indicato dalla tabella di routing, indicizzata

¹²⁰I NHS risiedono su dispositivi che sono anche router IP, e che quindi mantengono aggiornate le tabelle di instradamento che indicano il prossimo salto (*next hop*) verso destinazioni IP. Le richieste di risoluzione ATMARP per un certo indirizzo IP sono instradate mediante queste stesse tabelle, giungendo di salto in salto fino al router-NHS appartenente alla stessa LIS dell'IP di destinazione, che conosce la risposta. Quest'ultima ripercorre all'indietro il percorso fatto dalla richiesta, fino alla sorgente. I router attraversati dal *passa parola*, ricordano (per un pò) le risposte trasportate, riducendo il traffico NHRP.

¹²¹Il metodo si basa su di un meccanismo indicato come *flow detection*, attuato dal router IP-ATM prossimo alla sorgente, che è in grado di accorgersi di traffico non sporadico diretto verso una medesima destinazione. Questo router impersona allora un MPOA *Client* (MPC), ed interroga un MPOA *server* (MPS) per conoscere l'indirizzo ATM della destinazione, in modo da creare un collegamento diretto. Ogni MPS serve una o più ELAN, e gli MPS comunicano tra loro mediante il NHRP.

L'MPOA realizza la separazione tra il calcolo dell'instradamento e l'inoltro dei dati. A differenza di un router tradizionale, che svolge entrambi i compiti, l'MPC svolge solo l'inoltro verso l'indirizzo ATM di destinazione, mentre quest'ultimo è fornito dall'MPS, che si comporta quindi come un *route server*.

¹²²Nel routing IP tradizionale, una FEC coincide con l'instradamento individuato dal *longest match*.

dalla *label*¹²³. Nella stessa tabella, si trova anche la nuova *label* da assegnare al pacchetto, prima di consegnarlo all'LSR seguente. In tutti i LSR successivi, il pacchetto non è riclassificato, ma solo inoltrato verso il *next hop* con una nuova *label* come ordinato dalla tabella di routing. Pertanto, è il primo LSR a decidere tutto il tragitto, ed i pacchetti di una stessa FEC seguono tutti lo stesso *Label Switched Path* (LSP). In tal modo, gli switch possono essere più semplici, si possono stabilire instradamenti diversi per una stessa destinazione¹²⁴ in base al punto di ingresso, così come le FEC possono essere rese dipendenti non solo dalla destinazione, ma anche da altri parametri, come la classe di servizio richiesta.

L'associazione tra *label* e FEC (ossia il *next hop* per i pacchetti con quella *label*) è stabilita dal LSR di *destinazione*¹²⁵, e cioè un LSR indica agli LSR dai quali *si aspetta di ricevere* traffico, quale *label* usare in corrispondenza delle FEC per le quali conosce l'instradamento. Dato che la conoscenza di un instradamento è anche il prerequisito sulla cui base sono annunciate le informazioni di routing *hop-by-hop* in internet, il *Label Distribution Protocol* (LDP) può essere vantaggiosamente associato ai protocolli di distribuzione delle informazioni di routing già esistenti (es. BGP). Le associazioni tra FEC e *label* si propagano dunque fino ai nodi di ingresso, realizzando un reticolo di "alberi" di LSP, costituiti dagli LSP definiti da una stessa FEC, e che convergono verso uno stesso *egress* a partire da diversi *ingress*. Nel nodo in cui più LSP si riuniscono, è possibile effettuare il *label merging* assegnando la stessa *label* ai pacchetti uscenti, riducendo così la dimensione delle tabelle di routing.

L'etichetta *label* su cui si basa l'MPLS può genericamente consistere in un incapsulamento della PDU dello strato di rete, prima che questa sia passata allo strato di collegamento. Quando i LSR sono realizzati mediante switch ATM, la *label* è efficacemente realizzata usando la coppia VPI/VCI, realizzando i LSP come delle VCC. In questo caso però, sorgono problemi nel caso in cui si debba effettuare il *merge* di più LSP relative ad una stessa FEC, che passano da uno stesso ATM-LSR. Infatti, se un nodo adottasse in uscita una stessa *label-VCC* per differenti VCC entranti, le celle in cui sono segmentati i pacchetti IP, ed ora con uguale *label-VCC*, si alternerebbero, rendendo impossibile il riassetto dei pacchetti. Per questo motivo, MPLS può operare anche con LSR che non permettono il *merging*, e che possono quindi essere utilizzati assieme ad altri che ne sono capaci; in tal caso, l'LSR non-*merging* non è notificato automaticamente delle associazioni FEC-*label*, ma gli viene comunicata una (diversa) *label* ogni volta che ne chiede una (da associare ad una FEC), usando così più *label* del necessario. Una alternativa, è quella di codificare la FEC mediante il solo VPI, ed usare il VCI per indicare il nodo di partenza. In questo modo, il *merging* è per così dire *automatico*, senza problemi di alternanza temporale delle celle di diversi pacchetti IP, ed il metodo può essere applicato se è possibile coordinare l'assegnazione dei VCI tra sorgenti diverse, e se il numero delle *label* non oltrepassa la capacità di indirizzamento.

L'esposizione svolta è volutamente semplificata, e trascura per comodità alcune importanti caratteristiche di MPLS.

¹²³Nel routing IP convenzionale, per ogni router, la tabella di routing deve essere esaminata per intero per ogni pacchetto, alla ricerca del *longest match* tra le regole presenti.

¹²⁴Il routing IP tradizionale opera su di una base *hop-by-hop*, e per questo non può tenere conto della provenienza. Quando due pacchetti per una medesima destinazione passano da uno stesso router, proseguono per lo stesso percorso.

¹²⁵Infatti, è la *label* del pacchetto ricevuto che determina il *next hop*, e quindi è quest'ultimo a definire la semantica della *label* presso i propri vicini.

Capitolo 9

Densità spettrale e filtraggio

Si descrive l'effetto del passaggio di un segnale, o di un processo, attraverso un sistema fisico, in particolare per quanto riguarda le modifiche spettrali. Per i segnali periodici e di energia siamo già in grado (vedi § 3.6.1, 3.2 e 3.8.4) di determinare lo spettro di ampiezza dell'uscita come $Y(f) = H(f)X(f)$, e da questo ottenere lo spettro di densità di potenza o di energia come $|Y(f)|^2$; d'altra parte se $x(t)$ rappresenta una generica realizzazione di un processo, non abbiamo ancora approfondito la teoria che consente di definire lo spettro di densità potenza $\mathcal{P}_x(f)$ dei suoi membri: è proprio questo lo scopo della prima parte del capitolo, in cui la caratterizzazione probabilistica introdotta al cap. 7 viene estesa, giungendo alla definizione di *funzione di autocorrelazione*, la cui trasformata di Fourier è pari appunto alla densità cercata (teorema di WIENER).

Il resto del capitolo procede applicando ad esempi pratici la teoria fin qui sviluppata, affrontando la determinazione dello spettro di densità di potenza per alcuni segnali di interesse, e la sua stima basata su osservazioni sperimentali. Al § 9.4 si presenta un quadro riassuntivo sul filtraggio per segnali di energia, periodici e aleatori, mentre al 9.7 lo stesso schema è applicato alla loro somma e prodotto. Si passa quindi a descrivere un metodo di analisi e sintesi per una classe di filtri, ed alla caratterizzazione dei filtri stessi in termini delle modifiche apportate ai segnali in transito.

Il capitolo termina con una appendice in cui sono riferiti ulteriori risultati, e dimostrate alcune relazioni usate fin qui nel testo.

9.1 Correlazione e covarianza

Al § 7.3.5 abbiamo osservato come la caratterizzazione statistica del primo ordine $p_X(x)$ di un processo $\{x(t, \theta)\}$ stazionario ergodico, consenta il calcolo di valor medio m_x e varianza σ_X^2 , nonché della potenza $\mathcal{P}_X = E_X\{x^2\} = \sigma_X^2 + (m_x)^2$, valida per una qualunque realizzazione θ del processo. Definiamo ora una statistica *del secondo ordine* che permetterà di determinare anche lo spettro di densità di potenza delle realizzazioni del processo.

La statistica di secondo ordine si basa sulla considerazione di 2 istanti t_1 e t_2 , in corrispondenza dei quali estraiamo due variabili aleatorie $x_1 = x(t_1)$, $x_2 = x(t_2)$ a partire da una realizzazione $\bar{\theta}$ di un processo $x(t, \theta)$, come mostrato a sinistra di fig. 9.1. Al variare della realizzazione campionata $\theta \in \Theta$, tutte le coppie di valori estratti sono

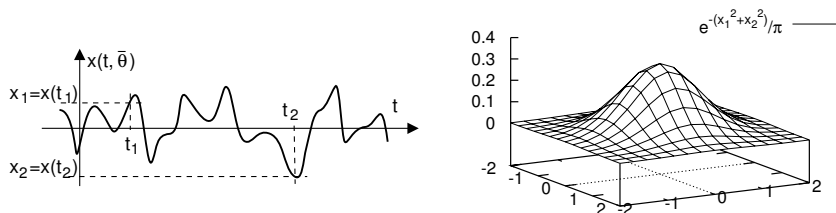


Figura 9.1: Estrazione di due variabili aleatorie da un membro di processo, e loro possibile d.d.p. congiunta

altrettante determinazioni di una variabile aleatoria *bidimensionale*, descritta da una densità di probabilità *congiunta* $p_{X_1 X_2}(x_1 x_2; t_1 t_2)$, che dipende anche dagli istanti t_1 e t_2 , ed è esemplificata nella parte destra di fig. 9.1, che mostra come questa sottenda un volume unitario, e descriva con il suo andamento le regioni del piano $x_1 x_2$ in cui cadono un maggior numero di coppie (ovvero dove la probabilità è più densa).

9.1.1 Correlazione

Questa grandezza dipende dalla definizione di un valore atteso, formalmente analogo a quanto già visto per il caso unidimensionale, che prende il nome di *momento misto di ordine* (i, j) e che risulta:

$$m_{X_1 X_2}^{(i, j)} = E_{X_1 X_2} \{x_1^i x_2^j\} = \int \int x_1^i x_2^j p_{X_1 X_2}(x_1 x_2; t_1 t_2) dx_1 dx_2$$

e che nel caso in cui $i = j = 1$ prende il nome di *correlazione* e si indica come

$$\mathcal{R}_X(x(t_1), x(t_2)) = m_{X_1 X_2}^{(1, 1)}$$

Prima di proseguire, soffermiamoci un istante per meglio comprendere il significato di $m_{X_1 X_2}^{(1, 1)}$. Il valore di correlazione fra due variabili aleatorie è indicativo del legame che esiste tra le due¹, nel senso di quanto l'una ha un valore che dipende da quello dell'altra, ed ha un valore assoluto tanto più elevato quanto più i valori di x_1 e x_2 sono legati in modo deterministico²: in effetti però, seguendo gli esempi riportati nella nota, non sempre questo accade, e pertanto è opportuno basarsi sull'uso di momenti centrati come descritto al punto successivo.

¹Inteso qui nel senso più generale, per esempio come temperatura e pressione in un punto ben preciso di un cilindro di un motore a scoppio, oppure come pressione e velocità in un circuito idraulico, o pneumatico... astruendo cioè dal caso particolare di due v.a. estratte da una stessa forma d'onda.

²Può tornare utile pensare $m_{X_1 X_2}^{(1, 1)}$ come una media pesata dei possibili valori del prodotto $x_1 x_2$; i termini di eguale ampiezza e segno opposto possono elidersi se equiprobabili. Negli esempi che seguono, riportiamo dei *diagrammi di scattering* per sei diversi casi di distribuzione delle coppie di valori x_1 e x_2 , assieme ai valori di correlazione $\mathcal{R}_{x_1 x_2}$ (corr), covarianza $\sigma_{x_1 x_2}$ (cov), e coefficiente di correlazione ρ (vedi § 9.10.1 per quest'ultimo parametro)

9.1.2 Indipendenza statistica e incorrelazione

Nel caso in cui le due v.a. siano *statisticamente indipendenti*, e cioè si possa scrivere $p_{X_1 X_2}(x_1, x_2; t_1, t_2) = p(x_1)p(x_2)$ ⁽³⁾, l'integrale che definisce la correlazione si fattorizza, e pertanto

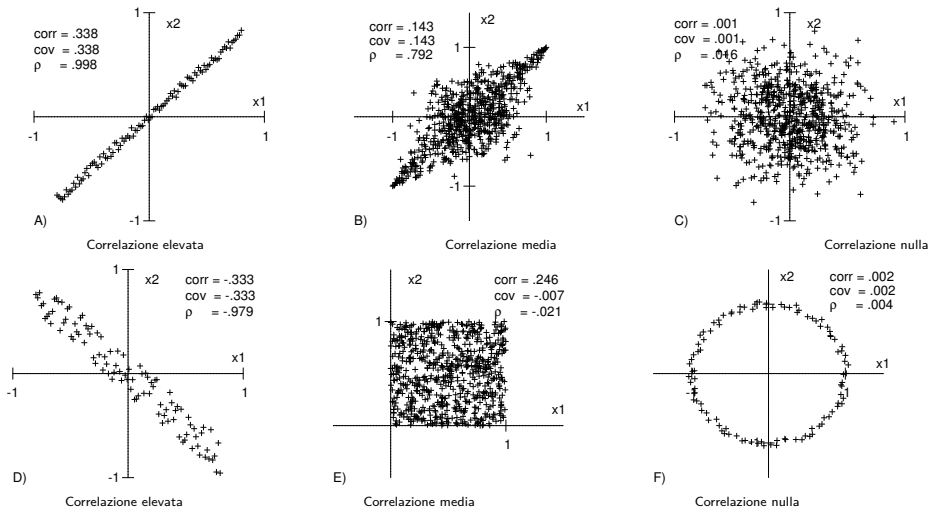
$$\begin{aligned} \mathcal{R}_X(x_1, x_2) &= \iint x_1 x_2 p(x_1) p(x_2) dx_1 dx_2 = \int x_1 p(x_1) dx_1 \cdot \int x_2 p(x_2) dx_2 = \\ &= E\{x_1\} E\{x_2\} = m_{X_1} m_{X_2} \end{aligned} \tag{9.1}$$

Covarianza E' indicata come $\sigma(x_1, x_2)$ e rappresenta il momento misto centrato tra le due v.a., di espressione⁴

$$\begin{aligned} \sigma(x_1, x_2) &= E\{(x_1 - m_{X_1})(x_2 - m_{X_2})\} = \\ &= E\{x_1 x_2\} - E\{x_1 m_{X_2}\} - E\{m_{X_1} x_2\} + E\{m_{X_1} m_{X_2}\} = \\ &= E\{x_1 x_2\} - m_{X_1} m_{X_2} = \mathcal{R}_X(x_1, x_2) - m_{X_1} m_{X_2} \end{aligned} \tag{9.2}$$

Incorrelazione Combinando i risultati (9.1) e (9.2) possiamo verificare che

*Se due variabili aleatorie x_1 ed x_2 sono statisticamente indipendenti, queste si dicono INCORRELATE, in quanto la covarianza $\sigma(x_1 x_2)$ è nulla*⁵.



In A) e F) le coppie di valori sono legate da una legge pressoché deterministica, mentre in B) e D) c'è più variabilità, ma si nota ancora una certa dipendenza tra le due. Infine nei casi C) ed E), osserviamo due v.a. *statisticamente indipendenti*, dato che $p_{X_1 X_2}(x_1, x_2)$ è fattorizzabile come $p_{X_1}(x_1) p_{X_2}(x_2)$, e per le quali risulta quindi $\mathcal{R}_X(x_1 x_2) = m_{x_1} m_{x_2}$.

³Omettiamo per brevità di indicare la variabile aleatoria a pedice della densità di probabilità.

⁴L'uguaglianza si ottiene ricordando che un valore atteso è in realtà un integrale, ed sfruttando la proprietà distributiva di quest'ultimo.

⁵Notiamo immediatamente che il termine più corretto sarebbe "incovarianzate"; l'uso (ormai storico e consolidato) dell'espressione *incorrelate* deriva probabilmente dal considerare usualmente grandezze a media nulla, per le quali le due espressioni coincidono.

La proprietà esposta ha valore in una sola direzione, in quanto se due v.a. esibiscono $\sigma(x_1, x_2) = 0$ non è detto che siano statisticamente indipendenti⁶. L'unico caso in cui l'incorrelazione tra variabili aleatorie ne implica l'indipendenza statistica, è quello relativo a v.a. estratte da un processo gaussiano, come mostrato in appendice 9.10.2.

9.1.3 Correlazione di processo ergodico

Nel caso in cui il processo da cui si estraggono x_1 e x_2 sia *stazionario*, si ottiene che

$$\mathcal{R}_X(x(t_1), x(t_2)) = \mathcal{R}_X(x(t_1), x(t_2 = t_1 + \tau)) = \mathcal{R}_X(\tau)$$

e cioè la correlazione *dipende solo dall'intervallo* $\tau = t_2 - t_1$. Infatti se un processo è stazionario, le proprietà statistiche non dipendono da traslazioni temporali.

Nel caso in cui il processo sia anche *ergodico*, allora le medie di insieme hanno lo stesso valore delle corrispondenti medie temporali, e dunque la correlazione (media di insieme) può essere calcolata in base alla sua media temporale equivalente, a partire da una qualunque realizzazione del processo

$$\mathcal{R}_x(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t) x(t + \tau) dt \quad (9.3)$$

Ovviamente è vero anche l'inverso, e cioè: il valor medio del prodotto tra due campioni, estratti (a caso) a distanza τ a partire da una specifica realizzazione⁷, ha un valore che è calcolabile come media di insieme a partire dalla conoscenza della densità di probabilità congiunta $p_{X_1 X_2}(x_1, x_2; t_1, t_1 + \tau)$.

9.1.4 Autocorrelazione e intercorrelazione

La media temporale $\mathcal{R}_x(\tau)$ appena introdotta prende il nome di *integrale di autocorrelazione*, che è definito anche per segnali di energia, come

$$\mathcal{R}_x(\tau) = \int_{-\infty}^{\infty} x^*(t) x(t + \tau) dt \quad (9.4)$$

in cui l'operatore di coniugato generalizza l'operazione anche al caso di segnali complessi. Simile, ma diversa, è la definizione di integrale di *intercorrelazione*, che esprime lo stesso calcolo, ma generalizzandone il concetto, in quanto relativo a due segnali $x(t)$ ed $y(t)$ estratti da processi diversi:

$$\mathcal{R}_{xy}(\tau) = \int_{-\infty}^{\infty} x^*(t) y(t + \tau) dt \quad (9.5)$$

che può essere interpretato come *prodotto scalare* o *energia mutua* (vedi § 3.2) tra $x(t)$ e una copia di $y(t)$ traslata nel tempo. L'*integrale* di autocorrelazione è anche detto *funzione* di autocorrelazione, in quanto il suo argomento è un tempo (l'intervallo tra due campioni) e dunque $\mathcal{R}_x(\tau)$ può essere visto come un segnale (funzione di τ anziché di t). Nello studio abbiamo già incontrato un integrale (di convoluzione) il

⁶Vedi ad esempio il caso F) della nota (2), in cui le variabili aleatorie risultano incorrelate, ma non sono per nulla indipendenti, in quanto l'una dipende strettissimamente dall'altra, dato che le coppie di valori si dispongono su di un cerchio.

⁷Si rifletta sulla descrizione ora data a parole dell'operazione di media temporale scritta sopra.

cui risultato è una funzione del tempo; la somiglianza tra i due è più profonda di una semplice analogia, in quanto si può scrivere

$$\mathcal{R}_x(\tau) = \int_{-\infty}^{\infty} x^*(t) x(t + \tau) dt = x^*(-t) * x(t) \tag{9.6}$$

in cui $*$ è il consueto simbolo di convoluzione⁸.

In base a quest'ultima osservazione otteniamo la costruzione grafica mostrata in fig. (9.2), che fornisce il risultato dell'integrale di autocorrelazione, e che è del tutto simile a quella già illustrata per la convoluzione (vedi § 3.5.3), con la differenza che ora non si effettuano ribaltamenti di asse. L'esempio mostrato ne illustra l'applicazione ad un caso noto, per il quale $x(t) = x^*(-t)$, e che fornisce quindi lo stesso risultato di $x(t) * x(t)$.

9.1.4.1 Proprietà dell'autocorrelazione

Elenchiamo ora alcuni aspetti *caratteristici* della funzione di autocorrelazione:

Traslazioni temporali: se consideriamo i segnali $x(t)$ e $y(t) = x(t + \theta)$, le rispettive autocorrelazioni $\mathcal{R}_x(\tau)$ ed $\mathcal{R}_y(\tau)$ sono identiche⁹. Questo risultato mostra come l'autocorrelazione non tenga conto dell'informazione legata alla fase dei segnali: infatti $x(t)$ e $y(t)$ hanno la stessa densità spettrale, a meno di un contributo di fase lineare, ed hanno uguale autocorrelazione.

Durata Limitata: la funzione di autocorrelazione di un segnale di durata limitata è anch'essa a durata limitata, e di estensione doppia rispetto alla durata del segnale originario.

Segnali Periodici: l'autocorrelazione di un segnale periodico di periodo T è anch'essa periodica, con lo stesso periodo. Infatti per $\tau = nT$ il secondo fattore integrando è traslato di un numero intero di periodi.

In particolare, le due proprietà seguenti sono da ritenersi *fondamentali* per la possibilità, che come vedremo offrono, di interpretare l'autocorrelazione in senso energetico:

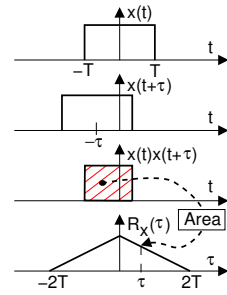


Figura 9.2: Autocorrelazione di due rettangoli

⁸Il risultato ottenuto si basa sul cambio di variabile $\theta = t + \tau$ che permette di scrivere

$$\mathcal{R}_x(\tau) = \int_{-\infty}^{\infty} x^*(\theta - \tau) x(\theta) d\theta = \int_{-\infty}^{\infty} x^*(-(\tau - \theta)) x(\theta) d\theta = x^*(-t) * x(t)$$

⁹Infatti otteniamo:

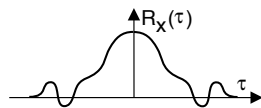
$$\begin{aligned} \mathcal{R}_y(\tau) &= \int_{-\infty}^{\infty} y(t) y(t + \tau) dt = \int_{-\infty}^{\infty} x(t + \theta) x(t + \theta + \tau) dt \\ &= \int_{-\infty}^{\infty} x(\alpha) x(\alpha + \tau) d\alpha = \mathcal{R}_x(\tau) \end{aligned}$$

Massimo nell'origine: la $\mathcal{R}_x(\tau)$ calcolata in $\tau = 0$ fornisce il valore di $\mathcal{R}_x(\tau)$ più grande di quello ottenibile per qualunque altro valore di τ . In particolare, $\mathcal{R}_x(\tau = 0)$ è uguale alla potenza del segnale $x(t)$, od all'energia se $x(t)$ è di energia, ossia

$$0 \leq \mathcal{R}_x(0) = \begin{cases} \int |x(t)|^2 dt = \mathcal{E}_x > |\mathcal{R}_x(\tau \neq 0)| & \text{se } x(t) \text{ è di energia} \\ \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |x(t)|^2 dt = \mathcal{P}_x \geq |\mathcal{R}_x(\tau \neq 0)| & \text{se } x(t) \text{ è di potenza} \end{cases}$$

e in particolare, se $x(t)$ è periodico, l'ultimo segno \geq è una eguaglianza per τ multiplo di un periodo.

Simmetria coniugata: è possibile verificare¹⁰ che risulta $\mathcal{R}_x(-\tau) = \mathcal{R}_x^*(\tau)$, e ciò consente (vedi § 3.3) di affermare che $\mathcal{F}\{\mathcal{R}_x(\tau)\}$ è reale. Nel caso in cui $x(t)$ è reale, si ottiene $\mathcal{R}_x(-\tau) = \mathcal{R}_x(\tau)$, ovvero l'autocorrelazione di un segnale reale è *reale pari*, alla stregua (come mostreremo ora) della sua trasformata di Fourier.



9.2 Densità spettrale

Come anticipato ad inizio capitolo, mostriamo il metodo con cui determinare lo spettro di densità di potenza nel caso di processi. La cosa decisamente gradevole è che questo stesso strumento è valido anche per gli altri tipi di segnale.

9.2.1 Teorema di Wiener¹¹

Lo spettro di densità di potenza $\mathcal{P}_x(f)$ (o di energia $\mathcal{E}_x(f)$) di $x(t)$ è uguale alla trasformata di Fourier della sua funzione di autocorrelazione $\mathcal{F}\{\mathcal{R}_x(\tau)\}$.

La dimostrazione del teorema è straordinariamente semplice, e per segnali di energia si scrive

$$\begin{aligned} \mathcal{R}_x(\tau) &= \int_{-\infty}^{\infty} x^*(t) x(t + \tau) dt = \int_{-\infty}^{\infty} X^*(f) X(f) e^{j2\pi f\tau} df = \\ &= \mathcal{F}^{-1}\{X^*(f) X(f)\} = \mathcal{F}^{-1}\{\mathcal{E}_x(f)\} \end{aligned}$$

in cui abbiamo prima applicato il teorema di Parseval, poi la proprietà di traslazione nel tempo, e quindi (vedi § 3.2) espresso $X^*(f) X(f)$ come $\mathcal{E}_x(f)$.

Come anticipato, questa proprietà è valida anche per segnali di potenza¹², comprese quindi le realizzazioni di processi. In particolare, se il processo è ergodico, la media di insieme $m_{X_1 X_2}^{(1,1)}$ risulta pari alla media temporale espressa dalla (9.3) e calcolata

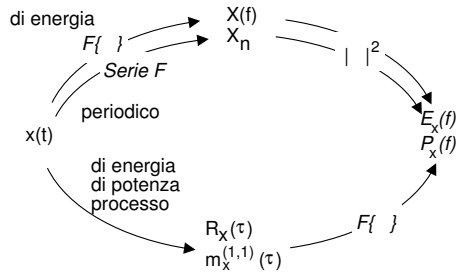
¹⁰ $\mathcal{R}_x(-\tau) = \int_{-\infty}^{\infty} x^*(t) x(t - \tau) dt = \int_{-\infty}^{\infty} x^*(\alpha + \tau) x(\alpha) d\alpha = \mathcal{R}_x^*(\tau)$, avendo operato il cambio di variabile $t - \tau = \alpha$, da cui $t = \alpha + \tau$ e $dt = d\alpha$.

¹¹In realtà le attribuzioni di questo risultato sono molteplici, comprendendo anche *Khinchin*, *Einstein* e *Kolmogorov* - fonte http://en.wikipedia.org/wiki/Wiener%E2%80%93Khinchin_theorem

¹²Per segnali di potenza la dimostrazione può essere fornita, senza troppe attenzioni di rigore analitico, partendo dalla stima della densità di potenza ottenuta mediante periodogramma (§ 9.3.1):

$\mathcal{P}_{x_T}(f) = \frac{|X_T(f)|^2}{T}$. Osserviamo che $|X_T(f)|^2$ rappresenta lo spettro di densità di energia $\mathcal{E}_{x_T}(f)$ di un segmento di segnale $x_T(t)$ di durata T estratto da $x(t)$, corrispondente alla funzione di autocorrelazione $\mathcal{R}_{x_T}(\tau) = \mathcal{F}^{-1}\{\mathcal{E}_{x_T}(f)\}$ calcolabile per $x_T(t)$ mediante la (9.4). Ma operando il

per una qualsiasi realizzazione del processo; pertanto lo spettro di densità di potenza di un processo si ottiene trasformando la funzione di autocorrelazione calcolata come media di insieme, oppure trasformando quella calcolata come media temporale per una delle sue realizzazioni. In virtù del teorema di Wiener, è dunque possibile ottenere $\mathcal{P}_x(f)$ anche per processi e segnali di potenza, oppure fare “la prova del nove” per segnali di energia o periodici.



Applichiamo ora questo metodo di valutazione della densità spettrale di un processo, ad alcuni casi particolari.

9.2.2 Processo armonico

E' definito in base ad una sua generica realizzazione

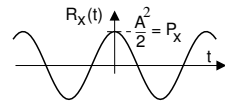
$$x(t, \theta) = A \cos(2\pi f_0 t + \theta)$$

che, se il parametro θ è una variabile aleatoria uniformemente distribuita tra $-\pi$ e π (ossia $p_\theta(\theta) = \frac{1}{2\pi} \text{rect}_{2\pi}(\theta)$), descrive un processo ergodico, ed in tal caso la d.d.p. $p_X(x) = \frac{1}{\pi\sqrt{A^2-x^2}}$ è graficata a pagina 160.

Sappiamo che una sua realizzazione (ad esempio quella con $\theta = 0$) ha una densità di potenza $\mathcal{P}_x(f) = \frac{A^2}{4} [\delta(f - f_0) + \delta(f + f_0)]$. Possiamo quindi ottenere l'autocorrelazione senza dover svolgere l'integrale:

$$\mathcal{R}_x(t) = \mathcal{F}^{-1} \{ \mathcal{P}_x(f) \} = \frac{A^2}{4} [e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}] = \frac{A^2}{2} \cos(2\pi f_0 t)$$

Il risultato ottenuto ci conferma che l'autocorrelazione di un segnale periodico è periodica; riflettiamo dunque sulla circostanza che anche un seno, od un coseno con qualunque altra fase, avrebbe avuto la stessa $\mathcal{R}_x(t)$. Ciò è d'altra parte evidente, avendo tutti questi segnali uguali densità $\mathcal{P}_x(f)$.



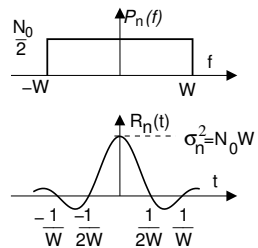
9.2.3 Processo gaussiano bianco limitato in banda

Un processo $n(t)$ è chiamato *bianco* qualora sia costante in frequenza (vedi cap. 16), e quindi descritto da una densità di potenza pari a

$$\mathcal{P}_n(f) = \frac{N_0}{2} \text{rect}_{2W}(f)$$

in cui W è l'occupazione di banda a frequenze positive. In tali ipotesi otteniamo

$$\mathcal{R}_n(t) = \mathcal{F}^{-1} \{ \mathcal{P}_n(f) \} = N_0 W \text{sinc}(2Wt)$$



passaggio al limite e la divisione $\lim_{T \rightarrow \infty} \frac{1}{T}$ di $\mathcal{R}_{x_T}(\tau)$, si ottiene l'autocorrelazione (9.3) dell'intero segnale, mentre nel dominio della frequenza il limite di $\mathcal{E}_{x_T}(f)$ tende allo spettro di densità di potenza $\mathcal{P}_{x_T}(f)$ del segnale $x(t)$ nella sua interezza.

da cui possiamo constatare che, con queste posizioni, si ottiene

$$\mathcal{R}_n(0) = \mathcal{P}_n = \int_{-\infty}^{\infty} \mathcal{P}_n(f) df = N_0 W = \sigma_n^2$$

in cui l'ultima eguaglianza sussiste in quanto l'assenza di impulsi nell'origine per $\mathcal{P}_n(f)$ corrisponde ad un $n(t)$ a media nulla (vedi eq. (7.1) a pag. 153). Inoltre, dato che $\mathcal{R}_n(1/2W) = 0$, osserviamo che campionando $n(t)$ con periodo $T_c = 1/2W$ si ottengono valori *incorrelati*, e se il processo è gaussiano, anche statisticamente indipendenti (vedi § 9.10.2). Questo risultato giustifica, almeno da un punto di vista teorico, una ipotesi che viene spesso fatta: quella di trovare sovrapposti ai campioni di segnale, dei campioni di rumore *statisticamente indipendenti*.

All'aumentare di W , $\mathcal{R}_n(t)$ tende a zero più rapidamente, cosicché il rumore si mantiene correlato per un tempo sempre minore, ovvero due campioni estratti ad una stessa distanza t hanno una correlazione sempre minore. Un risultato simile vale anche più in generale, in quanto l'autocorrelazione $\mathcal{R}_x(t)$ di un qualsiasi processo a media nulla (tranne nel caso periodico, riconducibile ad una combinazione di processi armonici) tende a 0 con $t \rightarrow \infty$, ovvero da un certo t in poi la correlazione è trascurabile.

Infine, se immaginiamo il rumore bianco limitato in banda come il risultato del filtraggio di un processo gaussiano a banda infinita (quindi, con $\mathcal{R}_n(t) = \delta(t)$) attraverso il filtro passa basso ideale con $H(f) = \text{rect}_{2W}(f)$ (vedi § 5.3.1), ci accorgiamo che la correlazione risultante per i campioni di rumore separati da un intervallo $\tau = 1/2W$ è una diretta conseguenza della memoria introdotta dalla risposta impulsiva $h(t) = 2W \text{sinc}(2Wt)$ sul segnale in transito, dato appunto che l'operazione di convoluzione rende i valori in uscita, una combinazione lineare di quelli passati in ingresso (vedi § 3.5.3).

9.2.4 Processo di segnale dati

Abbiamo già descritto al § 5.1.2 un generico segnale numerico come una somma di repliche di una funzione $g(t)$, con ampiezze a_n rappresentative dei valori da trasmettere:

$$x(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT + \theta)$$

La presenza della variabile aleatoria θ a distribuzione uniforme tra $\pm \frac{T}{2}$ (per cui $p_{\Theta}(\theta) = \frac{1}{T} \text{rect}_T(\theta)$), rende $x(t)$ un processo ergodico (vedi pag. 7.3.7).

Si mostrerà in appendice (al § 9.10.3) che, nelle ipotesi in cui le ampiezze a_n siano determinazioni di variabili aleatorie *indipendenti* ed *identicamente distribuite*, a media nulla e varianza σ_A^2 ¹³, l'autocorrelazione di $x(t)$ vale $\mathcal{R}_x(\tau) = \sigma_A^2 \frac{\mathcal{R}_g(\tau)}{T}$ in cui $\mathcal{R}_g(\tau)$ è l'autocorrelazione di $g(t)$, e dunque

$$\mathcal{P}_x(f) = \sigma_A^2 \frac{\mathcal{E}_g(f)}{T}$$

Osserviamo innanzitutto che è per questa via che al § 5.2.1 si sono caratterizzate le densità di potenza proprie dei codici di linea. Limitandoci a voler interpretare il

¹³Media m_A e varianza σ_A^2 sono qui riferite ai valori multilivello a_k (con $k = 1, 2, \dots, L$) che un generico simbolo a_n può assumere, pesati con le rispettive probabilità p_k , ossia $m_A = \sum_{k=1}^L p_k a_k$ e $\sigma_A^2 = \sum_{k=1}^L p_k (a_k - m_A)^2$

risultato, notiamo che $\mathcal{E}_g(f)$ è la densità di energia di una singola replica di $g(t)$. La sua ripetizione, con periodo T , fornisce una densità di potenza media $\frac{\mathcal{E}_g(f)}{T}$. Se ogni replica di $g(t)$ è moltiplicata per una v.a. indipendente a media nulla e varianza (potenza) σ_A^2 , la densità di potenza $\mathcal{P}_x(f)$ aumenta di egual misura (vedi § 9.7.1). Un'ultima avvertenza riguarda il fatto che, se gli a_n non sono indipendenti, il risultato è più complesso (vedi appendice 9.10.3).

9.3 Stima spettrale

Il teorema di Wiener ci aiuta qualora si desideri conoscere la densità di potenza per un processo, di cui siamo in grado di stimare o postulare un $m_X^{(1,1)}(\tau) = \mathcal{R}_X(\tau)$. Spesso però si ha a che fare con segnali di cui, pur ricorrendo le ipotesi di appartenenza ad un processo ergodico, si ignorano le statistiche di insieme. Un altro caso tipico è quello di un segnale che, pur se rappresentativo di molti altri, non presenta caratteristiche spettrali costanti nel tempo, e sono proprio le variazioni di queste ultime ad interessare¹⁴. In questi casi, tutto ciò che si può fare è di tentare una stima dello spettro di potenza del segnale, a partire da un suo segmento temporale. Esistono al riguardo tecniche differenti, come ad es. quella riportata al § 18.1.2; per ora ci limitiamo ad un caso semplice ma di rilievo teorico.

9.3.1 Periodogramma

Definiamo un intervallo T in cui isoliamo un segnale a durata limitata $x_T(t) = x(t, \theta) \text{rect}_T(t)$ da una realizzazione $x(t, \theta)$. Questo segmento di segnale è di energia, con $\mathcal{E}_{x_T}(f) = |X_T(f)|^2$, e sotto le ipotesi di stazionarietà, da questa si può stimare una densità di potenza

$$\mathcal{P}_{x_T}(f) = \frac{|X_T(f)|^2}{T}$$

che viene detta *periodogramma*¹⁵. Al tendere di T ad ∞ , il risultato trovato tende alla densità di potenza $\mathcal{P}_x(f) = \lim_{T \rightarrow \infty} \frac{|X_T(f, \theta)|^2}{T}$ della realizzazione $x_T(t, \theta)$ e, se questa appartiene ad un processo ergodico, a quella di un qualunque altro membro.

Nel caso più verosimile in cui T non tende ad infinito, si può mostrare¹⁶ che usando $\mathcal{P}_{x_T}(f)$ come una stima $\hat{\mathcal{P}}_x(f)$ della vera densità $\mathcal{P}_x(f)$ del processo, si commette un errore espresso come

$$\hat{\mathcal{P}}_x(f) = \mathcal{P}_x(f) * T(\text{sinc}(fT))^2$$

¹⁴Un esempio può essere un segnale sonoro, ad esempio una voce recitante, per il quale vogliamo studiare le caratteristiche spettrali dei diversi suoni della lingua (i fonemi), per confrontarle con quelle di un altro individuo, o per ridurre la quantità di dati necessaria a trasmettere il segnale in forma numerica (vedi § 18.1.2), o per realizzare un dispositivo di riconoscimento vocale.

¹⁵Il termine *periodogramma* trae origine dall'uso che ne fu inizialmente fatto, ossia per scoprire tracce di periodicità all'intero di segnali qualsiasi.

¹⁶Per una determinata frequenza f_0 , il valore $\frac{|X_T(f_0)|^2}{T}$ è una variabile aleatoria (dipende infatti da θ), che vorremmo avesse un valore atteso $m_T = E_\theta \{ \mathcal{P}_{x_T}(f_0) \}$ pari alla densità del processo ($m_X = \mathcal{P}_x(f_0)$) ed una varianza $\sigma_T^2 = E_\theta \{ (\mathcal{P}_{x_T}(f_0) - m_X)^2 \}$ che diminuisce al crescere di T . Per verificare se tali proprietà siano soddisfatte, valutiamo il valore atteso del periodogramma, a partire dalle relazioni fornite dal teorema di Wiener, applicato ad $X_T(f)$, e cioè $|X_T(f)|^2 =$

che è il risultato ottenuto al § 3.9.3 a riguardo del procedimento di *finestratura temporale*, e che mostra come lo stimatore è *polarizzato*¹⁷. In base allo scopo con cui è condotta la stima spettrale, può essere opportuno adottare al posto della finestra rettangolare un diverso andamento $w_T(t)$, in modo da calcolare $x_T(t, \theta) = x(t, \theta) \cdot w_T(t)$.

Il calcolo del periodogramma viene svolto mediante una *Trasformata Discreta di Fourier* (DCT), o meglio ancora con una FFT (vedi § 4.4). In tal caso, l'aumento di T corrisponde all'aumento del numero di campioni utilizzati, ed all'aumento della risoluzione in frequenza ottenibile. In base a queste considerazioni, si può mostrare che la varianza dello stimatore ottenuto per via numerica *non decresce* con T ¹⁸.

9.4 Filtraggio di segnali e processi

Ora che siamo in grado di descrivere da un punto di vista spettrale tutti i segnali, analizziamo le caratteristiche relative al segnale $y(t) = x(t) * h(t)$ uscente da un filtro con risposta impulsiva $h(t)$. Innanzitutto, individuiamo quanto valgono \mathcal{P}_y e $\mathcal{P}_y(f)$, oppure \mathcal{E}_y ed $\mathcal{E}_y(f)$ se $x(t)$ è di energia.

9.4.1 Densità spettrale in uscita da un filtro

Segnali di energia Sappiamo che per il teorema di Parseval risulta $\mathcal{E}_y(f) = Y(f)Y^*(f)$; dato che $Y(f) = X(f)H(f)$, allora

$$\mathcal{E}_y(f) = X(f)H(f)X^*(f)H^*(f) = |X(f)|^2 |H(f)|^2 = \mathcal{E}_x(f) |H(f)|^2$$

A questo punto, eseguendo l'antitrasformata di Fourier di ambo i membri, si ottiene:

$$\mathcal{R}_y(\tau) = \mathcal{F}^{-1} \{ \mathcal{E}_y(f) \} = \mathcal{F}^{-1} \{ \mathcal{E}_x(f) |H(f)|^2 \} = \mathcal{R}_x(\tau) * \mathcal{R}_h(\tau)$$

Il risultato ottenuto è posto in evidenza perché è valido anche per i due casi successivi di segnale periodico e di processo, e mostra come l'autocorrelazione dell'uscita di un filtro è pari alla convoluzione tra l'autocorrelazione dell'ingresso e quella della risposta impulsiva.

$\mathcal{E}_{x_T}(f) = \mathcal{F} \{ \mathcal{R}_{x_T}(\tau) \}$:

$$\begin{aligned} E_\theta \{ \mathcal{P}_{x_T}(f) \} &= E_\theta \left\{ \mathcal{F} \left\{ \frac{1}{T} \int_{-\infty}^{\infty} x(t) \text{rect}_T(t) x(t+\tau) \text{rect}_T(t+\tau) dt \right\} \right\} = \\ &= \mathcal{F} \left\{ \frac{1}{T} \int_{-\infty}^{\infty} E_\theta \{ x(t) x(t+\tau) \} \text{rect}_T(t) \text{rect}_T(t+\tau) dt \right\} = \\ &= \mathcal{F} \left\{ \mathcal{R}_x(\tau) \frac{1}{T} \int_{-\infty}^{\infty} \text{rect}_T(t) \text{rect}_T(t+\tau) dt \right\} = \mathcal{F} \{ \mathcal{R}_x(\tau) \cdot \text{tri}_{2T}(\tau) \} = \\ &= \mathcal{P}_x(f) * T(\text{sinc}(fT))^2 \end{aligned}$$

Osserviamo quindi come, all'aumentare di T , il nostro stimatore tende al valore vero, dato che $T(\text{sinc}(fT))^2$ tende ad un impulso.

¹⁷Ci consola verificare che, come commentato alla nota precedente, per $T \rightarrow \infty$ la polarizzazione tende a scomparire. Quando il valore atteso di uno stimatore tende al valore vero, si dice (vedi § 7.5.4) che lo stimatore è *non polarizzato* (o *unbiased*); se aumentando la dimensione del campione, la varianza della stima tende a zero, lo stimatore è detto *consistente*.

¹⁸Una possibile alternativa è quella di suddividere l'intervallo di osservazione in diversi sottointervalli, calcolare il periodogramma su ciascuno di essi, e mediare i risultati. In tal modo, all'aumentare dei dati a disposizione, viene mantenuta la stessa risoluzione in frequenza, ma si migliora la varianza della stima spettrale (vedi § 7.5.5).

A corollario di quanto esposto, sussistono le seguenti uguaglianze¹⁹, equivalenti ai fini del calcolo dell'energia totale:

$$\begin{aligned}\mathcal{E}_y &= \int_{-\infty}^{\infty} \mathcal{E}_y(f) df = \int_{-\infty}^{\infty} \mathcal{E}_x(f) |H(f)|^2 df = \int_{-\infty}^{\infty} \mathcal{R}_x(\tau) \mathcal{R}_h(\tau) d\tau = \\ &= \int_{-\infty}^{\infty} \mathcal{R}_x(\tau) \mathcal{R}_h^*(\tau) d\tau = \mathcal{R}_y(0)\end{aligned}$$

Pertanto è possibile utilizzare tutte queste come relazioni di equivalenza, quando si ha necessità di determinare una grandezza a partire da altre note.

Segnali periodici Se in ingresso ad un filtro è presente un segnale periodico, il segnale in uscita è anch'esso periodico²⁰, e per esso è valido lo sviluppo in serie di Fourier

$$y(t) = \sum_n Y_n e^{j2\pi n F t}$$

i cui coefficienti Y_n possono esprimersi in termini di quelli dell'ingresso X_n e dei valori della risposta in frequenza (vedi § 9.6) come $Y_n = X_n H(nF)$, ovvero in modulo e fase come

$$|Y_n| = |X_n| |H(nF)|; \quad \arg(Y_n) = \arg(X_n) + \arg(H(nF))$$

Per la densità di potenza di un segnale periodico possiamo scrivere (vedi § 2.3) $\mathcal{P}_y(f) = \sum_n |Y_n|^2 \delta(f - nF)$, ovvero

$$\mathcal{P}_y(f) = \sum_n |X_n|^2 |H(nF)|^2 \delta(f - nF) = |H(f)|^2 \mathcal{P}_x(f)$$

Di nuovo, antitrasformando si ottiene $\mathcal{R}_y(\tau) = \mathcal{R}_x(\tau) * \mathcal{R}_h(\tau)$.

Processi ergodici e segnali di potenza Anche in questo caso, si verifica (in appendice 9.10.5) che $m_Y^{(1,1)}(\tau) = m_X^{(1,1)}(\tau) * \mathcal{R}_h(\tau)$, e dunque

$$\mathcal{P}_y(f) = \mathcal{P}_x(f) |H(f)|^2 \quad (9.7)$$

Il risultato ovviamente si applica a qualunque membro del processo, per il quale come noto risulta $m_X^{(1,1)}(\tau) = \mathcal{R}_{xx}(\tau)$, e dunque la (9.7) è valida anche per un qualunque segnale di potenza

Risposta in frequenza di potenza (o energia) E' il nome che può essere dato alla grandezza $|H(f)|^2 = \mathcal{P}_y(f)/\mathcal{P}_x(f)$, o $\mathcal{E}_y(f)/\mathcal{E}_x(f)$, visto che ripropone in termini energetici il legame ingresso-uscita rappresentato da $H(f)$ (vedi § 3.6.1); ma per $|H(f)|^2$ è anche usato il termine di *guadagno di potenza* (§ 14.5.1).

9.4.2 Caratteristiche statistiche in uscita da un filtro

Passiamo quindi a calcolare le altre grandezze rappresentative:

¹⁹La terza uguaglianza sussiste in virtù del teorema di Parseval, mentre la quarta è valida se $\mathcal{R}_H(\tau)$ è reale, ossia $h(t)$ è idealmente realizzabile, proprietà quest'ultima definita al § 9.6.

²⁰Questa affermazione è vera, purché il filtro *non* presenti fenomeni di non linearità (definiti al § 9.6): in tal caso infatti, vale la sovrapposizione degli effetti, e *non possono* prodursi armoniche diverse da quelle in ingresso.

Media è a quella dell'ingresso, moltiplicata per il guadagno in continua del filtro:

$$\begin{aligned} m_Y &= E\{y(t)\} = E\{x(t) * h(t)\} = E\{x(t)\} * h(t) = \\ &= m_X \int_{-\infty}^{\infty} h(\tau) d\tau = m_X H(0) \end{aligned}$$

Potenza in linea generale, è sempre vero che $\mathcal{P}_y = \sigma_y^2 + (m_y)^2$; inoltre, valgono le relazioni

$$\begin{aligned} \mathcal{P}_y &= \mathcal{R}_y(0) = \int_{-\infty}^{\infty} \mathcal{P}_y(f) df = \int_{-\infty}^{\infty} \mathcal{P}_x(f) |H(f)|^2 df = \\ &= \int_{-\infty}^{\infty} \mathcal{R}_x(\tau) \mathcal{R}_h(\tau) d\tau \end{aligned}$$

Se ad esempio $x(t)$ è un processo bianco a media nulla e banda finita, con $\mathcal{P}_x(f) = \frac{N_0}{2} \text{rect}_{2B}(f)$ e quindi $\mathcal{R}_x(\tau) = N_0 B \cdot \text{sinc}(2Bt)$, si ottiene²¹

$$\sigma_y^2 = \mathcal{P}_y = \frac{N_0}{2} \int_{-B}^B |H(f)|^2 df = \frac{N_0}{2} \mathcal{R}_h(0)$$

e per la densità di potenza si ha $\mathcal{P}_y(f) = \frac{N_0}{2} |H(f)|^2$: pertanto, il processo in uscita dal filtro *non è più bianco*, ed in questo caso il processo si dice *colorato*. A questo fenomeno corrisponde anche una modifica della funzione di autocorrelazione: questa infatti non è più un *sinc*, ma vale $\mathcal{R}_y(\tau) = N_0 B \cdot \mathcal{R}_h(\tau) * \text{sinc}(2Bt)$; mentre prima quindi (per il processo bianco) due suoi valori estratti in istanti multipli di $1/2B$ erano comunque incorrelati, la colorazione introdotta dal filtro ha causato l'insorgenza di una dipendenza statistica tra i valori estratti a tali intervalli²².

Densità di probabilità A riguardo della $p_Y(y)$ non si può dire nulla di generale, tranne che essa dipende dalla $p_X(x)$ di ingresso e dalle operazioni compiute dal filtro; la sua espressione esatta va però determinata caso per caso. L'unico caso in cui la teoria fornisce una regola certa, è relativo ancora una volta al caso di processi gaussiani: posti in ingresso ad un filtro, producono in uscita un processo anch'esso gaussiano. Questo risultato è una diretta conseguenza della definizione stessa di processo gaussiano, come risultato della somma di infinite cause identicamente distribuite: dato che l'integrale di convoluzione effettivamente esegue una somma tra infiniti valori di ingresso, in linea di principio è lecito affermare che nel transito in un filtro la densità di probabilità dell'uscita *si gaussianizza*.

Esercizio Sia dato il filtro in figura, con

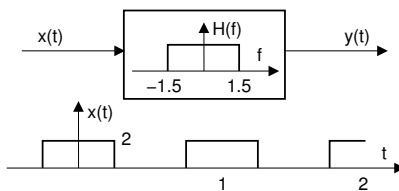
$$H(f) = \text{rect}_3(f)$$

ed al cui ingresso viene posto il segnale

$$x(t) = 2 \sum_{n=-\infty}^{\infty} \text{rect}_{\frac{1}{2}}(t - n)$$

Calcolare:

- 1) la potenza in ingresso \mathcal{P}_x ,
- 2) la potenza in uscita \mathcal{P}_y ,
- 3) l'espressione di $y(t)$.



²¹in realtà si fa l'ulteriore ipotesi che la banda passante di $H(f)$ sia minore di B

²²Il motivo di questo risultato può essere meglio compreso ricordando che l'integrale di convoluzione calcola i singoli valori in uscita da un filtro, come dipendenti da tutti gli ingressi passati, ognuno pesato con il valore della risposta impulsiva relativo al ritardo tra ingresso passato ed uscita presente. Pertanto, anche se i singoli valori in ingresso sono incorrelati, quelli di uscita (distanti tra loro per meno della durata della risposta impulsiva) condividono una porzione di storia comune, e quindi i loro valori non sono più indipendenti.

Risposte

1) Calcoliamo la media temporale: $\mathcal{P}_x = \frac{1}{T} \int_{-T/2}^{T/2} x^2(t) dt = \frac{1}{T} \int_{-1/4}^{1/4} 2^2(t) dt = \frac{4}{2} = 2$;

2) Sappiamo che $\mathcal{P}_y = \int_{-\infty}^{\infty} \mathcal{P}_y(f) df$, in cui $\mathcal{P}_y(f) = |Y(f)|^2$, ed a sua volta $Y(f) = X(f)H(f)$. Calcoliamo perciò innanzitutto

$$X(f) = F\{2\text{rect}_\tau(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT)\} = 2\tau \cdot \text{sinc}(f\tau) \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right)$$

essendo $\tau = \frac{1}{2}$ e $T = 1$, risulta

$$X(f) = \text{sinc}\left(\frac{f}{2}\right) \sum_{n=-\infty}^{\infty} \delta(f - n) = \sum_{n=-\infty}^{\infty} X_n \delta(f - n)$$

con $X_n = \text{sinc}\left(\frac{n}{2}\right)$. Dunque, dato che gli unici impulsi che cadono entro la risposta in frequenza $H(f)$ sono quelli per $f = -1, 0$ e 1 , si ha:

$$Y(f) = X(f)H(f) = \sum_{n=-1}^1 X_n H(n) \delta(f - n)$$

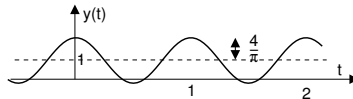
il cui modulo quadro fornisce $\mathcal{P}_y(f) = \sum_{n=-1}^1 |X_n|^2 |H(n)|^2 \delta(f - n)$, e pertanto si ottiene

$$\mathcal{P}_y = \int_{-\infty}^{\infty} \mathcal{P}_y(f) df = \left(\frac{\sin -\frac{\pi}{2}}{-\frac{\pi}{2}}\right)^2 + 1 + \left(\frac{\sin \frac{\pi}{2}}{\frac{\pi}{2}}\right)^2 = 1 + 2\left(\frac{2}{\pi}\right)^2 = 1.811$$

3) Risulta

$$y(t) = \sum_{n=-1}^1 X_n H(n) e^{j2\pi nt} = 1 + \frac{2}{\pi} (e^{j2\pi t} + e^{-j2\pi t}) = 1 + \frac{4}{\pi} \cos 2\pi t$$

Notiamo come il filtro lasci passare solamente la componente continua e la prima armonica.



Dopo aver illustrato le trasformazioni subite da un segnale generico che attraversa un filtro generico, affrontiamo lo studio del caso in cui il filtro è appositamente progettato per uno specifico segnale noto a priori.

9.5 Filtro adattato

Si tratta del filtro da utilizzare da parte di un *detettore di impulso*, ovvero un dispositivo che deve decidere per la presenza o assenza di una forma d'onda nota immersa nel rumore, in modo da rendere minima la probabilità di sbagliare. Un problema simile è già stato affrontato al § 7.6.2, ma ora ci riferiamo ad una *detezione coerente*, in cui il segnale è completamente specificato, compresa la sua temporizzazione o fase. Supponiamo quindi di trasmettere un segnale $x(t)$, ottenuto facendo transitare un impulso $\delta(t)$ in un filtro con risposta impulsiva (di durata limitata T) $h_T(t) = g\left(t - \frac{T}{2}\right)$, e di ricevere lo stesso segnale in presenza di rumore gaussiano a media nulla $n(t)$, con densità spettrale $\mathcal{P}_N(f) = \frac{N_0}{2}$.

Un ricevitore a filtro adattato effettua una decisione di *massima verosimiglianza* (vedi § 7.5.3) a riguardo della presenza (ipotesi H_1) o assenza (ipotesi H_0) del segnale $x(t)$ in base alla osservazione della grandezza $z(T)$ ottenuta (vedi figura 9.3) filtrando il segnale ricevuto

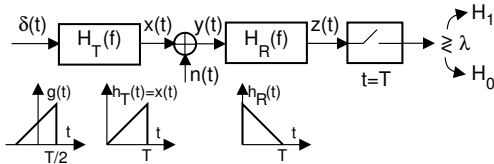
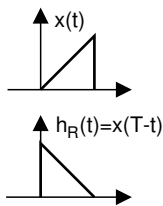


Figure 9.3: Detezione di impulso mediante filtro adattato $y(t)$ mediante il filtro (adattato) di ricezione $H_R(f)$, la cui uscita $z(t)$ è campionata all'istante $t = T$. Il valore osservato $z(T)$ è quindi confrontato con una soglia λ , determinando la decisione per H_1 o H_0 a seconda se λ sia superata o meno. Si commette un errore sia decidendo per H_1 in assenza di segnale, sia decidendo per H_0 in sua presenza²³. Si dimostra²⁴ che la probabilità di errore può essere resa minima se $H_R(f)$ è scelto in modo da rendere massimo il rapporto SNR all'istante di decisione, che corrisponde a scegliere²⁵

$$H_R(f) = H_T^*(f) e^{-j2\pi f T} = X^*(f) e^{-j2\pi f T} = G^*(f) e^{-j2\pi f \frac{T}{2}} \quad (9.8)$$

ovvero

$$h_R(t) = h_T^*(T-t) = x(T-t) = g^*\left(\frac{T}{2} - t\right) \quad (9.9)$$



Con tali scelte, nel caso H_0 in cui $x(t)$ è assente risulta $y(t) = n(t)$, e la grandezza di decisione $z(T)$ è una v.a. gaussiana²⁶ definita come

$$\begin{aligned} z(T) &= \int_{-\infty}^{\infty} h_R(\tau) y(T-\tau) d\tau = \\ &= \int_0^T x(T-\tau) n(T-\tau) d\tau = \mathcal{R}_{XN}(0) \end{aligned}$$

²³Indicando rispettivamente con P_{e0} e P_{e1} i due tipi di errore, pari a $P_{e0} = \int_{\lambda}^{\infty} p_Z(z/H_0) dz$ e $P_{e1} = \int_{-\infty}^{\lambda} p_Z(z/H_1) dz$, la probabilità di errore complessiva vale $P_e = P_{e0}P_0 + P_{e1}P_1$, in cui $P_0 = Pr(H_0)$ e $P_1 = Pr(H_1)$.

²⁴la dimostrazione è rimandata alla nota 31

²⁵Avendo definito $h_T(t) = g(t - \frac{T}{2})$, risulta che

$$H_T^*(f) = \left(G(f) e^{-j2\pi f \frac{T}{2}}\right)^* = G^*(f) e^{+j2\pi f \frac{T}{2}}$$

e dunque $H_T^*(f) e^{-j2\pi f T} = X^*(f) e^{-j2\pi f T} = G^*(f) e^{-j2\pi f \frac{T}{2}}$. D'altra parte, potendo scrivere $H_T^*(f) e^{-j2\pi f T} = \left(H_T(f) e^{j2\pi f T}\right)^*$ e ricordando ora la proprietà (3.3) espressa a pag. 34 $\mathcal{F}^{-1}\{X^*(f)\} = x^*(-t)$, otteniamo che

$$\begin{aligned} h_R(t) &= \mathcal{F}^{-1}\left\{H_T^*(f) e^{-j2\pi f T}\right\} = \mathcal{F}^{-1}\left\{\left(H_T(f) e^{j2\pi f T}\right)^*\right\} = \\ &= h_T^*(\theta + T)|_{\theta=-t} = h_T^*(T-t) = x^*(T-t) \end{aligned}$$

Infine, essendo $x(t) = g(t - T/2)$ si ottiene anche $h_R(t) = g^*(\theta - T/2)|_{\theta=T-t} = g^*(T/2 - t)$. La fig. 9.3 mostra l'esito di tali operazioni nel caso di una $g(t)$ triangolare.

²⁶Ricordiamo (vedi § 9.4.2) che l'uscita di un filtro al cui ingresso è posto un processo gaussiano, è anch'essa gaussiana.

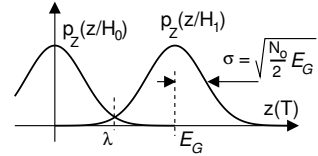
ossia pari all'intercorrelazione tra $x(t)$ e $n(t)$, calcolata nell'origine, e presenta valore atteso $m_{z(T)}^{H_0}$ nullo²⁷ e varianza²⁸ $\sigma_{z(T)}^2 = \frac{N_0}{2} \mathcal{E}_G$ in cui \mathcal{E}_G è l'energia dell'impulso $g(t)$.

Se invece il segnale è presente (ipotesi H_1), si ottiene

$$z(T) = \int_0^T x(T-\tau) [x(T-\tau) + n(T-\tau)] d\tau = \mathcal{R}_X(0) + \mathcal{R}_{XN}(0)$$

producendo in una grandezza di decisione $z(T)$ sempre gaussiana, con valor medio $m_{z(T)}^{H_1} = \mathcal{E}_G$ e varianza $\sigma_{z(T)}^2$ uguale al caso di assenza di segnale²⁹. Notiamo che $m_{z(T)}^{H_1} = \mathcal{E}_G$ non dipende dalla particolare $g(t)$ adottata, né dalla sua durata T , ma solo dalla sua energia, ed è per questo che il filtro di ricezione è detto *adattato*

La figura a fianco mostra la d.d.p. di $z(T)$ nelle ipotesi H_0 ed H_1 ; osserviamo ora (vedi nota 23) che nel caso in cui la probabilità a priori delle due ipotesi sia uguale (ovvero $P_o = Pr(H_0) = P_1 = Pr(H_1)$), fissando la soglia λ al valore $\frac{\mathcal{E}_G}{2}$ la probabilità di errore (calcolata nel seguito) risulta minima.



Il rapporto $\frac{(m_{z(T)}^{H_1})^2}{\sigma_{z(T)}^2}$ rappresenta³⁰ l'*SNR* all'istante di decisione, ed il suo valore

$$SNR = \frac{(\mathcal{E}_G)^2}{\frac{N_0}{2} \mathcal{E}_G} = \frac{2\mathcal{E}_G}{N_0} \quad (9.10)$$

costituisce il *massimo*³¹ che si può ottenere, adottando la (9.9) tra tutte le scelte possibili per il filtro di ricezione $h_R(t)$, di energia pari a \mathcal{E}_G . Notiamo che (9.10) è valida

²⁷ Infatti $m_{z(T)}^{H_0} = E\{\mathcal{R}_{XN}(0)\} = E\left\{\int_0^T x^*(t) n(t) dt\right\}$, che è pari a zero se $E\{n(t)\} = 0$.

²⁸ Risulta $\sigma_{z(T)}^2 = E\{z^2(T)\} = \mathcal{R}_Z(\tau)|_{\tau=0}$. Sappiamo che $\mathcal{R}_Z(\tau) = \mathcal{R}_N(\tau) * \mathcal{R}_{H_R}(\tau) = \frac{N_0}{2} \delta(\tau) * \mathcal{R}_{H_R}(\tau) = \frac{N_0}{2} \mathcal{R}_{H_R}(\tau)$; pertanto

$$\sigma_{z(T)}^2 = \frac{N_0}{2} \mathcal{R}_{H_R}(0) = \frac{N_0}{2} \int_{-\infty}^{\infty} h_R^*(t) h_R(t) dt = \frac{N_0}{2} \mathcal{E}_G$$

dato che $h_R(t)$ ha la stessa energia di $g(t)$.

²⁹ Infatti, ora risulta $m_{z(T)}^{H_1} = E\{\mathcal{R}_X(0) + \mathcal{R}_{XN}(0)\}$, in cui il contributo del secondo termine è nullo come già osservato, mentre quello del primo non è aleatorio, e vale $\mathcal{R}_X(0) = \int_0^T x^*(t) x(t) dt = \mathcal{E}_G$, in quanto il segnale $x(t)$ ha la stessa energia di $g(t)$. Per ciò che riguarda $\sigma_{z(T)}^2$, osserviamo che essendo il filtro di ricezione un operatore lineare, l'uscita si ottiene come sovrapposizione degli effetti delle due cause $x(t)$ ed $n(t)$, e la componente aleatoria dell'uscita è dovuta al solo $n(t)$; pertanto, la sua varianza è la stessa calcolata per il caso H_0 di segnale assente.

³⁰ Il significato fisico del rapporto indicato può essere meglio visualizzato considerandone le radici quadrate, ossia $\frac{m_{z(T)}^{H_1}}{\sigma_{z(T)}}$, che costituisce il rapporto tra l'uscita per $t = T$ in presenza di solo segnale, e la deviazione standard di tale valore introdotta dal rumore. In altre parole, è indicativo della separazione tra le gaussiane riportate in figura. Pertanto, maggiore è questo rapporto, e minore sarà la probabilità di errore.

³¹ Consideriamo il caso in cui si abbia una $H_R(f) = H(f)$ generica. In presenza di solo segnale, si ottiene

$$|z(T)|^2 = \left| F^{-1}\{Z(f)\} \Big|_{t=T} \right|^2 = \left| \int_{-\infty}^{\infty} H(f) X(f) e^{j2\pi f T} df \right|^2$$

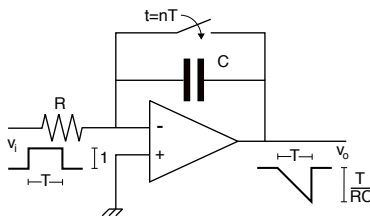
A questa espressione può essere applicata la disuguaglianza di Schwartz (a pag. 27 si enuncia la relazione $\left| \int_{-\infty}^{\infty} a(\theta) b^*(\theta) d\theta \right|^2 \leq \int_{-\infty}^{\infty} |a(\theta)|^2 d\theta \cdot \int_{-\infty}^{\infty} |b(\theta)|^2 d\theta$, con l'eguaglianza solo se $a(\theta) = k \cdot b(\theta)$), qualora si faccia corrispondere $H(f)$ ad $a(\theta)$, e $X(f) e^{j2\pi f T}$ a $b^*(\theta)$, ottenendo così

$$|z(T)|^2 = \left(m_{z(T)}^{H_1} \right)^2 \leq \int_{-\infty}^{\infty} |H(f)|^2 df \cdot \int_{-\infty}^{\infty} |X(f)|^2 df$$

solo in presenza di rumore bianco, mentre se questo è colorato, l' SNR diminuisce, ed il filtro ottimo va determinato nel modo specificato alla pagina seguente.

Integrate and dump Si tratta di una soluzione circuitale³² in grado di conseguire le prestazioni di detezione del filtro adattato nel caso di un $g(t) = \text{rect}_T(t)$, a cui corrisponde una $h_R(t)$ anch'essa rettangolare. Infatti, il segnale in uscita vale

$$v_o(t) = -\frac{1}{RC} \int_0^t v_i(t) dt + c$$



in cui c rappresenta $v_o(0)$ e quindi, a parte il segno ed il fattore $\frac{1}{RC}$, il valore $v_o(T)$ corrisponde a quello che si trova allo stesso istante in uscita da un filtro adattato³³ alla $g(t)$ rettangolare. L'interruttore che scarica (dump) la capacità C per $t = nT$ ha lo scopo di azzerare la costante c dopo la ricezione di ogni impulso, rendendo possibile l'uso dello schema per la detezione ottima di flussi binari a velocità $f_b = 1/T$.

Rumore colorato Nel caso in cui $\mathcal{P}_N(f)$ non sia pari ad una costante, la condizione per massimizzare (9.10) non è più la (9.8), bensì deve risultare³⁴

$$H_R(f) = \frac{X^*(f) e^{-j2\pi f T}}{\mathcal{P}_N(f)} \quad (9.11)$$

in modo che $H_R(f)$, oltre ad esaltare le frequenze per le quali lo spettro del segnale è maggiore, riesce anche ad attenuare quelle per le quali la potenza di rumore è più grande. Questa soluzione è nota come *filtro di WIENER*³⁵.

Assenza di rumore Se non fosse presente rumore, l'andamento dell'uscita del filtro adattato sarebbe proprio pari alla funzione di autocorrelazione di $g(t)$, che viene campionata in corrispondenza del suo massimo. Notiamo che la $H_R(f)$ non presenta modulo costante e fase lineare, dato che lo scopo qui non è quello di preservare la forma d'onda in transito, ma di massimizzare l' SNR all'istante di decisione.

con l'eguaglianza solo se $H(f) = kX^*(f) e^{-j2\pi f T}$, ovvero (vedi pag. 27) se $h(t) = kx(T-t)$, ossia se $H(f)$ è adattata a $X(f)$. Scegliendo $k = 1$, i due integrali a prodotto hanno lo stesso valore, pari a \mathcal{E}_G .

³²Vedi ad es. https://en.wikipedia.org/wiki/Op_amp_integrator

³³Il circuito non lineare mostrato non è un filtro adattato, dato che per $t \neq T$ non produce la stessa uscita (vedi <http://dsp.stackexchange.com/questions/9094/understanding-matched-filter>).

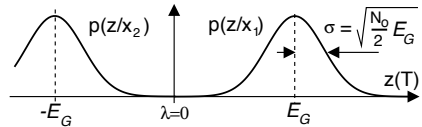
³⁴La condizione (9.11) si ottiene anche in questo caso imponendo la minimizzazione di $SNR = \frac{(m_z(T))^2}{\sigma_z^2(T)} = \frac{|\int_{-\infty}^{\infty} H(f) X(f) e^{j2\pi f T} df|^2}{\int_{-\infty}^{\infty} |H(f)|^2 \mathcal{P}_N(f) df}$ il cui denominatore tiene conto che $\sigma_z^2(T) = \int_{-\infty}^{\infty} \mathcal{P}_Z(f) df$ è dovuta al solo rumore. Applichiamo ora a SNR la disuguaglianza di Schwartz posta nella forma $\frac{|\int_{-\infty}^{\infty} a(\theta) b^*(\theta) d\theta|^2}{\int_{-\infty}^{\infty} |a(\theta)|^2 d\theta} \leq \int_{-\infty}^{\infty} |b(\theta)|^2 d\theta$ e identifichiamo $a(\theta)$ con $H(f) \sqrt{\mathcal{P}_N(f)}$ e $b^*(\theta)$ con $X(f) e^{j2\pi f T} / \sqrt{\mathcal{P}_N(f)}$. Imponendo di nuovo la condizione $a(\theta) = k \cdot b(\theta)$ con $k = 1$, otteniamo il massimo SNR come $SNR = \int_{-\infty}^{\infty} |b(\theta)|^2 d\theta = \int_{-\infty}^{\infty} \frac{|X(f)|^2}{\mathcal{P}_N(f)} df$, e quindi scrivendo $a(\theta) = b(\theta)$ ossia $H(f) \sqrt{\mathcal{P}_N(f)} = X^*(f) e^{-j2\pi f T} / \sqrt{\mathcal{P}_N(f)}$ si ottiene il risultato (9.11).

³⁵http://en.wikipedia.org/wiki/Wiener_filter

Probabilità di errore In base alla nota 23, nel caso di equiprobabilità delle due ipotesi H_1 e H_0 , la P_e è pari alla probabilità che una v.a. gaussiana con media nulla e varianza $\sigma^2 = \frac{N_0}{2} \mathcal{E}_G$ sia maggiore di $\frac{\mathcal{E}_G}{2}$; in accordo con la notazione introdotta in § 7.2.4, si ottiene $P_e = \frac{1}{2} \operatorname{erfc} \left\{ \frac{\mathcal{E}_G/2}{\sqrt{2} \sqrt{\frac{N_0}{2} \mathcal{E}_G}} \right\} = \frac{1}{2} \operatorname{erfc} \left\{ \frac{1}{2} \sqrt{\frac{\mathcal{E}_G}{N_0}} \right\}$. Ma questa non è ancora la *minima* P_e conseguibile, che si ottiene invece nel caso di *segnalazione antipodale*, molto più indicata nel caso in cui lo scopo del ricevitore non sia di individuare un solo impulso isolato, ma una sequenza simbolica binaria.

9.5.1 Segnalazione antipodale

Desiderando distinguere tra due possibili messaggi (ad es, x_1 ed x_2), e volendo rendere minima la probabilità di errore, la scelta *ottima* consiste nell'adottare $x_1(t) = g(t - \frac{T}{2})$ e $x_2(t) = -x_1(t)$, e di porre in ingresso al ricevitore un filtro adattato ad $x_1(t)$. In tal modo l'uscita del filtro all'istante di campionamento rappresenta una v.a. gaussiana z con media $m_z = \pm \mathcal{E}_G$ (a seconda se sia stato trasmesso x_1 o x_2) e varianza $\sigma_z^2 = \frac{N_0}{2} \mathcal{E}_G$. Pertanto ora la soglia di decisione è pari a zero, e nel caso di simboli equiprobabili, in base alla nota 23, la probabilità di errore P_e corrisponde a quella di osservare $z > \mathcal{E}_G$: in conseguenza del risultato (7.3) di pag. 155 si ottiene quindi



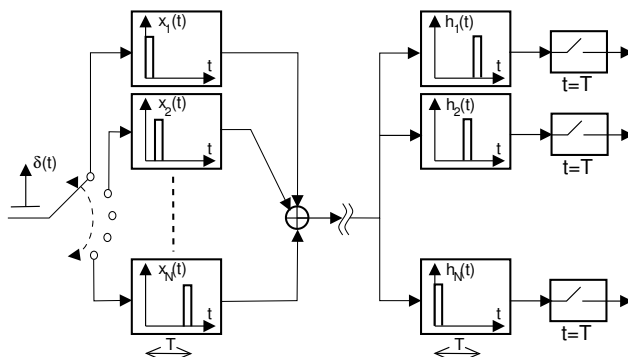
$$P_e = \frac{1}{2} \operatorname{erfc} \left\{ \frac{\mathcal{E}_G}{\sqrt{2} \sqrt{\frac{N_0}{2} \mathcal{E}_G}} \right\} = \frac{1}{2} \operatorname{erfc} \left\{ \sqrt{\frac{\mathcal{E}_G}{N_0}} \right\} \quad (9.12)$$

che, identificando in \mathcal{E}_G l'energia per bit E_b , fornisce un valore identico a quello di eq. (5.15) (pag. 89) e relativo al caso di trasmissione di un impulso di Nyquist a coseno rialzato con $L = 2$ e $\gamma = 0$, e con un filtro di ricezione passabasso. Ma se $\gamma = 0$, la trasmissione avviene a banda minima (§ 5.2.2.3), e dunque il filtro rettangolare passa-basso è proprio il filtro adattato!

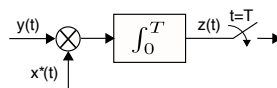
9.5.2 Segnalazione ortogonale

Dovendo trasmettere N diversi messaggi (x_1, x_2, \dots, x_N), associamo ad ognuno di essi una forma d'onda $x_i(t)$ tale che $\int x_i(t) x_j(t) dt = 0$ con $i \neq j$, ovvero in modo che i segnali $x_i(t)$ siano *ortogonali*. In tal caso il ricevitore ottimo è costituito da un banco di filtri, ognuno adattato ad una diversa $x_i(t)$, e in assenza di rumore la ricezione di una delle forme d'onda $x_i(t)$ non produce nessuna uscita sui filtri del banco per $j \neq i$. In presenza di rumore, la decisione su cosa sia stato trasmesso viene presa valutando quale dei filtri presenta il valore massimo in corrispondenza dell'istante di campionamento, realizzando così un *ricevitore a correlazione* (vedi § 13.5 a pag. 322).

Esempio L'impulso $\delta(t)$ entra in uno di filtri mostrati nella figura seguente, le cui risposte impulsive $x_i(t)$ realizzano una famiglia di funzioni ortogonali, dato che le rispettive forme d'onda non si sovrappongono nel tempo. In ricezione, solo uno dei filtri adattati con risposta impulsiva $h_i(t)$ produce una uscita diversa da zero per $t = T$, come verificabile ricordando la costruzione grafica dell'operazione di convoluzione mostrata a pag. 38.



Correlatore Si tratta di un modo alternativo di realizzare un filtro adattato, derivante dall'osservazione che il suo ruolo essenzialmente si riduce al calcolo della *intercorrelazione* (eq. 9.5) tra il segnale ricevuto e quello atteso. Tale funzione può essere realizzata anche ricorrendo allo schema in figura, dove un integratore (implementato ad es. mediante il circuito *integrate and dump*, pag. 232) opera sul prodotto tra il segnale in arrivo ed una copia locale del coniugato³⁶ della forma d'onda trasmessa. Un caso di applicazione di questo schema si trova a pag. 322 a proposito del ricevitore FSK ortogonale, mentre una variante idonea a stimare l'*autocorrelazione* di $y(t)$ è proposta al § 9.7.3.



9.6 Caratteristiche dei sistemi fisici

Abbiamo già osservato come il legame $\mathcal{T}[x(t)] = y(t)$ definito al § 1.7.1 tra ingresso $x(t)$ ed uscita $y(t)$ di un sistema fisico sia definito, in base alla conoscenza della risposta impulsiva, dall'integrale di convoluzione $y(t) = x(t) * h(t)$, che descrive il risultato di una operazione di filtraggio. Alla luce di questo risultato, torniamo ad analizzare le proprietà enunciate al § 1.7.1, assieme ad altre due.

Linearità Un sistema è lineare se, in presenza di una combinazione lineare di ingressi, l'uscita è la combinazione lineare delle uscite, ossia sussiste la legge di sovrapposizione degli effetti, ovvero

$$\mathcal{T} \left[\sum_i a_i x_i(t) \right] = \sum_i a_i \mathcal{T}[x_i(t)]$$

Ad esempio, il legame ingresso-uscita descritto dall'integrale di convoluzione è di tipo *lineare*, in virtù della distributività dell'integrale. Al contrario, un operatore basato sulla elevazione a potenza è *non lineare*, come approfondito al § 14.7.

Memoria Notiamo che un sistema descritto da una risposta impulsiva $h(t)$ con estensione temporale non nulla, è detto *con memoria*, in quanto i singoli valori di uscita dipendono da tutti i valori di ingresso "raccolti" dalla risposta impulsiva.

³⁶ Anche se nel caso di banda base il segnale trasmesso è reale, volendo applicare la teoria esposta ad un involuppo complesso (§ 10.2.1) si rende necessario tener conto dell'operazione di coniugato.

Permanenza Un sistema è permanente se la risposta ad un ingresso ritardato è anch'essa ritardata. Se la risposta impulsiva non cambia nel tempo, il sistema è permanente.

Realizzabilità ideale Un sistema è idealmente realizzabile se $h(t)$ è reale.

Realizzabilità fisica La proprietà di osservare valori di uscita che dipendono solo dagli ingressi passati, è automaticamente verificata se $h(t) = 0$ con $t < 0$. Abbiamo già osservato (nota 14 a pag. 81) come sistemi non realizzabili fisicamente possano essere approssimati da sistemi realizzabili accettando un ritardo dell'uscita.

Stabilità Si può mostrare che la proprietà di stabilità equivale alla condizione $\int |h(t)| dt < \infty$, ovvero che $h(t)$ sia un segnale impulsivo. Notiamo che questa circostanza garantisce l'esistenza della sua trasformata $H(f)$.

Risposta in frequenza Se un sistema, oltre che stabile, è anche idealmente realizzabile, allora $H(f) = H^*(-f)$, e dunque è sufficiente conoscere la parte a frequenze positive indicata con $H^+(f)$, dato che l'altra metà è ottenibile mediante una operazione di coniugazione. Questo fatto permette di misurare

$$H(f) = M(f) e^{j\varphi(f)}$$

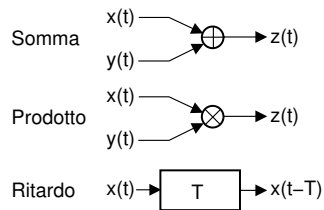
(ossia modulo e fase di $H(f)$), che prende il nome di *risposta in frequenza*, utilizzando come ingresso una funzione sinusoidale con ampiezza A e fase θ note: $x(t) = A \cos(2\pi f_0 t + \theta)$. Il segnale in uscita sarà ancora una cosinusoide³⁷ e avrà ampiezza $A \cdot M(f_0)$ e fase $\varphi(f_0) + \theta$; pertanto potremo ricavare

$$M(f_0) = \frac{\max\{y(t)\}}{\max\{x(t)\}}, \quad \text{e} \quad \varphi(f_0) = \arg\{y(t)\} - \arg\{x(t)\}$$

Ripetendo il procedimento per diverse f_0 , possiamo "campionare" $H(f)$.

9.7 Unità di elaborazione dei segnali

Mentre un filtro viene definito, in base alla precedente classificazione, un operatore lineare con memoria, può essere utile considerare altri *operatori elementari*, che rivestono la funzione di "mattoni" per realizzare dispositivi più complessi. Essi rappresentano le operazioni di somma, prodotto, e ritardo, come indicato in figura. Analizziamo ora nel dettaglio il risultato della combinazione di processi e segnali certi mediante gli operatori introdotti, mentre alla prossima sezione descriviamo l'utilizzo di questi operatori per la realizzazione di elementi filtranti.



³⁷Svolgiamo i calcoli nel dominio della frequenza:

$$X(f) = \frac{A}{2} \left(e^{j\theta} \delta(f - f_0) + e^{-j\theta} \delta(f + f_0) \right);$$

$$Y(f) = X(f) H(f) = \frac{A}{2} M(f_0) \left(e^{j\theta} e^{j\varphi(f_0)} \delta(f - f_0) + e^{-j\theta} e^{-j\varphi(f_0)} \delta(f + f_0) \right)$$

e antitrasformando si ottiene

$$y(t) = A \cdot M(f_0) \cos(2\pi f_0 t + \varphi(f_0) + \theta)$$

9.7.1 Prodotto

Nel caso in cui un fattore sia un processo, e l'altro un segnale certo, il risultato (in generale) è un processo *non stazionario*. Infatti ora le medie d'insieme dipendono, istante per istante, dal valore che il segnale certo assume in quell'istante (tranne il caso in cui sia una costante)³⁸.

Se uno dei due fattori (ad es $y(t) = \tilde{y}$) è una costante, $z(t)$ è un processo della stessa natura di $x(t)$, con media $m_z = m_x \cdot \tilde{y}$, potenza $\mathcal{P}_z = \mathcal{P}_x \cdot \tilde{y}^2$, e autocorrelazione $\mathcal{R}_z(\tau) = \mathcal{R}_x(\tau) \cdot \tilde{y}^2$.

Se i fattori $x(t)$ ed $y(t)$ sono processi statisticamente indipendenti³⁹, stazionari e congiuntamente⁴⁰ ergodici, per il loro prodotto si ottiene:

Valor medio

$$m_z = E\{z(t)\} = E\{x(t)y(t)\} = E\{x(t)\}E\{y(t)\} = m_x \cdot m_y$$

Potenza totale

$$\mathcal{P}_z = E\{z^2(t)\} = E\{x^2(t)y^2(t)\} = E\{x^2(t)\}E\{y^2(t)\} = \mathcal{P}_x \cdot \mathcal{P}_y$$

Varianza

$$\sigma_z^2 = E\{(z(t) - m_z)^2\} = \mathcal{P}_z - (m_z)^2 = \mathcal{P}_x \cdot \mathcal{P}_y - (m_x \cdot m_y)^2$$

Funzione di autocorrelazione

$$\begin{aligned} \mathcal{R}_z(\tau) &= E\{z(t)z(t+\tau)\} = E\{x(t)y(t)x(t+\tau)y(t+\tau)\} = \\ &= E\{x(t)x(t+\tau)\}E\{y(t)y(t+\tau)\} = \mathcal{R}_x(\tau) \cdot \mathcal{R}_y(\tau) \end{aligned}$$

In particolare, notiamo che l'incorrelazione di uno dei due processi, per un certo valore di τ , provoca l'incorrelazione del prodotto.

Spettro di densità di potenza

$$\mathcal{P}_z(f) = \mathcal{F}\{\mathcal{R}_z(\tau)\} = \mathcal{F}\{\mathcal{R}_x(\tau) \cdot \mathcal{R}_y(\tau)\} = \mathcal{P}_x(f) * \mathcal{P}_y(f)$$

ossia è pari alla convoluzione tra le densità spettrali dei fattori. Notiamo quindi che la densità di potenza del prodotto presenta una occupazione di banda maggiore di quella dei singoli fattori.

³⁸Se il segnale certo è periodico, il risultato della moltiplicazione per un processo stazionario dà luogo ad un processo detto *ciclostazionario*, in quanto le statistiche variano nel tempo, ma assumono valori identici con periodicità uguale a quella del segnale certo.

³⁹Compreso il caso di processi armonici!

⁴⁰La proprietà di ergodicità congiunta corrisponde a verificare le condizioni ergodiche anche per i momenti misti $m_{XY}^{(1,1)}(x, y)$ relativi a coppie di valori estratti da realizzazioni di due differenti processi.

Densità di probabilità Si calcola con le regole per il cambiamento di variabile, illustrate al § 7.4. Nel caso in cui i due processi siano statisticamente indipendenti, il risultato è

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(\theta) p_Y\left(\frac{z}{\theta}\right) \frac{d\theta}{\theta}$$

In Appendice (pag. 243), troviamo l'applicazione di questi risultati al calcolo della densità di potenza di un segnale dati.

9.7.2 Somma

Anche in questo caso, se un termine è un processo e l'altro un segnale certo, la somma è (in generale) un segnale non stazionario. Se il segnale certo è costante, si torna al caso stazionario⁴¹. Procediamo ora nel calcolo delle solite grandezze, con l'ipotesi che $x(t)$ ed $y(t)$ siano processi *statisticamente indipendenti*.

Valore medio

$$m_z = E\{x(t) + y(t)\} = E\{x(t)\} + E\{y(t)\} = m_x + m_y$$

Potenza totale

$$\begin{aligned} \mathcal{P}_z &= E\{(x(t) + y(t))^2\} = E\{x^2(t)\} + E\{y^2(t)\} + 2E\{x(t) \cdot y(t)\} \\ &= \mathcal{P}_x + \mathcal{P}_y + 2m_x m_y \end{aligned}$$

Pertanto, se almeno uno dei due processi è a media nulla, le potenze dei due processi si sommano.

Varianza

$$\begin{aligned} \sigma_z^2 &= E\{(z(t) - m_z)^2\} = \mathcal{P}_z - (m_z)^2 = \mathcal{P}_x + \mathcal{P}_y + 2m_x m_y - (m_x + m_y)^2 = \\ &= \mathcal{P}_x - (m_x)^2 + \mathcal{P}_y - (m_y)^2 = \sigma_x^2 + \sigma_y^2 \end{aligned}$$

Autocorrelazione

$$\begin{aligned} \mathcal{R}_z(\tau) &= E\{z(t)z(t+\tau)\} = E\{x(t)x(t+\tau)\} + \\ &E\{y(t)y(t+\tau)\} + E\{x(t)y(t+\tau)\} + E\{x(t+\tau)y(t)\} = \\ &= \mathcal{R}_x(\tau) + \mathcal{R}_y(\tau) + 2m_x m_y \end{aligned}$$

Osserviamo come per $\tau = 0$ si ritrovi il valore della potenza totale.

Spettro di densità di potenza

$$\mathcal{P}_z(f) = \mathcal{F}\{\mathcal{R}_z(\tau)\} = \mathcal{P}_x(f) + \mathcal{P}_y(f) + 2m_x m_y \delta(f)$$

⁴¹Come per il prodotto, se il segnale certo è periodico, la somma si dice *ciclostazionaria* perché la dipendenza temporale non è assoluta, ma periodica.

Densità di probabilità Applicando le regole del cambiamento di variabile (§ 7.4), o passando per il calcolo della funzione caratteristica (§ 7.6.1), nel caso di $x(t)$ ed $y(t)$ indipendenti, si ottiene

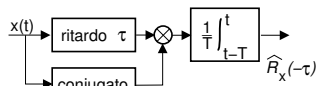
$$p_Z(z) = \int_{-\infty}^{\infty} p_X(\theta) p_Y(z - \theta) d\theta = p_X(x) * p_Y(y)$$

La relazione esprime l'importante risultato che la densità di probabilità della somma di variabili aleatorie è la *convoluzione* tra le densità dei termini della somma.

Esempio Se x ed y sono ad es. due v.a. a distribuzione uniforme tra $\pm\Delta$, la loro somma ha densità di probabilità triangolare con base 2Δ . Pertanto, nel lancio di 2 dadi il risultato più probabile è 7. Infatti può essere ottenuto come $6+1$, $5+2$, $4+3$, $3+4$, $2+5$, $1+6$, ovvero in 6 modi diversi, ognuno con probabilità $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$ e dunque $Pr\{7\} = 6 \frac{1}{36} = \frac{1}{6}$.

9.7.3 Stima della autocorrelazione

Come primo esempio dell'uso degli operatori elementari, la figura mostra l'architettura di uno schema di elaborazione idoneo a misurare una stima⁴² $\widehat{\mathcal{R}}_x(\tau)$ della funzione di autocorrelazione (§ 9.1.4) di un segnale $x(t)$ per un ritardo τ assegnato. Variando quest'ultimo si ottiene $\widehat{\mathcal{R}}_x(\tau)$ per i diversi valori di τ , e nel caso di segnali stazionari sarà poi possibile calcolare $\widehat{\mathcal{P}}_x(f) = \mathcal{F}\{\widehat{\mathcal{R}}_x(\tau)\}$; se infine $x(t)$ è un membro di un processo ergodico, $\widehat{\mathcal{P}}_x(f)$ rappresenta una stima della densità di potenza per una qualunque realizzazione.



9.8 Filtri digitali

Sono una particolare classe di filtri, per i quali l'integrale di convoluzione si riduce in una sommatoria. Sebbene nel seguito sia analizzato il solo caso di segnali tempo-continui, i filtri digitali sono di particolare rilievo perché permettono di svolgere le operazioni di filtraggio operando direttamente sui campioni dei segnali (vedi § 4.5.1), e dunque possono essere realizzati via software o hardware logico⁴³.

9.8.1 Filtro trasversale

E' una particolare architettura di filtro digitale, mostrata nella parte sinistra di fig. 9.4, e per la quale è facile verificare come la risposta impulsiva abbia espressione

$$h(t) = \sum_{n=0}^N c_n \delta(t - nT) \quad (9.13)$$

in cui N è l'ordine del filtro⁴⁴, che essendo finito, fa rientrare il filtro trasversale nella categoria dei filtri FIR (*finite impulse response*). E' altrettanto facile verificare che la

⁴²Si tratta di una *stima* (vedi § 7.5.4) in quanto l'intervallo di integrazione T è limitato.

⁴³Il tema delle realizzazioni numeriche dei filtri digitali non è al momento qui svolto, ma citiamo come fonte di approfondimento <http://www.dspguide.com/ch14/6.htm>.

⁴⁴I coefficienti c_n vengono indicati nei testi inglesi come *taps* (rubinetti) in quanto possono essere pensati "spillare" frazioni del segnale. Per effetto di un processo di trasposizione linguistica, gli stessi coefficienti in italiano vengono a volte indicati discorsivamente come *tappi* (!).

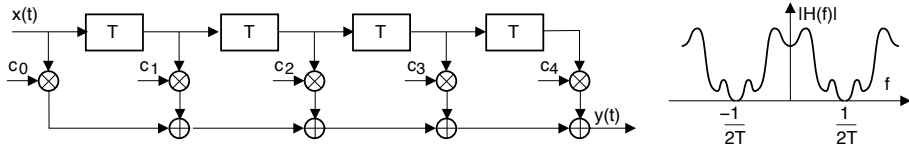


Figura 9.4: Schema simbolico di un filtro trasversale e sua risposta in frequenza

risposta in frequenza associata alla (9.13) risulta

$$H(f) = \mathcal{F}\{h(t)\} = \sum_{n=0}^N c_n e^{-j2\pi f n T}$$

Una tale architettura può essere rappresentativa di un effettivo fenomeno naturale, come ad esempio la presenza di “echi” nel segnale ricevuto. Come mostrato a destra in fig. 9.4, $H(f)$ risulta periodica (in frequenza) con periodo $F = \frac{1}{T}$, dato che tutti gli esponenziali $e^{-j2\pi f n T}$ lo sono. Pertanto questa architettura può essere adottata per sintetizzare una risposta in frequenza desiderata in una certa gamma di frequenze, e “qualsiasi” altrove. Infatti, se il segnale di ingresso $x(t)$ è limitato in una banda (base o traslata) di estensione minore di $\frac{1}{T}$, l’azione filtrante ha luogo appunto nella sola banda del segnale.

Per sintetizzare il filtro a partire dall’andamento desiderato di $H(f)$ nella banda di interesse, si calcolano i coefficienti c_k mediante la formula

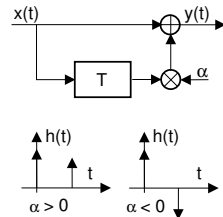
$$c_k = T \int_{-1/2T}^{1/2T} H(f) e^{j2\pi f k T} df \tag{9.14}$$

che è del tutto analoga alla eq. (4.7) di pag. 60 od (a parte il segno di f) alla eq. (2.2) di pag. 17. Ovviamente, se $H(f)$ è qualsiasi, occorrerebbero un numero infinito di coefficienti c_k ; usandone un numero inferiore (finito) si produce una approssimazione della $H(f)$ desiderata⁴⁵.

9.8.2 Filtro trasversale del 1° ordine

E’ descritto dalla architettura mostrata a lato, a cui corrisponde una risposta impulsiva

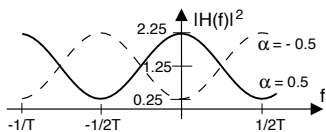
$$h(t) = \delta(t) + \alpha \delta(t - T)$$



⁴⁵Il troncamento (§ 2.2.2) della serie di coefficienti c_k avviene in modo simmetrico rispetto a c_0 , prendendo cioè sia gli indici positivi che quelli negativi. Viceversa, nello schema di filtro trasversale si usano solo coefficienti con indici ≥ 0 . Nel caso in cui l’ $H(f)$ da cui partiamo sia reale pari, allora i c_k sono una serie reale (pari), garantendo un filtro idealmente realizzabile, ma la cui $h(t) = \sum_{k=-N/2}^{N/2} c_k \delta(t - kT)$ necessita di una traslazione temporale per essere anche fisicamente realizzabile. Se invece $H(f)$ ha un andamento qualunque, non si può dire nulla a riguardo di eventuali simmetrie per i coefficienti c_k .

la cui trasformata è $H(f) = 1 + \alpha e^{-j2\pi fT}$ ⁽⁴⁶⁾, e quindi

$$\begin{aligned} |H(f)|^2 &= (1 + \alpha \cos 2\pi fT)^2 + (\alpha \sin 2\pi fT)^2 = \\ &= 1 + 2\alpha \cos 2\pi fT + \alpha^2 (\cos^2 2\pi fT + \sin^2 2\pi fT) = \\ &= 1 + \alpha^2 + 2\alpha \cos 2\pi fT \end{aligned}$$

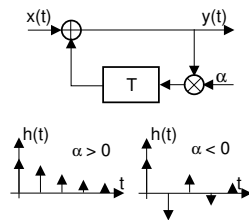


A lato è riportato l'andamento di $|H(f)|^2$ per due valori di $\alpha = \pm 0.5$, ed osserviamo che nell'intervallo di frequenze $|f| < \frac{1}{2T}$ può assumere un comportamento passa-alto oppure passa-basso⁴⁷, in funzione del segno di α . Notiamo inoltre che ponendo $\alpha = -1$ si ottiene un *differenziatore*, in grado di rimuovere dall'ingresso segnali periodici di periodo T .

9.8.3 Filtro a risposta impulsiva *infinita* del 1° ordine

L'architettura rappresentata in figura identifica un filtro di tipo *ricorsivo*, ovvero il cui valore in uscita dipende dalle *uscite precedenti*. La risposta impulsiva corrispondente ha una *durata infinita*⁴⁸, ed è pari a

$$h(t) = \sum_{n=0}^{\infty} \alpha^n \delta(t - nT)$$



La funzione di trasferimento risulta quindi pari a $H(f) = \sum_{n=0}^{\infty} \alpha^n e^{-j2\pi f nT}$ e, ricorrendo la formula dello sviluppo in serie geometrica $\sum_{n=0}^{\infty} \beta^n = \frac{1}{1-\beta}$, si può scrivere

$$H(f) = \sum_{n=0}^{\infty} \left(\alpha e^{-j2\pi fT} \right)^n = \frac{1}{1 - \alpha e^{-j2\pi fT}}$$

Notiamo subito che il filtro è stabile purché $|\alpha| < 1$, altrimenti si ha una uscita anche senza ingresso, ovvero uscita infinita con ingresso limitato. Per ciò che riguarda $|H(f)|^2$, otteniamo

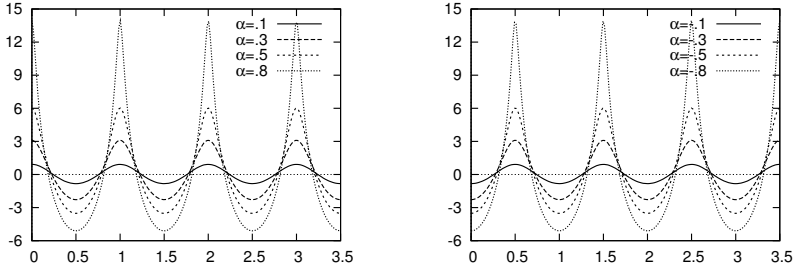
$$|H(f)|^2 = \frac{1}{(1 - \alpha \cos 2\pi fT)^2 + (\alpha \sin 2\pi fT)^2} = \frac{1}{1 + \alpha^2 - 2\alpha \cos 2\pi fT}$$

Le curve mostrate appresso rappresentano $10 \log_{10} |H(f)|^2$, con $T = 1$ e diversi valori di α , positivi a sinistra e negativi a destra. Osserviamo infine che il caso $\alpha = 1$ corrisponde ad avere un integratore perfetto che, ad esempio, produce una rampa in uscita, se in ingresso c'è un gradino.

⁴⁶In questo caso pur risultando $H(f)$ a simmetria coniugata ($H(f) = H^*(-f)$), è complessa. Pertanto, i coefficienti c_k ottenibili dalla (9.14) sono reali, ma non necessariamente pari. Svolgendo i calcoli, si ha: $c_k = T \int_{-1/2T}^{1/2T} (1 + \alpha e^{-j2\pi fT}) e^{j2\pi f k T} df = T \int_{-1/2T}^{1/2T} e^{j2\pi f k T} df + \alpha T \int_{-1/2T}^{1/2T} e^{j2\pi f(k-1)T} df$. Il primo integrale è nullo per $k \neq 0$, mentre il secondo per $k \neq 1$, in quanto le funzioni integrande hanno media nulla sull'intervallo $1/T$; pertanto $c_0 = 1$ e $c_1 = \alpha$, esattamente come è definita la risposta impulsiva.

⁴⁷In altre parole, l'andamento ondulatorio di $|H(f)|^2$ rende il filtro idoneo a diversi utilizzi, in funzione dell'andamento in frequenza del segnale di ingresso.

⁴⁸In questo caso si parla di filtro IIR o *infinite impulse response*.



9.9 Filtri analogici

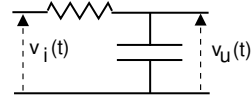
Sono ottenuti mediante componenti elettrici a costanti concentrate come condensatori, induttori e resistori. Applicando la trasformata di Laplace alle equazioni differenziali che descrivono la relazione ingresso-uscita, si ottiene una funzione di trasferimento razionale del tipo

$$H(s) = \frac{\sum_{i=0}^N a_i s^i}{\sum_{j=0}^M b_j s^j}$$

(in cui $N \leq M$), definita su di un piano complesso $s = \sigma + j2\pi f$. Ponendo $s = j2\pi f$ si ottiene la funzione di trasferimento in f : $H(f) = H(s = j2\pi f)$. Questo procedimento è valido solo se il filtro è stabile, che nel dominio di Laplace equivale a richiedere che tutti i poli di $H(s)$ siano a sinistra dell'asse immaginario.

9.9.1 Filtro analogico ad un polo

La tensione $v_u(t)$ in uscita dal filtro RC di tipo passa basso mostrato in figura può essere espressa come convoluzione tra quella di ingresso $v_i(t)$ ed una risposta impulsiva



$$h(t) = \frac{1}{RC} e^{-\frac{t}{RC}}$$

L'analisi del circuito mostra che la funzione di trasferimento risulta

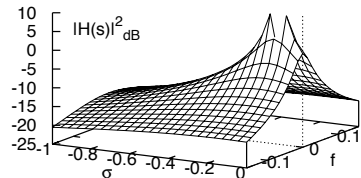
$$H(f) = \mathcal{F}\{h(t)\} = \frac{1/j\omega C}{R + 1/j\omega C} = \frac{1}{1 + j2\pi f RC}$$

ovvero, nel dominio di Laplace

$$H(s) = \frac{1}{1 + sRC}$$

Pertanto, $H(s)$ presenta un polo in $s = -\frac{1}{RC}$ che fa sì che $H(s)|_{s=-\frac{1}{RC}} = \infty$.

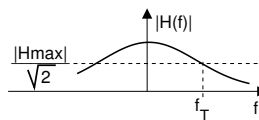
A lato è raffigurato l'andamento di $|H(s)|^2$, espresso in decibel, e con $RC = 8$. Come evidente, $|H(s)|^2$ può essere pensata come una sorta di *cono vulcanico* attorno al polo, le cui falde, quando intersecate dal piano verticale infisso sull'asse $j2\pi f$, individuano la funzione di trasferimento in frequenza $H(f) = H(s = j2\pi f)$. Come si vede dalla figura, $H(f)$ risulta di tipo passa basso, con fianchi tanto più ripidi quanto più il polo è vicino all'origine.



9.9.2 Frequenza di taglio

Definiamo *frequenza di taglio* di un filtro la frequenza f_T per la quale

$$|H(f_T)| = \frac{|H_{Max}|}{\sqrt{2}}$$



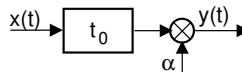
Nel caso del filtro RC, si ha $|H_{Max}| = H(0) = 1$ e dunque scriviamo

$$|H(f)| = \frac{1}{\sqrt{1 + (2\pi fRC)^2}} = \frac{1}{\sqrt{1 + \left(\frac{f}{f_T}\right)^2}}$$

in cui $f_T = \frac{1}{2\pi RC}$, pari quindi alla frequenza di taglio (infatti $|H(f_T)| = \frac{1}{\sqrt{1+1}} = \frac{1}{\sqrt{2}}$). Notiamo infine che $|H(f_T)|^2 = \frac{1}{2}$ e dunque $|H(f_T)|^2|_{dB} = -3$ dB; per questo la frequenza di taglio è indicata anche come frequenza a 3 dB.

9.9.3 Assenza di distorsione lineare

Quali proprietà devono essere verificate da un filtro affinché l'uscita non differisca dall'ingresso per più di un fattore di scala ed un ritardo, ovvero si verifichi la proprietà di *canale perfetto* espressa a pag. 349? La condizione cercata si esprime come $y(t) = \alpha x(t - t_0)$, che corrisponde a $Y(f) = \alpha X(f) e^{-j2\pi f t_0}$, e quindi la risposta in frequenza di tale filtro deve essere del tipo



$$H(f) = \frac{Y(f)}{X(f)} = \alpha e^{-j2\pi f t_0}$$

ovvero la sua risposta impulsiva è pari a $h(t) = \alpha \delta(t - t_0)$. Pertanto le condizioni poste nel tempo, si riflettono su di una risposta in frequenza con *modulo costante* e *fase lineare*, quantomeno, nella banda del segnale.

9.10 Appendici

9.10.1 Coefficiente di correlazione

I diagrammi di esempio presentati alla nota (2) a pag. 218 basano la valutazione di quanto una coppia di v.a. x ed y siano correlate, anche sul calcolo del coefficiente di correlazione ρ_{xy} , che ha valori compresi tra +1 e -1, ed è definito come

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

In tal modo, si opera una normalizzazione del valore della covarianza σ_{xy} , rispetto alle deviazioni standard σ_x e σ_y delle due v.a., rendendo così il valore di ρ indipendente dalla dinamica dei valori assunti da x ed y .

Il coefficiente ρ si presta ad una interessante interpretazione geometrica, una volta identificate (vedi § 2.4.1) la deviazione standard σ_x come la norma $\|\bar{x}\|$ di x , e la covarianza σ_{xy} come il prodotto scalare (\bar{x}, \bar{y}) tra x ed y ⁴⁹. In tale contesto, possiamo

⁴⁹L'analogia non è poi troppo peregrina, considerando che se x è estratta da un processo ergodico a media nulla, la sua varianza σ_x^2 coincide con la potenza del segnale da cui è estratta, mentre se x ed y sono estratte da segnali congiuntamente ergodici, la covarianza σ_{xy} coincide con la funzione di intercorrelazione (eq. (9.3)), ovvero con la loro *potenza mutua*.

indicare due v.a. come *ortogonali* se risulta $\sigma_{xy} = \rho_{xy} = 0$, mentre un valore $\rho_{xy} = \pm 1$ indica che una delle due v.a. è costantemente uguale all'altra, a meno di un coefficiente costante. Notiamo che l'ortogonalità $\rho_{xy} = 0$ esprime unicamente l'assenza di legami di tipo *lineare* tra x ed y , come esemplificato dal caso F) della nota (2) a pag. 218.

Per ultima citiamo l'estensione formale del risultato noto come *diseguaglianza di Schwartz* (pag. 27), una volta che al coefficiente di correlazione ρ_{xy} sia stato associato il concetto di coseno tra x ed y : tale posizione deriva dall'essere $-1 < \rho_{xy} < 1$, e permette di asserire che $|\sigma_{xy}| \leq \sigma_x \sigma_y$.

9.10.2 Gaussiana multidimensionale ed indipendenza statistica

In fondo al § 9.1.2 si è affermato che, *unicamente nel caso di v.a. congiuntamente gaussiane*, il sussistere di incorrelazione tra le stesse ne implica l'indipendenza statistica: mostriamo qui il motivo. Iniziamo con il definire la d.d.p. di una v.a. gaussiana multidimensionale \mathbf{X} come

$$p_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_x}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m}_x) \Sigma_x^{-1} (\mathbf{x} - \mathbf{m}_x)^T \right\} \quad (9.15)$$

in cui $\mathbf{x} = [x_1, x_2, \dots, x_n]$ è un vettore riga che rappresenta le n v.a. marginali, \mathbf{m}_x è il vettore dei rispettivi valori medi, e Σ_x è la *matrice di covarianza* i cui $n \times n$ elementi risultano pari a $\sigma_{x_i, x_j} = E \{ (x_i - m_{x_i})(x_j - m_{x_j}) \}$. Notiamo che la conoscenza di \mathbf{m}_x e Σ_x definisce in modo *completo* la densità di probabilità.

Osserviamo ora che nel caso in cui le v.a. siano incorrelate, ossia $\sigma_{x_i, x_j} = 0$ con $i \neq j$, la matrice di covarianza Σ_x risulta essere *diagonale*, e così la sua inversa, i cui elementi risultano in tal caso essere pari a $1/\sigma_{x_i}^2$. Sotto queste ipotesi la (9.15) si esprime come

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_{x_i}}} \exp \left\{ -\frac{1}{2} \left[\sum_{i=1}^n \frac{(x_i - m_{x_i})^2}{\sigma_{x_i}^2} \right] \right\}$$

e dunque equivale al *prodotto* delle singole d.d.p. marginali⁵⁰

$$p(x_i) = \frac{1}{\sqrt{2\pi} \sigma_{x_i}} \exp \left\{ -\frac{1}{2} \frac{(x_i - m_{x_i})^2}{\sigma_{x_i}^2} \right\}$$

Ma dato che questo risultato è proprio la definizione di indipendenza statistica (§ 9.1.2) tra le v.a. marginali, abbiamo ottenuto la dimostrazione cercata.

9.10.3 Densità spettrale per onda PAM

L'acronimo PAM sta per *Pulse Amplitude Modulation* (pag. 5.1.2), e individua una classe di segnali realizzati ripetendo indefinitivamente uno stesso *impulso* elementare $g(t)$ con periodo T , ognuno moltiplicato (o *modulato in ampiezza*) per un diverso coefficiente a_n . Entro questa generica definizione rientra sia il segnale dati, che la conversione D/A per segnali campionati, di cui in questa sezione si fornisce una trattazione unificata.

⁵⁰Si verifichi per esercizio che nel caso di una coppia di v.a. congiuntamente gaussiane, a media nulla ed uguale varianza, si ottiene l'espressione (7.8) di pag. 163.

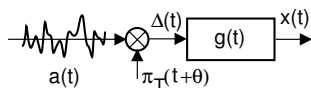
Accingiamoci pertanto a verificare quanto affermato al § 9.2.4, ossia che al segnale dati

$$x(t) = \sum a_n g(t - nT - \theta) \quad (9.16)$$

corrisponde uno spettro di densità di potenza pari a

$$\mathcal{P}_x(f) = \sigma_A^2 \frac{|G(f)|^2}{T} \quad (9.17)$$

Il segnale (9.16) rappresenta un processo stazionario ergodico qualora θ sia una v.a. aleatoria uniformemente distribuita tra $\pm \frac{T}{2}$, ed i valori a_n realizzazioni di v.a. statisticamente indipendenti, a media nulla ed identicamente distribuite, e varianza $\sigma_A^2 = E\{a_n^2\}$.



Per verificare la (9.17) adottiamo il modello di generazione del segnale (9.16) rappresentato in figura: un processo gaussiano bianco a media nulla $a(t)$ è moltiplicato per un processo impulsivo $\pi_T(t - \theta) = \sum_n \delta(t - nT - \theta)$, dando luogo ad un nuovo processo impulsivo $\Delta(t) = \sum_n a_n \delta(t - nT - \theta)$ in cui i coefficienti a_n sono (per costruzione) statisticamente indipendenti e quindi incorrelati, in modo che $\mathcal{R}_a(\tau) = E\{a_n a_{n+\tau}\} = \begin{cases} \sigma_A^2 & \text{per } \tau = 0 \\ 0 & \text{altrimenti} \end{cases}$, e quindi

$$\mathcal{R}_\Delta(\tau) = \mathcal{R}_\pi(\tau) \cdot \mathcal{R}_a(\tau) = \begin{cases} \sigma_A^2 \cdot \mathcal{R}_\pi(0) & \text{per } \tau = 0 \\ 0 & \text{altrimenti} \end{cases} \quad (9.18)$$

Per quanto riguarda $\mathcal{R}_\pi(\tau)$, osserviamo che

- deve risultare periodica con periodo T , come lo è $\pi_T(t)$;
- gli estremi dell'integrale che ne definisce il valore possono limitarsi⁵¹ alla durata di un periodo;
- l'ergodicità consente di eseguire il calcolo su di una particolare realizzazione di θ , ad es. pari a zero

permettendo di scrivere

$$\begin{aligned} \mathcal{R}_\pi(\tau) &= \frac{1}{T} \int_{-T/2}^{T/2} \sum_n \delta(t - nT) \sum_m \delta(t - mT + \tau) dt \\ &= \frac{1}{T} \int_{-T/2}^{T/2} \delta(t) \delta(t + \tau) dt \end{aligned}$$

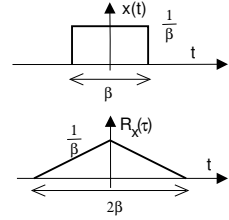
in quanto gli altri termini delle sommatorie cadono al di fuori dell'intervallo di integrazione.

Notiamo ora che $\mathcal{R}_\pi(\tau)$ risulta pari a zero se $\tau \neq 0$, mentre per valutare $\mathcal{R}_\pi(0) = \frac{1}{T} \int_{-T/2}^{T/2} [\delta(t)]^2 dt$ occorre riscrivere $\delta(t)$ in base alla sua definizione, ossia il limite a cui tende una successione di funzioni; ponendo (ad es.) $\delta(t) = \lim_{\beta \rightarrow 0} \frac{1}{\beta} \text{rect}_\beta(t)$, si

⁵¹La definizione di autocorrelazione prescrive il calcolo $\mathcal{R}_y(\tau) = \lim_{\alpha \rightarrow \infty} \frac{1}{\alpha} \int_{-\alpha/2}^{\alpha/2} y(t) y(t + \tau) dt$ ma, se $y(t)$ è periodico, il valore dell'integrale su di un intervallo pari ad un periodo fornisce lo stesso valore per qualunque possibile traslazione temporale di $y(t)$

ottiene⁵² (per $-\frac{T}{2} \leq \tau < \frac{T}{2}$)

$$\begin{aligned} \mathcal{R}_\pi(\tau) &= \lim_{\beta \rightarrow 0} \frac{1}{T} \int_{-T/2}^{T/2} \frac{1}{\beta^2} \text{rect}_{\beta}(t) \text{rect}_{\beta}(t + \tau) dt = \\ &= \lim_{\beta \rightarrow 0} \frac{1}{T} \frac{1}{\beta} \text{tri}_{2\beta}(\tau) = \frac{1}{T} \delta(\tau) \end{aligned}$$



e quindi in definitiva, dovendo risultare $\mathcal{R}_\pi(\tau)$ periodico, otteniamo

$$\mathcal{R}_\pi(\tau) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(\tau - nT) \tag{9.19}$$

Sostituendo ora (9.19) in (9.18), arriviamo al risultato $\mathcal{R}_\Delta(\tau) = \frac{\sigma_A^2}{T} \delta(\tau)$ e $\mathcal{P}_\Delta(f) = \mathcal{F}\{\mathcal{R}_\Delta(\tau)\} = \frac{\sigma_A^2}{T}$. Pertanto, si è dimostrato che l'onda PAM $x(t)$ ha densità di potenza

$$\mathcal{P}_x(f) = \mathcal{P}_\Delta(f) |G(f)|^2 = \sigma_A^2 \frac{|G(f)|^2}{T} \tag{9.20}$$

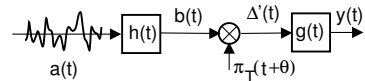
Mediante una diversa trattazione⁵³ si mostra che, nel caso in cui il valor medio $m_A = E\{a_n\}$ dei valori a_n non sia nullo, risulta

$$\mathcal{P}_x(f) = \sigma_A^2 \frac{|G(f)|^2}{T} + \left(\frac{m_A}{T}\right)^2 \sum_{n=-\infty}^{\infty} \left|G\left(\frac{n}{T}\right)\right|^2 \delta\left(f - \frac{n}{T}\right)$$

che, rispetto alla (9.20), evidenzia l'aggiunta di un termine *a righe*.

Infine, prendiamo in considerazione il caso in cui ai simboli a_n si sostituiscono dei b_n ancora a media nulla, ma *non più indipendenti* tra loro, ovvero la cui autocorrelazione $\mathcal{R}_b(\tau) = E\{b_n b_{n+\tau}\}$ sia diversa da zero per $\tau = nT$, adottando lo schema funzionale disegnato sotto per realizzare un modello di ciò che accade.

Mentre $a(t)$ è lo stesso di prima, il suo passaggio attraverso il filtro $h(t)$ introduce una correlazione tra i valori estratti da $b(t)$ mediante moltiplicazione per $\pi_T(t + \theta)$, ottenendo infatti



$\mathcal{R}_b(\tau) = \mathcal{R}_H(\tau) * \mathcal{R}_a(\tau)$ ⁵⁴, indicando ora l'ingresso al filtro $g(t)$ come $\Delta'(t) = \sum b_n \delta(t - nT + \theta)$, si ottiene

$$\mathcal{R}_{\Delta'}(\tau) = \mathcal{R}_\pi(\tau) \cdot \mathcal{R}_b(\tau) = \sigma_A^2 \cdot \mathcal{R}_\pi(\tau) \cdot \mathcal{R}_H(\tau)$$

e dunque⁵⁵

$$\begin{aligned} \mathcal{P}_{\Delta'}(f) &= \mathcal{F}\{\mathcal{R}_{\Delta'}(\tau)\} = \sigma_A^2 \cdot \mathcal{P}_\pi(f) * |H(f)|^2 = \\ &= \frac{\sigma_A^2}{T^2} \cdot \sum_n \delta\left(f - \frac{n}{T}\right) * |H(f)|^2 = \frac{\sigma_A^2}{T^2} \cdot \sum_n \left|H\left(f - \frac{n}{T}\right)\right|^2 \end{aligned}$$

⁵²E' facile verificare dalla figura mostrata che sia $\frac{1}{\beta} \text{rect}_{\beta}(t)$ che $\frac{1}{\beta} \text{tri}_{2\beta}(\tau)$ possiedono area unitaria, e dunque per $\beta \rightarrow 0$ convergono ad un impulso.

⁵³S. Barbarossa, <http://infocom.uniroma1.it/sergio/PAM.pdf>

⁵⁴Più precisamente, l'autocorrelazione tra simboli b_n è valutabile solo per distanze temporali multiple del periodo di simbolo, ed in virtù della presenza di $h(t)$, risulta $\mathcal{R}_b((n-m)T) = E\{b_n b_m\} = \sigma_A^2 \mathcal{R}_H((n-m)T)$.

⁵⁵Nell'ultima serie di passaggi si può valutare $\mathcal{P}_\pi(f)$ come $\frac{1}{T^2} \sum_n \delta\left(f - \frac{n}{T}\right)$, in quanto $\Pi(f) = \mathcal{F}\{\pi(t)\} = \mathcal{F}\{\sum_n \delta(t - nT)\} = \frac{1}{T} \sum_n \delta\left(f - \frac{n}{T}\right)$ (vedi § 3.8.3) e quindi $\mathcal{P}_\pi(f) = |\Pi(f)|^2$, oppure applicando la relazione $\mathcal{P}_\pi(f) = \mathcal{F}\{\mathcal{R}_\pi(\tau)\}$ a partire dalla (9.19).

ma dato che $\mathcal{P}_y(f) = |G(f)|^2 \mathcal{P}_{\Delta'}(f)$, in definitiva si ottiene

$$\mathcal{P}_y(f) = \frac{\sigma_A^2}{T^2} |G(f)|^2 \sum_n \left| H\left(f - \frac{n}{T}\right) \right|^2 \quad (9.21)$$

ovvero la densità spettrale dell'onda PAM viene a dipendere da quella dei dati trasportati. Notiamo che esiste un caso particolare in cui ciò non avviene, e si verifica se $\sum_n |H(f - \frac{n}{T})|^2 = \text{cost}$, condizione simile a quella (5.4) di Nyquist per l'assenza di ISI. Seguendo gli sviluppi analitici svolti in quel caso, otteniamo che la (9.21) può *non dipendere* da $H(f)$ nel caso in cui

$$\mathcal{F}^{-1} \left\{ \left| H\left(f - \frac{n}{T}\right) \right|^2 \right\} = \mathcal{R}_H(\tau) = \begin{cases} 1/T & \text{per } \tau = 0 \\ 0 & \text{per } \tau = nT \\ \vee & \text{altrimenti} \end{cases}$$

che corrisponde alla incorrelazione tra i valori b presi a distanza multipla di T .

Osserviamo infine che la (9.21) costituisce una diversa derivazione della (4.4) di pag. 55, nel caso in cui i b_n siano campioni di segnale.

9.10.4 Potenza di un segnale dati

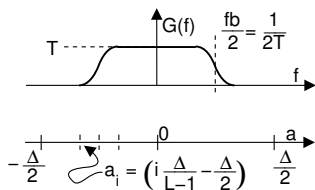
Al § 5.3.3 si è affermato che la potenza di un segnale dati

$$s(t) = \sum_n a_n g(t - nT)$$

in cui $g(t)$ è una caratteristica di Nyquist a coseno rialzato con roll-off γ , e gli a_i sono una sequenza di v.a. discrete, statisticamente indipendenti, a media nulla ed uniformemente distribuite su L livelli in una dinamica $-\frac{\Delta}{2} \leq a_i \leq \frac{\Delta}{2}$, ha valore

$$\mathcal{P}_s = \frac{\Delta^2}{12} \frac{L+1}{L-1} \left(1 - \frac{\gamma}{4}\right)$$

Mostriamo ora i passi necessari ad arrivare a questo risultato. Nella precedente appendice, si è mostrato che per lo stesso segnale risulta $\mathcal{P}_s(f) = \sigma_A^2 \frac{|G(f)|^2}{T}$, e dunque



$$\mathcal{P}_s = \int \mathcal{P}_s(f) df = \int \sigma_A^2 \frac{|G(f)|^2}{T} df$$

Svolgendo i relativi calcoli, si può mostrare⁵⁶ che $\int |G(f)|^2 df = T \left(1 - \frac{\gamma}{4}\right)$, e quindi $\mathcal{P}_s = \sigma_A^2 \left(1 - \frac{\gamma}{4}\right)$; resta pertanto da calcolare σ_A^2 :

$$\begin{aligned} \sigma_A^2 &= (a_i \text{ a media nulla}) = E_X \{a_i^2\} = \sum_{i=0}^{L-1} p_{a_i} \cdot a_i^2 = \quad (57) = \\ &= \frac{1}{L} \sum_{i=0}^{L-1} \left(i \frac{\Delta}{L-1} - \frac{\Delta}{2}\right)^2 = \frac{\Delta^2}{L} \sum_{i=0}^{L-1} \left(\frac{i^2}{(L-1)^2} + \frac{1}{4} - \frac{i}{L-1}\right) = \end{aligned}$$

⁵⁶La dimostrazione sarà sperabilmente sviluppata in una prossima edizione... è una delle poche a mancare in questo libro! ...al momento, la fonte che trovo più in accordo con questa tesi, è ancora una volta https://en.wikipedia.org/wiki/Raised-cosine_filter

⁵⁷Consideriamo valori a_i equiprobabili, ovvero $p(a_i) = 1/L$, e con dinamica bilanciata attorno allo zero, ossia valori compresi tra $-\Delta/2$ e $\Delta/2$

$$\begin{aligned}
&= \frac{\Delta^2}{L} \left(\frac{1}{(L-1)^2} \sum_{i=0}^{L-1} (i)^2 + \frac{L}{4} - \frac{1}{L-1} \sum_{i=0}^{L-1} i \right) = \text{58} = \\
&= \frac{\Delta^2}{L} \left(\frac{L}{4} - \frac{1}{L-1} \frac{L(L-1)}{2} + \frac{1}{(L-1)^2} \frac{(L-1)L(2(L-1)+1)}{6} \right) = \Delta^2 \left(\frac{1}{4} - \frac{1}{2} + \frac{2L-2+1}{6(L-1)} \right) = \\
&= \Delta^2 \frac{6L-6-12L+12+8L-8+4}{24(L-1)} = \Delta^2 \frac{2L+2}{24(L-1)} = \frac{\Delta^2}{12} \frac{L+1}{L-1}.
\end{aligned}$$

9.10.5 Autocorrelazione dell'uscita di un filtro

Al § 9.4.1 si è affermato che, quando un processo attraversa un filtro, il processo di uscita è caratterizzato da $\mathcal{R}_y(\tau) = \mathcal{R}_x(\tau) * \mathcal{R}_h(\tau)$. Mostriamo che è vero.

$$\begin{aligned}
\mathcal{R}_y(t, t + \tau) &= E \{ y(t) y(t + \tau) \} = \\
&= E \left\{ \int h(\alpha) x(t - \alpha) d\alpha \int h(\beta) x(t + \tau - \beta) d\beta \right\} = \\
&= \int \int h(\alpha) h(\beta) E \{ x(t - \alpha) x(t + \tau - \beta) \} d\alpha d\beta = \\
&= \int h(\alpha) \int h(\beta) \mathcal{R}_x(\tau + \alpha - \beta) d\beta d\alpha = \\
&= \int h(\alpha) \mathcal{R}_{xy}(\tau + \alpha) d\alpha = \mathcal{R}_{xy}(\tau) * h(-\tau)
\end{aligned}$$

nel cui terzultimo passaggio si è assunto che $x(t)$ sia stazionario, e $\mathcal{R}_{xy}(\tau) = \mathcal{R}_x(\tau) * h(\tau)$ è l'intercorrelazione tra $x(t)$ ed $y(t)$ ⁵⁹.

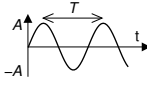
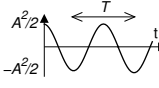

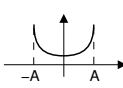
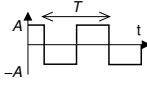
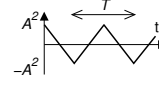
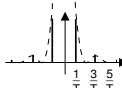

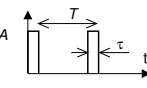
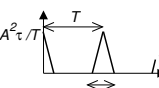
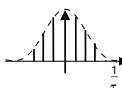
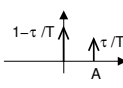
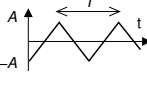
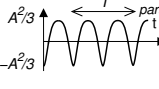
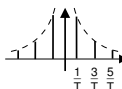
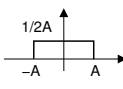
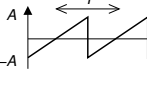

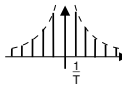
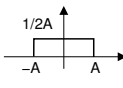
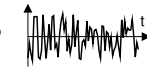
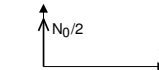
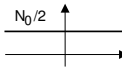



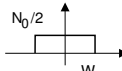
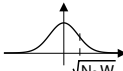
Scritto in altra forma: $\mathcal{R}_y(\tau) = \mathcal{R}_x(\tau) * h(\tau) * h(-\tau)$, e dunque antitrasformando si ottiene $\mathcal{P}_y(f) = \mathcal{P}_x(f) \cdot H(f) \cdot H^*(f) = \mathcal{P}_x(f) \cdot |H(f)|^2 = \mathcal{F} \{ \mathcal{R}_x(\tau) * \mathcal{R}_h(\tau) \}$.

9.10.6 Grafici di esempio

Alla pagina seguente sono riportati i grafici della forma d'onda, dell'autocorrelazione, della densità spettrale e della densità di probabilità, per alcuni segnali tipici.

⁵⁸Facciamo uso delle relazioni $\sum_{n=1}^N n = \frac{N(N+1)}{2}$ e $\sum_{n=1}^N n^2 = \frac{N(N+1)(2N+1)}{6}$

⁵⁹Infatti, $\mathcal{R}_x(\tau) * h(\tau) = x^*(-\tau) * x(\tau) * h(\tau) = x^*(-\tau) * y(\tau)$, che è appunto la definizione di $\mathcal{R}_{xy}(\tau)$, vedi anche eq. (9.6) a pag. 221.

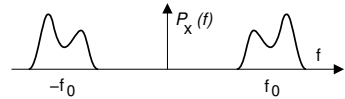
Segnale	Forma d'onda	Autocorrelazione	Densità spettrale	Densità di probabilità
Sinusoide				
Onda quadra				
Impulso rettangolare				
Triangolare				
Dente di sega				
Rumore gaussiano bianco				
Rumore gaussiano limitato in banda				

Capitolo 10

Segnali modulati

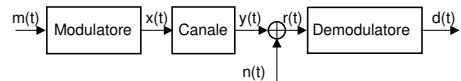
10.1 Caratteristiche ed applicazioni

I segnali *modulati*, di cui ci occuperemo ora, sono anche indicati come segnali in *banda traslata*, per la particolarità di esibire una occupazione spettrale limitata ad una banda ristretta attorno ad una frequenza f_0 chiamata *portante*.



Questi segnali sono prodotti mediante un *processo di modulazione*, che consiste nell'operare una trasformazione su di un segnale $m(t)$ *modulante* in modo da ottenerne un secondo $x(t)$ *modulato*, allo scopo di renderne il contenuto spettrale idoneo alla trasmissione mediante il canale a disposizione.

Il segnale *ricevuto* $r(t)$ (a cui è sovrapposto un processo di rumore $n(t)$) deve quindi essere *demodulato* (opera-



zione che, se applicata ad $x(t)$, restituirebbe $m(t)$) per ottenere $d(t)$, che rappresenta il segnale trasmesso, più eventuali *distorsioni* $\varepsilon(t)$. Per evidenziare la situazione, scriviamo il segnale demodulato come $d(t) = m(t) + \varepsilon(t)$, in cui $\varepsilon(t) = Dem\{n(t)\} + Dem\{y(t) - x(t)\}$: evidentemente, le distorsioni hanno origine sia dal risultato della demodulazione del rumore in ingresso, sia dagli effetti prodotti dalla demodulazione delle alterazioni introdotte dal canale sul segnale in transito.

Il processo di modulazione è spesso associato ad una trasmissione radio, ma può rendersi necessario e/o utile anche per trasmissioni in cavo. In generale, individuiamo almeno tre situazioni in cui è necessario l'impiego di segnali modulati:

1. il canale non permette la trasmissione di frequenze contigue all'origine
2. il canale presenta un comportamento ideale (modulo costante e fase lineare) solo in determinati intervalli di frequenza
3. il canale presenta disturbi additivi solo in determinate regioni di frequenza.

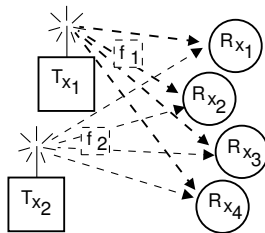
10.1.1 Multiplazione a divisione di frequenza - FDM

Come intuibile dal titolo, si tratta di una tecnica di trasmissione in cui più comunicazioni avvengono in contemporanea, condividendo lo stesso mezzo fisico, ed impegnando ognuna una diversa regione di frequenza, per il semplice motivo che se utilizzassero

tutte la stessa, costituirebbero termini di *interferenza* reciproca. Molto spesso tutti i segnali multiplati sono di natura simile, ed ognuno è il risultato di una modulazione operata con una diversa frequenza portante. Portiamo ad esempio tre casi:

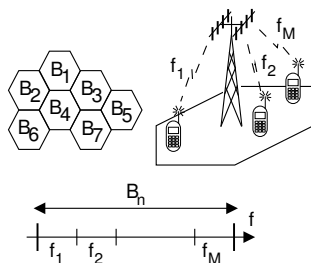
10.1.1.1 Collegamenti punto-multipunto

E' lo topologia adottata ad esempio nel caso di trasmissioni televisive o radiofoniche (dette trasmissioni *broadcast*), in cui ogni emittente (in figura indicata come T_x) trasmette a tutti i ricevitori (R_x) sintonizzati sulla propria portante (i famosi “canali” della TV), mentre altre emittenti utilizzano contemporaneamente lo stesso mezzo trasmissivo, occupando canali centrati ad altre frequenze.



10.1.1.2 Accesso multiplo

E' la strategia tipica di accesso alla rete (pag. 120) per le *comunicazioni mobili*, ovvero telefonini e WiFi. In tal caso il territorio è suddiviso in *celle*, per ognuna delle quali viene utilizzata una diversa banda (B_n) di frequenze radio, dedicata alla comunicazione tra i terminali ed un unico ripetitore. All'interno della cella, la banda a disposizione è suddivisa tra più canali, ognuno associato ad una diversa portante (f_i), che vengono usate a turno dai terminali che desiderano comunicare¹.



Sotto certi aspetti, questo caso è in qualche modo antitetico rispetto al precedente, e potrebbe essere indicato come collegamento *multipunto-punto*. In realtà la situazione è un pò più complessa, e gli aspetti qualificanti da un punto di vista sistemistico sono i protocolli di rete, necessari per consentire le fasi di richiesta di accesso, la localizzazione dei radiomobili, e la corretta gestione dell'*handover* (il cambio di cella)².

¹Senza entrare nei dettagli, specifichiamo semplicemente che celle limitrofe adottano regioni di frequenza differenti, onde evitare interferenze tra celle; inoltre (per il sistema GSM) nell'ambito di una stessa portante è realizzata una struttura di trama, in modo da permettere l'utilizzo dello stesso canale da parte di più terminali contemporaneamente, multiplati a divisione di tempo.

²Un minimo di approfondimento (sempre per il GSM) però ci sta bene... Aggiungiamo quindi che la scelta del canale su cui comunicare avviene in base alle condizioni di ricezione del singolo radiomobile che, per effetto dei cammini multipli (§ 15.3.3.5) sul segnale del ripetitore, può ricevere meglio certe portanti che non altre.

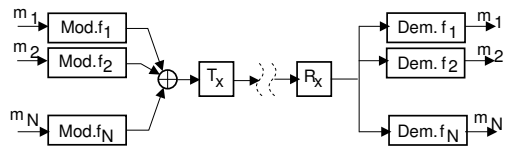
La trasmissione che ha luogo su di una portante, inoltre, può aver origine da più terminali, che si ripartiscono la medesima banda a divisione di tempo, in accordo ad una suddivisione di trama dell'asse dei tempi. Pertanto, dopo che un terminale si è aggiudicato una portante ed un intervallo temporale, la trasmissione (attuata mediante una modulazione numerica) ha luogo solo per brevi periodi, in corrispondenza del time-slot di propria pertinenza.

Dato che i singoli terminali si trovano a distanze diverse dal ripetitore di cella, diversi sono i tempi di propagazione del segnale di sincronismo di trama e di time-slot, e dunque l'intervallo temporale che viene “riempito” da ogni terminale giunge al ripetitore con un ritardo variabile. Per questo motivo, i time-slot della trama sono separati da piccoli periodi di inattività, chiamati *intervalli di guardia*, che garantiscono l'assenza di sovrapposizioni temporali delle trasmissioni originate dai diversi terminali.

10.1.1.3 Collegamenti punto-punto

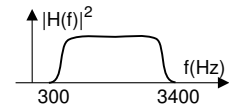
E' una forma di moltiplicazione più simile a quella già illustrata nel caso della TDM, in cui il collegamento condiviso si sviluppa tra due località ben determinate. Un insieme di N segnali m_i , $i = 1, 2, \dots, N$, transita su di uno stesso mezzo trasmissivo, occupando ognuno una differente banda, centrata su di una diversa portante f_i , $i = 1, 2, \dots, N$, e può essere individualmente demodolato e separato in ricezione.

La trasmissione può avvenire sia mediante un collegamento in cavo, che mediante una trasmissione radio; in questa seconda evenienza, il collegamento è spesso indicato come *ponte radio*.

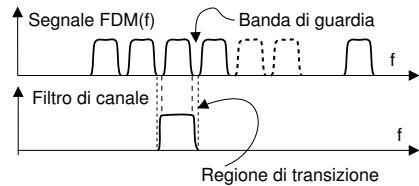


10.1.2 Canale telefonico

Le caratteristiche del collegamento offerto dalla comune linea telefonica (§ 6.9.1) rivestono molteplici aspetti. Uno di questi, forse il principale³, è la limitazione della banda del canale, che rende la trasmissione garantita solo in un intervallo di frequenze comprese tra i 300 ed i 3400 Hz, mentre la banda nominalmente occupata è definita come pari a 4000 Hz⁴: discutiamo brevemente le origini storiche di tali limitazioni.



L'assenza della regione $-300 \div 300$ Hz è legata alla presenza, all'interno del telefono, di un componente (detto *ibrido*⁵) che di fatto impedisce la trasmissione di frequenze molto basse. Invece la limitazione per le frequenze da 3400 Hz in poi, è legata al *vecchio* metodo di trasmissione mediante moltiplicazione FDM punto-punto, con i singoli canali modulati AM-BLU (vedi § 11.1.2), che pure impone di rimuovere le componenti frequenziali più basse. Inoltre, la necessità di separare tra loro i canali modulati FDM mediante i *filtri di canale* che, per essere economicamente realizzabili, devono presentare una regione di transizione di estensione apprezzabile, ha determinato l'esigenza di prevedere tra due canali contigui un intervallo di frequenze detto *banda di guardia* (pari a 900 Hz) che determina la limitazione a 3400 Hz per la massima frequenza di segnale, in modo da ottenere



³Un altro fattore particolarmente rilevante è la *limitazione della potenza* che è possibile immettere su di un singolo collegamento telefonico che, associato al precedente, identifica il canale telefonico come limitato sia in banda che in potenza, e dunque con capacità (§ 17.2.4) $C = W \log_2 \left(1 + \frac{P_s}{N_0 W} \right)$ dipendente solo dal livello del rumore. La limitazione in potenza è storicamente motivata da problemi di *diafonia* (pag. 382) dovuti a fenomeni di induzione elettromagnetica, mentre attualmente è determinata dalla limitata dinamica del segnale che viene campionato e trasmesso in forma numerica.

⁴Questo valore massimo nominale determina che la frequenza di campionamento del PCM telefonico è pari a $2 \cdot 4000 = 8000$ campioni al secondo. Utilizzando 8 bit/campione, si ottiene la velocità binaria $f_b = 64000$ campioni/secondo. Velocità inferiori si possono conseguire adottando metodi di codifica di sorgente per il segnale vocale.

⁵L'ibrido telefonico è un trasformatore con quattro porte, che realizza la separazione tra le due vie di comunicazione che viaggiano sullo stesso cavo (vedi § 6.9.1). Nel caso di una linea ISDN, invece, il telefono stesso effettua la conversione numerica, ed i campioni di voce viaggiano nei due sensi (tra utente e centrale) secondo uno schema a divisione di tempo (vedi § 6.9.2).

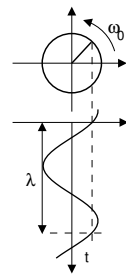
$300 + (4000 - 3400) = 900$ Hz. In caso contrario infatti, all'uscita del filtro di canale si troverebbe anche parte del segnale di un canale contiguo producendo una interferenza tra comunicazioni diverse.

La limitazione in banda di un canale telefonico tra 300 e 3400 Hz è dunque il motivo per il cui la connessione telefonica di un computer ad un fornitore di connettività Internet pre-ADSL richiedeva l'uso di un dispositivo (modem), che effettuava una forma di modulazione sul segnale da trasmettere, che arrivava in tale forma fino al provider. Al contrario, nel caso dell'accesso ADSL (vedi § 6.9.4) la connettività numerica inizia direttamente nella centrale del chiamante; d'altra parte, il segnale prodotto dal modem ADSL occupa ora una banda *disgiunta* da quella del canale telefonico, usando tutta la capacità del doppino che è ad uso esclusivo dell'utente.

10.1.3 Antenne e lunghezza d'onda

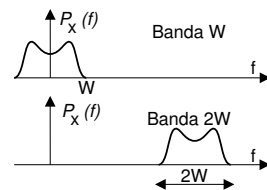
La trasmissione di un segnale via onda radio necessita di una antenna di dimensioni comparabili con quelle della lunghezza d'onda. Quest'ultima quantità (indicata con λ) è pari allo spazio percorso dall'onda in un tempo pari ad un periodo: dato che *spazio* = *velocità* · *tempo*, e considerando che le onde elettromagnetiche si propagano alla velocità della luce ($c = 3 \cdot 10^8$ m/s), si ha $\lambda = c \cdot T = \frac{c}{f}$.

Nel caso di segnali modulati, il valore di f è quello della portante, in quanto in genere il segnale modulato occupa una banda ristretta attorno ad essa. Trasmissioni con frequenze più elevate necessitano di antenne di dimensioni ridotte; d'altra parte se per assurdo trasmettissimo con portante di 300 Hz, occorrerebbe una antenna di dimensioni $\lambda = \frac{c}{f} = \frac{3 \cdot 10^8}{300} = 10^6$ m = 1000 Km!⁶



10.1.4 Banda di segnale

La banda occupata da un segnale è la regione di frequenze al di fuori della quale non vi sono componenti energetiche; la sua misura in Hz è indicata come *larghezza di banda*. Per segnali reali l'occupazione di banda è espressa in termini del solo contenuto a frequenze positive; dato che in tal caso lo spettro di potenza è una funzione pari di f , la banda totale è doppia. Tale definizione è pertanto non ambigua, ed in accordo alla comune accezione di frequenza (positiva); pertanto, viene spesso indicata come *banda a frequenze positive*⁷.

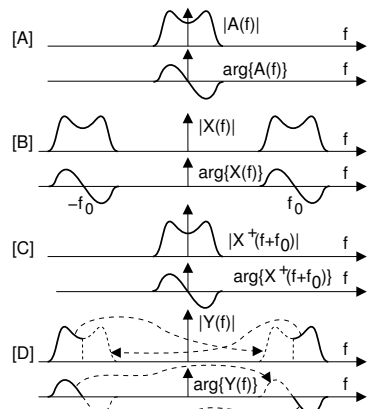


10.1.5 Trasmissione a banda laterale unica

⁶Antenne più corte hanno una efficienza ridotta, ma sono ancora buone. Altrimenti la radio AM (540 - 1600 KHz) avrebbe bisogno di $\frac{3 \cdot 10^8}{1000 \cdot 10^3} = 300$ metri! Al § 15.5.2 è riportata una tabella dei valori di λ per i diversi servizi di TLC.

⁷Se un segnale è strettamente limitato in banda, deve avere durata infinita, e viceversa. E' pratica comune, invece, parlare di limitazione in banda anche per segnali di durata finita. Nel fare questo, si considera un $X(f)$ pari a zero per le frequenze f tali che $|X(f)| < \epsilon$, ovvero considerare anziché $X(f)$ a banda illimitata, una sua finestra *in frequenza* $X_W(f) = X(f)W(f)$ a banda limitata, la cui antitrasformata $x_W(t)$ è diversa da $x(t)$ (sappiamo infatti che si ha $x_W(t) = x(t) * w(t)$), ma ne costituisce una approssimazione.

Consideriamo un segnale $a(t)$ reale e limitato in banda, con $A(f) = A^*(-f)$ (grafico [A]), a simmetria coniugata. In virtù delle proprietà di simmetria coniugata per segnali reali, la conoscenza del solo contenuto a frequenze positive $f > 0$, ovvero di $A^+(f) = A(f) \text{rectw}(f - \frac{W}{2})$, è sufficiente a definire $a(t)$ in modo completo. Se definiamo il segnale modulato $x(t) = a(t) \cos \omega_0 t$, anch'esso reale, otteniamo che $X(f)$ [B], oltre ad essere a simmetria coniugata rispetto all'origine, ha simmetria coniugata anche rispetto ad f_0 : $X^+(f + f_0) = \{X^+(-f + f_0)\}^*$ [C].



Questo risultato mostra come sia *praticamente* possibile (con una fotocopiatrice ed un paio di forbicini!) produrre un segnale $Y(f)$ eliminando da $X(f)$ tutta la banda $|f| < f_0$ [D], e quindi da quel che resta, ri-ottenere il segnale $X(f)$ a partire da un $Y(f)$. La ricostruzione di $X(f)$ avviene infatti (freccie tratteggiate) spostando le copie duplicate di $Y^+(f)$ e $Y^-(f)$ come indicato dalle freccie.

Una volta verificata la *correttezza ipotetica* di questo procedimento che ci consente di ricevere per intero $X(f)$ *trasmettendone solo metà* (cioè $Y(f)$), osserviamo che anche $Y(f)$ è a simmetria coniugata rispetto a zero (ossia $Y(f) = Y^*(-f)$), e quindi la sua antitrasformata $y(t)$ è *reale*, e dunque può essere realmente trasmesso.

A parte il “dettaglio” di come ricostruire “veramente” $X(f)$ a partire da $Y(f)$, ci chiediamo: esiste una formula per ottenere $y(t)$ in modo *diretto* a partire da $a(t)$? La risposta è positiva; per descriverla però occorre prima affrontare alcune pagine di teoria, che illustrano un metodo di rappresentazione (nel dominio del tempo) per segnali modulati.

10.2 Rappresentazione dei segnali modulati

Questa sezione è dedicata alla teoria che permette di scrivere un qualunque segnale $x(t)$ nella forma

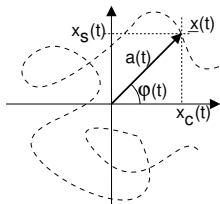
$$x(t) = x_c(t) \cos 2\pi f_0 t - x_s(t) \sin 2\pi f_0 t \tag{10.1}$$

che assume una particolare rilevanza nel caso in cui $x(t)$ sia un segnale modulato attorno ad f_0 , perché allora $x_c(t)$ e $x_s(t)$ sono segnali limitati in banda con banda contigua all'origine, e le alterazioni prodotte sul segnale modulato, compresa l'estrazione del messaggio modulante $m(t)$, possono essere descritte mediante operazioni condotte su $x_c(t)$ ed $x_s(t)$.

10.2.1 Involuppo complesso

Introduciamo l'argomento ricordando (vedi § 2.1.3) come un segnale $x(t) = a \cos(\omega_0 t + \varphi)$ può essere rappresentato per mezzo del fasore $\underline{x} = ae^{j\varphi}$, mediante la relazione $x(t) = \Re\{\underline{x}e^{j\omega_0 t}\}$. Estendiamo ora il concetto, definendo l'*inviluppo complesso* $\underline{x}(t)$ come un fasore per il quale il modulo a e la fase φ siano funzioni del tempo

$$\underline{x}(t) = a(t) e^{j\varphi(t)} \tag{10.2}$$



rappresentato nella figura a fianco assieme ad una sua potenziale traiettoria temporale. Ad $\underline{x}(t)$ possiamo quindi associare un segnale reale

$$x(t) = \Re \left\{ \underline{x}(t) e^{j\omega_0 t} \right\} = a(t) \cos(\omega_0 t + \varphi(t)) \quad (10.3)$$

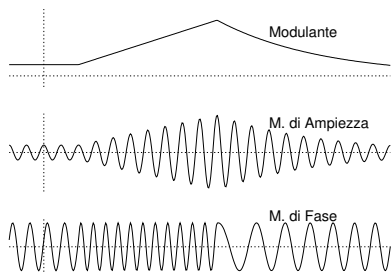
corrispondente ad imprimere al piano dell'involuppo complesso una rotazione antioraria a velocità angolare ω_0 . D'altra parte, indicando con $x_c(t) = a(t) \cos \varphi(t)$ ed $x_s(t) = a(t) \sin \varphi(t)$ la parte reale ed immaginaria dell'involuppo complesso $\underline{x}(t)$ ⁸, la (10.3) è equivalente alla (10.1), dato che⁹

$$\begin{aligned} x(t) &= \Re \left\{ \underline{x}(t) e^{j\omega_0 t} \right\} = a(t) \cos(\omega_0 t + \varphi(t)) = \\ &= a(t) [\cos \omega_0 t \cos \varphi(t) - \sin \omega_0 t \sin \varphi(t)] \\ &= x_c(t) \cos 2\pi f_0 t - x_s(t) \sin 2\pi f_0 t \end{aligned}$$

10.2.2 Modulazione di ampiezza e/o angolare

L'involuppo complesso è un potente strumento che permette di descrivere il processo di modulazione in modo semplice ed omogeneo. Ad esempio, il caso (già noto) di traslazione in frequenza del segnale $a(t)$ mediante moltiplicazione per un coseno, corrisponde ad un involuppo complesso $\underline{x}(t) = a(t)$ a fase nulla: ad esso si dà il nome di *modulazione di ampiezza*¹⁰ per (l'evidente) ragione che l'ampiezza del coseno varia in funzione del segnale $a(t)$; la frequenza $f_0 = \frac{\omega_0}{2\pi}$ prende il nome di *frequenza portante*. Se al contrario consideriamo un involuppo complesso con modulo costante $\underline{x}(t) = e^{j\varphi(t)}$, l'andamento della fase $\varphi(t)$ imprime alla portante non modulata un diverso tipo di modulazione, detto *modulazione di fase*¹¹ o *angolare* in quanto il segnale modulante ($\varphi(t)$ in questo caso) altera l'argomento del coseno.

Prima di proseguire, riflettiamo sul risultato mostrato alla figura seguente, in cui si considera un segnale modulante $m(t)$ prima costante, poi a rampa lineare, e quindi decrescente. Ponendo $\underline{x}(t) = m(t)$ si ottiene una portante modulata in ampiezza, mentre con $\underline{x}(t) = e^{jm(t)}$ la portante modulata angularmente $x(t) = \cos(2\pi f_0 t + m(t))$ presenta una ampiezza costante, ed una frequenza che (nell'intervallo in cui $m(t)$ aumenta linearmente) cambia in un valore più elevato, per poi diminuire. In pratica, se $m(t) = \alpha t$, allora l'argomento del coseno diviene $2\pi f_0 t + \alpha t = 2\pi \left(f_0 + \frac{\alpha}{2\pi}\right) t$.



Per meglio descrivere il caso di modulazione angolare, definiamo una *fase istantanea*

$$\psi(t) = 2\pi f_0 t + \varphi(t)$$

ed una *frequenza istantanea*

$$f_i(t) = \frac{1}{2\pi} \frac{d}{dt} \psi(t) = f_0 + \frac{1}{2\pi} \frac{d}{dt} \varphi(t)$$

⁸ $x_c(t)$ e $x_s(t)$ si ottengono a partire dalla rappresentazione polare $\underline{x}(t) = a(t) e^{j\varphi(t)}$ di $\underline{x}(t)$, semplicemente sviluppando la stessa come $\underline{x}(t) = a(t) e^{j\varphi(t)} = a(t) \cos \varphi(t) + ja(t) \sin \varphi(t) = x_c(t) + jx_s(t)$

⁹Si faccia uso della relazione $\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta$.

¹⁰Indicata anche come AM (*amplitude modulation*).

¹¹Indicata anche come PM (*phase modulation*).

In questi termini, la modulazione angolare viene distinta in *modulazione di fase* propriamente detta quando

$$\varphi(t) = k_\varphi m(t)$$

mentre viene detta *modulazione di frequenza* quando

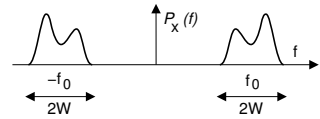
$$\varphi(t) = 2\pi k_f \int_{-\infty}^t m(\tau) d\tau$$

dato che in questo caso è la frequenza istantanea a dipendere direttamente dal segnale modulante: $f_i(t) = f_0 + k_f m(t)$.

10.2.3 Componenti analogiche di bassa frequenza

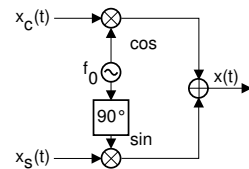
Abbiamo mostrato come un generico segnale modulato $x(t)$ possa essere rappresentato per mezzo di un diverso segnale chiamato *inviluppo complesso* $\underline{x}(t) = a(t) e^{j\varphi(t)}$, le cui parti reale ed immaginaria $x_c(t) = a(t) \cos \varphi(t)$ ed $x_s(t) = a(t) \sin \varphi(t)$, che prendono il nome di *componenti analogiche di bassa frequenza* di $x(t)$ per un motivo presto chiaro, permettono di ri-scrivere $x(t)$ come $x(t) = x_c(t) \cos 2\pi f_0 t - x_s(t) \sin 2\pi f_0 t$.

Mentre il risultato ottenuto è valido per un qualunque segnale, esso riveste una importanza particolare nel caso in cui $x(t)$ sia di tipo limitato in banda con banda $2W$ centrata attorno ad f_0 , con $W < f_0$ (fig. a lato): in tal caso infatti, sia $x_c(t)$ che $x_s(t)$ risultano *limitate in banda tra $\pm W$ e contigue all'origine*.



Che sia vero anche il viceversa, può essere verificato in modo intuitivo, partendo da $x_c(t)$ e $x_s(t)$ limitate in banda, e moltiplicandole per coseno e seno, come mostrato alla figura seguente.

L'ultima osservazione ci mostra una via per *sin-tetizzare* un segnale modulato in ampiezza, od angolarmente, od entrambe le cose, mediante il semplice schema circuitale disegnato, che si basa sulla conoscenza delle componenti analogiche di bassa frequenza, che a loro volta sono ottenibili a partire da $a(t)$ e $\varphi(t)$. Restano comunque (per ora) aperti i seguenti problemi:



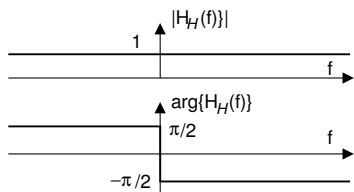
- noto $x(t)$, come ottenere $x_c(t)$ ed $x_s(t)$?
- noto lo spettro di densità di potenza $\mathcal{P}_x(f)$, che dire di $\mathcal{P}_{\underline{x}}(f)$?

Mentre la prima domanda può trovare una risposta di tipo operativo nella *demodulazione omodina* (vedi § 10.5), dal punto di vista formale occorre prima definire il *segnale analitico* (vedi § 10.2.5), e prima ancora, definire il...

10.2.4 Filtro di Hilbert

Il *filtro di Hilbert* è caratterizzato da una risposta in frequenza descritta come

$$H_{\mathcal{H}}(f) = -j \cdot \text{sgn}(f) \tag{10.4}$$



Filtro di Hilbert

ed il cui andamento è rappresentato nella figura a lato, che permette di evidenziare l'andamento costante del modulo $|H_{\mathcal{H}}(f)| = 1$, e quello discontinuo della fase $\angle H_{\mathcal{H}}(f)$, che passa da $\frac{\pi}{2}$ per $f < 0$ a $-\frac{\pi}{2}$ per $f > 0$.

Il risultato del passaggio di un segnale $x(t)$ attraverso¹² il filtro di Hilbert è un secondo segnale, identificato come la *trasformata di Hilbert* del primo, indicato come $\hat{x}(t) = \mathcal{H}\{x(t)\}$,

ed il cui andamento in frequenza ha espressione

$$\hat{X}(f) = \mathcal{F}\{\hat{x}(t)\} = H_{\mathcal{H}}(f) X(f) = -j \cdot \text{sgn}(f) \cdot X(f)$$

ossia differisce da $X(f)$ per uno sfasamento di $\mp \frac{\pi}{2}$ per frequenze rispettivamente positive o negative.

10.2.4.1 Estrazione delle componenti analogiche di bassa frequenza

Senza soffermarci ora sulle proprietà¹³ della trasformata di Hilbert, introduciamo direttamente il risultato¹⁴

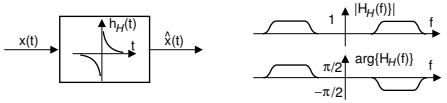
$$\begin{cases} \mathcal{H}\{x_c(t) \cos \omega_0 t\} = x_c(t) \sin \omega_0 t \\ \mathcal{H}\{x_s(t) \sin \omega_0 t\} = -x_s(t) \cos \omega_0 t \end{cases}$$

che, assieme all'espressione (10.1) di $x(t)$ in funzione di $x_c(t)$ ed $x_s(t)$, permette di esprimere la trasformata di Hilbert di un segnale modulato, e quindi di impostare un sistema di due equazioni nelle due incognite $x_c(t)$ e $x_s(t)$:

$$\begin{cases} x(t) = x_c(t) \cos \omega_0 t - x_s(t) \sin \omega_0 t \\ \hat{x}(t) = x_c(t) \sin \omega_0 t + x_s(t) \cos \omega_0 t \end{cases}$$

¹²L'antitrasmformata di Fourier di $H_{\mathcal{H}}(f)$ è calcolata al § 10.7.1, e fornisce l'espressione della risposta impulsiva del filtro di Hilbert $h_{\mathcal{H}}(t) = \mathcal{F}^{-1}\{H_{\mathcal{H}}(f)\} = \frac{1}{\pi t}$, permettendo di scrivere la trasformata di Hilbert nella forma di un integrale di convoluzione: $\hat{x}(t) = \mathcal{H}\{x(t)\} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau = x(t) * \frac{1}{\pi t}$. La realizzazione di un filtro di Hilbert con una risposta in frequenza esattamente descritta dalla (10.4) può risultare molto ardua, a causa della brusca transizione della fase in corrispondenza di $f = 0$. In realtà, il filtro di Hilbert si usa principalmente per segnali modulati, che non presentano componenti spettrali a frequenze prossime allo zero.

Pertanto, lo stesso scopo può essere svolto da un diverso filtro $H_{\mathcal{H}}(f)$, con andamento più dolce della fase, e che presenti gli stessi valori nominali del filtro di Hilbert solamente per le frequenze comprese nella banda di segnale.



¹³Accenniamo brevemente alle principali proprietà della trasformata di Hilbert:

- $\mathcal{H}\{x(t) = x_0\} = 0$: una costante ha trasformata di Hilbert nulla, e la trasformata di Hilbert è definita a meno di una costante. Il valore medio di $x(t)$ non si ripercuote su $\hat{x}(t)$;
- $\mathcal{H}\{\mathcal{H}\{x(t)\}\} = \hat{\hat{x}}(t) = -x(t)$: infatti una rotazione di fase pari a π radianti corrisponde ad una inversione di segno;
- $\int_{-\infty}^{\infty} x(t) \hat{x}(t) dt = 0$: ortogonalità tra un segnale e la sua trasformata di Hilbert;
- $\mathcal{H}\{x(t) * h(t)\} = \hat{x}(t) * \hat{h}(t) = x(t) * \hat{\hat{h}}(t)$: la trasformata di Hilbert di una convoluzione (cioè dell'uscita di un filtro) è la convoluzione tra un operando trasformato e l'altro no.

¹⁴Il risultato è valido solamente se $x_c(t)$ ed $x_s(t)$ sono limitate in banda $\pm W$ con $W < f_0$, come mostrato in appendice 10.7.2.

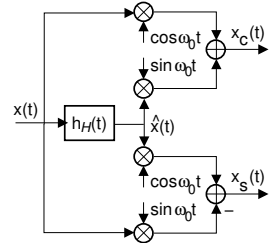
Il sistema può essere risolto¹⁵, permettendo in definitiva di esprimere le componenti analogiche di bassa frequenza in termini di $x(t)$ e di $\hat{x}(t)$:

$$\begin{cases} x_c(t) = x(t) \cos \omega_0 t + \hat{x}(t) \sin \omega_0 t \\ x_s(t) = \hat{x}(t) \cos \omega_0 t - x(t) \sin \omega_0 t \end{cases}$$

Pertanto, le componenti analogiche di bassa frequenza possono essere estratte direttamente da $x(t)$, utilizzando un filtro di Hilbert per ottenere $\hat{x}(t)$, e combinando i due segnali per mezzo di oscillatori in quadratura, in accordo allo schema circuitale mostrato nella figura a fianco.

Infine, $x_c(t)$ e $x_s(t)$ permettono di risalire alle modulazioni di ampiezza ed angolare:

$$\begin{cases} a(t) = |\underline{x}(t)| = \sqrt{x_c^2(t) + x_s^2(t)} \\ \varphi(t) = \angle \underline{x}(t) = \arctan \frac{x_s(t)}{x_c(t)} \end{cases}$$



Prima di procedere a calcolare $\mathcal{P}_{\underline{x}}(f)$, occorre introdurre l'ulteriore concetto di *segnale analitico*.

10.2.5 Segnale analitico

Il segnale analitico associato ad $x(t)$ corrisponde al suo *contenuto a frequenze positive* $x^+(t)$, introdotto al § 10.1.5; si può mostrare che $x^+(t)$ è esprimibile nei termini di $x(t)$ e $\hat{x}(t)$, secondo l'espressione¹⁶:

$$x^+(t) = \frac{1}{2} (x(t) + j\hat{x}(t)) \tag{10.5}$$

Molto utile è anche la relazione che lega il segnale analitico all'involuppo complesso:

$$x^+(t) = \frac{1}{2} \underline{x}(t) e^{j\omega_0 t} \tag{10.6}$$

che si ottiene tenendo conto dalla (10.5), come illustrato alla nota¹⁷. Effettivamente, l'ultima relazione rappresenta il contenuto a frequenze positive di $x(t)$, a patto che

¹⁵Potremmo notare come la matrice dei coefficienti costituisca una rotazione di assi (vedi ad es. a pag. 361), rotazione che "ruota" letteralmente a velocità angolare ω_0 . Tale rotazione stabilisce che le coppie di segnali $(x_c(t), x_s(t))$ e $(x(t), \hat{x}(t))$ rappresentano comunque l'evoluzione dell'involuppo complesso $\underline{x}(t) = a(t) e^{j\varphi(t)}$: mentre $x_c(t)$ e $x_s(t)$ lo rappresentano su due assi ad esso solidali, $x(t)$ e $\hat{x}(t)$ sono definiti su assi ruotanti che tengono conto della frequenza portante.

¹⁶L'eguaglianza si dimostra valutandola nel dominio nella frequenza, ricordando la definizione di filtro di Hilbert, in quanto risulta:

$$X^+(f) = \frac{1}{2} (X(f) + j\hat{X}(f)) = \begin{cases} \frac{1}{2} \{X(f) + j[-jX(f)]\} = X(f) & \text{con } f > 0 \\ \frac{1}{2} \{X(f) + j[jX(f)]\} = 0 & \text{con } f < 0 \end{cases}$$

infatti, a frequenze negative il prodotto $j \cdot j = -1$ costituisce uno sfasamento di π radianti per tutte le frequenze, provocando l'elisione tra $X(f)$ e $-X(f)$ per tutti i valori $f < 0$.

¹⁷Sviluppando il secondo membro di (10.6) si ottiene:

$$\begin{aligned} \frac{1}{2} \underline{x}(t) e^{j\omega_0 t} &= \frac{1}{2} (x_c(t) + jx_s(t)) (\cos \omega_0 t + j \sin \omega_0 t) = \frac{1}{2} [(x_c(t) \cos \omega_0 t - x_s(t) \sin \omega_0 t) + \\ &+ j(x_c(t) \sin \omega_0 t + x_s(t) \cos \omega_0 t)] = \frac{1}{2} (x(t) + j\hat{x}(t)) \end{aligned}$$

che corrisponde al secondo membro di (10.5), e quindi a $x^+(t)$.

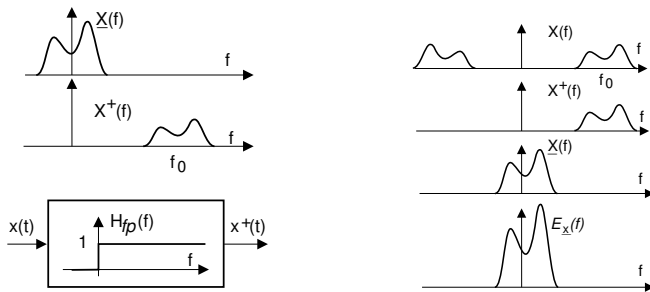


Figura 10.1: Derivazione di $x^+(t)$ mediante filtraggio e densità spettrale di $\underline{x}(t)$

$\underline{x}(t)$ sia di banda base con frequenza massima $W < f_0$; in tal caso infatti, trasformando la (10.6), risulta

$$X^+(f) = \mathcal{F}\{x^+(t)\} = \frac{1}{2}\underline{X}(f - f_0)$$

che giace completamente nel semipiano $f > 0$.

Alternativamente alla (10.5), si può ottenere $x^+(t)$ senza utilizzare $\hat{x}(t)$, pensandolo come il risultato del passaggio di $x(t)$ attraverso un filtro $H_{fp}(f)$ ¹⁸ (vedi fig. 10.1) con funzione di trasferimento a gradino unitario:

$$x^+(t) = x(t) * h_{fp}(t)$$

10.2.6 Densità spettrale di segnali passa-banda

Invertendo la (10.6) otteniamo che $\underline{x}(t) = 2x^+(t)e^{-j\omega_0 t}$, la cui trasformata ci consente di valutare l'espressione di $\underline{X}(f)$:

$$\underline{X}(f) = 2X^+(f + f_0)$$

Quest'ultima espressione consente di osservare, con riferimento alla fig. 10.1, che in linea di principio $\underline{X}(f)$ non gode di simmetria rispetto allo zero, come peraltro prevedibile per $\underline{x}(t)$ complesso. Ricordando ora che $\mathcal{E}_x(f) = |X(f)|^2$, otteniamo

$$\mathcal{E}_{\underline{x}}(f) = 4|X^+(f + f_0)|^2 = 4\mathcal{E}_{x^+}(f + f_0)$$

Un risultato del tutto simile può essere ottenuto¹⁹ per segnali di potenza, ovvero

$$\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{x^+}(f + f_0) \quad (10.7)$$

Pertanto, la densità di potenza di $\underline{x}(t)$ si ottiene da quella a frequenze positive di $x(t)$, traslata nell'origine e moltiplicata per 4. Infine, la discussione riportata al § 10.6 mostra come un risultato del tutto simile sia valido anche nel caso in cui $x(t)$ è membro di un processo ergodico.

¹⁸Il pedice f_p sta per *frequenze positive*.

¹⁹La (10.7) può essere motivata seguendo le stesse linee guida indicate alla nota 12 a pag. 222.

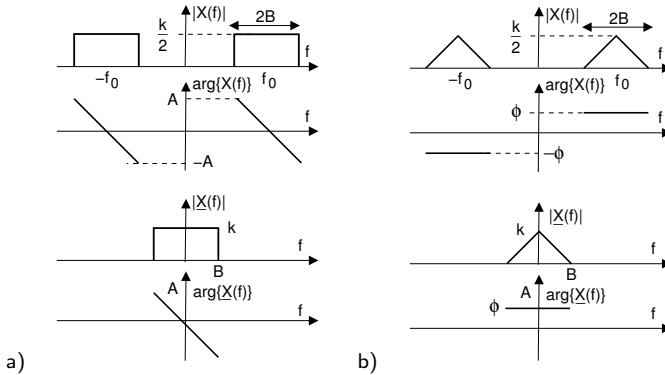


Figura 10.2: Densità spettrali utilizzate negli esempi

Esempi

- Sia dato il segnale $x(t)$ la cui trasformata $X(f)$ è riportata nel lato sinistro in alto di Fig. 10.2. Quali sono le sue componenti analogiche di bassa frequenza, espresse nel dominio della frequenza?
- Notiamo che $|X^+(f)| = \frac{k}{2} \text{rect}_{2B}(f - f_0)$, e dunque

$$|\underline{X}(f)| = 2|X^+(f + f_0)| = k \text{rect}_{2B}(f)$$

Per la fase si opera una traslazione analoga, ma senza moltiplicare per il fattore 2 che, in quanto fattore, incide solo sul modulo.

Osserviamo ora che $\underline{X}(f)$ ha modulo pari e fase dispari, e dunque la sua antitrasformata è un segnale reale: $\underline{x}(t) = x_c(t) + jx_s(t) = x_c(t)$, ovvero la componente in quadratura $x_s(t)$ è nulla. Pertanto, risulta²⁰ $\begin{cases} X_c(f) = k \text{rect}_{2B}(f) e^{-j2\pi \frac{A}{2\pi B} f} \\ X_s(f) = 0 \end{cases}$, ed effettuando l'antitrasformata di $X_c(f)$ si ottiene

$$x_c(t) = 2kB \text{sinc} \left[2B \left(t - \frac{A}{2\pi B} \right) \right]$$

in cui la traslazione nel tempo è dovuta alla fase lineare presente in $\underline{X}(f)$.

- Lo stesso problema precedente, ma applicato al segnale b), la cui trasformata $X(f)$ è mostrata al lato destro in alto di Fig. 10.2.
- Eseguendo di nuovo le operazioni di traslazione si ottiene l'involuppo complesso riportato in basso. Questa volta la fase di $\underline{X}(f)$ non è dispari, e dunque non si verificano le condizioni di simmetria coniugata, quindi $\underline{x}(t)$ è complesso. Si ha: $\underline{x}(t) = kB \left(\frac{\sin \pi B t}{\pi B t} \right)^2 e^{j\phi}$ e dunque

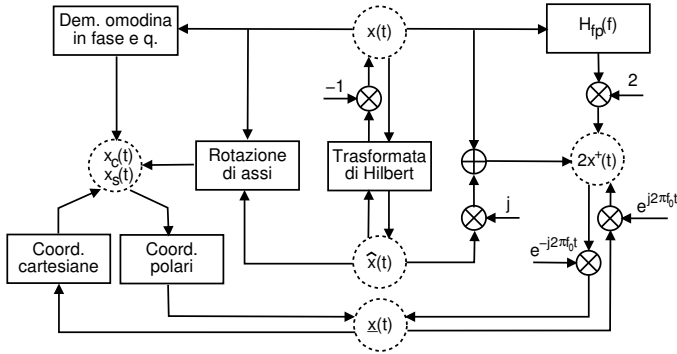
$$\begin{cases} x_c(t) = kB \left(\frac{\sin \pi B t}{\pi B t} \right)^2 \cos \phi \\ x_s(t) = kB \left(\frac{\sin \pi B t}{\pi B t} \right)^2 \sin \phi \end{cases} \Rightarrow \begin{cases} X_c(f) = k \left(1 - \frac{|f|}{B} \right) \cos \phi \\ X_s(f) = k \left(1 - \frac{|f|}{B} \right) \sin \phi \end{cases}$$

con $|f| < B$.

²⁰ Approfittiamo dell'occasione per notare che, pur potendo scrivere $\underline{X}(f) = X_c(f) + jX_s(f)$, non è assolutamente lecito dire che $\Re \{ \underline{X}(f) \} = X_c(f)$ e $\Im \{ \underline{X}(f) \} = X_s(f)$; infatti sia $X_c(f)$ che $X_s(f)$ possono a loro volta essere complessi (mentre $x_c(t)$ e $x_s(t)$ sono necessariamente reali).

10.2.7 Schema delle trasformazioni

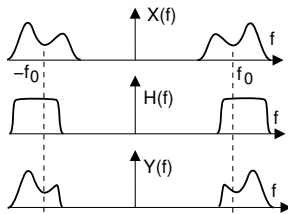
La figura che segue riassume le relazioni esistenti tra le grandezze $x(t)$, $\hat{x}(t)$, $\underline{x}(t)$, $x^+(t)$, $x_c(t)$ e $x_s(t)$.



10.3 Transito dei segnali modulati nei sistemi fisici

Trattiamo ora delle trasformazioni subite da un segnale modulato nel passaggio attraverso un sistema fisico, in funzione dell'involucro complesso del segnale $\underline{x}(t)$ e di quello della risposta impulsiva $\underline{h}(t)$. L'ovvia utilità di questa forma di descrizione è che ci consente di svolgere i calcoli maneggiando segnali di banda base.

10.3.1 Filtraggio



Poniamoci nelle condizioni generali di un segnale $x(t)$ che attraversa un filtro $h(t)$, che per entrambi valgano le condizioni di limitazione in banda, e che si adotti una stessa f_0 di riferimento. Anche il segnale in uscita $y(t)$ è dello stesso tipo, e può essere mostrato²¹ che il suo involucro complesso vale:

$$\underline{y}(t) = \frac{1}{2} \underline{x}(t) * \underline{h}(t)$$

²¹Per dimostrare il risultato, mostriamo innanzitutto che il segnale analitico in uscita vale $y^+(t) = x^+(t) * h^+(t)$. Infatti, omettendo di indicare nei passaggi la variabile (t) per compattezza di notazione, risulta

$$x^+(t) * h^+(t) = [x * h_{fp}] * [h * h_{fp}] = [x * h] * [h_{fp} * h_{fp}] = y * h_{fp} = y^+$$

in cui $h_{fp}(t)$ è la risposta impulsiva del filtro necessario ad estrarre il segnale analitico. Non resta ora che mostrare lo sviluppo per il risultato anticipato:

$$\begin{aligned} \frac{1}{2} \underline{x}(t) * \underline{h}(t) &= \frac{1}{2} [2x^+(t) e^{-j\omega_0 t}] * [2h^+(t) e^{-j\omega_0 t}] = \\ &= 2 \int_{-\infty}^{\infty} x^+(\tau) e^{-j\omega_0 \tau} h^+(t - \tau) e^{-j\omega_0 (t - \tau)} d\tau = \\ &= 2e^{-j\omega_0 t} \int_{-\infty}^{\infty} x^+(\tau) h^+(t - \tau) d\tau = 2e^{-j\omega_0 t} y^+(t) = \underline{y}(t) \end{aligned}$$

da cui è facile ottenere l'espressione di $y_c(t)$ ed $y_s(t)$ in funzione delle c.a. di b.f. di $x(t)$ e di quelle del filtro:

$$\begin{aligned} \underline{y} &= \frac{1}{2} \underline{x} * \underline{h} = \frac{1}{2} [x_c + jx_s] * [h_c + jh_s] = \\ &= \frac{1}{2} [x_c * h_c - x_s * h_s] + j \frac{1}{2} [x_s * h_c + x_c * h_s] \end{aligned}$$

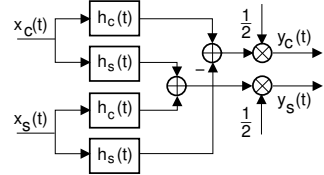
Dunque, per le componenti reale e immaginaria del segnale modulato, dopo il filtraggio, sussistono le relazioni

$$\begin{cases} y_c(t) = \frac{1}{2} [x_c(t) * h_c(t) - x_s(t) * h_s(t)] \\ y_s(t) = \frac{1}{2} [x_s(t) * h_c(t) + x_c(t) * h_s(t)] \end{cases} \quad (10.8)$$

e lo schema riportato appresso raffigura un circuito equivalente alle (10.8), ossia operante su $x_c(t)$ e $x_s(t)$, e che determina lo stesso risultato.

10.3.2 Intermodulazione tra componenti analogiche

Osservando il risultato (10.8) notiamo che sia $y_c(t)$ che $y_s(t)$ dipendono in generale da entrambe le componenti $x_c(t)$ e $x_s(t)$: questo fenomeno prende il nome di *intermodulazione* tra componenti analogiche di bassa frequenza, ed è fonte di una distorsione *ineliminabile* in banda base. Infatti, le informazioni contenute in $x_c(t)$ ed $x_s(t)$ sono ora mescolate in modo tale che, anche disponendo sia di $y_c(t)$ che di $y_s(t)$, non possono essere separate. Gli unici casi in cui ciò non si verifica sono relativi all'evenienza che $\underline{x}(t)$ oppure $\underline{h}(t)$ presentino *una sola* delle due C.A. di B.F., ossia almeno uno dei due sia solo reale o solo immaginario.



10.3.3 Equalizzazione di banda base

Nel caso in cui *non avvenga* il fenomeno di intermodulazione suddetto, come ad esempio se $\underline{h}(t) = h_c(t)$, e quindi risulti

$$\begin{cases} y_c(t) = \frac{1}{2} x_c(t) * h_c(t) \\ y_s(t) = \frac{1}{2} x_s(t) * h_c(t) \end{cases}$$

allora $y_c(t)$ e $y_s(t)$ sono affette unicamente da distorsione lineare (§ 14.5), e quindi le componenti *trasmesse* $x_c(t)$ e $x_s(t)$ possono essere ri-ottenute a partire da quelle *ricevute* $y_c(t)$ e $y_s(t)$ mediante un procedimento di *equalizzazione*, che consiste nel loro passaggio attraverso un filtro con risposta impulsiva $g_{eq}(t)$, la cui risposta in frequenza risulta pari a $G_{eq}(f) = ae^{j2\pi f\tau} / H_c(f)$, permettendo di ottenere

$$\begin{cases} x_c(t) = 2y_c(t) * g_{eq}(t) \\ x_s(t) = 2y_s(t) * g_{eq}(t) \end{cases}$$

Infatti in questo modo la risposta in frequenza complessiva, risultato del passaggio del segnale modulato prima nel canale, e quindi nell'equalizzatore, risulta essere il prodotto delle due risposte in frequenza $H_c(f) G_{eq}(f)$, e quindi pari a quella di un canale perfetto $H(f) = ae^{j2\pi f\tau}$ (vedi § 14).

Se al contrario sono presenti entrambe $h_c(t)$ e $h_s(t)$, o entrambe $x_c(t)$ e $x_s(t)$, allora non è più possibile eseguire il processo di equalizzazione operando sulle c.a. di b.f., mentre è teoricamente possibile operare direttamente sul segnale modulato, a patto di non incontrare ostacoli tecnologici dovuti alle elevate frequenze in gioco.

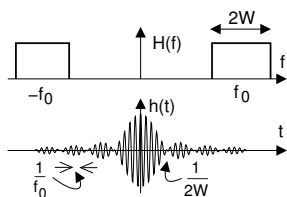
10.4 Condizioni per involuppo complesso reale

Data la rilevanza dei segnali con involuppo complesso ad una sola componente, determiniamo quali condizioni si debbano verificare per dar luogo ad una simile circostanza, iniziando da un esempio.

10.4.1 Filtro passa banda ideale

E' descritto da una risposta in frequenza $H(f)$ nulla ovunque, tranne che negli intervalli di frequenze $f_0 - W \leq |f| \leq f_0 + W$ dove ha valore unitario. Pertanto, risulta $H(f) = \text{rect}_{2W}(f - f_0) + \text{rect}_{2W}(f + f_0)$, da cui si ottiene facilmente

$$\begin{aligned} h(t) &= \mathcal{F}^{-1}\{H(f)\} = 2W \text{sinc}(2Wt) (e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}) = \\ &= 4W \text{sinc}(2Wt) \cos 2\pi f_0 t \end{aligned}$$



D'altra parte, l'andamento di $H(f)$ è quello tipico dei segnali modulati, e quindi per $h(t)$ vale la sua rappresentazione in termini di involuppo complesso $\underline{h}(t) = h_c(t) + jh_s(t)$, per cui possiamo scrivere

$$h(t) = h_c(t) \cos 2\pi f_0 t - h_s(t) \sin 2\pi f_0 t$$

Confrontando questa espressione con quella trovata prima, si osserva che deve necessariamente risultare $h_s(t) = 0$ ed $h_c(t) = 4W \frac{\sin 2\pi W t}{2\pi W t}$, per cui $\underline{h}(t)$ è reale.

10.4.2 Simmetria coniugata attorno ad f_0

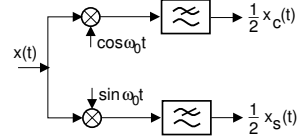
Il filtro passa banda ideale presenta $\underline{h}(t)$ reale, in quanto $\underline{H}(f) = 2H^+(f + f_0) = 2\text{rect}_{2W}(f)$ esibisce simmetria coniugata attorno all'origine. E' proprio questa la condizione cercata, che ci permette di enunciare

Un segnale modulato $x(t)$ possiede un involuppo complesso $\underline{x}(t)$ reale, se lo spettro di quest'ultimo $\underline{X}(f)$ ha simmetria coniugata attorno all'origine $\underline{X}(f) = \underline{X}^(-f)$, ovvero il segnale analitico $X^+(f)$ ha simmetria coniugata attorno ad f_0 : $X^+(f_0 + \phi) = [X^+(f_0 - \phi)]^*$ ($|\phi| < W$).*

In altre parole, $\underline{x}(t) = x_c(t)$ se $X^+(f)$ ha modulo pari e fase dispari rispetto ad f_0 .

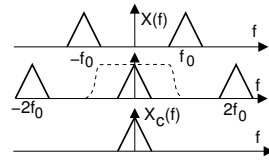
10.5 Demodulazione delle componenti analogiche di b. f.

Lo schema adottato al § 10.2.4 per ottenere le C.A. di B.F., basato sul filtro di Hilbert, non è l'unico. Un secondo metodo molto pratico prevede ancora l'uso di portanti *sincrone* (od *omodina*)²², e di due filtri passa basso anziché del filtro di Hilbert, come mostrato in figura. Il suo funzionamento è basato sul fatto che, considerando $x(t)$ espresso in termini delle sue C.A. di B.F., si ha²³:



$$\begin{aligned} x(t) \cos \omega_0 t &= [x_c(t) \cos \omega_0 t - x_s(t) \sin \omega_0 t] \cos \omega_0 t = \\ &= x_c(t) \cos^2 \omega_0 t - x_s(t) \sin \omega_0 t \cos \omega_0 t = \\ &= \frac{1}{2} x_c(t) + \frac{1}{2} x_c(t) \cos 2\omega_0 t - \frac{1}{2} x_s(t) \sin 2\omega_0 t \end{aligned}$$

I termini in cui compaiono $\cos 2\omega_0 t$ e $\sin 2\omega_0 t$ rappresentano componenti di segnale centrate attorno a $2f_0$, ed il filtro passa basso²⁴ (la cui $H(f)$ è tratteggiata in figura) provvede ad eliminarle: la banda del filtro deve quindi essere maggiore di W ma inferiore a $2f_0 - W$. Pertanto, il filtro non deve necessariamente essere ideale, e se $f_0 \gg W$ non sussistono particolari problemi realizzativi. Procedendo in maniera simile, per il ramo in quadratura si ottiene:



$$x(t) \sin \omega_0 t = \frac{1}{2} x_s(t) - \frac{1}{2} x_c(t) \cos 2\omega_0 t - \frac{1}{2} x_s(t) \sin 2\omega_0 t$$

e dunque, anche in questo caso, il filtro passa-basso rimuove le componenti a frequenza doppia.

Se i filtri non sono ideali, ma hanno ad esempio una fase lineare, saranno equivalenti ad un ritardo; se presentano distorsioni più severe (modulo non costante o fase non lineare), allora introducono distorsioni aggiuntive; per ridurre al minimo gli effetti di queste ultime, si tenta almeno di realizzare i due filtri quanto più identici tra loro.

10.6 Rappresentazione dei processi in banda traslata

Finora abbiamo trattato i casi di segnali di energia e di potenza; per ottenere una rappresentazione adeguata anche dei processi, occorre ancora un pò di teoria. Il lettore impaziente, o timoroso di perdersi tra i calcoli (che sono effettivamente intricati), può saltare direttamente alle conclusioni (§ 10.6.1), che sono le uniche che ci serviranno per il resto del testo. Altrimenti, armiamoci di pazienza e iniziamo.

Siamo ora interessati ad ottenere, una volta noto uno spettro di densità di potenza $\mathcal{P}_x(f)$ limitato in banda attorno ad f_0 , delle rappresentazioni utili per gli spettri di densità di potenza delle componenti analogiche di bassa frequenza, ovvero espressioni per le loro funzioni di autocorrelazione. Infatti, come abbiamo visto, il passaggio di un segnale in banda traslata attraverso un filtro può essere scomposto in 4 filtraggi in

²²Con queste parole si indica l'uso in ricezione della stessa identica portante usata per la trasmissione, senza errori né di fase né di frequenza, e le cui modalità sono indicate al § 11.2.1.

²³Si fa uso delle relazioni $\cos^2 \alpha = \frac{1}{2} (1 + \cos 2\alpha)$ e $\sin \alpha \cos \alpha = \frac{1}{2} \sin 2\alpha$

²⁴Il simbolo $\boxed{\approx}$ rappresenta un filtro passa-basso, poiché viene cancellata l'ondina superiore.

Nello stesso stile, sono a volte indicati un passa-alto $\boxed{\approx}$ ed un passa-banda $\boxed{\approx}$.

banda base: pertanto la rappresentazione delle C.A. di B.F. è sufficiente per ottenere tutte le altre grandezze di interesse.

Osserviamo innanzitutto che, se un processo aleatorio presenta una $\mathcal{P}_x(f)$ limitata in banda attorno ad f_0 , allora la funzione di autocorrelazione $\mathcal{R}_x(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_x(f)\}$ può essere espressa in termini delle componenti analogiche di bassa frequenza della funzione di autocorrelazione stessa:

$$\mathcal{R}_x(\tau) = \mathcal{R}_c(\tau) \cos \omega_0 \tau - \mathcal{R}_s(\tau) \sin \omega_0 \tau$$

Pertanto, non si ottengono direttamente le C.A. di B.F. del processo, come invece accade per segnali di energia di cui è noto $\underline{X}(f)$. D'altra parte, è innegabile che una realizzazione di $x(t)$ sia limitata in banda centrata a f_0 , e che quindi per essa debba esistere la rappresentazione $x(t) = x_c(t) \cos \omega_0 t - x_s(t) \sin \omega_0 t$; data la natura aleatoria di $x(t)$, gli stessi $x_c(t)$ ed $x_s(t)$ sono realizzazioni di processi. Questi ultimi in generale non sono indipendenti tra loro, in quanto la loro combinazione deve produrre un $x(t)$ che appartiene al processo originario. Si pensi ad esempio al segnale $x(t) = x_c(t) \cos \omega_0 t$, in cui $x_c(t)$ è stazionario ed ergodico: come già osservato al § 9.7.1, $x(t)$ è solamente ciclostazionario.

Come prima cosa, proviamo a calcolare la funzione di autocorrelazione dell'involuppo complesso di una generica realizzazione:

$$\begin{aligned} \mathcal{R}_{\underline{x}}(\tau) &= E\{\underline{x}^*(\tau) \underline{x}(t+\tau)\} = \\ &= E\{[x_c(\tau) - jx_s(\tau)][x_c(t+\tau) + jx_s(t+\tau)]\} = \\ &= E\{x_c(\tau)x_c(t+\tau) + x_s(\tau)x_s(t+\tau) + \\ &\quad + j[x_c(\tau)x_s(t+\tau) - x_s(\tau)x_c(t+\tau)]\} = \\ &= \mathcal{R}_{x_c}(\tau) + \mathcal{R}_{x_s}(\tau) + j[\mathcal{R}_{x_c x_s}(\tau) - \mathcal{R}_{x_s x_c}(\tau)] \end{aligned}$$

e queste 4 quantità sono calcolate in appendice 10.7.3. Il risultato finale dei calcoli, nel caso in cui $x_c(t)$ e $x_s(t)$ siano stazionari ed ergodici, fornisce le espressioni

$$\begin{cases} \mathcal{R}_{x_c}(\tau) &= \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x(\tau) \cos \omega_0 \tau + \widehat{\mathcal{R}}_x(\tau) \sin \omega_0 \tau \\ \mathcal{R}_{x_c x_s}(\tau) &= -\mathcal{R}_{x_s x_c}(\tau) = \widehat{\mathcal{R}}_x(\tau) \cos \omega_0 \tau - \mathcal{R}_x(\tau) \sin \omega_0 \tau \end{cases}$$

in cui $\widehat{\mathcal{R}}_x(\tau) = \mathcal{H}\{\mathcal{R}_x(\tau)\}$ è la trasformata di Hilbert di $\mathcal{R}_x(\tau)$. Osserviamo quindi come risulti $\mathcal{R}_{\underline{x}}(\tau) = 2[\mathcal{R}_{x_c}(\tau) + j\mathcal{R}_{x_c x_s}(\tau)]$, e pertanto $\mathcal{P}_{\underline{x}}(f) = 2[\mathcal{P}_{x_c}(f) + j\mathcal{P}_{x_c x_s}(f)]$.

- Da quest'ultima espressione, sembrerebbe che $\mathcal{P}_{\underline{x}}(f)$ possa assumere valori complessi, perdendo il senso fisico di potenza: ma non è così. Osserviamo infatti che $\mathcal{R}_{x_c x_s}(\tau)$ è un segnale reale dispari²⁵; pertanto $\mathcal{P}_{x_c x_s}(f) = \mathcal{F}\{\mathcal{R}_{x_c x_s}(\tau)\}$ è completamente immaginario, e dunque $\mathcal{P}_{\underline{x}}(f)$ è reale;
- se risultasse $\mathcal{R}_{x_c x_s}(\tau) = 0$ per ogni τ allora $\mathcal{P}_{x_c x_s}(f) = 0$ e $\mathcal{P}_{\underline{x}}(f) = 2\mathcal{P}_{x_c}(f)$ sarebbe reale pari; la presenza di $\mathcal{P}_{x_c x_s}(f)$ lo può invece rendere asimmetrico, permettendo di ottenere ancora $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_x^+(f + f_0)$ (non dimostrato);
- corollario del punto precedente è che, se $\mathcal{P}_x(f)$ è simmetrico rispetto ad f_0 , allora $\mathcal{R}_{x_c x_s}(\tau) = 0$ e le due C.A. di B.F. sono incorrelate; *se* inoltre queste sono congiuntamente gaussiane, allora risultano anche statisticamente indipendenti;

²⁵ Infatti $\mathcal{R}_{x_c x_s}(\tau) = \widehat{\mathcal{R}}_x(\tau) \cos \omega_0 \tau - \mathcal{R}_x(\tau) \sin \omega_0 \tau$, in cui $\mathcal{R}_x(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_x(f)\}$ è pari e $\sin \omega_0 t$ è dispari, mentre $\widehat{\mathcal{R}}_x(\tau)$ è dispari (non è stato dimostrato, ma vale per le trasformate di Hilbert di segnali pari) e $\cos \omega_0 \tau$ è pari. Inoltre, essendo $x_c(t)$ ed $x_s(t)$ reali, $\mathcal{R}_{x_c x_s}(\tau)$ è reale.

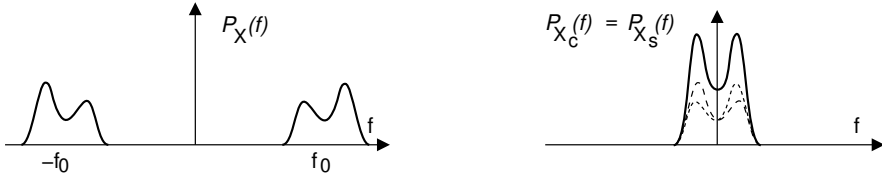


Figura 10.3: Segnale modulato e densità di potenza delle componenti analogiche di b.f.

- dato che $\mathcal{R}_{x_c x_s}(\tau) = -\mathcal{R}_{x_s x_c}(\tau)$ sono dispari, deve risultare che $\mathcal{R}_{x_c x_s}(0) = -\mathcal{R}_{x_s x_c}(0) = 0$; se i processi sono a media nulla, allora la potenza è pari al valore dell'autocorrelazione per $\tau = 0$, e quindi

$$\mathcal{P}_{\underline{x}} = \sigma_{\underline{x}}^2 = \mathcal{R}_{\underline{x}}(0) = 2\mathcal{R}_{x_c}(0) = 2\mathcal{R}_{x_s}(0) = 2\sigma_{x_c}^2 = 2\sigma_{x_s}^2 = 2\mathcal{P}_{x_c} = 2\mathcal{P}_{x_s}$$

In definitiva, le componenti analogiche di bassa frequenza hanno entrambe potenza metà di quella dell'involuppo complesso;

- $x_c(t)$ e $x_s(t)$ hanno (ciascuna) potenza pari a quella di $x(t)$, ovvero $\mathcal{P}_{x_c} = \mathcal{P}_{x_s} = \mathcal{P}_x$; infatti, ricordando che $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_x^+(f + f_0)$, si ottiene $\mathcal{P}_{\underline{x}} = 4\mathcal{P}_x^+$. Dovendo chiaramente risultare $\mathcal{P}_x = \mathcal{P}_x^+ + \mathcal{P}_x^-$, si ottiene

$$\mathcal{P}_x = \frac{1}{4} [\mathcal{P}_{\underline{x}} + \mathcal{P}_{\underline{x}}] = \frac{1}{2} \mathcal{P}_{\underline{x}} = \mathcal{P}_{x_c} = \mathcal{P}_{x_s}$$

- è possibile mostrare che, volendo esprimere l'autocorrelazione di $x(t)$ in termini delle sue C.A. di B.F. $\mathcal{R}_x(\tau) = \mathcal{R}_c(\tau) \cos \omega_0 \tau - \mathcal{R}_s(\tau) \sin \omega_0 \tau$, risulta

$$\begin{cases} \mathcal{R}_c(\tau) = \mathcal{R}_{x_c}(\tau) \\ \mathcal{R}_s(\tau) = -\mathcal{R}_{x_s}(\tau) \end{cases}$$

da cui è possibile mostrare che $\mathcal{R}_x(\tau) = \mathcal{R}_{\hat{x}}(\tau)$;

- volendo valutare $\mathcal{P}_{x_c}(f)$, questo risulta identico a $\mathcal{P}_{x_s}(f)$, in quanto (come già visto) $\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x(\tau) \cos \omega_0 \tau + \widehat{\mathcal{R}}_x(\tau) \sin \omega_0 \tau$; applicando ora la formula di Eulero per seno e coseno si ottiene

$$\begin{aligned} \mathcal{R}_{x_c}(\tau) &= \mathcal{R}_{x_s}(\tau) = \\ &= \mathcal{R}_x(\tau) \frac{e^{j\omega_0 \tau} + e^{-j\omega_0 \tau}}{2} + \widehat{\mathcal{R}}_x(\tau) \frac{e^{j\omega_0 \tau} - e^{-j\omega_0 \tau}}{2j} \\ &= \frac{1}{2} \left[\mathcal{R}_x(\tau) - j\widehat{\mathcal{R}}_x(\tau) \right] e^{j\omega_0 \tau} + \frac{1}{2} \left[\mathcal{R}_x(\tau) + j\widehat{\mathcal{R}}_x(\tau) \right] e^{-j\omega_0 \tau} \\ &= \mathcal{R}_x^-(\tau) e^{j\omega_0 \tau} + \mathcal{R}_x^+(\tau) e^{-j\omega_0 \tau} \end{aligned}$$

infatti i termini tra parentesi quadre corrispondono alla definizione di componenti a frequenze positive e negative ottenute tramite trasformata di Hilbert.

10.6.1 Conclusioni

Abbiamo mostrato che $\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x^-(\tau) e^{j\omega_0 \tau} + \mathcal{R}_x^+(\tau) e^{-j\omega_0 \tau}$. Risulta quindi:

$$\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f) = \mathcal{P}_x^-(f - f_0) + \mathcal{P}_x^+(f + f_0)$$

e dunque lo spettro di densità di potenza delle componenti analogiche di un processo si ottiene traslando e sovrapponendo (vedi fig. 10.3) le componenti a frequenze positive e negative del $\mathcal{P}_x(f)$ di partenza.

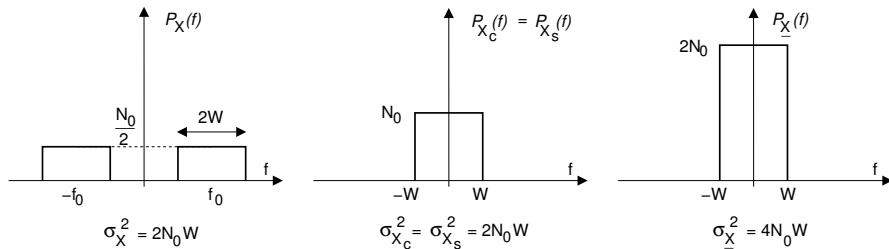


Figura 10.4: Densità di potenza per rumore passabanda

10.6.2 Processo gaussiano bianco limitato in banda

Se $x(t)$ è gaussiano stazionario ergodico e bianco, con $\mathcal{P}_x(f) = \frac{N_0}{2}$ limitato in banda $\pm W$ attorno ad f_0 ed a media nulla, allora (vedi fig. 10.4):

- $x_c(t)$ e $x_s(t)$ sono congiuntamente gaussiane, stazionarie, ergodiche e indipendenti, con $\mathcal{P}_{x_c} = \mathcal{P}_{x_s} = \mathcal{P}_x = 2N_0W$, e pari alle varianze σ_x^2 , $\sigma_{x_c}^2$ e $\sigma_{x_s}^2$. Le rispettive densità di potenza valgono:

$$\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f) = \mathcal{P}_x^+(f + f_0) + \mathcal{P}_x^-(f - f_0) = N_0 \text{rect}_{2W}(f)$$

- l'involuppo complesso ha potenza doppia:

$$\mathcal{P}_{\underline{x}} = 2\mathcal{P}_{x_c} = 4N_0W; \quad \mathcal{P}_{\underline{x}}(f) = 2N_0 \text{rect}_{2W}(f)$$

10.7 Appendici

10.7.1 Risposta impulsiva del filtro di Hilbert

Al § 10.2.4 si è affermato che $h_{\mathcal{H}}(t) = \frac{1}{\pi t}$, ed ora passiamo a dimostrarlo. Innanzitutto osserviamo che $H_{\mathcal{H}}(f)$ può essere scritta in termini di due gradini in frequenza $g(f)$, con quello relativo alle frequenze negative, di tipo anticausale:

$$H_{\mathcal{H}}(f) = -j \cdot g(f) + j \cdot g(-f)$$

Per calcolare l'antitrasformata di Fourier di un gradino in frequenza, utilizziamo il risultato noto per la trasformata di un gradino nel tempo (vedi § 3.9.4) $G(f) = \mathcal{F}\{g(t)\} = \frac{1}{2} \left(\delta(f) - \frac{j}{\pi f} \right)$, a cui applichiamo la proprietà di dualità (vedi pag. 32) che asserisce che se $G(f) = \mathcal{F}\{g(t)\}$, allora $\mathcal{F}^{-1}\{g(f)\} = G(-t)$, per ottenere

$$\mathcal{F}^{-1}\{g(f)\} = \frac{1}{2} \left(\delta(-t) - \frac{j}{\pi(-t)} \right) = \frac{1}{2} \left(\delta(t) + \frac{j}{\pi t} \right) \quad (10.9)$$

Per scrivere l'antitrasformata del gradino anticausale in frequenza $g(-f)$, occorre tenere conto della proprietà del cambiamento di scala (vedi pag. § 34), che asserisce che $\mathcal{F}^{-1}\{X(-f)\} = x(-t)$, e che applicata alla (10.9), consente di ottenere

$$\mathcal{F}^{-1}\{g(-f)\} = \frac{1}{2} \left(\delta(-t) + \frac{j}{\pi(-t)} \right) = \frac{1}{2} \left(\delta(t) - \frac{j}{\pi t} \right)$$

Ora possiamo quindi scrivere

$$\begin{aligned} h_{\mathcal{H}}(t) &= \mathcal{F}^{-1}\{H_{\mathcal{H}}(f)\} = -j \cdot \frac{1}{2} \left(\delta(t) + \frac{j}{\pi t} \right) + j \frac{1}{2} \left(\delta(t) - \frac{j}{\pi t} \right) \\ &= -j \frac{1}{2} \delta(t) - j^2 \frac{1}{2\pi t} + j \frac{1}{2} \delta(t) - j^2 \frac{1}{2\pi t} \\ &= \frac{1}{\pi t} \end{aligned}$$

10.7.2 Trasformata di Hilbert di un segnale modulato

Limitiamoci a dimostrare che

$$\mathcal{H}\{x_c(t) \cos 2\pi f_0 t\} = x_c(t) \sin 2\pi f_0 t \quad (10.10)$$

Iniziamo con il considerare che dopo aver \mathcal{F} -trasformato la (10.10), possiamo evidenziarne le componenti a frequenza positiva e negativa $X_c(f - f_0)$ e $X_c(f + f_0)$

$$x_c(t) \cos 2\pi f_0 t = \frac{x_c(t)}{2} \left(e^{j2\pi f_0 t} + e^{-j2\pi f_0 t} \right) \xrightarrow{\mathcal{F}} \frac{1}{2} [X_c(f - f_0) + X_c(f + f_0)] \quad (10.11)$$

che, se $x_c(t)$ ha una banda minore di f_0 , possono essere facilmente \mathcal{H} -trasformate semplicemente aggiungendo lo sfasamento introdotto a frequenze positive e negative dal filtro di Hilbert

$$\frac{1}{2} [X_c(f - f_0) + X_c(f + f_0)] \xrightarrow{\mathcal{H}} \frac{1}{2} [X_c(f - f_0) e^{-j\frac{\pi}{2}} + X_c(f + f_0) e^{j\frac{\pi}{2}}]$$

e quindi \mathcal{F} -antitrasformando questa espressione si ottiene la \mathcal{H} -trasformata del segnale (10.10)

$$\frac{1}{2} [X_c(f - f_0) e^{-j\frac{\pi}{2}} + X_c(f + f_0) e^{j\frac{\pi}{2}}] \xrightarrow{\mathcal{F}^{-1}} \frac{x_c(t)}{2} \left(e^{j2\pi f_0 t} e^{-j\frac{\pi}{2}} + e^{-j2\pi f_0 t} e^{j\frac{\pi}{2}} \right)$$

risultato che, anche se non ancora nella forma anticipata, poteva comunque essere ottenuto anche direttamente a partire dal secondo membro di (10.11), invocando subito la limitazione ad una semibanda di $x_c(t) e^{\pm j2\pi f_0 t}$. Per ottenere la (10.10) è ora sufficiente moltiplicare e dividere per $j = e^{j\frac{\pi}{2}}$, ossia

$$\begin{aligned} \frac{x_c(t)}{2} \left(e^{j2\pi f_0 t} e^{-j\frac{\pi}{2}} + e^{-j2\pi f_0 t} e^{j\frac{\pi}{2}} \right) \cdot \frac{e^{j\frac{\pi}{2}}}{e^{j\frac{\pi}{2}}} &= \frac{x_c(t)}{2j} \left(e^{j2\pi f_0 t} + e^{-j2\pi f_0 t} e^{j\pi} \right) = \\ &= \frac{x_c(t)}{2j} \left(e^{j2\pi f_0 t} - e^{-j2\pi f_0 t} \right) = x_c(t) \sin 2\pi f_0 t \end{aligned}$$

in quanto $e^{j\pi} = -1$.

10.7.3 Autocorrelazione di processi passa-banda

Svolgiamo qui il calcolo relativo al valore di $\mathcal{R}_{x_c}(\tau)$, $\mathcal{R}_{x_s}(\tau)$, $\mathcal{R}_{x_c x_s}(\tau)$ e $\mathcal{R}_{x_s x_c}(\tau)$. Ricordando che $x_c(t) = x(t) \cos \omega_0 t + \hat{x}(t) \sin \omega_0 t$, iniziamo a svolgere i calcoli per $\mathcal{R}_{x_c}(\tau)$:

$$\begin{aligned}
\mathcal{R}_{x_c}(\tau) &= E\{x_c(\tau)x_c(t+\tau)\} = \\
&= E\{[x(t)\cos\omega_0t + \hat{x}(t)\sin\omega_0t] \cdot \\
&\quad \cdot [x(t+\tau)\cos\omega_0(t+\tau) + \hat{x}(t+\tau)\sin\omega_0(t+\tau)]\} = \\
&= E\{x(t)x(t+\tau)\} \cdot \cos\omega_0t \cdot \cos\omega_0(t+\tau) + \\
&+ E\{x(t)\hat{x}(t+\tau)\} \cdot \cos\omega_0t \cdot \sin\omega_0(t+\tau) + \\
&+ E\{\hat{x}(t)x(t+\tau)\} \cdot \sin\omega_0t \cdot \cos\omega_0(t+\tau) \\
&+ E\{\hat{x}(t)\hat{x}(t+\tau)\} \cdot \sin\omega_0t \cdot \sin\omega_0(t+\tau)
\end{aligned}$$

Valutiamo quindi i quattro valori attesi singolarmente, indicando con $\overline{x(t)}$ la media temporale di $x(t)$, ossia $\overline{x(t)} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t) dt$:

$$\begin{aligned}
E\{x(t)x(t+\tau)\} &= \overline{x(t)x(t+\tau)} = \mathcal{R}_x(\tau) \\
E\{x(t)\hat{x}(t+\tau)\} &= \overline{x(t)\hat{x}(t+\tau)} = \mathcal{R}_{x\hat{x}}(\tau) = x(-\tau) * \hat{x}(\tau) = \\
&= x(-\tau) * x(\tau) * \frac{1}{\pi\tau} = \mathcal{R}_x(\tau) * \frac{1}{\pi\tau} = \widehat{\mathcal{R}}_x(\tau) \\
E\{\hat{x}(t)x(t+\tau)\} &= \overline{\hat{x}(t)x(t+\tau)} = \mathcal{R}_{\hat{x}x}(\tau) = \hat{x}(-\tau) * x(\tau) = \\
&= x(-\tau) * \left(-\frac{1}{\pi\tau}\right) * x(\tau) = x(-\tau) * x(\tau) * \left(-\frac{1}{\pi\tau}\right) = \\
&= \mathcal{R}_x(\tau) * \left(-\frac{1}{\pi\tau}\right) = -\widehat{\mathcal{R}}_x(\tau) \\
E\{\hat{x}(t)\hat{x}(t+\tau)\} &= \overline{\hat{x}(t)\hat{x}(t+\tau)} = \mathcal{R}_{\hat{x}\hat{x}}(\tau) = \hat{x}(-\tau) * \hat{x}(\tau) = \\
&= x(-\tau) * \left(-\frac{1}{\pi\tau}\right) * x(\tau) * \frac{1}{\pi\tau} = \\
&= x(-\tau) * x(\tau) * \left(-\frac{1}{\pi\tau}\right) * \frac{1}{\pi\tau} = -\widehat{\widehat{\mathcal{R}}}_x(\tau) = \mathcal{R}_x(\tau)
\end{aligned}$$

Sostituendo le relazioni ora trovate nella espressione di $\mathcal{R}_{x_c}(\tau)$, si ottiene

$$\begin{aligned}
\mathcal{R}_{x_c}(\tau) &= \mathcal{R}_x(\tau) \cdot \cos\omega_0t \cdot \cos\omega_0(t+\tau) + \widehat{\mathcal{R}}_x(\tau) \cdot \cos\omega_0t \cdot \sin\omega_0(t+\tau) + \\
&- \widehat{\widehat{\mathcal{R}}}_x(\tau) \cdot \sin\omega_0t \cdot \cos\omega_0(t+\tau) + \mathcal{R}_x(\tau) \cdot \sin\omega_0t \cdot \sin\omega_0(t+\tau) = \\
&= \frac{1}{2}\mathcal{R}_x(\tau) [\cos\omega_0(-\tau) + \cos\omega_0(2t+\tau)] + \\
&+ \frac{1}{2}\widehat{\mathcal{R}}_x(\tau) [\sin\omega_0(\tau) + \sin\omega_0(2t+\tau)] + \\
&- \frac{1}{2}\widehat{\widehat{\mathcal{R}}}_x(\tau) [\sin\omega_0(-\tau) + \sin\omega_0(2t+\tau)] + \\
&+ \frac{1}{2}\mathcal{R}_x(\tau) [\cos\omega_0(-\tau) - \cos\omega_0(2t+\tau)] = \\
&= \mathcal{R}_x(\tau) \cdot \cos\omega_0\tau + \widehat{\mathcal{R}}_x(\tau) \cdot \sin\omega_0\tau
\end{aligned}$$

che costituisce il risultato anticipato. Per l'espansione dei termini trigonometrici, si è fatto uso delle relazioni:

$$\begin{aligned}
\cos\alpha \cdot \cos\beta &= \frac{1}{2} [\cos(\alpha - \beta) + \cos(\alpha + \beta)] \\
\sin\alpha \cdot \sin\beta &= \frac{1}{2} [\cos(\alpha - \beta) - \cos(\alpha + \beta)] \\
\sin\alpha \cdot \cos\beta &= \frac{1}{2} [\sin(\alpha - \beta) + \sin(\alpha + \beta)]
\end{aligned}$$

I calcoli relativi al valore di $\mathcal{R}_{x_s}(\tau)$ sono del tutto simili, ed il loro svolgimento porta al risultato $\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau)$.

Per quanto riguarda $\mathcal{R}_{x_c x_s}(\tau)$, applichiamo la relazione $x_s(t) = \widehat{x}(t) \cos \omega_0 t - x(t) \sin \omega_0 t$, per ottenere:

$$\begin{aligned}
 \mathcal{R}_{x_c x_s}(\tau) &= E \{x_c(\tau) x_s(t + \tau)\} = \\
 &= E \{[x(t) \cos \omega_0 t + \widehat{x}(t) \sin \omega_0 t] \cdot \\
 &\quad \cdot [\widehat{x}(t + \tau) \cos \omega_0(t + \tau) - x(t + \tau) \sin \omega_0(t + \tau)]\} = \\
 &= E \{x(t) \widehat{x}(t + \tau)\} \cdot \cos \omega_0 t \cdot \cos \omega_0(t + \tau) + \\
 &\quad - E \{x(t) x(t + \tau)\} \cdot \cos \omega_0 t \cdot \sin \omega_0(t + \tau) + \\
 &\quad + E \{\widehat{x}(t) \widehat{x}(t + \tau)\} \cdot \sin \omega_0 t \cdot \cos \omega_0(t + \tau) + \\
 &\quad - E \{\widehat{x}(t) x(t + \tau)\} \cdot \sin \omega_0 t \cdot \sin \omega_0(t + \tau)
 \end{aligned}$$

I valori attesi che vediamo comparire sono stati già calcolati, e quindi possiamo scrivere direttamente lo sviluppo dei calcoli, in cui si applicano nuovamente le identità trigonometriche note:

$$\begin{aligned}
 \mathcal{R}_{x_c x_s}(\tau) &= \widehat{\mathcal{R}}_x(\tau) \cdot \cos \omega_0 t \cdot \cos \omega_0(t + \tau) - \mathcal{R}_x(\tau) \cdot \cos \omega_0 t \cdot \sin \omega_0(t + \tau) + \\
 &\quad + \mathcal{R}_x(\tau) \cdot \sin \omega_0 t \cdot \cos \omega_0(t + \tau) - \widehat{\mathcal{R}}_x(\tau) \cdot \sin \omega_0 t \cdot \sin \omega_0(t + \tau) = \\
 &= \frac{1}{2} \widehat{\mathcal{R}}_x(\tau) [\cos \omega_0(-\tau) + \cos \omega_0(2t + \tau)] + \\
 &\quad - \frac{1}{2} \mathcal{R}_x(\tau) [\sin \omega_0(\tau) + \sin \omega_0(2t + \tau)] + \\
 &\quad - \frac{1}{2} \mathcal{R}_x(\tau) [\sin \omega_0(-\tau) + \sin \omega_0(2t + \tau)] + \\
 &\quad - \frac{1}{2} \widehat{\mathcal{R}}_x(\tau) [\cos \omega_0(-\tau) - \cos \omega_0(2t + \tau)] = \\
 &= -\mathcal{R}_x(\tau) \cdot \sin \omega_0(2t + \tau) + \widehat{\mathcal{R}}_x(\tau) \cdot \cos \omega_0(2t + \tau)
 \end{aligned}$$

Per quanto riguarda gli argomenti delle funzioni trigonometriche, il valore di t è lasciato non specificato. Pertanto, visto che il processo è stazionario per ipotesi, può senz'altro essere posto a zero, e dunque ottenere il risultato previsto.

I calcoli relativi al valore di $\mathcal{R}_{x_s x_c}(\tau)$ sono del tutto simili, ed il loro svolgimento porta al risultato $\mathcal{R}_{x_s x_c}(\tau) = -\mathcal{R}_{x_c x_s}(\tau)$.

Capitolo 11

Modulazione per segnali analogici

Analizziamo ora le tecniche adottate per modulare (in ampiezza o angularmente) una portante con l'informazione di un messaggio informativo di natura analogica¹, studiando allo stesso tempo le caratteristiche spettrali del segnale ottenuto. Sono inoltre discussi i diversi approcci di demodulazione, i circuiti che li realizzano, e l'influenza di eventuali inaccurattezze. Infine, si fornisce un breve dettaglio del sistema di broadcast radiotelevisivo analogico.

11.1 Modulazione di ampiezza - AM

Al § 10.2 si è mostrato che un segnale modulato $x(t)$ può essere rappresentato nei termini delle sue *componenti analogiche di bassa frequenza* $x_c(t)$ e $x_s(t)$: nel caso in cui queste rappresentino due segnali indipendenti, la loro trasmissione congiunta sulla medesima portante realizza un segnale QAM (*quadrature amplitude modulation*). Al contrario, nei casi più tipici x_c ed x_s non sono qualsiasi, ma sussiste tra loro una relazione, in base alla quale si distinguono le seguenti classi di segnali *modulati in ampiezza*:

- *banda laterale doppia*: la componente $x_s(t)$ è nulla, cosicché $\mathcal{P}_x(f)$ è simmetrico rispetto ad f_0 . Si tratta del caso a noi già noto, ed è indicato dagli acronimi BLD o DSB (*double side band*).
- *banda laterale unica*: sono presenti sia $x_c(t)$ che $x_s(t)$, e risulta $x_s(t) = \hat{x}_c(t)$. Questo fa sì che (come vedremo) la densità $\mathcal{P}_x(f)$ del segnale modulato giaccia tutta all'*esterno* (od all'*interno*) di $\pm f_0$ (BLU o SSB - *single side band*).
- *banda laterale ridotta*: è una via di mezzo tra i due casi precedenti, e cioè $\mathcal{P}_x(f)$ non è simmetrica rispetto ad f_0 , ma comunque giace da ambo i lati (BLR o VSB - *vestigial side band*²).

Per completare la classificazione, per ognuna delle possibilità precedenti può verificarsi uno tra tre sottocasi, che si riferiscono alla presenza o meno, in $\mathcal{P}_x(f)$, di una concentrazione di potenza (ossia un impulso) a frequenza f_0 , corrispondente alla trasmissione di potenza non associata al messaggio $m(t)$, ma solamente alla portante, e quindi priva

¹Per i segnali numerici si usano tecniche peculiari, esposte al capitolo 13.

²Come sarà più chiaro nel seguito, l'acronimo VSB simboleggia che, anziché sopprimere completamente una delle due bande laterali, se ne mantengono *delle vestigia*.

di contenuto informativo ai fini della trasmissione. I tre sottocasi citati sono indicati come:

- portante soppressa (PS o SC - *suppressed carrier*);
- portante intera (PI o LC - *large carrier*);
- portante parzialmente soppressa (PPS).

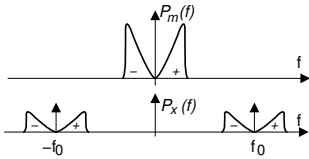
11.1.1 Banda laterale doppia - BLD

In questo caso l'involuppo complesso $\underline{x}(t)$ del segnale modulato presenta *una sola* componente analogica di bassa frequenza, che per convenzione è posta pari a $x_c(t)$ ³. La dipendenza di $x_c(t)$ da $m(t)$ è posta nella forma generale $x_c(t) = a_p + k_a m(t)$, e quindi

$$x_{BLD}(t) = (a_p + k_a m(t)) \cos \omega_0 t$$

L'involuppo complesso perciò risulta $\underline{x}(t) = a_p + k_a m(t)$, mentre la sua densità di potenza vale

$$\mathcal{P}_{\underline{x}}(f) = a_p^2 \delta(f) + k_a^2 \mathcal{P}_m(f)$$



e quindi, dato che $\mathcal{P}_x(f) = \mathcal{P}_{x^+}(f) + \mathcal{P}_{x^-}(f)$ e che, in base alla (10.7) risulta $\mathcal{P}_{x^+}(f) = \frac{1}{4} \mathcal{P}_{\underline{x}}(f - f_0)$ e $\mathcal{P}_{x^-}(f) = \frac{1}{4} \mathcal{P}_{\underline{x}}(f + f_0)$, si ottiene una densità di potenza per il segnale modulato pari a

$$\mathcal{P}_x(f) = \frac{a_p^2}{4} [\delta(f - f_0) + \delta(f + f_0)] + \frac{k_a^2}{4} [\mathcal{P}_m(f - f_0) + \mathcal{P}_m(f + f_0)] \quad (11.1)$$

La potenza totale di $x(t)$ vale perciò $\mathcal{P}_x = \frac{a_p^2}{2} + \frac{k_a^2}{2} \mathcal{P}_m$, mentre la sua densità spettrale è riportata in figura, dove si è posto $k_a = 1$.

11.1.1.1 Portante soppressa - PS

Esaminando l'espressione trovata per \mathcal{P}_x per AM-BLD, notiamo che il termine $\frac{a_p^2}{2}$ rappresenta la potenza della portante non modulata (concentrata per metà ad f_0 e per metà a $-f_0$), che quindi svanisce per $a_p = 0$, dando luogo in quest'ultima circostanza al sottocaso di *portante soppressa*.

La demodulazione di AM-BLD-PS si effettua in modo coerente (§ 11.2.1), dopo aver ricostruito la portante per quadratura (§ 11.4.2), oppure mediante demodulatore ad involuppo (§ 11.2.3), dopo aver elaborato la portante ricostruita come spiegato al § 11.1.1.3.

11.1.1.2 Portante intera - PI

Nel caso in cui si ponga $a_p \neq 0$ si può scegliere $a_p \geq k_a \cdot \max\{|m(t)|\}$, in modo che risulti sempre $x_c(t) = a_p + k_a m(t) \geq 0$; la disequaglianza può essere scritta in modo equivalente come

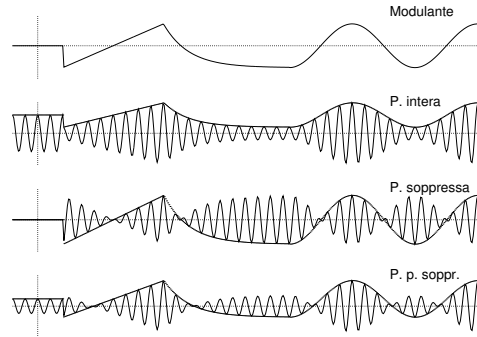
$$a_p^2 \geq k_a^2 m^2(t) \quad \text{per } \forall t \quad (11.2)$$

³Considerando che la portante di modulazione può avere una fase iniziale arbitraria, e che con una traslazione temporale ci si può sempre ricondurre ad usare una funzione $\cos \omega_0 t$, la convenzione posta tratta il caso di un segnale modulato $x(t) = a(t) \cos(\omega_0 t + \varphi)$ generico, con φ costante.

che rappresenta la condizione che caratterizza il caso di *portante intera*, in cui $x_c(t)$ appunto *non inverte* mai il segno; il rapporto $\left(\frac{a_p}{k_a}\right)^2$ identifica quindi la massima *potenza istantanea*⁴ (o di *picco*) P_I^{Max} per $m(t)$ affinché la (11.2) si verifichi. Viceversa, la conoscenza di P_I^{Max} consente di dimensionare i parametri a_p e k_a in modo da realizzare le condizioni di portante intera⁵, nel cui caso il rapporto $I_a = \frac{k_a \cdot \max\{|m(t)|\}}{a_p}$. 100 rappresenta *l'indice di modulazione*⁶.

D'altra parte, nel caso di portante soppressa $x_c(t)$ ha invece media nulla (se $m(t)$ ha media nulla) e la portante cambia frequentemente segno, cosicché per $f = f_0$ non compaiono impulsi in $\mathcal{P}_x(f)$.

La ragione principale dell'utilizzo della portante intera è che in tal caso il processo di decodifica non richiede la conoscenza di f_0 , e può svolgersi facendo uso di un semplice demodulatore di involuppo, descritto in 11.2.3.



11.1.1.3 Portante parzialmente soppressa - PPS

Se a_p è inferiore al valore necessario per avere la portante intera, ma non è nullo, si ottiene il caso della portante *parzialmente soppressa*, che ci permette di risparmiare potenza (vedi §11.1.1.4). Il residuo di portante presente può essere usato per la sua *ri-generazione* al lato ricevente, utilizzando un PLL (§11.2.1.2), in modo da *sommarla* al segnale ricevuto, ri-producendo così il termine $a_p \cos \omega_0 t$. In tal modo ci si riconduce al caso *PI*, e si può effettuare la demodulazione di involuppo (§11.2.3).

11.1.1.4 Efficienza di PI-PPS

Nell'espressione della potenza totale $\mathcal{P}_x = \frac{1}{2}(a_p^2 + k_a^2 \mathcal{P}_m)$ per un generico segnale AM, notiamo che solo $\mathcal{P}_u = \frac{k_a^2}{2} \mathcal{P}_m$ è relativa al segnale utile, mentre $\frac{a_p^2}{2}$ viene spesa sulla portante, che non trasporta informazione. Pertanto, si definisce una *efficienza energetica*

$$\eta = \frac{\mathcal{P}_u}{\mathcal{P}_x} = \frac{\frac{1}{2} k_a^2 \mathcal{P}_m}{\frac{1}{2} (a_p^2 + k_a^2 \mathcal{P}_m)} = \frac{1}{1 + \frac{a_p^2}{k_a^2 \mathcal{P}_m}}$$

⁴Si definisce come potenza istantanea di $m(t)$, il segnale $\mathcal{P}_I(t) = m^2(t)$, e la potenza di picco P_I^{Max} ne rappresenta il massimo valore per qualunque t .

⁵Ad esempio, nel caso in cui $m(t)$ sia un processo con densità di probabilità uniforme tra $\pm \frac{\Delta}{2}$, la potenza di picco risulta essere $\frac{\Delta^2}{4} = 3\sigma_M^2$, dato che (come mostrato al § 7.2.3) in quel caso risulta $\sigma_M^2 = \frac{\Delta^2}{12}$; se invece $m(t) = a \sin 2\pi f_M t$, allora si ha una potenza di picco $a^2 = 2\sigma_M^2$ (dato che $\mathcal{P}_M = \sigma_M^2 = \frac{a^2}{2}$). Oppure ancora, se $m(t)$ è gaussiano la potenza di picco (e dunque a_p^2/k_a^2 per ottenere la portante intera) risulta *infinita*. E cosa accade allora? Si avrà necessariamente una portante ridotta...

⁶Un indice I_a pari al 100% indica che si sta sfruttando *tutta* la dinamica della portante, mentre se $I_a > 100$ ci troviamo in condizioni di *sovramodulazione*, e non più di portante intera.

che rappresenta la frazione di potenza trasmessa utile ai fini della ricostruzione del messaggio⁷.

11.1.2 Banda laterale unica - BLU

Mentre con la modulazione BLD si determina una occupazione di banda per il segnale modulato $x(t)$ doppia di quella del segnale modulante, la tecnica BLU impegna invece una banda uguale a quella di $m(t)$. Un tale risultato è ottenuto rendendo le componenti analogiche $x_c(t)$ ed $x_s(t)$ del segnale modulato $x(t)$ dipendenti tra loro, ed in particolare imponendo che $x_c(t) = m(t)$ e $x_s(t) = \hat{m}(t)$: infatti, in tal modo si ottiene

$$\begin{aligned} x_{BLU}(t) &= m(t) \cos \omega_0 t - \hat{m}(t) \sin \omega_0 t = \\ &= m(t) \frac{e^{j\omega_0 t} + e^{-j\omega_0 t}}{2} - \hat{m}(t) \frac{e^{j\omega_0 t} - e^{-j\omega_0 t}}{2j} = \\ &= e^{j\omega_0 t} \frac{1}{2} [m(t) + j\hat{m}(t)] + e^{-j\omega_0 t} \frac{1}{2} [m(t) - j\hat{m}(t)] \end{aligned}$$

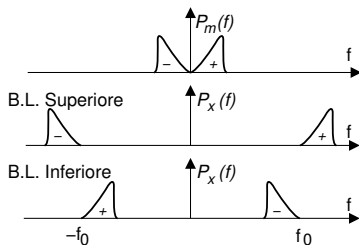
Ricordando ora che $\frac{1}{2} [m(t) \pm j\hat{m}(t)] = m^\pm(t)$ (vedi eq. (10.5)) è proprio il contenuto a frequenze positive (negative), allora se $x(t)$ è di energia, effettuando la trasformata di Fourier di ambo i membri si ottiene

$$\begin{aligned} X_{BLU}(f) &= \delta(f - f_0) * M^+(f) + \delta(f + f_0) * M^-(f) = \\ &= M^+(f - f_0) + M^-(f + f_0) \end{aligned}$$

e quindi il segnale modulato AM-BLU è costituito dal contenuto a frequenze positive e negative di $m(t)$, traslato ai lati della portante f_0 .

Qualora invece si consideri il segnale modulante $m(t)$ come una realizzazione di un processo ergodico, si può dimostrare (passando dalla trasformata di $\mathcal{R}_x(\tau)$) un risultato del tutto analogo, ovvero

$$\mathcal{P}_x(f) = \mathcal{P}_{m+}(f - f_0) + \mathcal{P}_{m-}(f + f_0)$$



Nel caso descritto abbiamo considerato soppressa la portante, ed il segnale modulato (considerato nel dominio della frequenza) risulta *esterno* ad f_0 : questa circostanza è indicata con il termine di *banda laterale superiore*. Il caso opposto (*banda laterale inferiore*) si ottiene cambiando segno a $x_s(t)$. Scriviamo dunque

$$x_{BLU}(t) = \frac{k_a}{\sqrt{2}} m(t) \cos \omega_0 t \mp \frac{k_a}{\sqrt{2}} \hat{m}(t) \sin \omega_0 t$$

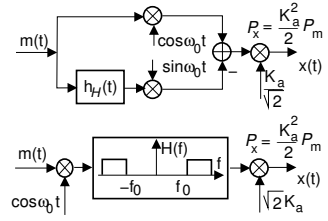
con $-$ e $+$ rispettivamente per ottenere un segnale BLU con banda superiore o inferiore. In entrambi i casi il segnale modulato BLU ha una potenza $\mathcal{P}_x = 2 \cdot (\frac{k_a^2}{2} \cdot \mathcal{P}_m \cdot \frac{1}{2}) = \frac{k_a^2}{2} \mathcal{P}_m$ (vedi § 11.4.1), eguale a quella di un segnale AM-BLD in cui $x_c(t) = k_a m(t)$ e $x_s(t) = 0$.

I vantaggi di un tale metodo di modulazione sono subito evidenti: consente infatti di risparmiare banda, permettendo la trasmissione di più messaggi in divisione di frequenza FDM, vedi pag. 251.

⁷Ad esempio, se $m(t) = \sin 2\pi f_M t$ si ha $\mathcal{P}_M = 1/2$ e, nel caso di portante intera, deve risultare $a_p = k_a$ e dunque $\eta = \frac{1}{1+2} = 0.33$. Ovvero solo 1/3 della potenza trasmessa è utile al ricevitore!

11.1.2.1 Generazione di segnali BLU

Un segnale BLU può essere generato in due diversi modi, rappresentati alla figura seguente. Il primo consiste nell'uso di un filtro di Hilbert per calcolare $\hat{m}(t)$, da usare assieme ad $m(t)$ in un modulatore in fase ed in quadratura. E' subito evidente come si possano presentare problemi se $m(t)$ ha contenuti energetici prossimi a frequenza zero, che rendono assai stringenti le specifiche per approssimare il filtro di Hilbert.

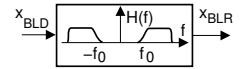


Un problema simile si presenta anche con il secondo metodo di generazione del segnale BLU, in cui viene prima prodotto un segnale BLD, e quindi lo si filtra in modo da eliminare una delle bande laterali: la necessità di trasmettere frequenze di $m(t)$ prossime allo zero, complica infatti la realizzazione del filtro.

La trasmissione FDM di segnali BLU è stata lungamente usata per i ponti radio telefonici (vedi § 10.1.1.3). Pertanto, la limitazione sulla minima frequenza di una canale telefonico a 300 Hz è motivata anche dalla necessità di effettuare modulazioni BLU.

11.1.3 Banda laterale ridotta - BLR

Si può verificare il caso in cui non si possa assolutamente fare a meno di componenti di segnale a frequenza molto bassa, come avviene, ad esempio, nel segnale televisivo⁸ (vedi appendice 11.4.4). Si ricorre allora alla modulazione a *banda laterale ridotta* (BLR), ottenuta facendo transitare il segnale modulato BLD attraverso uno specifico filtro, che presenta una regione di transizione tra la banda passante e quella attenuata *più dolce* di quella di un passa-banda ideale, e che si estende oltre f_0 .



11.1.4 Potenza di un segnale AM

Qui sotto è mostrato uno schema riassuntivo dell'espressione del segnale modulato per i diversi tipi di modulazione AM, assieme ai valori k_a ed a_p tali da determinare una potenza totale \mathcal{P}_x .

	Segnale modulato $x(t)$	Potenza \mathcal{P}_x	k_a per \mathcal{P}_x dato
BLD-PS	$k_a m(t) \cos(\omega_0 t)$	$\frac{k_a^2}{2} \mathcal{P}_m$	$\sqrt{\frac{2\mathcal{P}_x}{\mathcal{P}_m}}$
BLU-PS	$\frac{k_a}{\sqrt{2}} m(t) \cos(\omega_0 t) - \frac{k_a}{\sqrt{2}} \hat{m}(t) \sin(\omega_0 t)$	$\frac{k_a^2}{2} \mathcal{P}_m$	$\sqrt{\frac{2\mathcal{P}_x}{\mathcal{P}_m}}$
BLD-PI	$[a_p + k_a m(t)] \cos(\omega_0 t)$ con $a_p \geq k_a \cdot \max\{ m(t) \}$	$\frac{a_p^2}{2} + \frac{k_a^2}{2} \mathcal{P}_m$	$\sqrt{\frac{2\mathcal{P}_x - a_p^2}{\mathcal{P}_m}}$

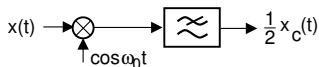
11.2 Demodulazione di ampiezza

Un segnale AM può essere demodulato mediante diverse tecniche, ossia *omodina*, *involuppo*, *in fase e quadratura*, *eterodina*; ognuna di esse ha il suo campo di applicazione, ed i suoi pregi e difetti.

⁸Nel caso ad esempio di ampie zone di immagine uniformi ed a luminosità costante, il segnale è praticamente costante.

11.2.1 Demodulazione coerente o omodina

Si tratta del circuito già noto (vedi § 10.5) di estrazione della componente in fase $x_c(t)$ mediante moltiplicazione⁹ di $x(t)$ per una portante di demodulazione $\cos \omega_0 t$, e di rimozione delle componenti a frequenza $2f_0$ mediante un filtro passa-basso, come mostrato in figura. La portante generata localmente deve presentare la stessa fase e frequenza di quella in arrivo¹⁰, e per questo lo schema viene indicato anche con il nome di demodulazione *omodina*, *sincrona* o *coerente*. La portante di demodulazione viene allora ri-generata al ricevitore mediante un circuito PLL (§ 11.2.1.2) oppure un quadratore (§ 11.4.2). Il metodo è applicabile a tutti i tipi di modulazione di ampiezza, in quanto in tutti la componente in fase dipende da $m(t)$; nella pratica, nei casi di BLD-PI ed in quelli ad esso riconducibili, viene invece adottato il demodulatore di inviluppo (§ 11.2.3).



11.2.1.1 Errori di fase e frequenza

Esaminiamo ora cosa accade nel caso in cui tra la portante del segnale in arrivo e quella di demodulazione $\cos [2\pi (f_o + \Delta f) t + \theta]$ siano presenti errori di fase θ e/o di frequenza Δf : in tali condizioni il risultato della demodulazione coerente risulta¹¹:

$$\begin{aligned} y(t) &= x_c(t) \cos \omega_0 t \cos [(\omega_o + \Delta \omega) t + \theta] \\ &= \frac{1}{2} x_c(t) [\cos (\Delta \omega t + \theta) + \cos ((2\omega_o + \Delta \omega) t + \theta)] \end{aligned}$$

Pertanto, mentre il termine a frequenza (circa) doppia viene eliminato come al solito da un filtro apposito, per il segnale utile si riscontrano le seguenti distorsioni:

- per errori di fase: si ottiene $\frac{1}{2} x_c(t) \cos \theta \leq \frac{1}{2} x_c(t)$ e che... può annullarsi per $\theta = \pm \frac{\pi}{2}$!
- per errori di frequenza: si ottiene $\frac{1}{2} x_c(t) \cos (\Delta \omega t)$ e dunque il segnale demodulato, oltre ad invertire periodicamente polarità, presenta una notevole oscillazione di ampiezza che, ad esempio, nel caso di segnale audio può rendere il risultato inintelligibile già con Δf pari a pochi Hertz.

11.2.1.2 Phase Locked Loop - PLL

Approfondiamo il tema di come generare una portante di demodulazione *coerente* con quella del segnale ricevuto. Una possibile soluzione¹² adotta un approccio a *controreazione*, noto come *circuito ad aggancio di fase*, che basa il suo funzionamento su di un dispositivo chiamato *oscillatore controllato in tensione* (VCO, VOLTAGE CONTROLLED OSCILLATOR) il quale genera una sinusoide

$$y(t) = \sin \left(\omega_0 t + 2\pi k_f \int_{-\infty}^t x(\tau) d\tau \right)$$

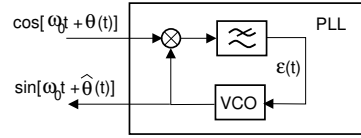
⁹In appendice 11.4.3 sono illustrate due tecniche di realizzazione del moltiplicatore.

¹⁰Dato che un qualunque canale presenta un ritardo di propagazione τ , la portante del segnale ricevuto sarà nella forma $\cos 2\pi f_0 (t - \tau) = \cos (2\pi f_0 t - 2\pi f_0 \tau) = \cos (2\pi f_0 t - \theta)$, ovvero sarà sempre presente una fase *incognita*. Nel caso poi di un collegamento radiomobile, sarà anche presente un errore di frequenza, dovuto all'*effetto doppler*.

¹¹Si applichi $\cos \alpha \cos \beta = \frac{1}{2} [\cos (\alpha + \beta) + \cos (\alpha - \beta)]$.

¹²Per un approccio alternativo, vedi § 11.4.2.

la cui fase varia nel tempo con l'integrale del segnale di ingresso¹³. Il circuito ad aggancio di fase opera secondo lo schema indicato in figura: supponiamo di disporre di un segnale AM in cui sia presente un residuo di portante, come ad esempio nel caso di *portante parzialmente soppressa*. Indicando con $\hat{\theta}(t) = 2\pi k_f \int_{-\infty}^t x(\tau) d\tau$ la stima locale della fase, già integrata dal VCO fino all'istante t , all'uscita del moltiplicatore è presente un segnale¹⁴:



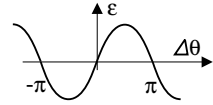
$$\frac{1}{2} \sin [2\omega_0 t + \theta(t) + \hat{\theta}(t)] + \frac{1}{2} \sin [\theta(t) - \hat{\theta}(t)]$$

Il termine centrato a frequenza doppia ($2\omega_0$) viene eliminato dal filtro passa basso, e dunque rimane

$$\varepsilon(t) = \frac{1}{2} \sin [\theta(t) - \hat{\theta}(t)] = \frac{1}{2} \sin(\Delta\theta(t))$$

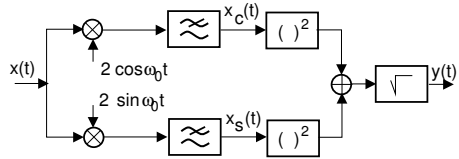
in cui $\Delta\theta(t)$ rappresenta l'errore di fase, e $\varepsilon(t)$ è la grandezza in ingresso al VCO.

Pensiamo ora di avere una $\theta(t)$ di ingresso costante: nel caso in cui risulti $\Delta\theta = 0$, si ottiene che anche $\varepsilon = 0$, ed il VCO non altera la fase (esatta) della portante generata. Se invece $\Delta\theta \geq 0$ (e $|\Delta\theta| < \pi$)¹⁵, allora $\varepsilon \geq 0$, e dunque il VCO è portato ad aumentare (diminuire) la fase della propria portante, riducendo di conseguenza l'errore di fase¹⁶. Nel caso in cui, infine, la fase $\theta(t)$ del segnale in arrivo vari nel tempo, allora il PLL *insegue* tali variazioni tanto più da vicino, quanto più è elevato il coefficiente di proporzionalità k_f tra $\hat{\theta}(t)$ e l'integrale di $\varepsilon(t)$ k_f ¹⁷.



11.2.2 Demodulazione incoerente in fase e quadratura

In questo caso la coerenza di fase tra portante ricevuta e di demodulazione viene deliberatamente *trascurata*, realizzando una architettura che comprende anche il ramo *in quadratura*, disegnato in basso nella figura. Dato che ruotando $\underline{x}(t)$ di $\pi/2$ le c.a. di b.f. si scambiano di ruolo,



¹³Se quest'ultimo ad esempio è costante ($x(t) = \Delta$), allora si avrà: $y(t) = \sin(2\pi f_0 t + 2\pi \Delta t) = \sin[2\pi(f_0 + \Delta)t]$, ovvero la frequenza si è alterata di una quantità pari a Δ . Il lettore più attento avrà riconosciuto che il VCO realizza il processo di *modulazione di frequenza*.

¹⁴Si utilizzi $\cos \alpha \sin \beta = \frac{1}{2} [\sin(\alpha + \beta) + \sin(\alpha - \beta)]$.

¹⁵La grandezza di controllo $\varepsilon(t)$ proporzionale a $\sin(\Delta\theta)$ si azzerava per $\Delta\theta = k\pi$ con k intero, positivo o negativo. Per k *dispari* si hanno condizioni di instabilità, in quanto ad es. per $\Delta\theta$ che *aumenta o diminuisce* rispetto a $\Delta\theta = \pi$, il segno di ε è rispettivamente *negativo e positivo*, causando un ulteriore ritardo o aumento di $\hat{\theta}(t)$ che causa un ulteriore aumento o diminuzione di $\Delta\theta$, finché questo non raggiunge il valore 0 o 2π , corrispondenti a condizioni di stabilità.

¹⁶Notiamo che un moltiplicatore, seguito da un filtro passabasso, esegue il calcolo dell'intercorrelazione tra gli ingressi del moltiplicatore (vedi §9.7.3), che nel nostro caso è una sinusoide.

¹⁷In particolare, le realizzazioni pratiche del PLL dipendono fortemente dalla banda e dall'ordine del filtro di loop, in quanto è quest'ultimo che limita la velocità di variazione di $\varepsilon(t)$ e l'estensione dell'intervallo di aggancio. Lo studio teorico delle prestazioni si basa sull'uso della trasformata di Laplace e sulla approssimazione $\sin(\Delta\theta) \simeq \Delta\theta$, in quanto così il PLL può essere studiato come un sistema di controllo *linearizzato*, sommariamente descritto al §11.3.1.1. Per approfondimenti, vedi http://it.wikipedia.org/wiki/Phase-locked_loop.

potremmo aggirare il problema illustrato al § 11.2.1.1 scegliendo quale dei due rami è meno attenuato; lo stesso schema può essere utile nella *fase di ricerca* di una regione di frequenza in cui è presente un segnale¹⁸, oppure ancora se si desidera solamente verificare la presenza o meno di un segnale ad una determinata frequenza, come nel caso del radar¹⁹. Infatti, nel caso in cui il segnale AM-BLD-PS ricevuto

$$x(t) = m(t) \cos(\omega_0 t + \theta)$$

presenti una fase θ incognita, l'involuppo complesso risulta

$$\underline{x}(t) = m(t) e^{j\theta} = m(t) \cos \theta + jm(t) \sin \theta$$

e quindi $x_c(t) = m(t) \cos \theta$ e $x_s(t) = m(t) \sin \theta$. Ma i due rami del demodulatore in fase e quadratura estraggono proprio $x_c(t)$ ed $x_s(t)$, e dunque il segnale $y(t)$ risulta pari a:

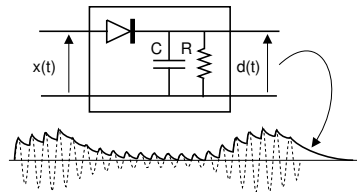
$$y(t) = \sqrt{x_c^2(t) + x_s^2(t)} = |m(t)| \sqrt{\cos^2 \theta + \sin^2 \theta} = |m(t)|$$

Pertanto, nonostante l'ignoranza della fase θ , siamo ancora in grado di individuare la *presenza* di un segnale modulante. L'operazione di modulo impedisce l'uso dello schema per demodulare generici segnali BLD-PS (mentre il caso PI sarebbe perfettamente demodulabile, ma in tal caso è più che sufficiente un demodulatore di involuppo (§ 11.2.3)). Al § 7.6.2 si discute di come usare il demodulatore incoerente al fine di decidere per la presenza o meno di una sinusoide immersa nel rumore, e della relativa probabilità di errore.

11.2.3 Demodulatore di involuppo per AM-BLD-PI

Si tratta del semplice circuito *non lineare* riportato in figura²⁰. Durante i periodi in cui il segnale in ingresso $x(t)$ è positivo rispetto alla tensione $d(t)$ accumulata dal condensatore, quest'ultimo si carica, inseguendo l'andamento dell'ingresso. Quando diviene $x(t) < d(t)$, il condensatore si scarica sulla resistenza con una costante di tempo $\tau = RC$, abbastanza grande rispetto ad $\frac{1}{f_0}$,

e tale da permettere la ricostruzione dell'andamento di $x_c(t)$. Le oscillazioni a frequenza f_0 (e sue armoniche) possono quindi essere rimosse da un successivo filtro passa-basso, mentre la costante a_p è rimossa mediante un passa alto. D'altra parte,



¹⁸La ricerca dell'emittente, che può essere banalmente l'azione di sintonizzare la propria radio sul programma preferito, può richiedere interventi automatici, qualora si tratti ad esempio di dover compensare le variazioni di frequenza dovute al movimento reciproco di trasmettitore e ricevitore (*effetto doppler*), come per il caso delle comunicazioni con mezzi mobili.

In questi casi, prima della comunicazione vera e propria, è necessario prevedere una fase di *acquisizione della portante*, svolta ad esempio mediante un circuito del tipo di quello che stiamo discutendo, in cui vengono provate diverse portanti di demodulazione, finché non si produce un segnale in uscita.

¹⁹Un radar trasmette ad elevata potenza per periodi molto brevi, e stima la presenza di oggetti basandosi sul ritardo con cui il segnale, riflesso da questi, torna indietro. Per questo, il ritardo di fase rappresenta proprio la grandezza che fornirà l'informazione relativa alla distanza, e può essere qualsiasi. Prima di iniziare a stimare tale informazione, è essenziale per il sistema accertarsi che *ci sia* un segnale da stimare.

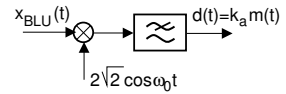
²⁰Il simbolo $\text{---}\triangleright\text{---}$ rappresenta un *diodo*, costituito da un bipolo di materiale semiconduttore drogato, che ha la particolarità di condurre in un solo verso (quello della freccia).

il valore di τ deve essere scelto ne troppo piccolo ne troppo grande, per evitare una eccessiva seghettatura, oppure di non riuscire ad inseguire il messaggio abbastanza rapidamente²¹.

La semplicità del circuito è tale da farlo usare nel maggior numero di casi possibili, anche se il suo uso prevalente è per la demodulazione di segnali a *portante intera*. D'altra parte, la presenza di altri segnali modulati, oltre a quello desiderato, rendono obbligatoria l'adozione di ulteriori elaborazioni, come discusso nel § 11.2.5 relativo alla demodulazione eterodina.

11.2.4 Demodulazione per segnali a banda laterale unica e *ridotta*

Nel caso di segnali *BLU*, il segnale modulante può essere riottenuto a partire da $x(t)$ utilizzando un demodulatore omodina, in quanto la componente in fase $x_c(t)$ dell'involuppo complesso è proprio pari al messaggio modulante $m(t)$. In questo caso, occorre prestare particolare attenzione ad eventuali errori di frequenza e di fase (Δf e θ) della portante di demodulazione perché, essendo presenti entrambe le componenti $x_c(t)$ ed $x_s(t)$, in uscita al demodulatore si ottiene (per il caso di banda laterale superiore):



$$d(t) = k_a m(t) \cos(\Delta\omega t + \theta) - k_a \hat{m}(t) \sin(\Delta\omega t + \theta)$$

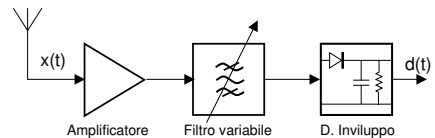
Pertanto si nota come la modulazione *BLU* sia più sensibile di quella *BLD* agli errori della portante di demodulazione, in quanto ora anche un semplice errore di fase θ produce non solo un affievolimento, ma una vera *intermodulazione* tra $m(t)$ e $\hat{m}(t)$.

Per i segnali *BLU* a portante intera è possibile anche usare un demodulatore di involuppo: in tal caso è necessario che l'ampiezza della portante sia sufficiente a non far invertire $x(t)$ neanche per i picchi di segnale più ampi. L'analisi di questa esigenza determina una efficienza inferiore a quella del caso *BLD*.

Anche nel caso *BLR*, è possibile ricorrere ad un demodulatore di tipo omodina, purché il filtro $H(f)$ usato in trasmissione per rimuovere parte di una banda laterale presenti alcune condizioni di simmetria attorno a f_0 ²².

11.2.5 Demodulatore eterodina

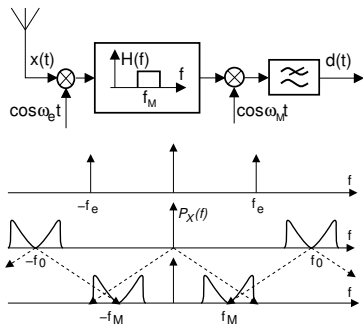
Questa tecnica utilizza una frequenza di demodulazione *differente* da quella della portante, ed è particolarmente idonea (ma non solo) alla ricezione di una tra diverse trasmissioni operanti in bande vicine tra loro, come nel caso della diffusione broadcast (vedi § 10.1.1.1). Infatti, nel caso in cui si volesse usare un demodulatore di involuppo (a patto che le emittenti trasmettano a portante intera) occorrerebbe *sintonizzare* l'emittente con un filtro passa banda *variabile* (rappresentato in figura da una freccia), la cui realizzazione a radio fre-



²¹Presso http://it.wikipedia.org/wiki/Rivelatore_d'involuppo qualche linea guida di progetto.

²²Si può dimostrare che per l'involuppo complesso $\underline{H}(f)$ di $H(f)$ deve risultare: $\underline{H}(f) + \underline{H}^*(-f) = \text{cost}$ perché in tal modo il residuo di banda parzialmente soppressa si combina esattamente con ciò *che manca* alla banda laterale *non* soppressa.

quenza può presentare difficoltà non trascurabili²³. D'altro canto, adottando un demodulatore *omodina* oltre al problema di realizzare il filtro passa-banda centrato sull'emittente desiderata, si introduce anche quello di mantenere coerenza tra la portante di demodulazione e quella ricevuta, essendo la realizzazione di oscillatori variabili e di precisione, via via più problematica all'aumentare delle frequenze in gioco.



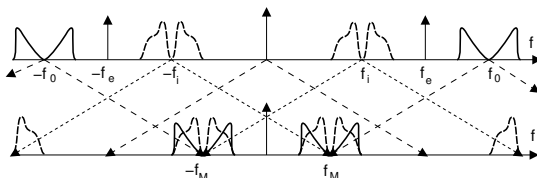
portante $f_0 = f_M + f_e$ risulta ora centrata in f_M (detta anche *media frequenza MF*), e può essere isolata dalle altre mediante un filtro *passa banda*, centrato proprio su f_M , e quindi demodulata da un demodulatore *omodina fisso*.

Per sintonizzare una diversa emittente a frequenza f'_0 , è sufficiente porre $f'_e = f'_0 - f_M$, mentre il resto del ricevitore (con i suoi amplificatori e filtri) opera su di un segnale centrato sempre alla stessa media frequenza f_M , indipendentemente dall'emittente.

11.2.5.1 Frequenze immagine

In realtà un ricevitore eterodina prevede la presenza di un ulteriore filtro posto *prima* del mixer con f_e , necessario ad evitare che in ingresso al filtro a media frequenza si presenti, oltre all'emittente centrata a $\pm f_0 = \pm(f_e + f_M)$, anche quella a portante $\pm f_i = \pm(f_e - f_M)$, per la quale cioè $f_e - f_i = f_M$.

La frequenza f_i prende il nome di *frequenza immagine*, in quanto è l'immagine speculare di f_0 rispetto ad f_e ; in altre parole, l'utilizzo di una portante eterodina f_e provoca la traslazione a media frequenza sia della emittente desiderata e centrata in $f_0 = f_e + f_M$, che della sua immagine a distanza $2f_M$, centrata in $f_i = f_e - f_M$. Pertanto, in ingresso al ricevitore va anteposto un filtro, che elimini dal segnale di ingresso le frequenze immagine, ovvero, una volta nota la gamma di frequenze che si vuole sintonizzare, elimini tutte quelle a distanza $2f_M$ dalla banda di interesse.



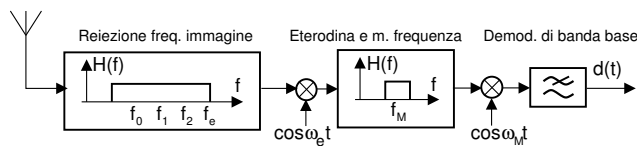
²³Le difficoltà nascono sia dall'esigenza di *accordare* il filtro attorno alla frequenza portante desiderata, sia dalla necessità di attenuare sufficientemente le trasmissioni che avvengono su frequenze limitrofe, determinando la necessità di realizzare un filtro con regione di transizione molto ripida, problema che può divenire insormontabile se il rapporto tra banda del segnale e portante (la cosiddetta *banda frazionaria*) è particolarmente ridotto.

Esempio La maggior parte dei ricevitori di trasmissioni AM Broadcast (540-1600 KHz) utilizza un ricevitore detto *supereterodina*, con $f_e > f_0$ anziché $f_e < f_0$ come discusso sopra²⁴, e $f_M = 455$ KHz. Volendo ad esempio sintonizzare una stazione con $f_p = 600$ KHz, occorre una $f_e = f_M + f_p = 1055$ KHz. Ma allo stesso tempo anche l'emittente a portante $f_i = f_e + f_M = 1510$ KHz viene traslata nella stessa banda del filtro MF. Dunque, prima del mixer occorre un filtro che elimini le stazioni centrate su portanti $f_i > f_e$.

La scelta $f_M = 455$ KHz, inferiore alla minima frequenza di 510 KHz, permette di utilizzare una regione di frequenze libera da altre trasmissioni, che altrimenti potrebbero essere amplificate dagli stadi ad alto guadagno posti dopo il filtro MF. La scelta di $f_e > f_0$ permette di posizionare il *filtro di reiezione* delle frequenze immagine al disotto della f_e , rendendo più semplice la realizzazione del filtro.

11.2.5.2 Supereterodina

La figura seguente mostra lo schema finale del ricevitore con $f_e > f_0$.



Elenchiamo di seguito i vantaggi conseguiti:

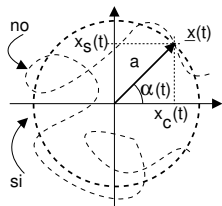
- la sintonia avviene mediante la variazione di f_e , ed il resto non cambia;
- la distanza tra f_0 ed f_M scongiura il rischio di instabilità, che potrebbe verificarsi se parte del segnale uscente dal filtro di media frequenza, amplificato, fosse ricaptato dallo stadio di ingresso, mentre ora invece l'amplificazione può aver luogo proprio nello stadio a media frequenza;
- la scelta di f_M è una opportunità di progetto, che consente di realizzare il filtraggio della emittente desiderata a frequenza sufficientemente bassa da non porre grossi problemi;
- lo stadio di eterodina può essere ulteriormente ripartito in due conversioni di frequenza successive (vedi ad es. la fig. 15.9 a pag. 415), di cui la seconda conversione opera la sintonia, mentre la prima ha il solo scopo di traslare le frequenze in gioco in un intervallo più basso, più idoneo ad esempio alla sua trasmissione via cavo.

11.3 Modulazione angolare

In questo tipo di modulazione l'informazione contenuta nel messaggio $x(t)$ è impressa sulla portante modificandone la fase: $x(t) = a \cos(\omega_0 t + \alpha(t))$, ed il suo involuppo complesso (vedi eq. (10.2)) risulta

$$\underline{x}(t) = ae^{j\alpha(t)} = a [\cos \alpha(t) + j \sin \alpha(t)] = x_c(t) + jx_s(t)$$

²⁴ Scegliere $f_e < f_0$ oppure $f_e > f_0$ praticamente ha l'effetto di scambiare i ruoli tra la portante desiderata, e la sua immagine. In particolare, il risultato non cambia se la modulazione è a banda laterale doppia, ma dato che nel secondo caso, $X^+(f)$ si rialloca sull'asse delle frequenze negative (ed il contrario per $X^-(f)$), per la modulazione BLU o BLR questo fenomeno di inversione delle bande laterali deve essere tenuto in conto nel demodulatore omodina finale.



Notiamo subito che, a differenza della AM, il modulo di $\underline{x}(t)$ è rigorosamente *costante* e pari ad a , mentre la fase $\alpha(t)$ varia continuamente. Si è già mostrato al § 10.2.2 come sia possibile definire 2 diversi tipi di legame tra messaggio $m(t)$ e fase dell'involuppo complesso $\alpha(t)$, indicati con PM (modulazione di fase) ed FM (modulazione di frequenza), riassunti nella tabella seguente:

	$\alpha(t)$	$f_i(t)$
PM	$k_\phi m(t)$	$f_0 + \frac{k_\phi}{2\pi} \frac{d}{dt} m(t)$
FM	$2\pi k_f \int_{-\infty}^t m(\tau) d\tau$	$f_0 + k_f m(t)$

In particolare, i due tipi di modulazione sono esprimibili anche nei termini della *frequenza istantanea*, definita come la derivata della *fase istantanea* $\psi(t) = 2\pi f_0 t + \alpha(t)$:

$$f_i(t) = \frac{1}{2\pi} \frac{d}{dt} \psi(t) = f_0 + \frac{1}{2\pi} \frac{d}{dt} \alpha(t)$$

Le due alternative (PM e FM) sono analizzate assieme, in quanto reciprocamente intercambiabili qualora si effettu

- una PM con $m(t)$ pari all'integrale del messaggio informativo *oppure*
- una FM con $m(t)$ pari alla derivata del messaggio informativo.

Illustriamo subito alcune particolarità della modulazione angolare, prima di applicarci al problema della ricezione, ed alla determinazione della densità di potenza del segnale modulato.

Non linearità La caratteristica *fondamentale* della modulazione angolare è che il segnale modulato dipende da $m(t)$ in modo fortemente *non lineare*, e pertanto lo spettro di densità di potenza $\mathcal{P}_x(f)$ del segnale modulato non può essere calcolato con le tecniche tradizionali. Infatti, l'involuppo complesso di un segnale modulato angolarmente può essere espresso²⁵ come:

$$\underline{x}(t) = a e^{j\alpha(t)} = a \left[1 + j\alpha(t) - \frac{\alpha^2(t)}{2} - j\frac{\alpha^3(t)}{3!} + \dots \right] \quad (11.3)$$

da cui risulta evidente che, anche se $\mathcal{P}_\alpha(f)$ è esprimibile a partire da $\mathcal{P}_m(f)$, nulla può essere detto in generale per $\mathcal{P}_x(f)$ (e dunque per $\mathcal{P}_x(f) = \frac{1}{4}\mathcal{P}_x(f-f_0) + \frac{1}{4}\mathcal{P}_x(f+f_0)$). Infatti, la presenza delle potenze della fase modulante $\alpha(t)$ impedisce l'applicabilità del principio di sovrapposizione degli effetti, ovvero, anche se sono noti i risultati della modulazione per due diversi messaggi $x_1(t) = FM\{m_1(t)\}$ e $x_2(t) = FM\{m_2(t)\}$, il risultato ottenibile modulando la loro somma, non è quello della somma dei risultati individuali: $FM\{m_1(t) + m_2(t)\} \neq FM\{m_1(t)\} + FM\{m_2(t)\}$.

Ampiezza costante La circostanza che $\underline{x}(t) = a e^{j\alpha(t)}$ presenti un modulo costante pari ad a , indipendentemente dalle ampiezze del segnale modulante, è particolarmente utile qualora per $m(t)$ siano da aspettarsi forti variazioni di dinamica, come ad

²⁵Si fa qui uso della espansione in serie di potenze dell'esponenziale: $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots$

es. nel caso del segnale FDM (pag. 251) utilizzato per trasmettere più canali telefonici²⁶. Infatti in questo caso, non essendo noto a priori il numero di canali effettivamente impegnati, la potenza del segnale $y(t) = \sum_{n=1}^N BLU \{m_n(t), f_n\}$ ottenuto sommando i diversi canali (ognuno a modulazione BLU su di una diversa portante) può variare di molto: allora, il segnale complessivo $y(t)$ viene applicato all'ingresso di un modulatore FM e trasmesso come tale.

Generazione di un segnale a modulazione angolare Come anticipato, per effettuare una modulazione PM $x(t) = k_\phi m(t)$ si può usare un modulatore FM, in cui $\alpha(t) = 2\pi k_f \int_{-\infty}^t m'(\tau) d\tau$, ponendo $m'(t) = \frac{1}{2\pi} \frac{k_\phi}{k_f} \frac{d}{dt} m(t)$. Pertanto, consideriamo nel seguito solo le operazioni di modulazione/demodulazione FM. Un metodo *diretto* di generare un segnale FM è quello di utilizzare un VCO (introdotto al § 11.2.1.2), ossia un oscillatore controllato in tensione, che produce il segnale $x(t) = a \sin \left(\omega_0 t + 2\pi k_f \int_{-\infty}^t m(\tau) d\tau \right)$, e dunque realizza proprio la funzione desiderata. Un secondo metodo verrà illustrato per un caso particolare in appendice § 11.4.5. Infine, è sempre valido il modulatore in fase e quadratura, in cui si pone $x_c(t) = \cos \alpha(t)$ e $x_s(t) = \sin \alpha(t)$.

11.3.1 Ricezione di un segnale a modulazione angolare

In linea di principio, una volta ottenute $x_c(t)$ ed $x_s(t)$ del segnale modulato (ad esempio mediante un demodulatore *coerente* in fase e quadratura) è sempre valida la relazione $\alpha(t) = \arctan \frac{x_s(t)}{x_c(t)}$. D'altra parte, tale soluzione si presta esclusivamente a realizzazioni digitali, in quanto è difficile realizzare un dispositivo che presenti esattamente la relazione non lineare di tipo arcotangente. Illustriamo allora i due metodi più comunemente usati:

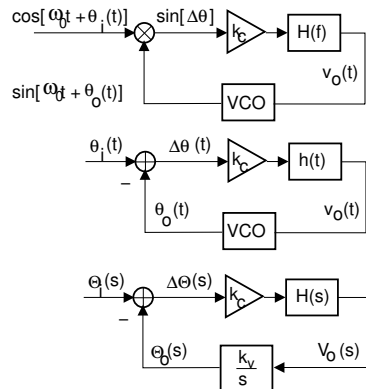
11.3.1.1 Ricevitore a PLL

Al § 11.2.1.2 si è già mostrato l'uso del circuito PLL per l'aggancio della fase della portante di modulazione. Lo stesso schema può essere usato per inseguire l'andamento temporale della fase di una portante modulata angolarmente, ottenendo in tal modo l'informazione desiderata.

La figura a lato riporta lo schema generale di un PLL, in cui il VCO genera un segnale pari a $\sin(\omega_0 t + \theta_o(t))$, con $\theta_o(t) = k_v \int_{-\infty}^t v_o(\tau) d\tau$, mentre il segnale ricevuto ha la forma $x(t) = \cos(\omega_0 t + \theta_i(t))$. Lo schema può essere analizzato con i metodi dei controlli automatici, in quanto rappresenta un sistema che tenta di mantenere nullo l'errore $\sin \Delta\theta$, con $\Delta\theta(t) = \theta_i(t) - \theta_o(t)$; l'analisi si basa quindi sulla linearizzazione $\sin \Delta\theta \simeq \Delta\theta$, valida per $\Delta\theta$ piccolo.

L'analisi di Laplace fornisce allora il risultato

$$\Theta_o(s) = \frac{k_c k_v H(s)}{s + k_c k_v H(s)} \Theta_i(s)$$



²⁶Un altro caso di multiplex FDM è quello del downlink di un trasponder DVB-S, introdotto al § 15.5.1

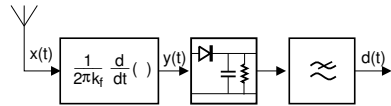
che, antitrasformato, permette di esprimere $\theta_o(t)$ (fase del VCO) come una versione filtrata della fase della portante modulata $\theta_i(t)$, da parte della funzione di trasferimento ad anello chiuso

$$H(f) = \frac{k_c k_v H(s)}{s + k_c k_v H(s)} \Big|_{s=j2\pi f}$$

Inoltre, dato che il VCO produce $\theta_o(t) = k_v \int_{-\infty}^t v_o(\tau) d\tau$, si riconosce subito che l'uscita $v_o(t)$ del filtro di loop $H(s)$ corrisponde alla ricostruzione del messaggio modulante $m(t)$ nel caso di modulazione FM. Pertanto, l'uscita del filtro di loop del PLL realizza la demodulazione di frequenza.

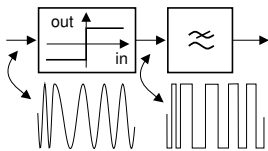
11.3.1.2 Ricevitore a discriminatore

E' realizzato mediante il circuito di figura, in cui il derivatore effettua una conversione FM-AM, e quindi viene demodulato l'involuppo di ampiezza di $y(t)$. Infatti quest'ultimo segnale risulta pari a



$$\begin{aligned} y(t) &= \frac{1}{2\pi k_f} \frac{d}{dt} a \cos \left(\omega_0 t + 2\pi k_f \int_{-\infty}^t m(\tau) d\tau \right) = \\ &= \frac{1}{2\pi k_f} (2\pi f_0 + 2\pi k_f m(t)) a \sin \left(\omega_0 t + 2\pi k_f \int_{-\infty}^t m(\tau) d\tau \right) \end{aligned}$$

e corrisponde dunque ad un segnale modulato sia angolarmente che in ampiezza, in cui $a(t) = a \left(\frac{f_0}{k_f} + m(t) \right)$. Siamo dunque in presenza di una modulazione di ampiezza BLD-PI (§ 11.1.1.2) e quindi, con una scelta opportuna²⁷ di $\frac{f_0}{k_f}$, il messaggio $m(t)$ può essere estratto mediante un demodulatore d'involuppo (§ 11.2.3).



Il risultato ottenuto è valido purché $x(t)$ sia privo esso stesso di variazioni di ampiezza: per questo, spesso il derivatore è preceduto da un blocco *squadratore*, che produce una versione, appunto, "squadrata" del segnale ricevuto e quindi priva di modulazioni di ampiezza. Essendo lo squadratore fortemente non lineare, in uscita saranno presenti, oltre al segnale originario, anche componenti centrate a frequenze multiple di quella della portante, che vengono rimosse dal filtro passa basso a valle dello squadratore.

11.3.2 Densità spettrale di segnali a modulazione angolare

Riprendiamo la relazione (11.3) che esprime l'involuppo complesso di un segnale modulato angolarmente nei termini di una serie di potenze:

$$\underline{x}(t) = a e^{j\alpha(t)} = a \sum_{n=0}^{\infty} \frac{[j\alpha(t)]^n}{n!} \quad (11.4)$$

²⁷ Per utilizzare il demodulatore involuppo, deve risultare sempre $\frac{f_0}{k_f} + m(t) > 0$, e dunque occorre scegliere $\frac{f_0}{k_f} > \max_t \{|m(t)|\}$.

Innanzitutto osserviamo che la potenza totale di $\underline{x}(t)$ vale sempre $\mathcal{P}_{\underline{x}} = a^2$, indipendentemente da $\alpha(t)$, e dunque $\mathcal{P}_x = \frac{a^2}{2}$. Per ciò che riguarda $\mathcal{P}_{\underline{x}}(f)$, in linea di principio non si potrebbe neanche affermare che $\underline{x}(t)$ sia limitato in banda, vista la presenza delle potenze di qualunque ordine di $\alpha(t)$. D'altro canto, la presenza dei fattoriali a denominatore fa sì che la serie possa essere troncata ad un certo ordine $\nu < \infty$. Se poniamo ora $\alpha(t) = k_\phi m(t)$, osserviamo che quanto più $|k_\phi m(t)|$ è piccolo, tanto prima può essere troncata, con errori trascurabili. In particolare, se $\alpha(t)$ si mantiene sempre *molto piccolo*, la (11.4) può essere troncata al primo termine ($n = 1$), dando luogo ad un comportamento praticamente lineare.

Se invece $\alpha(t)$ assume valori *molto elevati*, e quindi (11.4) comprende parecchi termini, subentra un secondo aspetto peculiare dell'FM, indicato come *conversione ampiezza \rightarrow frequenza*, che può essere descritto tenendo conto che in base alla relazione $f_i(t) = f_0 + k_f m(t)$, la frequenza istantanea presenta scostamenti rispetto ad f_0 completamente dipendenti dalle ampiezze di $m(t)$, e quindi l'andamento della densità di potenza $\mathcal{P}_x(f)$ risulta strettamente dipendente da quello della densità di probabilità di $p_M(m)$ che descrive le ampiezze di $m(t)$.

Per valori intermedi della dinamica di $\alpha(t)$, invece, la $\mathcal{P}_x(f)$ risultante sarà una via di mezzo tra i due casi estremi discussi, che pertanto possono essere pensati come casi limite tra cui porre la densità di potenza effettiva.

Come anticipato, la natura non lineare della modulazione angolare rende necessario studiare ogni caso individualmente; pertanto la determinazione di $\mathcal{P}_x(f)$ viene svolta per due casi particolari, considerando per questi le due possibilità estreme di $\alpha(t)$ molto piccolo o molto grande, ed i risultati estrapolati per approssimare altre situazioni; i due casi esaminati sono:

- $m(t)$ sinusoidale e
- $m(t)$ membro di un processo stazionario ergodico.

11.3.2.1 Segnale modulante sinusoidale

Ponendo $m(t) = \cos(2\pi wt)$, si ottiene che la fase modulante $\alpha(t)$ e la frequenza istantanea $f_i(t)$ per i due casi PM ed FM, relativi al segnale $x(t) = a \cos(2\pi f_0 t + \alpha(t))$, risultano:

	$\alpha(t)$	$f_i(t)$	$\Delta\alpha$	Δf
PM	$k_\phi \cos(2\pi wt)$	$f_0 + wk_\phi \sin(2\pi wt)$	k_ϕ	wk_ϕ
FM	$2\pi k_f \int_{-\infty}^t m(\tau) d\tau = \frac{k_f}{w} \sin(2\pi wt)$	$f_0 + k_f \cos(2\pi wt)$	$\frac{k_f}{w}$	k_f

in cui si è anche indicata la massima deviazione di fase $\Delta\alpha = \max\{|\alpha(t)|\}$ e di frequenza $\Delta f = \max\{|f_i(t) - f_0|\}$. Notiamo subito che, in entrambi i casi, sia la fase $\alpha(t)$ che la frequenza istantanea $f_i(t)$ variano sinusoidalmente con periodo $\frac{1}{w}$; nel caso PM l'entità di Δf aumenta con w , mentre nell'FM la $\Delta\alpha$ diminuisce con w . Nel seguito si farà riferimento all'*indice di modulazione angolare* β , corrispondente alla massima escursione della fase $\Delta\alpha$, che risulta:

$$\beta = \begin{cases} k_\phi & \text{(PM)} \\ \frac{k_f}{w} & \text{(FM)} \end{cases}$$

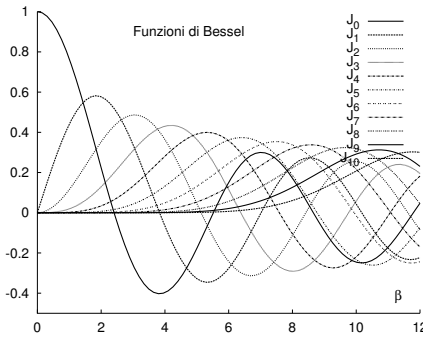
Con questa convenzione, possiamo trattare congiuntamente entrambi i casi PM ed FM riscrivendo l'involuppo complesso come ²⁸

$$\underline{x}(t) = ae^{j\beta \sin(2\pi wt)}$$

Notiamo ora che $\underline{x}(t)$ è periodico di periodo $\frac{1}{w}$, e dunque per esso vale lo sviluppo in serie di Fourier $\underline{x}(t) = a \sum_{n=-\infty}^{\infty} X_n e^{j2\pi nwt}$, i cui coefficienti risultano

$$X_n = w \int_{-\frac{1}{2w}}^{\frac{1}{2w}} e^{j\beta \sin(2\pi wt)} e^{-j2\pi nwt} dt = \mathcal{J}_n(\beta)$$

ovvero sono pari²⁹ alle *funzioni di Bessel* del primo tipo, ordine n ed argomento β . Queste hanno l'andamento mostrato in figura, a fianco della quale sono riportate le proprietà che le caratterizzano:



- $\mathcal{J}_n(\beta)$ è reale con $\forall n, \beta$
- $\begin{cases} \mathcal{J}_n(\beta) = \mathcal{J}_{-n}(\beta) & n \text{ pari} \\ \mathcal{J}_n(\beta) = -\mathcal{J}_{-n}(\beta) & n \text{ dispari} \end{cases}$
- $\sum_{n=-\infty}^{+\infty} \mathcal{J}_n^2(\beta) = 1$
- $\mathcal{J}_n(\beta) \simeq 0$ con $n > \beta$ se $\beta \gg 1$

e quindi i valori di X_n si ottengono tracciando una linea verticale nel diagramma di figura in corrispondenza del valore β adottato, e individuando il valore di ciascuna \mathcal{J}_n per quel β . Osserviamo ora che l'ultima proprietà mostra come, in presenza di un valore di β elevato, le funzioni di Bessel di ordine $n > \beta$ siano praticamente nulle: è quindi lecito in tal caso limitare lo sviluppo in serie di Fourier di $\underline{x}(t)$ ai primi β termini (positivi e negativi), ovvero: $\underline{x}_{FM}(t) \simeq a \sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta) e^{j2\pi nwt}$. Pertanto, il segnale modulato $x(t) = \Re\{\underline{x}(t)e^{j\omega_0 t}\}$ risulta:

$$x(t) \simeq a \sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta) \cos 2\pi(f_0 + nw)t$$

e quindi lo spettro di densità di potenza di $x(t)$ ha espressione

$$\mathcal{P}_x(f) \simeq \frac{a^2}{4} \sum_{n=-\beta}^{\beta} |\mathcal{J}_n(\beta)|^2 [\delta(f - f_0 - nw) + \delta(f + f_0 + nw)]$$

ed è formato da impulsi centrati a frequenze $f = \pm f_0 \pm nw$ (³⁰).

²⁸Si è sostituito *cos* con *sin* nel caso PM per omogeneità di formulazione, senza alterare la sostanza delle cose.

²⁹Le *funzioni di Bessel* del primo tipo, ordine n ed argomento β sono definite come $\mathcal{J}_n(\beta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j(\beta \sin x - nx)} dx$.

³⁰Integrando l'espressione di $\mathcal{P}_x(f)$, e ricordando che $\sum_{n=-\infty}^{+\infty} \mathcal{J}_n^2(\beta) = 1$, si ottiene ancora un risultato già noto, e cioè che la potenza totale del segnale modulato risulta pari a quella della portante non modulata, e pari a $\mathcal{P}_x = \frac{a^2}{2}$, indipendentemente da β .

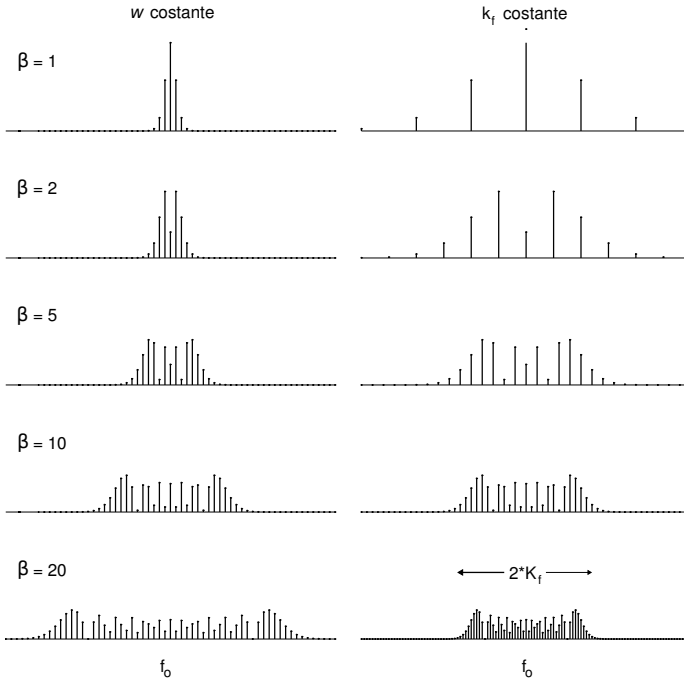


Figura 11.1: Spettro di ampiezza per segnale FM a modulazione sinusoidale

La fig. 11.1 mette a confronto $|X(f)|$ per f vicino ad f_0 , ovvero mostra $|X^+(f)| = \frac{\alpha}{2} \sum_{n=-\beta}^{\beta} |\mathcal{J}_n(\beta)| \delta(f - f_0 - nw)$, calcolato per diversi valori di β , mantenendo fisso w oppure k_f (a sinistra e destra rispettivamente), e ci aiuta a comprendere i ragionamenti che seguono.

Modulazione a basso indice In questo caso $\beta \ll 1$, e tale da rendere trascurabili le funzioni di Bessel $\mathcal{J}_n(\beta)$ con $n > 1$. Allora, $x(t)$ occupa una banda pari a $2w$, in modo del tutto simile all'AM-BLD.

Modulazione ad alto indice In tal caso sono presenti più funzioni di Bessel, ed il comportamento non lineare tende a legare $\mathcal{P}_x(f)$ ai valori assunti da $f_i(t) = k_f m(t)$, realizzando una conversione ampiezza \rightarrow frequenza.

In Fig. 11.1 sono evidenziati gli effetti dell'aumento di $\beta = \frac{k_f}{w}$ nelle due circostanze:

1. Si mantiene w fisso, aumentando k_f . Il numero di righe spettrali a frequenza $f_0 \pm nw$ aumenta, occupando una banda crescente, e per β molto grande si verifica che $\mathcal{J}_n(\beta) \simeq 0$ per $n > \beta$. Pertanto, la banda occupata tende a $B = 2\beta w = 2 \frac{k_f}{w} \cdot w = 2k_f$;
2. Si mantiene k_f fisso, diminuendo w . La banda occupata tende a ridursi, mentre le nuove righe spettrali a frequenza $f_0 \pm nw$ si infittiscono. Per $\beta \rightarrow \infty$, la spaziatura w tra le righe spettrali tende ad annullarsi, producendo una $\mathcal{P}_x(f)$ praticamente

continua, e con una banda $B = 2k_f$, ossia pari alla massima deviazione di frequenza istantanea Δf .

Notiamo che in entrambi i casi, all'aumentare di β la $\mathcal{P}_x(f)$ tende ad assumere la densità di ampiezza tipica del processo armonico, descritta dalla (7.4) a pag. 160.

Regola di Carson Come mostrato, nei due casi a basso ed alto indice la banda occupata da $x(t)$ varia tra $2w$ e $2k_f$ rispettivamente. Nei casi intermedi, è pratica comune ricorrere all'espressione

$$B_C \simeq 2(k_f + w) = 2w(\beta + 1) \quad (11.5)$$

nota come *regola di Carson*³¹, che tiene conto di entrambi i fattori che concorrono alla determinazione della banda, e che fornisce i valori esatti sia per $\beta \ll 1$, che per $\beta \rightarrow \infty$, in entrambi i casi in cui $k_f \rightarrow \infty$ o $w \rightarrow 0$.

Sebbene la determinazione approssimata della banda mediante la *regola di Carson* sia stata ottenuta nel caso di $m(t) = \cos(2\pi wt)$, la stessa espressione è spesso adottata come una buona approssimazione anche per segnali non sinusoidali, ma limitati in banda tra $-W$ e W , e contraddistinti da una $\Delta f = k_f \cdot \max\{|m(t)|\}$. In tal caso, la regola di Carson si applica ponendo ora $B_C \simeq 2W(\beta + 1)$ con $\beta = \frac{\Delta f}{W}$. Per un approfondimento della questione, si veda l'appendice 11.4.6 e la sottosezione 11.3.3.1.

A prima vista, l'estensione del risultato per $m(t) = \cos(2\pi wt)$ al caso qualunque appare più che ragionevole; il comportamento non lineare della modulazione angolare impedisce però una sua verifica analitica. D'altra parte, i risultati sperimentali mostrano che l'approssimazione fornita dalla banda di Carson può effettivamente costituire una stima plausibile della banda occupata per segnali modulanti qualsiasi.

In base alla regola di Carson, notiamo ora che la banda occupata dal segnale modulato può risultare $\beta + 1$ volte più estesa di quella ottenibile mediante modulazione AM. Nonostante questo aumento di banda possa apparire un fatto negativo, vedremo nel capitolo 12 che ciò può produrre un SNR dopo la demodulazione migliore di quanto ottenibile nel caso AM. Al contrario, se $\beta \ll 1$, il comportamento si avvicina molto a quello lineare (vedi appendice 11.4.5).

11.3.3 Densità spettrale FM con processo aleatorio modulante

Riprendiamo il ragionamento iniziato in 11.3.2, relativo all'influenza di $p_M(m)$ su $\mathcal{P}_x(f)$. Considerando che la frequenza istantanea ha espressione $f_i = f_0 + k_f m(t)$, la frazione di potenza tra f_1 ed f_2 sarà pari alla frazione di tempo che il segnale $m(t)$ si trova tra $m_1 = \frac{f_1 - f_0}{k_f} \leq m(t) \leq m_2 = \frac{f_2 - f_0}{k_f}$. Nel caso in cui $m(t)$ sia sinusoidale, con fase iniziale aleatoria a distribuzione uniforme, $m(t)$ è una realizzazione di un processo armonico (pag. 160), e la frazione di tempo su indicata equivale alla $\text{Prob}\{m_1 \leq m(t) \leq m_2\}$. Pertanto le righe spettrali, addensandosi, tendono a disporsi in accordo all'andamento della densità $p_M(m)$ ³².

Il risultato a cui siamo pervenuti nel caso di modulante sinusoidale è generale, e pertanto si può affermare che qualora si generi un segnale FM ad alto indice, a

³¹J. R. Carson fu uno dei primi ad investigare la modulazione negli anni '20, vedi ad es. http://en.wikipedia.org/wiki/John_Renshaw_Carson

³²In particolare, per $\beta \rightarrow \infty$ risulterà $\mathcal{P}_x(f) = \frac{a^2}{1 - (f/k_f)^2}$, che è l'andamento a cui tendono (per $\beta \rightarrow \infty$) i grafici in basso di fig. 11.1.

partire da un processo con densità di probabilità nota, lo spettro di densità di potenza del segnale modulato acquisisce l'andamento proprio della densità di probabilità del processo modulante, indipendentemente dal suo spettro di densità di potenza.

La conclusione riportata si mantiene valida purché $\beta \gg 1$; nel caso contrario, sono validi i ragionamenti sviluppati alla sezione 11.3.3.2.

Esempio un processo uniforme $m(t)$ limitato in banda $\pm W$, con densità di probabilità $p_M(m) = \frac{1}{\Delta_M} \text{rect}_{\Delta_M}(m)$, modula ad alto indice la frequenza di una portante, con frequenza f_0 ed ampiezza a , con un coefficiente di modulazione k_f . Determinare $\mathcal{P}_x(f)$ del segnale modulato.

Notiamo subito che la frequenza istantanea f_i rimane limitata tra $f_0 - \frac{\Delta_M}{2} k_f$ e $f_0 + \frac{\Delta_M}{2} k_f$.

Inoltre, la potenza totale deve risultare ancora pari a $\frac{a^2}{2}$. Pertanto si ottiene³³:

$$\mathcal{P}_x(f) = \frac{a^2}{4\Delta_M k_f} \left[\text{rect}_{\Delta_M k_f}(f - f_0) + \text{rect}_{\Delta_M k_f}(f + f_0) \right]$$

11.3.3.1 Indice di modulazione per processi

Ai fini dell'applicazione della regola di Carson, si è posto l'indice di modulazione $\beta = \frac{\Delta f}{W}$, con $\Delta f = k_f \max\{|m(t)|\}$. Nel caso di processi, può accadere che $m(t)$ non sia limitata in ampiezza, come ad esempio nel caso gaussiano, rendendo problematica la quantificazione di β . Per risolvere la questione, l'indice di modulazione β è ridefinito ancora una volta, e nel caso in cui $m(t)$ sia un generico processo si pone

$$\beta' = \begin{cases} \frac{\sigma_\alpha}{W} & \text{(PM)} \\ \frac{\sigma_f}{W} & \text{(FM)} \end{cases}$$

in cui W è la banda a frequenze positive del segnale modulante, $\sigma_f = k_f \sqrt{\mathcal{P}_m}$ rappresenta la deviazione standard della frequenza istantanea³⁴, e $\sigma_\alpha = k_\phi \sqrt{\mathcal{P}_m}$ è la deviazione standard della fase modulante³⁵. L'applicazione della regola di Carson con il nuovo valore di β' , fornisce per la banda un risultato che non indica più la banda *totale* occupata, ma individua una *banda efficace* entro cui $\mathcal{P}_x(f)$ è in larga parte (ma non completamente) contenuta (vedi anche 11.4.6).

Nel caso in cui *non* risulti $\beta \gg 1$, lo spettro di potenza del segnale modulato FM torna a dipendere da quello del segnale modulante, e si ricade nella trattazione che segue.

11.3.3.2 Modulazione a basso indice

In questo caso si suppone l'indice di modulazione β piccolo a sufficienza, in modo che lo sviluppo in serie dell'involuppo complesso del segnale modulato possa essere arrestato ai primi termini.

³³Volendo applicare la regola di Carson per calcolare la banda, si avrebbe (considerando $\beta \gg 1$) $B_C = 2W(\beta + 1) \simeq 2 \frac{\Delta f}{W} W = 2\Delta f$, in cui $\Delta f = k_f \frac{\Delta_M}{2}$. Pertanto risulta $B_C = 2k_f \frac{\Delta_M}{2} = k_f \Delta_M$, in accordo al risultato previsto nel caso di modulazione ad alto indice.

Qualora si fosse invece posto $\beta = \frac{\sigma_f}{W}$ (vedi 11.3.3.1) si sarebbe ottenuto $B_C = 2W(\beta + 1) \simeq 2 \frac{\sigma_f}{W} W = 2\sigma_f = 2k_f \sqrt{\mathcal{P}_m} = 2k_f \sqrt{\frac{\Delta_M^2}{12}} = 2k_f \frac{\Delta_M}{2\sqrt{3}} = \frac{\Delta_M k_f}{\sqrt{3}}$, un risultato che è circa pari a 0.58 volte quello precedente. Data la particolarità di $p_M(m)$ uniforme, in questo caso è da preferire il primo risultato.

³⁴Infatti, dalla definizione $f_i(t) = f_0 + k_f m(t)$ si ottiene che $\sigma_f^2 = k_f^2 \sigma_M^2$, in cui $\sigma_M^2 = \mathcal{P}_M$ se $m(t)$ è un processo stazionario ergodico a media nulla.

³⁵Come sopra, partendo dalla relazione $\alpha(t) = k_\phi m(t)$.

Sotto opportune ipotesi, si può mostrare che vale il risultato

$$\mathcal{P}_{\underline{x}}(f) \simeq a^2 e^{-\sigma_\alpha^2} \left[\delta(f) + \mathcal{P}_\alpha(f) + \frac{1}{2} \mathcal{P}_\alpha(f) * \mathcal{P}_\alpha(f) + \frac{1}{3!} \mathcal{P}_\alpha(f) * \mathcal{P}_\alpha(f) * \mathcal{P}_\alpha(f) + \dots \right]$$

avendo indicando con σ_α^2 la varianza della fase modulata e con $\mathcal{P}_\alpha(f)$ il relativo spettro di densità di potenza, pari rispettivamente a

	$\mathcal{P}_\alpha(f)$	σ_α^2
PM	$k_\phi^2 \mathcal{P}_m(f)$	$k_\phi^2 \mathcal{P}_m$
FM	$k_f^2 \frac{\mathcal{P}_m(f)}{f^2}$	$k_f^2 \int_{-w}^w \frac{\mathcal{P}_m(f)}{f^2} df$

Osserviamo che se k_ϕ (o k_f) tende a zero, $\mathcal{P}_{\underline{x}}(f)$ si riduce ad un impulso, corrispondente alla portante non modulata. All'aumentare di k_ϕ (o k_f), aumenta anche σ_α^2 e dunque il termine $e^{-\sigma_\alpha^2}$ diminuisce, riducendo la concentrazione di potenza a frequenza portante. Dato che risulta comunque $\mathcal{P}_{\underline{x}} = a^2$, la potenza residua si distribuisce sugli altri termini, rappresentati da $\mathcal{P}_\alpha(f)$ e delle sue *autoconvoluzioni*. E' immediato notare come, al crescere di k_ϕ (o k_f), cresca la banda.

In appendice 11.4.5 è illustrata una tecnica di modulazione per segnali FM modulati a basso indice.

11.4 Appendici

11.4.1 Calcolo della potenza di un segnale AM BLU

Mostriamo che se $X_{BLU}(t) = \frac{k_a}{\sqrt{2}} (m(t) \cos \omega_0 t - \hat{m}(t) \sin \omega_0 t)$, allora $\mathcal{P}_x = \frac{k_a^2}{2} \mathcal{P}_m$. Possiamo innanzitutto scrivere che

$$\mathcal{P}_x = \mathcal{P}_{x^+} + \mathcal{P}_{x^-} = 2\mathcal{P}_{x^+}$$

in quanto le componenti a frequenza positiva e negativa di $x(t)$ sono ortogonali (infatti $\int_{-\infty}^{\infty} X^+(f) X^-(f) df = 0$), e lo spettro di densità di potenza è una funzione pari della frequenza: $\mathcal{P}_x(f) = \mathcal{P}_x(-f)$. Inoltre, invertendo la relazione $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{x^+}(f - f_0)$ valida per la densità di potenza dell'involuppo complesso, otteniamo $\mathcal{P}_{x^+}(f) = \frac{1}{4} \mathcal{P}_{\underline{x}}(f + f_0)$, e quindi

$$\mathcal{P}_{x^+} = \frac{1}{4} \int_{-\infty}^{\infty} \mathcal{P}_{\underline{x}}(f + f_0) df = \frac{1}{4} \mathcal{P}_{\underline{x}}$$

che, sostituita nella prima relazione mostrata, fornisce $\mathcal{P}_x = 2\mathcal{P}_{x^+} = \frac{1}{2} \mathcal{P}_{\underline{x}}$.

Nel caso AM-BLU si ha inoltre $\underline{x}(t) = \frac{k_a}{\sqrt{2}} [m(t) + j\hat{m}(t)]$, tenendo ora conto che $\mathcal{P}_{\underline{x}} = \mathcal{R}_{\underline{x}}(0)$, si ottiene

$$\mathcal{P}_x = \frac{1}{2} \left(\frac{k_a}{\sqrt{2}} \right)^2 [\mathcal{R}_{MM}(0) + \mathcal{R}_{\widehat{M}\widehat{M}}(0) + 2j\mathcal{R}_{M\widehat{M}}(0)]$$

Osserviamo ora che $\mathcal{R}_{M\widehat{M}}(0) = \int_{-\infty}^{\infty} m(t) \hat{m}(t) dt = 0$ in quanto $m(t)$ ed $\hat{m}(t)$ sono ortogonali; inoltre, $\mathcal{R}_{MM}(0) = \mathcal{P}_M = \mathcal{R}_{\widehat{M}\widehat{M}}(0)$ (non dimostrato ma intuibile). Pertanto si ottiene

$$\mathcal{P}_x = \frac{1}{2} \frac{k_a^2}{2} [\mathcal{P}_m + \mathcal{P}_m] = \frac{1}{4} k_a^2 \cdot 2\mathcal{P}_m = \frac{k_a^2}{2} \mathcal{P}_m$$

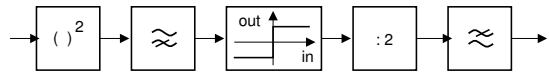
11.4.1.1 Calcolo della potenza di segnali BLD-PI, PS, PPS

Vale lo stesso procedimento adottato sopra, in cui ora

$$\mathcal{P}_{\underline{x}} = \mathcal{P}_{x_c} = \begin{cases} k_a^2 \mathcal{P}_m & \text{(BLD-PS)} \\ a_p^2 + k_a^2 \mathcal{P}_m & \text{(BLD-PI, PPS)} \end{cases}$$

11.4.2 Ricostruzione della portante mediante quadratura

Nel caso di una trasmissione BLD-PS, la portante di demodulazione può essere ottenuta mediante la seguente elaborazione. Il segnale $x(t) = m(t) \cos(\omega_0 t + \varphi)$ viene elevato al quadrato producendo



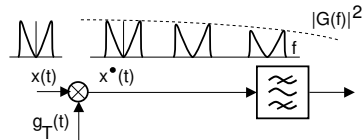
$$\frac{1}{2} m^2(t) [1 + \cos(2\omega_0 t + 2\varphi)]$$

il cui termine di banda base $\frac{1}{2} m^2(t)$ viene rimosso dal filtro passa alto. Successivamente, lo squadratore produce un'onda quadra a frequenza $2f_0$ che viene divisa per 2 da un contatore binario. Infine, un passa basso provvede ad eliminare le armoniche dell'onda quadra a frequenza $n f_0$, fornendo così la portante desiderata, a meno di una *ambiguità di segno*.

11.4.3 Il mixer

Il dispositivo moltiplicatore, presente negli schemi di mo-demodulazione, viene anche chiamato *mixer*, in quanto miscela tra loro due segnali.

Non è strettamente necessario disporre di un oscillatore sinusoidale per realizzare il prodotto di un segnale con una portante: ridotto ai minimi termini... è sufficiente un'onda quadra ed un filtro! Infatti, un qualunque segnale periodico



$$g_T(t) = g(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT)$$

di periodo $T = k/f_0$ (con k intero) possiede una densità di potenza

$$\mathcal{P}_{G_T}(f) = |G(f)|^2 \cdot \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{k} f_0\right)$$

Il prodotto di tale segnale per $x(t)$ produce un $x^\bullet(t)$ con densità di potenza

$$\mathcal{P}_{x^\bullet}(f) = \mathcal{P}_x(f) * \mathcal{P}_{G_T}(f) = \frac{|G(f)|^2}{T} \sum_{n=-\infty}^{\infty} \mathcal{P}_x\left(f - \frac{n}{k} f_0\right)$$

Pertanto, il desiderato spettro di potenza si ottiene inserendo dopo il moltiplicatore un filtro passa banda centrato su f_0 , ossia sulla k -esima replica spettrale di $\mathcal{P}_{x^\bullet}(f)$. Lo stesso dispositivo può essere usato anche per i moltiplicatori di ricezione: in tal caso, il filtro da usare sarà un passa basso.

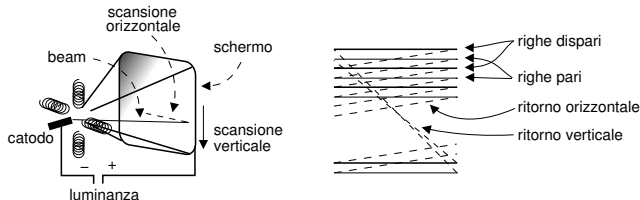
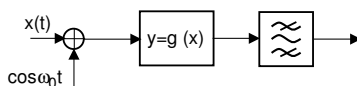


Figura 11.2: Modalità di scansione interlacciata dell'immagine televisiva

Un secondo metodo di realizzare il mixer è con un sommatore, un oscillatore, un dispositivo non lineare, e di nuovo un filtro passa-banda. Il dispositivo non lineare è del tipo



$$y = a_1 x + a_2 x^2 + a_3 x^3 + \dots$$

e quando in ingresso viene applicata la somma di due segnali $x(t) + \cos \omega_0 t$, produce in uscita

$$y(t) = a_1 (x(t) + \cos \omega_0 t) + a_2 (x^2(t) + \cos^2 \omega_0 t + 2x(t) \cos \omega_0 t) + a_3 (\dots) + \dots$$

in cui, osservando che i termini $\cos^n \omega_0 t$ sono relativi a termini a frequenza $n f_0$, è possibile ancora una volta estrarre il termine che ci interessa.

11.4.4 Trasmissione televisiva

Illustriamo molto brevemente le principali caratteristiche delle trasmissioni televisive broadcast *analogiche*, con riferimento agli standard *dismessi* in Italia.

Codifica dell'immagine Una trasmissione televisiva avviene riproducendo 25 diverse immagini (dette *quadri*) al secondo. Ogni immagine è scomposta in 625 linee orizzontali, che vengono trasmesse in due fasi: prima le linee dispari, poi quelle pari. In questo modo un singolo quadro è riprodotto due volte³⁶ ogni $\frac{1}{25} = 0.04$ secondi (seppure in modo alternato) portando così a 50 semiquadri/secondo³⁷ la frequenza di rinfresco, in modo da impedire i fenomeni di *sfarfallamento* (FLICKER) ottico³⁸.

La riproduzione di un quadro avviene (vedi fig. 11.2) mediante un *tubo catodico*, il quale dispone posteriormente di un catodo che emette elettroni, che sono accelerati da un segnale di luminanza positivo applicato all'anodo, e che terminano alla loro corsa contro lo strato di fosforo distribuito sulla parte anteriore (schermo) del tubo. Il *fascio* (BEAM) di elettroni è focalizzato elettronicamente, e viene deflesso ciclicamente sia in orizzontale alla frequenza di $625 \frac{\text{linee}}{\text{quadro}} \cdot 25 \frac{\text{quadri}}{\text{secondo}} = 15625 \text{ Hz}$ (*frequenza di riga*), sia verticalmente con velocità di $50 \frac{\text{semiquadri}}{\text{secondo}}$.

³⁶La riproduzione di metà quadro alla volta è chiamata *scansione interlacciata* dell'immagine. Nulla vieta al costruttore del ricevitore di prevedere una *memoria di quadro* e di riprodurre le immagini in modo non interlacciato; il segnale trasmesso invece presenta sempre le righe in formato interlacciato.

³⁷La frequenza di 50 semiquadri/secondo è stata scelta di proposito uguale alla frequenza di funzionamento della rete elettrica, in modo che eventuali disturbi elettrici avvengano sempre *nello stesso punto* dell'immagine, riducendo gli effetti fastidiosi.

³⁸Il *flicker* si manifesta nel caso in cui la frequenza di rinfresco è inferiore al tempo di persistenza delle immagini sulla retina, pari a circa $\frac{1}{40}$ di secondo.

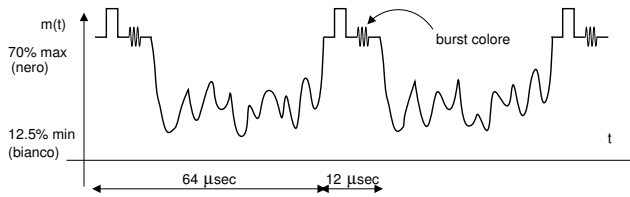


Figura 11.3: Forma d'onda del segnale televisivo

Segnale televisivo in bianco e nero Il segnale televisivo contiene sia le informazioni di temporizzazione necessarie a sincronizzare la scansione dell'immagine, che l'informazione di luminanza che pilota la tensione anodica, e quindi la forza con cui l'elettrone urta lo schermo.

Durante la trasmissione di ogni semiquadro, ogni riga dispone di $\frac{1}{15625} = 64$ μ secondi. Il segnale modulante è sempre positivo (vedi fig. 11.3), ed associa ai valori più piccoli la maggiore luminanza³⁹, trasmettendo in logica negata, in modo che gli impulsi di sincronismo orizzontale siano di ampiezza superiore al livello del nero (il 70 % dell'ampiezza). Il tempo dedicato alla trasmissione della luminanza di una riga è di 52 μ sec, mentre nei restanti 12 il segnale oltrepassa il livello del nero (in modo da rendere invisibile il beam) e quindi un impulso determina il ritorno orizzontale. In figura è anche mostrato un *burst colore* che è presente nelle trasmissioni a colori per sincronizzare la *portante di colore* (vedi di seguito).

Formato dell'immagine Ogni singolo quadro è realizzato con un rapporto di aspetto 4:3 (che rappresenta il rapporto tra le dimensioni orizzontale e verticale), e solo 575 delle 625 linee vengono mostrate (infatti 25 linee per ogni semiquadro cadono al di fuori dello schermo⁴⁰).

Occupazione spettrale Diverse considerazioni⁴¹ hanno portato a stabilire che la banda del segnale televisivo sia di circa ± 5 Mhz, e nell'ultima versione del sistema PAL questa è stata portata a 6 MHz. In particolare, dato che le immagini presentano spesso ampie zone uniformi, corrispondenti ad un segnale di luminanza pressoché costante, la densità spettrale del segnale televisivo è piuttosto concentrata nella regione delle basse frequenze. Per questo motivo, si è deciso di trasmettere il segnale mediante modulazione di ampiezza a banda laterale ridotta, conseguendo un risparmio di banda e contemporaneamente preservando le componenti del messaggio a frequenze più basse.

La figura 11.4 mostra la situazione, in cui solo parte (1.75 MHz) della banda inferiore del segnale di luminanza viene trasmessa, mentre il filtro di ricezione provvede a realizzare un filtraggio complessivo tale che $\underline{H}(f) + \underline{H}^*(-f) = \text{cost}$ (vedi nota 22 a pag. 279).

³⁹In questo modo si riduce mediamente la potenza trasmessa, dato che sono più frequenti scene chiare che scure.

⁴⁰Nel tempo destinato alle linee che non sono mostrate, vengono comunque trasmesse altre informazioni, come ad esempio i dati che compaiono nelle pagine del televideo.

⁴¹Ad esempio, si può stabilire di realizzare la stessa risoluzione orizzontale e verticale. A fronte delle 625 linee, il rapporto di aspetto di $\frac{4}{3}$ determina l'esigenza di individuare $625 \cdot \frac{4}{3} = 833$ $\frac{\text{punti}}{\text{linea}}$, e quindi $833 \cdot 625 = 520625$ $\frac{\text{punti}}{\text{quadro}}$, ossia circa $13 \cdot 10^6$ $\frac{\text{punti}}{\text{secondo}}$. Per il teorema del campionamento, il segnale deve avere una banda minore od uguale di $\frac{f_c}{2} = 6.5$ MHz.

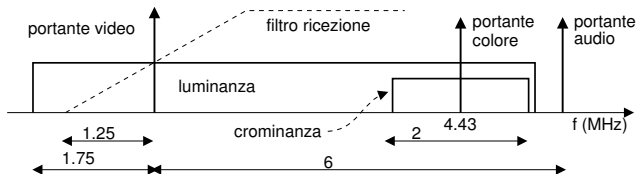


Figura 11.4: Occupazione spettrale di un segnale televisivo

Segnale audio In figura 11.4 è rappresentata anche una portante audio, che viene trasmessa oltre la banda del segnale video, mediante una modulazione FM con $\Delta f_{Max} = 25$ KHz.

Segnale di crominanza Il requisito che più di altri ha determinato quale soluzione adottare per effettuare trasmissioni a colori, è che queste dovessero essere correttamente visibili anche da parte di ricevitori in bianco e nero.

Un risultato di colorimetria è che ogni colore è scomponibile nella somma di tre colori fondamentali (verde, rosso e blu), effettivamente operata dagli apparati di acquisizione. La somma⁴² della tre componenti fornisce il segnale di luminanza L , che viene utilizzato esattamente come per il bianco e nero. Il segnale di crominanza è invece costruito da una coppia di segnali differenza $\begin{cases} \Delta_R = R - L \\ \Delta_B = B - L \end{cases}$, che sono usati per modulare in ampiezza, portante soppressa, una portante di colore, usando Δ_R come componente in fase e Δ_B come componente in quadratura⁴³. Una analisi più precisa, è fornita al § 18.2.2.

L'occupazione spettrale del segnale di crominanza è ridotta (± 1 MHz) rispetto a quello di luminanza, in quanto la *risoluzione spaziale* dell'occhio umano è ridotta per stimoli colorati, e quindi Δ_R e Δ_B possono variare più lentamente di L .

Sincronizzazione Per impedire fenomeni di interferenza tra C.A. di B.F. nella ricezione del segnale di crominanza, occorre effettuare una demodulazione omodina, e l'oscillatore del ricevitore si mantiene coerente con la portante di colore, grazie ai *burst di colore* presenti dopo l'impulso di sincronizzazione orizzontale, costituiti da 8 cicli di portante. Questo segnale ha inoltre lo scopo di segnalare la *presenza* della componente di crominanza: in caso contrario infatti (trasmissione B/N) il ricevitore deve disattivare il circuito del colore, per non produrre deterioramenti dell'immagine.

Interferenza La presenza di entrambi i segnali di luminanza e crominanza nella stessa banda sembrerebbe dare luogo a difficili problemi di interferenza. Innanzitutto osserviamo che, come anticipato, il segnale di luminanza è concentrato attorno alla portante video, e dunque arreca un disturbo ridotto⁴⁴ alla crominanza. Quest'ultima

⁴²In realtà ogni componente è pesata mediante un opportuno coefficiente che tiene conto della diversa sensibilità dell'occhio ai tre colori fondamentali. Infatti, per ottenere il bianco, i tre colori non devono essere mescolati in parti uguali, bensì 59% di verde, 30% di rosso e 11% di blu.

⁴³Le ampiezze delle componenti in fase e quadratura del segnale di crominanza devono essere opportunamente scalate, per impedire al segnale complessivo (luminanza più crominanza) di assumere valori troppo elevati.

⁴⁴Possiamo riflettere su quali siano le circostanze che producono la massima interferenza della luminanza sulla crominanza: ciò avviene in corrispondenza di scene molto definite, relative ad

quindi, prima di essere demodolata, viene filtrata per rimuovere il segnale di luminanza fuori della banda di crominanza, ed il disturbo è generalmente trascurabile. Viene inoltre adottata una soluzione che riduce anche l'interferenza di crominanza su luminanza. Quest'ultima presenta infatti una spiccata periodicità, legata alla frequenza di riga f_r ed alla presenza degli impulsi di sincronismo ogni $64 \mu\text{sec}$, che determina uno spettro con energia concentrata alle armoniche di $f_r = 15625 \text{ Hz}$. Pertanto, si colloca la portante di colore a metà tra due armoniche del segnale di luminanza, in modo che le densità spettrali risultino, pur se sovrapposte, intercalate. L'uso di *filtri a pettine*⁴⁵ nel ricevitore può quindi ridurre notevolmente l'interferenza.

11.4.5 Modulazione FM a basso indice

Riprendiamo qui il caso in cui $\beta \ll 1 \Rightarrow \Delta\alpha \ll 1$, e quindi l'espansione in serie di potenze di $\underline{x}(t)$ può arrestarsi al primo ordine; se il segnale modulante è cosinusoidale, il segnale FM risulta

$$x_{FM}(t) = a \cos \left(\omega_0 t + 2\pi k_f \int_{-\infty}^t \cos(2\pi w\tau) d\tau \right) = a \cos(\omega_0 t + \beta \sin(2\pi w t))$$

Ricordando che $\cos(\alpha + \beta) = \cos\alpha \cos\beta - \sin\alpha \sin\beta$, $x_{FM}(t)$ può essere riscritto come

$$x_{FM}(t) = a \cos \omega_0 t \cos(\beta \sin 2\pi w t) - a \sin \omega_0 t \sin(\beta \sin 2\pi w t)$$

che, se $\beta \ll 1$, diviene

$$x_{FM}(t) = a \cos \omega_0 t - \beta a \sin \omega_0 t \sin 2\pi w t$$

che confrontiamo con l'espressione

$$x_{AM}(t) = a_p \cos \omega_0 t + k_a \cos \omega_0 t \cos 2\pi w t$$

che si otterrebbe per modulazione a portante intera, o ridotta, dello stesso $m(t)$.

Il confronto rivela che mentre nell'AM il segnale modulante opera *in fase* alla portante, nell'FM a basso indice opera *in quadratura*. Il risultato esposto costituisce ad ogni modo uno *schema di modulazione* per segnali FM a basso indice, realizzabile *sommando* alla portante, un segnale modulato AM su di una portante in quadratura.

Resta il fatto che uno schema di modulazione del genere produce anche una modulazione AM parassita: quest'ultima è eliminata in ricezione dall'azione congiunta di uno squadratore e di un filtro passa basso.

11.4.6 FM broadcast

Illustriamo brevissimamente i parametri delle trasmissioni FM ricevibili mediante *la radio di casa*. Nella banda 88-108 MHz operano le radio FM, con spaziatura di 200 KHz l'una dall'altra. Ad ogni emittente è concessa una deviazione massima della frequenza istantanea pari a $\Delta f = 75 \text{ KHz}$.

Il trasmettitore viene tarato mediante un $m(t)$ sinusoidale a frequenza di 15 KHz, e k_f regolato in modo da ottenere

$$\Delta f = 75 \text{ KHz}$$

In queste condizioni, risulta $\beta = \frac{k_f}{w} = \frac{75}{15} = 5$, e la regola di Carson fornisce

$$B_C = 2(k_f + w) = 2(75 + 15) = 180 \text{ KHz}$$

Un esame degli andamenti riportati in Fig. 11.1 mostra che per $\beta = 5$, le $\mathcal{J}_n(\beta) \neq 0$ sono le prime 8, e dunque la "vera" banda ha una estensione

immagini con elevato contenuto di frequenze spaziali elevate, ad esempio nel caso di righe fitte; il disturbo è più appariscente nel caso in cui la zona ad elevato contrasto sia povera di componenti cromatiche. Avete mai notato cravatte a righe bianche e nere, divenire cangianti?

⁴⁵Introdotta al § 9.8, l'argomento può essere approfondito presso http://it.wikipedia.org/wiki/Filtro_comb

$$B = 2 \cdot 8w = 16 \cdot 15 \cdot 10^3 = 240 \text{ KHz}$$

mostrando l'approssimazione della regola di Carson. D'altra parte, risulta che

$$2 \sum_{n=6}^8 |\mathcal{J}_n(5)|^2 = 2 [(.13)^2 + (.05)^2 + (.02)^2] = 2 \cdot 0.0198 = 0.0396$$

e dunque l'errore commesso esclude circa il 4% della potenza totale.

Qualora il segnale sinusoidale venga sostituito da un messaggio limitato in banda con $\pm W = \pm 15 \text{ KHz}$, con potenza eguale a quella del seno e cioè $P_m = \frac{1}{2}$, la Δf non è più definita con esattezza, e conviene ricorrere alla definizione di

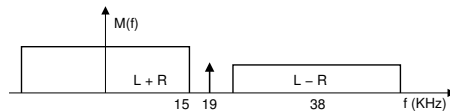
$$\beta' = \frac{\sigma_f}{W} = \frac{k_f \sqrt{P_m}}{W} = \beta \sqrt{P_m} = \beta \frac{1}{\sqrt{2}} = 0.707 \cdot \beta$$

a cui corrisponde una banda *efficace*

$$B = 2W (\beta' + 1) = 2 \cdot 15 \cdot 10^3 \cdot (0.707 \cdot 5 + 1) = 136 \text{ KHz}.$$

Nell'FM stereo, il segnale trasmesso deve essere compatibile con i ricevitori mono, ed allora il segnale modulante è un segnale multiplato FDM, e "composto" da tre "canali":

- La somma di Left + Right (L+R) come segnale di banda base, che consente la compatibilità con gli apparati "mono";
- Il segnale L-R è centrato a frequenza di 38 KHz mediante modulazione AM-BLD;
- A 19 KHz, è presente una portante a cui si concede il 10% di \mathcal{P}_M , mentre il restante 90% di \mathcal{P}_M è condiviso tra L+R e L-R. Il tutto è poi modulato FM.



La portante a 19 KHz può essere impiegata per sincronizzare il ricevitore, e generare la portante (a frequenza doppia, di 38 KHz) necessaria a demodulare il canale L-R. Se assente, indica la ricezione di un canale mono.

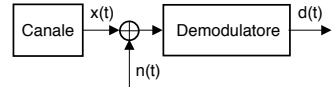
A prima vista, sembrerebbe che la presenza del canale L-R possa aumentare la massima deviazione di frequenza. In realtà non è così, per due motivi:

- quando L+R è grande, vuol dire che i due canali sono simili, e dunque L-R è piccolo, e viceversa;
- Il canale L-R, trovandosi a frequenze più elevate, è caratterizzato da un indice di modulazione inferiore. Infatti, la massima deviazione di frequenza istantanea dipende dalle *ampiezze* di $m(t)$, e non dalla sua banda.

Capitolo 12

Prestazioni delle trasmissioni modulate

Analizziamo ora l'effetto prodotto dal rumore $n(t)$ presente in ricezione sul risultato del processo di demodulazione, per i due casi analizzati di modulazione di ampiezza ed angolare, e descritto quantitativamente dal rapporto segnale-rumore osservabile per il segnale demodolato.

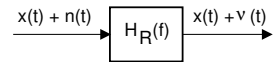


12.1 Il rumore nei segnali modulati

Consideriamo un segnale modulato, affetto da un rumore additivo gaussiano bianco $n(t)$ (§ 9.2.3), con densità di potenza

$$\mathcal{P}_n(f) = \frac{N_0}{2}$$

Similmente a quanto già illustrato al § 5.3.1 per trasmissioni di banda base, prima di essere demodolato il segnale ricevuto attraversa un *filtro di ricezione* passa-banda



$H_R(f)$, allo scopo di limitare la banda del rumore ricevuto, e quindi ridurre l'entità della potenza di rumore in ingresso al ricevitore. Il filtro $H_R(f)$ presenta un modulo costante nella banda del segnale, mentre tende a zero al di fuori di tale banda. In questo modo, il segnale utile $x(t)$ transita inalterato, ed il rumore $n(t)$ viene limitato in banda, producendo $\nu(t)$.

12.1.1 Rapporto segnale-rumore e banda di rumore

Considerando la densità di potenza del segnale modulato $\mathcal{P}_x(f)$ e quella del rumore $\mathcal{P}_\nu(f)$, la qualità di ricezione (in funzione della frequenza) è descritta dal rapporto

$$SNR_{RF}(f) = \frac{\mathcal{P}_x(f)}{\mathcal{P}_\nu(f)}$$

in cui $\mathcal{P}_x(f)$ dipende dal tipo di modulazione (cap. 11), mentre dopo il filtraggio passa-banda, $\mathcal{P}_\nu(f)$ risulta pari a

$$\mathcal{P}_\nu(f) = \mathcal{P}_n(f) |H_R(f)|^2 = \frac{N_0}{2} |H_R(f)|^2$$

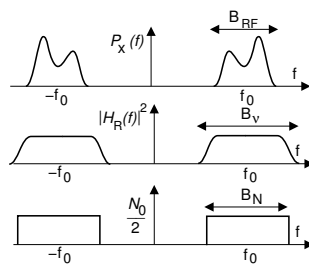
D'altra parte, ha senso valutare l'*SNR complessivo*, ovvero il rapporto tra le potenze di segnale e di rumore *totali*: la potenza di segnale risulta allora pari a $\mathcal{P}_x = \int_{-\infty}^{\infty} \mathcal{P}_x(f) df$, mentre per quella di rumore si ottiene $\mathcal{P}_\nu = \int_{-\infty}^{\infty} \frac{N_0}{2} |H_R(f)|^2 df$, valutando cioè la potenza che attraversa il filtro di ricezione $H_R(f)$. A questo proposito, data l'impossibilità pratica di realizzare un filtro passa-banda ideale (rettangolare), $H_R(f)$ è caratterizzata da una banda (a frequenze positive, vedi § 10.1.4) B_ν , più estesa di B_{RF} (che è la banda di segnale). La potenza totale del rumore uscente da $H_R(f)$ risulta pertanto pari a

$$\begin{aligned} \mathcal{P}_\nu &= \frac{N_0}{2} \cdot \int_{-\infty}^{\infty} |H_R(f)|^2 df = \\ &= N_0 \int_0^{\infty} |H_R(f)|^2 df = N_0 B_N |H_R(f_0)|^2 \end{aligned}$$

Il termine $B_{RF} \leq B_N \leq B_\nu$ rappresenta la cosiddetta *banda di rumore* definita come

$$B_N = \frac{\int_0^{\infty} |H_R(f)|^2 df}{|H_R(f_0)|^2}$$

ossia come la banda di un filtro ideale (rettangolare) che lascia passare la stessa quantità di rumore.



12.1.2 Demodulazione di un processo di rumore

Il rumore $\nu(t)$ che esce dal filtro di ricezione $H_R(f)$ è di tipo *passa-banda*, e può quindi essere descritto nei termini delle sue componenti analogiche di bassa frequenza:

$$\nu(t) = \nu_c(t) \cos \omega_0 t - \nu_s(t) \sin \omega_0 t$$

in cui, ricordando i risultati ottenuti al § 10.6.2, $\nu_c(t)$ e $\nu_s(t)$ sono entrambe gaussiane se $\nu(t)$ lo è; osserviamo inoltre che nel caso in cui la banda di $\nu(t)$ sia *stretta rispetto a* f_0 , l'involuppo complesso $\underline{\nu}(t) = \nu_c(t) + j\nu_s(t)$ evolve *lentamente* rispetto alla velocità di rotazione di $\underline{\nu}(t) e^{j\omega_0 t}$.

La figura a lato rappresenta la situazione in cui il rumore $n(t)$ in ingresso a $H_R(f)$ sia un processo ergodico bianco con densità di potenza $\mathcal{P}_n(f) = \frac{N_0}{2}$, ed il filtro di ricezione possenga una risposta in frequenza unitaria $|H_R(f_0)|^2 = 1$; in questo caso si ottiene che

$$\mathcal{P}_\nu(f) = \frac{N_0}{2} \text{rect}_{B_N}(f - f_0) + \frac{N_0}{2} \text{rect}_{B_N}(f + f_0)$$

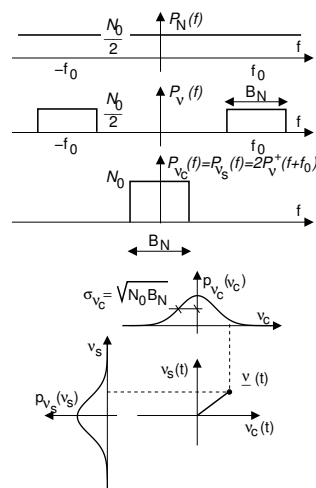
e quindi

$$\mathcal{P}_{\nu_c}(f) = \mathcal{P}_{\nu_s}(f) = 2\mathcal{P}_\nu^+(f + f_0) = N_0 \text{rect}_{B_N}(f)$$

Pertanto, $\nu_c(t)$ e $\nu_s(t)$ risultano essere due processi congiuntamente gaussiani, ergodici, a media nulla ed uguale varianza (e potenza)

$$\sigma_{\nu_c}^2 = \sigma_{\nu_s}^2 = \mathcal{P}_\nu = N_0 B_N$$

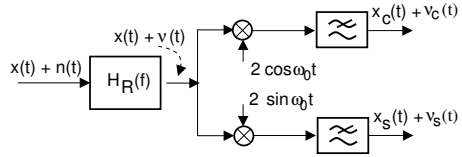
In definitiva quindi, operando una demodulazione coerente in fase ed in quadratura del segnale ricevuto, nelle componenti analogiche risultanti saranno presenti i termini additivi $\nu_c(t)$ e $\nu_s(t)$, entrambi di potenza $\mathcal{P}_\nu = N_0 B_N$.



12.2 Prestazioni delle trasmissioni AM

Per valutare il rapporto SNR per le diverse tecniche di modulazione AM, esprimiamo il segnale modulato nei termini delle sue componenti analogiche

$$x_{AM}(t) = x_c(t) \cos \omega_0 t - x_s(t) \sin \omega_0 t$$



Operando una demodulazione coerente del segnale $x(t)$ affetto da rumore bianco $n(t)$ e filtrato da $H_R(f)$ come in figura, si ottengono le due componenti analogiche per il segnale demodulato:

$$\begin{cases} d_c(t) = x_c(t) + v_c(t) \\ d_s(t) = x_s(t) + v_s(t) \end{cases} \text{ . Tra la potenza del segnale ricevuto e quella delle sue C.A. di B.F. sussiste}^1 \text{ la relazione: } \mathcal{P}_x = \frac{1}{2}\mathcal{P}_{x_c} + \frac{1}{2}\mathcal{P}_{x_s}.$$

12.2.1 Potenza di segnale e di rumore dopo demodulazione ed SNR

Nel caso di modulazione AM, siamo interessati alla sola componente in fase, che è sufficiente a fornire $m(t)$; il rapporto tra \mathcal{P}_{x_c} e \mathcal{P}_{v_c} fornirà dunque il valore di SNR che stiamo cercando. La tabella che segue mostra i valori delle componenti di segnale e di rumore, assieme alle rispettive potenze espresse in funzione di una medesima potenza di messaggio \mathcal{P}_m , per tre casi di modulazione AM, di cui discutiamo individualmente.

	$x_c(t)$	$x_s(t)$	\mathcal{P}_{x_c}	\mathcal{P}_x	B_N	\mathcal{P}_{v_c}
BLD-PS	$m(t)$	0	\mathcal{P}_m	$\frac{1}{2}\mathcal{P}_m$	$2W$	$2WN_0$
BLD-PI	$\sqrt{\eta}(a_p + m(t))$	0	$\eta \cdot \alpha$	$\frac{1}{2}\eta \cdot \alpha = \frac{1}{2}\mathcal{P}_m$	$2W$	$2WN_0$
BLU-PS	$\frac{1}{\sqrt{2}}m(t)$	$\mp \frac{1}{\sqrt{2}}\hat{m}(t)$	$\frac{1}{2}\mathcal{P}_m$	$\frac{1}{2}\mathcal{P}_m$	W	WN_0

Rispetto alla notazione adottata al capitolo 11, si considera ora il termine k_a inglobato in $m(t)$, e quindi non evidenziato in tabella; inoltre, si pone $\alpha = (a_p^2 + \mathcal{P}_m)$. Inoltre, la banda di rumore B_N presa in considerazione nella tabella è la minima possibile, pari a quella del segnale modulato B_{RF} , direttamente legata (nella modulazione AM) a quella ($\pm W$) del segnale modulante. Pertanto, i risultati che otterremo sono i migliori possibili: infatti, se $B_N > B_{RF}$, l' SNR risulterà peggiore. Precisiamo infine che nella valutazione dell' SNR che segue, ci si riferisce sempre ad una medesima potenza ricevuta \mathcal{P}_x e ad una densità di rumore $\mathcal{P}_n(f) = \frac{N_0}{2}$, allo scopo di rendere confrontabili i risultati.

12.2.1.1 Modulazione BLD-PS

In questo caso si ha $\mathcal{P}_{x_c} = \mathcal{P}_m$ e $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$; dopo demodulazione il segnale $d_c(t) = x_c(t) + v_c(t)$ presenta dunque una potenza di segnale utile $\mathcal{P}_{x_c} = \mathcal{P}_m = 2\mathcal{P}_x$ (con \mathcal{P}_x pari alla potenza del segnale ricevuto) ed una potenza di rumore $\mathcal{P}_{v_c} = 2WN_0$; dunque un rapporto segnale-rumore pari a

$$SNR = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{v_c}} = \frac{\mathcal{P}_m}{2WN_0} = \frac{\mathcal{P}_x}{WN_0} = SNR_0$$

in cui nell'ultima eguaglianza si definisce:

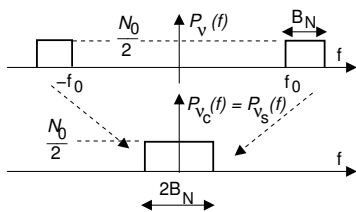
¹Infatti i segnali $x_c(t) \cos \omega_0 t$ e $x_s(t) \sin \omega_0 t$ risultano ortogonali, e le potenze si sommano.

SNR di riferimento La grandezza SNR_0 è denominata rapporto segnale-rumore *di riferimento*, ed è il rapporto tra la potenza di segnale *ricevuto* e la potenza di rumore *in una banda pari a quella del messaggio di banda base*.²

Osserviamo dunque che la modulazione BLD-PS *non altera* il rapporto SNR_0 di banda base, ovvero è come se il processo di modulazione fosse *assente*. Notiamo infine (e questo è valido anche per i casi che seguono) che SNR può riferirsi indifferentemente sia alle potenze di segnale che a quelle disponibili (vedi § 14.2.2.1), in quanto $SNR_0 = \frac{\mathcal{P}_x}{WN_0} = \frac{\mathcal{P}_x}{WN_0} \frac{4R_g}{4R_g} = \frac{W_{d_x}}{W_{d_N}}$.

12.2.1.2 Modulazione BLU-PS

In questo caso, per ottenere una $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$ uguale al caso BLD-PS, le componenti $x_c(t)$ ed $x_s(t)$ devono essere poste rispettivamente pari a $\frac{1}{\sqrt{2}}m(t)$ e $\frac{1}{\sqrt{2}}\hat{m}(t)$ (vedi § 11.1.4).



A seguito del processo di demodulazione, si ottiene un rumore in banda base che occupa ancora una banda $\pm B_N$ come nel caso AM-BLD, ma possiede una densità *uguale* a quella del rumore a radio frequenza (e non *doppia* come al § 12.1.2), in quanto i contenuti a frequenze positive e negative non si sovrappongono, come mostrato in figura. Pertanto risulta:

$$SNR = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} = \frac{\frac{1}{2}\mathcal{P}_m}{WN_0} = \frac{\mathcal{P}_x}{WN_0} = SNR_0$$

Dunque, si ottengono prestazioni *identiche* a quelle del caso AM-BLD. Si noti che il risultato è valido solo se $\nu(t)$ è effettivamente limitato alla sola banda B_{RF} del segnale BLU: se infatti si fosse adottato un filtro con banda più larga, come ad esempio un $H_R(f)$ con $B_N = 2W$, si sarebbe ottenuto $\mathcal{P}_{\nu_c}(f) = N_0$, ed SNR risulterebbe dimezzato.

12.2.1.3 Modulazione BLD-PI

Per ottenere una potenza di segnale ricevuto $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$ uguale ai due casi precedenti, si considera la ricezione di un segnale

$$x(t) = \sqrt{\eta}(a_p + m(t)) \cos \omega_0 t \quad \text{in cui} \quad \eta = \frac{\mathcal{P}_m}{a_p^2 + \mathcal{P}_m}$$

dove η è proprio pari all'efficienza della BLD-PI introdotta al § 11.1.1.4. Nel valutare l' SNR , faremo riferimento alla sola componente di messaggio $\sqrt{\eta}m(t)$ del segnale demodulato \mathcal{P}_x , che ha potenza $\eta\mathcal{P}_m = 2\eta\mathcal{P}_x$: la quantità a_p^2 si riferisce infatti alla portante non modulata, e non trasporta informazione. Si ha pertanto:

$$SNR = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} = \frac{2\eta\mathcal{P}_x}{2WN_0} = \eta \frac{2\mathcal{P}_x}{2WN_0} = \eta SNR_0$$

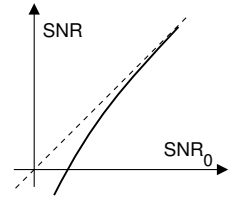
²In virtù di questa definizione, SNR_0 è una grandezza che caratterizza le condizioni operative (\mathcal{P}_x , $\mathcal{P}_m(f) = \frac{N_0}{2}$ e W sono grandezze *indipendenti*) ma non è legata alla particolare tecnica di modulazione adottata. Pertanto, esprimere SNR in funzione di SNR_0 permette il confronto tra i diversi tipi di modulazione *a parità di condizioni operative*.

Dunque in questo caso constatiamo che la presenza della portante comporta una riduzione di prestazioni proprio pari all'efficienza $\eta = \frac{P_m}{a_p^2 + P_m}$.

L'analisi fin qui esposta si riferisce però al caso di demodulazione coerente: invece per BLD-PI si usa il demodulatore di involuppo (§ 11.2.3) ! In tal caso, il segnale demodulato è il modulo dell'involuppo complesso, ovvero

$$d(t) = |\underline{x}(t) + \underline{\nu}(t)| = \sqrt{[\sqrt{\eta}(a_p + m(t)) + \nu_c(t)]^2 + \nu_s^2(t)}$$

Nel caso in cui $\underline{\nu}(t)$ sia piccolo, si può ottenere una approssimazione che ci riconduce al caso precedente. In caso contrario, sorgono termini *prodotto* tra $m(t)$ e $\nu_c(t)$, ed in definitiva l'*SNR* risulta peggiore (per bassi SNR_0) del caso di demodulazione sincrona omodina, come illustrato nella curva riportata a fianco.



12.3 Prestazioni delle trasmissioni FM

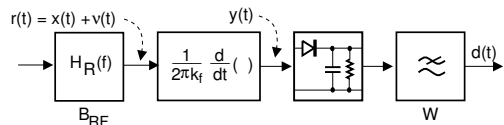
Quando si è analizzata la tecnica di modulazione FM, si è fatta più volte notare la sua natura *non lineare* (pag. 282). E' lecito aspettarsi che questa caratteristica determini dei risvolti "bizzarri" per quanto riguarda l'*SNR* del segnale demodulato, e difatti è proprio così. In particolare, nel caso in cui la potenza del rumore in ingresso al ricevitore non sia eccessiva, avviene che:

- la potenza del rumore dopo demodulazione *diminuisce* all'aumentare della potenza ricevuta;
- l'*SNR* dopo demodulazione *migliora* all'aumentare della banda occupata.

Sviluppiamo quindi l'analisi che porta a questi risultati, valutando prima ciò che accade nella ricezione di una portante non modulata, e quindi discutendo come lo scenario si modifica in presenza di segnale. Infine, illustriamo i motivi che determinano il rapido degrado di prestazioni nel caso di rumore elevato.

12.3.1 Rumore dopo demodulazione FM

Analizziamo innanzitutto il comportamento di un demodulatore a *discriminatore*³, quando in ingresso è presente una portante *non modulata* di ampiezza⁴ $A = \sqrt{2P_x}$ sovrapposta ad un rumore gaussiano bianco limitato in banda:

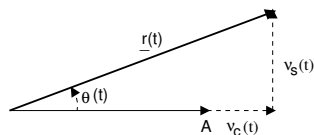


$$r(t) = A \cos \omega_0 t + \nu_c(t) \cos \omega_0 t - \nu_s(t) \sin \omega_0 t$$

A differenza del caso AM, ora il filtro $H_R(f)$ deve lasciar passare una banda di frequenze di estensione almeno pari alla banda che sarebbe occupata se la portante fosse modulata FM, e stimata applicando ad esempio la *regola di Carson* 11.5, ossia $B_C \simeq 2W(\beta + 1)$.

³Discusso al §11.3.1.2, e che si basa sul demodulatore di involuppo introdotto al § 11.2.3.

⁴Con questa posizione, la potenza della portante risulta $\frac{(\sqrt{2P_x})^2}{2} = \frac{2P_x}{2} = P_x$.



E' allora immediato verificare come nel caso di portante non modulata le componenti analogiche di bassa frequenza di

$$\underline{r}(t) = r_c(t) + jr_s(t)$$

valgono $\begin{cases} r_c(t) = A + \nu_c(t) \\ r_s(t) = \nu_s(t) \end{cases}$ come illustrato in

figura. Notiamo quindi che, per piccoli valori (rispetto ad A) di $\nu_c(t)$ e $\nu_s(t)$, l'involuppo complesso ricevuto $\underline{r}(t)$ rimane *prossimo* a quello (A) della portante non modulata. Come noto, $\nu_c(t)$ e $\nu_s(t)$ appartengono a due processi congiuntamente gaussiani, a media nulla e deviazione standard $\sigma_{\nu_c} = \sigma_{\nu_s} = \sqrt{N_0 B_N}$, in cui $B_N \geq B_{RF}$ è la banda di rumore del ricevitore, ed $N_0/2$ è la densità di potenza del rumore in ingresso.

Ricordiamo ora che nel caso FM il *segnale informativo* è legato alla *derivata* della fase $\theta(t)$: esprimiamo dunque $r(t)$ mettendo $\theta(t)$ in evidenza:

$$r(t) = \Re \left\{ \underline{r}(t) e^{j\omega_0 t} \right\} = \Re \left\{ |\underline{r}(t)| e^{j\theta(t)} e^{j\omega_0 t} \right\} = |\underline{r}(t)| \cos(\omega_0 t + \theta(t))$$

Osserviamo quindi che il termine $|\underline{r}(t)|$ viene rimosso dal limitatore (vedi § 11.3.1.2) che usualmente è anteposto al discriminatore. Il segnale $y(t)$ in uscita dal derivatore è quindi dato da

$$y(t) = \left(\frac{f_0}{k_f} + \frac{1}{2\pi k_f} \frac{d}{dt} \theta(t) \right) \sin(\omega_0 t + \theta(t))$$

che viene a sua volta elaborato da parte del demodulatore di involuppo come fosse un segnale BLD-PI, fornendo in definitiva un segnale demodulato

$$d(t) = \frac{1}{2\pi k_f} \frac{d}{dt} \theta(t) \quad (12.1)$$

12.3.2 Caso di basso rumore

Se $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma_{\nu_c}^2 = \sigma_{\nu_s}^2 = N_0 B_N$ allora, come osservato, l'involuppo complesso del rumore ha modulo *abbastanza* più piccolo di A . Pertanto si può scrivere

$$\theta(t) = \arctan \frac{\nu_s(t)}{A + \nu_c(t)} \simeq \arctan \frac{\nu_s(t)}{A} \simeq \frac{\nu_s(t)}{A}$$

e dunque

$$\mathcal{P}_\theta(f) = \frac{1}{A^2} \mathcal{P}_{\nu_s}(f) = \frac{N_0}{A^2} \quad (12.2)$$

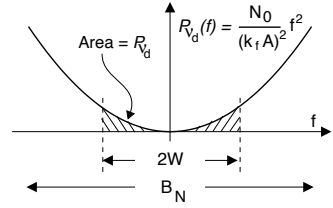
Ricordiamo ora (vedi § 3.7) che l'operazione di derivata (svolta dal discriminatore) equivale a moltiplicare lo spettro di ampiezza del segnale che si deriva per $j2\pi f$, ovvero moltiplicare la sua densità di potenza per $(2\pi f)^2$: applicando questo risultato alla densità di potenza di (12.1) e tenendo poi conto della (12.2), si ottiene che la densità di potenza dopo demodulazione (dovuta al solo rumore sovrapposto alla portante non modulata) risulta pari a

$$\mathcal{P}_{\nu_d}(f) = \frac{1}{(2\pi k_f)^2} (2\pi f)^2 \mathcal{P}_\theta(f) = \left(\frac{f}{k_f} \right)^2 \frac{N_0}{A^2} = \frac{N_0}{(k_f A)^2} f^2$$

Infine, troviamo che la potenza totale di rumore dopo demodulazione risulta pari a

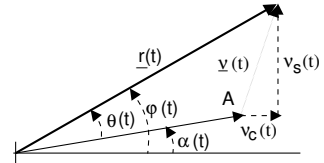
$$\mathcal{P}_{\nu_d} = \sigma_{\nu_d}^2 = 2 \int_0^W \frac{N_0}{(k_f A)^2} f^2 df = \frac{2}{3} \frac{N_0}{(k_f A)^2} W^3 \quad (12.3)$$

in cui W è la banda del segnale modulante (se ci fosse), ed il rumore è limitato in tale banda in virtù del filtro passa basso posto a valle del discriminatore. Si noti che invece le potenze $\sigma_{\nu_c}^2$ e $\sigma_{\nu_s}^2$ sono relative alla banda B_N , legata a quella del segnale modulato.



Notiamo subito la veridicità della prima affermazione fatta ad inizio sezione: la potenza complessiva del rumore dopo demodulazione FM *diminuisce* all'aumentare della potenza del segnale ricevuto $\mathcal{P}_x = \frac{A^2}{2}$. Una seconda osservazione molto importante è che, per effetto della derivata, la densità di potenza del rumore demodolato ha un andamento *parabolico*.

Segnale presente Continuando ad ipotizzare $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma_{\nu_c}^2 = \sigma_{\nu_s}^2 = N_0 B_N$ possiamo osservare che, in presenza di una fase modulante $\alpha(t)$, la fase $\varphi(t)$ dell'involuppo complesso del segnale ricevuto $r(t)$ è costituita dalla somma di $\alpha(t)$ con l'angolo $\theta(t)$ dovuto al rumore sovrapposto alla portante di ampiezza A . Pertanto all'uscita del discriminatore si ottiene



$$d(t) = \frac{1}{2\pi k_f} \frac{d}{dt} (\alpha(t) + \theta(t))$$

Il rapporto SNR è ora definito come $SNR = \frac{\mathcal{P}_d}{\mathcal{P}_{\nu_d}}$, in cui \mathcal{P}_{ν_d} è la potenza del rumore demodolato, calcolata alla (12.3), e \mathcal{P}_d è la potenza di segnale utile demodolato, pari a $d(t) = \frac{1}{2\pi k_f} \frac{d}{dt} \alpha(t)$. Sappiamo che $\alpha(t) = 2\pi k_f \int_{-\infty}^t m(\tau) d\tau$, e pertanto la potenza di segnale utile demodolato risulta proprio pari a $\mathcal{P}_d = \mathcal{P}_m = \int_{-W}^W \mathcal{P}_m(f) df$. Quindi:

$$SNR = \frac{\mathcal{P}_d}{\mathcal{P}_{\nu_d}} = \frac{\mathcal{P}_m}{\frac{2}{3} \frac{N_0}{(k_f A)^2} W^3} = 3 \frac{\mathcal{P}_m k_f^2}{W^2 N_0 W} \frac{A^2}{2} = 3 \frac{\sigma_{f_d}^2}{W^2} \frac{\mathcal{P}_x}{N_0 W} = 3\beta^2 SNR_0$$

Il risultato ottenuto conferma la seconda affermazione di inizio sezione: si ha un *miglioramento* rispetto all' SNR di riferimento SNR_0 (e dunque rispetto all'AM) *tanto maggiore* quanto più è grande l'indice di modulazione β , ovvero *quanto maggiore* è la *banda occupata* dal segnale modulato.

Discussione dei passaggi Innanzi tutto osserviamo che $\frac{A^2}{2} = \mathcal{P}_x$ (la potenza ricevuta non cambia, ed è sempre uguale a quella della portante non modulata) e che $\frac{\mathcal{P}_x}{N_0 W} = SNR_0$, ovvero il rapporto tra potenza ricevuta e potenza di rumore *nella banda del messaggio*. Per mostrare che $\mathcal{P}_m k_f^2 = \sigma_{f_d}^2$, indichiamo con $f_d(t) = f_i(t) - f_0$ la *deviazione* della frequenza istantanea $f_i(t)$ rispetto ad f_0 , ovvero la differenza tra la derivata della fase istantanea diviso 2π , e la frequenza portante:

$$f_d(t) = \frac{1}{2\pi} \frac{d}{dt} \left(2\pi f_0 t + 2\pi k_f \int_{-\infty}^t m(\tau) d\tau \right) - f_0 = (f_0 + k_f m(t)) - f_0 = k_f m(t)$$

Pertanto si ha $\sigma_{f_d}^2 = k_f^2 \sigma_m^2 = k_f^2 \mathcal{P}_m$ se $m(t)$ è a media nulla: praticamente, σ_{f_d} rappresenta la deviazione standard della frequenza istantanea, e per questo è una

grandezza rappresentativa della larghezza di banda del segnale modulato. Infine, il rapporto $\frac{\sigma_{fd}}{W} = \beta$ è identificato con l'indice di modulazione β , in quanto rappresenta appunto una misura del rapporto tra l'occupazione di banda del segnale modulato, e la massima frequenza W presente nel segnale modulante.

Discussione del risultato $SNR = 3\beta^2 SNR_0$. Notiamo innanzitutto che se $\beta < \frac{1}{\sqrt{3}} = 0,57$ non si ha miglioramento, anzi si peggiora. Ma con bassi indici di modulazione abbiamo già visto che FM ha un comportamento che può avvicinarsi a quello lineare dell'AM, e dunque ci possiamo non-sorprendere. D'altra parte, SNR può migliorare (e di molto) con $\beta > \frac{1}{\sqrt{3}}$: ad esempio, se $\beta = 5 \Rightarrow 3\beta^2 = 75$ volte meglio, ovvero 17,75 dB di miglioramento! In compenso, la regola di Carson ci dice che la banda occupata aumenta di circa $2(\beta + 1) = 12$ volte quella di banda base... dunque il miglioramento di SNR avviene a spese dell'occupazione di banda⁵. Ma ad esempio, non è una cosa tanto grave nel caso di trasmissioni via satellite, in quanto c'è riuscito di frequenze in diversità di spazio.

Potrebbe ora sembrare che si possa aumentare indefinitamente β (nei limiti della banda disponibile) per migliorare a piacere l' SNR . In realtà non è così, dato che ad un certo punto l'analisi effettuata *perde validità*. Questo accade perché se β è troppo elevato, occorre che la banda di rumore del ricevitore sia più ampia (essendo aumentata la banda del segnale modulato) e perciò non si verifica più che $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma_{v_c}^2 = \sigma_{v_s}^2 = N_0 B_N$. Le conseguenze di questo fatto sono illustrate alla sottosezione successiva.

Esercizio

Sia dato un trasmettitore FM con potenza trasmessa 1 Watt e segnale modulante $m(t)$ con banda $\pm B = \pm 10$ MHz. Un collegamento con attenuazione disponibile $A_d = 100$ dB lo interfaccia ad un ricevitore con temperatura di sistema $T_{ei} = 2900$ °K. Desiderando un $SNR = 40$ dB, calcolare:

- 1) Il fattore di rumore del ricevitore in dB;
- 2) Il minimo valore dell'indice di modulazione e la banda occupata a radiofrequenza B_{RF} ;
- 3) Se il valore di β trovato in 2) non sia troppo piccolo, e quale sia il suo massimo valore;
- 4) Il nuovo valore β' , volendo dotare il collegamento di un margine pari a 25 dB.

Soluzione

- 1) Questa domanda va affrontata dopo lo studio del capitolo 16, dove è mostrato che $T_{ei} = T_0(F - 1) + T_g = T_0 F$ se $T_g = T_0$; assumiamo quest'ipotesi per vera e dunque $F = \frac{T_{ei}}{T_0} = 10$; pertanto $F_{dB} = 10$ dB. Per proseguire l'esercizio con le nozioni fin qui acquisite, esplicitiamo che $\mathcal{P}_n(f) = \frac{N_0}{2} = \frac{1}{2} k T_{ei} = \frac{1}{2} \cdot 1.38 \cdot 10^{-23} \cdot 2900 \simeq 2 \cdot 10^{-20}$ Watt/Hz.
- 2) $SNR = 3\beta^2 SNR_0 = 3\beta^2 \frac{W_R}{N_0 W} = 3\beta^2 \frac{W_T G_d}{N_0 B}$; il valore numerico di SNR risulta $10^{\frac{SNR_{dB}}{10}} = 10^4$, mentre quello di A_d è $10^{\frac{A_d(dB)}{10}} = 10^{10}$ e quindi $G_d = 1/A_d = 10^{-10}$. Sostituendo i valori, ed invertendo la relazione, si ottiene $\beta_{min} = \sqrt{\frac{SNR \cdot N_0 B}{3 \cdot W_T G_d}} = \sqrt{\frac{10^4 \cdot 4 \cdot 10^{-20} \cdot 10^7}{3 \cdot 10^{-10}}} = 3.65$. Applicando la regola di Carson per la banda: $B_{RF} \simeq 2B \cdot (\beta + 1) = 2 \cdot 10^7 \cdot 4.65 = 9.3 \cdot 10^7 = 93$ MHz.

⁵In questo senso, il fenomeno può essere visto come una manifestazione del *compromesso velocità-distorsione*, vedi § 17.1.2.

- 3) Perché l'analisi svolta abbia valore, deve risultare $W_R \gg \sigma_{n_c}^2 = \sigma_{n_s}^2 = N_0 B_N = N_0 B_{RF} = 4 \cdot 10^{-20} \cdot 9.3 \cdot 10^7 = 3.72 \cdot 10^{-12}$ Watt, ma poiché $W_R = \frac{W_T}{A_d} = \frac{1}{10^{10}} = 10^{-10}$, si ha $\frac{W_R}{\sigma_{n_c}^2} = \frac{10^{-10}}{3.72 \cdot 10^{-12}} = 26$. Il valore di 26 soddisfa quindi pienamente l'esigenza di *grande segnale*. Per trovare β_{Max} , scriviamo allora $\beta_{Max} \Rightarrow W_R = 10 \cdot \sigma_{n_c}^2 = 10 \cdot N_0 \cdot B_{RF} = 10 \cdot N_0 \cdot 2B \cdot (\beta_{Max} + 1)$, e dunque $\beta_{Max} = \frac{W_R}{10 \cdot N_0 \cdot 2B} - 1 = \frac{10^{-10}}{8 \cdot 10^{-12}} - 1 = 12.5 - 1 = 11.5$, a cui corrisponde una banda $B_{RF} = 2B \cdot (\beta_{Max} + 1) = 2 \cdot 10^7 \cdot 12.5 = 250$ MHz, ed un guadagno di $SNR = 10 \lg_{10} 3\beta_{Max}^2 \simeq 26$ dB, mentre con β_{min} nominale si sarebbe ottenuto $10 \lg_{10} (3 \cdot 3.65^2) = 16$ dB.
- 4) Il concetto di margine è introdotto al § 15.1; un margine di 25 dB equivale a far fronte ad una attenuazione supplementare $A'_d = 10^{2.5} = 316$ volte. Proviamo ad ottenere lo stesso SNR con un nuovo valore β' : $SNR = 10^4 = 3\beta'^2 \frac{W_T G_d G'_d}{N_0 B} = 3\beta^2 \frac{W_T G_d}{N_0 B} \frac{\beta'^2}{\beta^2} G'_d$; dunque deve risultare $\frac{\beta'^2}{\beta^2} G'_d = 1$ e quindi $\beta'^2 = \beta^2 \sqrt{\frac{1}{G'_d}} = 3.65 \sqrt{316} = 3.65 \cdot 17.7 = 64.88$ non ce la facciamo. Infatti, al più (con $\beta = \beta_{Max} = 11.5$) si ha un margine di 10 dB.

12.3.3 Caso di elevato rumore

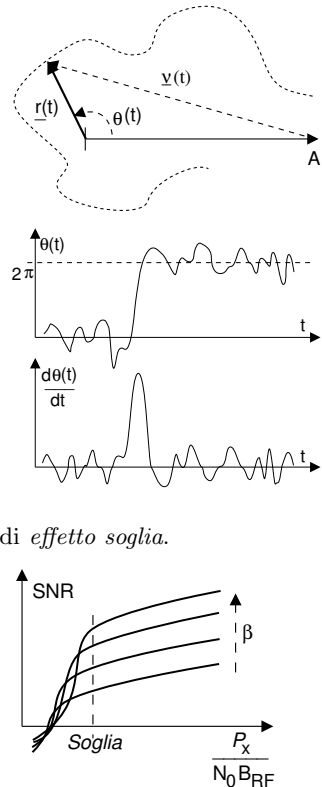
Qualora il valore efficace del rumore in ingresso al discriminatore sia confrontabile con quello del segnale utile ricevuto, si verifica un *effetto soglia* all'aumentare del rumore, e l' SNR degrada molto rapidamente.

Riprendendo lo schema che mostra l'inviluppo complesso della portante non modulata A , del rumore in ingresso $\underline{v}(t)$, e del segnale ricevuto $\underline{r}(t)$, notiamo che se i valori efficaci dei primi due sono comparabili, può verificarsi il caso che $\underline{r}(t)$ ruoti attorno all'origine.

Quando ciò si verifica, a valle del derivatore che è presente nel discriminatore si determina un *click*, ovvero un segnale impulsivo di area pari a 2π , come illustrato nella figura a lato. Questo fatto è facilmente verificabile, ascoltando una radio FM broadcast, che in condizioni di cattiva ricezione manifesta la comparsa di un rumore, appunto, impulsivo.

All'aumentare della potenza di rumore, aumenta la frequenza con la quale $\underline{r}(t)$ "aggira" l'origine, e pertanto aumenta la frequenza dei *click*, che tendono a produrre un crepitio indistinto. Si è trovato che questo effetto si manifesta a partire da un SNR di ingresso pari a 10 dB, e per SNR peggiori di tale valore l'effetto aumenta molto rapidamente, cosicché si parla di *effetto soglia*.

Nel grafico a lato è riportato un tipico andamento dell' SNR dopo demodulazione, con l'indice di modulazione β che svolge il ruolo di parametro, e possiamo osservare come con un SNR di ingresso inferiore alla soglia, le prestazioni degradino rapidamente. Si è trovato che demodulando con un PLL, anziché con un discriminatore, la soglia si riduce a circa 7 dB.



Nella pratica comune il segnale di rumore può essere costituito da una *interferenza* dovuta ad una emittente adiacente (ossia con una portante prossima alla nostra) che *sovramodula*, ovvero adotta un indice di modulazione troppo elevato, ed invade la banda delle emittenti contigue.

12.3.4 Enfasi e de-enfasi

Abbiamo osservato che, in presenza di rumore bianco in ingresso, il rumore dopo demodulazione ha un andamento parabolico. Questo comporta che, se il messaggio modulante $m(t)$ avesse un $\mathcal{P}_m(f)$ a sua volta bianco, l' $SNR(f)$ alle frequenze più elevate sarebbe molto peggiore, rispetto al suo valore per frequenze inferiori. Nella pratica, si possono verificare (ad esempio) i seguenti problemi:

1. Nelle trasmissioni FDM-FM (vedi § 10.1.1.3), in cui più canali vengono modulati AM-BLU, multiplati in frequenza, e ri-modulati congiuntamente in FM a basso indice, i canali agli estremi della banda FDM sono più rumorosi;
2. nell'FM commerciale (vedi § 11.4.6), il segnale modulante è molto più ricco di energia alle basse frequenze, dunque il problema del rumore elevato in alta frequenza è aggravato dal “basso segnale”.

Il rimedio a tutto ciò consiste nel modificare $m(t)$ mediante un circuito detto *di enfasi*, in modo tale che anch'esso presenti uno spettro “parabolico”, e poi aggiungere una rete di *de-enfasi* in ricezione (praticamente un integratore, ovvero un passa-basso) tale da ripristinare l'originale sagoma spettrale del segnale e rendere la densità di potenza del rumore nuovamente uniforme.

Con un po di riflessione, ci si accorge che l'uso di una coppia enfasi-deenfasi equivale ad effettuare una trasmissione PM (vedi pag. 283) ! In realtà, la rete di enfasi non è un derivatore perfetto (altrimenti annullerebbe le componenti del segnale a frequenza prossima allo zero), ed esalta le frequenze solo se queste sono maggiori di un valore minimo. Pertanto, si realizza un metodo di modulazione “misto”, FM in bassa frequenza e PM a frequenze (di messaggio) più elevate.

Capitolo 13

Modulazione numerica

Sono qui discusse le tecniche adottate qualora il segnale da trasmettere in forma modulata non sia analogico, ma costituito da una sequenza numerica, o simbolica. Il contesto applicativo può variare su un ampio ventaglio di casi, come le forme di broadcast numerico (terrestre o satellitare), le reti di accesso WIFI o di telefonia cellulare, i modem e l'ADSL, le comunicazioni satellitari dallo spazio profondo, i ponti radio numerici per flussi dati ottenuti come moltiplicazione temporale di più sorgenti, di tipo multimediale e/o provenienti da reti a pacchetto... in pratica, la gran parte delle comunicazioni dati che non viaggiano su fibra ottica.

In tutti questi casi ci si trova in presenza di un canale trasmissivo di tipo *passa-banda*, quindi inadatto a trasportare un segnale dati realizzato mediante codici di linea di *banda base* (vedi § 5.2), e dunque è necessario produrre un segnale modulato (vedi cap. 10) per trasporre la banda del segnale in accordo ai vincoli imposti dal canale. Ora non vengono però semplicemente applicate le tecniche esposte al cap. 11, ma queste vengono rese specifiche alla caratteristica del segnale dati di essere costituito da sequenze di simboli appartenenti ad un alfabeto finito, da mappare su un insieme finito di punti dello spazio, che per segnali modulati è lo spazio dell'involuppo complesso.

13.1 Modulazione di ampiezza

In questo caso l'informazione numerica è impressa sulla portante alterando le ampiezze di una delle componenti analogiche di b. f.

13.1.1 Modulazione BPSK

E' l'acronimo di *Bi-Phase Shift Keying*¹, e individua una tecnica per il trasporto dell'informazione basata sull'utilizzo di 2 possibili fasi per la portante:

$$x_{BPSK}(t) = a \sin(\omega_0 t + \varphi(t)) \quad \text{dove} \quad \varphi(t) = \sum_{k=-\infty}^{\infty} \varphi_k \text{rect}_{T_b}(t - kT_b) \quad (13.1)$$

con i valori φ_k pari a $\pm \frac{\pi}{2}$ per rappresentare le cifre binarie 0 ed 1 trasmesse agli istanti kT_b . Sebbene l'operazione così definita corrisponda ad una modulazione di fase (§ 11.3), è facile mostrare come possa essere realizzata mediante una comune

¹Letteralmente, *slittamento di tasto a due fasi*.

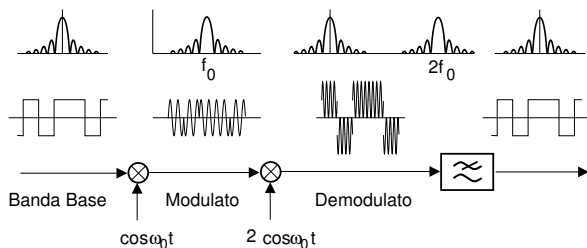


Figura 13.1: Architettura di mo-demodulazione BPSK, forme d'onda, e densità spettrali

modulazione di ampiezza BLD (§ 11.1.1). Se definiamo infatti un segnale $m(t)$ come un codice di linea NRZ bipolare (§ 5.2.1), che assume valori ± 1 in corrispondenza delle cifre binarie 0 ed 1, allora il segnale

$$x_{BPSK}(t) = m(t) \cos \omega_0 t$$

è equivalente al precedente, e la sua mo-demodulazione coerente avviene mediante l'architettura mostrata alla fig. 13.1. Il segnale uscente dal moltiplicatore di demodulazione² ha espressione

$$y(t) = x(t) \cdot 2 \cos \omega_0 t = 2m(t) \cdot \cos^2 \omega_0 t = m(t) + m(t) \cdot \cos 2\omega_0 t$$

e dunque il codice di linea $m(t)$ può essere riottenuto mediante filtraggio passa-basso.

Una buona caratteristica di questa tecnica è l'andamento *costante* dell'ampiezza della portante modulata, che permette di utilizzare la massima potenza del trasmettitore, appena inferiore al valore che inizia a produrre fenomeni di distorsione (§ 14.7). L'aspetto negativo è l'elevata occupazione di banda³, dovuta all'uso di forme d'onda rettangolari per $m(t)$ che, nel caso di trasmissione su canali con limitazioni di banda, causano vincoli sulla massima frequenza binaria. Per questi motivi, il metodo è particolarmente indicato nel caso di collegamenti in cui la potenza di trasmissione è limitata, ma non la banda⁴.

Gli aspetti ora illustrati possono essere *sovertiti* qualora il segnale $m(t)$ sia generato utilizzando forme d'onda $g(t)$ con occupazione di banda limitata, come ad esempio la famiglia a coseno rialzato (§ 5.2.2.3). In quest'ultimo caso la banda occupata a frequenze positive risulta pari a $f_b(1 + \gamma)$, doppia rispetto al caso di banda base, a causa della modulazione BLD, e l'ampiezza del segnale modulato *non è più costante*. Mentre infatti in corrispondenza degli istanti kT_b l'ampiezza di $x_{BPSK}(t)$ assume esattamente uno dei valori (± 1) del segnale dati $m(t)$, nell'intervallo tra due istanti $kT_b < t < (k + 1)T_b$ questa dipende della somma di tutte le code delle funzioni $g(t)$ relative ai simboli trasmessi (vedi fig. 5.5 a pag. 84).

²Qui e nel seguito assumiamo di disporre di una portante di demodulazione omodina o coerente (§ 11.2.1), ossia priva di errori di fase e frequenza, così come di una perfetta temporizzazione di simbolo; le considerazioni al riguardo di quest'ultimo aspetto sono svolte all'appendice 13.10.1.

³Come discusso al § 11.1.1, la modulazione AM-BLD produce la traslazione della densità spettrale del segnale modulante alla frequenza portante f_0 . A sua volta (§ 5.1.2.1), in base all'eq. (5.2) il segnale NRZ ha densità spettrale di tipo $\text{sinc}^2(f\tau)$, il cui andamento in dB è mostrato in fig. 3.5, riportata in fig. 13.1.

⁴Come ad esempio i collegamenti satellitari.

13.1.2 Modulazione L-ASK

Ci riferiamo ora al caso in cui si operi una classica AM-BLD (da cui il termine *Amplitude Shift Keying* - ASK) a partire da un segnale dati $m(t)$ multilivello (§ 5.1.2.4), producendo un segnale modulato di espressione

$$x_{L-ASK}(t) = m(t) \cos(2\pi f_0 t) \quad \text{dove} \quad m(t) = \sum_{k=-\infty}^{\infty} a_k \cdot \text{rect}_{T_s}(t - kT_s)$$

in cui $m(t)$ agli istanti kT_s assume valori a_k distribuiti uniformemente, in un intervallo Δ , su L livelli di ampiezza centrati sullo zero⁵.

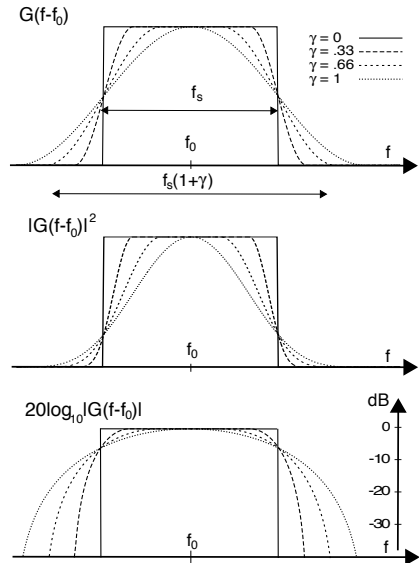
L'ampiezza di L-ASK subisce dunque variazioni, come mostrato nella figura a fianco per un caso con $L = 8$, in cui è rappresentato anche un diagramma



detto *costellazione*, che rappresenta i valori assunti dall'involuppo complesso in corrispondenza degli istanti di simbolo, che in virtù della AM-BLD presenta la sola c.a. di b.f. $x_c(t)$. Ogni a_k rappresenta dunque $M = \log_2 L$ bit, ed il periodo di simbolo $T_s = MT_b$ ha durata multipla di T_b , pertanto la banda occupata da L-ASK è *minore* rispetto a quella del BPSK di un fattore pari a $M = \log_2 L$ ⁽⁶⁾.

Anche qui, nel caso in cui $m(t)$ sia generato mediante forme d'onda con caratteristica a coseno rialzato (§ 5.2.2.3) anziché con un codice di linea NRZ, la densità spettrale assume il noto andamento (fig. 5.4) riportato a fianco, assieme al suo andamento in dB, e la banda a frequenze positive occupata da $x_{L-ASK}(t)$ risulta pari a

$$B_{L-ASK} = f_s(1 + \gamma) = \frac{f_b}{\log_2 L} (1 + \gamma) \tag{13.2}$$



Efficienza spettrale o densità di informazione E' definita come il rapporto ρ tra la frequenza binaria e la banda occupata

$$\rho = \frac{f_b}{B}$$

e si esprime in *bit/sec/Hz*, rappresentando appunto quanti bit/sec sono trasmessi per ogni Hz utilizzato. Nel caso di L-ASK con impulsi a banda minima ($\gamma = 0$) si trova allora $\rho_{L-ASK} = \frac{f_b}{B} = \log_2 L$, mentre per altre forme di modulazione e/o di impulsi si ottengono altri valori, confrontando i quali si valuta la bontà di un metodo rispetto all'altro nei termini dell'utilizzo di banda. Ad

⁵Per chi si sta chiedendo quanto valgono questi livelli, diciamo che il livello i -esimo (con $i = 0, 1, \dots, L-1$) corrisponde al valore $a^i = i \cdot \frac{\Delta}{L-1} - 1$. Verificare per esercizio con $\Delta = 2$ ed $L = 4$.

⁶Ad esempio: se $L = 32$ livelli, la banda si riduce di 5 volte, ed infatti con $M = 5$ bit si individuano $L = 2^M = 32$ configurazioni. Dato che il numero M di bit/simbolo deve risultare un intero, si ottiene che i valori validi di L sono le potenze di 2.

esempio, se paragoniamo il risultato per ρ_{L-ASK} con quello relativo ad una trasmissione numerica di banda base (vedi eq. (5.7)), notiamo un *peggioramento* di un fattore 2, dovuto all'uso di una AM-BLD. Come per il caso analogico, la banda può essere dimezzata adottando una AM-BLU, ma vedremo invece tra breve che si preferiscono seguire approcci diversi, come ad esempio PSK e QAM.

Natura di E_b/N_0 L'analisi delle prestazioni delle tecniche di modulazione numerica che affronteremo esprime la probabilità di errore per simbolo P_e (*simbolo*) o per bit P_e (*bit*) in funzione della grandezza $\frac{E_b}{N_0}$, introdotta a pag. 5.3.3, che rappresenta l'equivalente del rapporto segnale rumore *di riferimento* $SNR_0 = \frac{P_x}{N_0 W}$ definito al §12.2.1.1⁷, e quindi consente il confronto tra tecniche diverse. Infatti, considerando che la potenza ricevuta può essere espressa come $P_x = \frac{E_b}{T_b} = E_b \cdot f_b$, si ottiene

$$SNR_0 = \frac{P_x}{N_0 f_b} = \frac{E_b}{N_0} \quad (13.3)$$

che dipende unicamente da grandezze che identificano la qualità del collegamento (P_x ed N_0) e la velocità del messaggio (f_b) senza alcun riferimento al metodo di modulazione. A pag. 5.3.3 si deriva la relazione esatta tra $\frac{E_b}{N_0}$ e SNR *effettivo* per un segnale dati a coseno rialzato e multilivello.

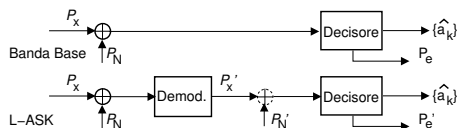
13.1.3 Prestazioni di L-ASK

Valutiamo ora la probabilità di errore per un segnale L-ASK in funzione di E_b/N_0 , al variare del numero di livelli. Abbiamo osservato al § 13.1.2 come l'L-ASK sia ottenibile come modulazione AM BLD di un segnale dati di banda base, ed al § 12.2.1.1 si è mostrato che nel caso AM-BLD, dopo demodulazione, risulta $SNR = SNR_0$. Pertanto in base alla (13.3) si ottiene che le prestazioni per un segnale multilivello a coseno rialzato di banda base sono le stesse⁸ di quelle ricavate al § 5.3.5, ovvero una probabilità di errore *per simbolo* pari a

$$P_e^{L-ASK}(\text{simbolo}) = \left(1 - \frac{1}{L}\right) \operatorname{erfc} \left\{ \sqrt{3 \frac{E_b}{N_0} \frac{\log_2 L}{L^2 - 1}} \right\} \quad (13.4)$$

⁷ricordiamo che P_x esprime la potenza ricevuta, N_0 rappresenta il doppio della $\mathcal{P}_n(f)$ presente al decisore, e W è la banda del segnale modulante.

⁸Forniamo qui dimostrazione forse inutilmente elaborata. Con riferimento alla figura seguente, il calcolo della P_e per l'L-ASK si imposta definendo valori di E_b ed N_0 equivalenti a quelli di banda base, ma ottenuti dopo demodulazione, e cioè $E'_b = P'_x T_b$ e $N'_0 = P'_N / W$ (infatti, $P'_N = \frac{N'_0}{2} 2W$, con $W = \frac{f_s}{2} = \frac{f_b}{2 \log_2 L}$).



L'equivalenza è presto fatta, una volta tarato il demodulatore in modo che produca in uscita la componente in fase $x_c(t)$ limitata in banda tra $\pm W$. Infatti in tal caso (vedi § 12.2.1) $P'_x = P_{x_c} = k_a^2 P_M = 2P_x$ e quindi $E'_b = P'_x T_s = 2P_x T_s = 2E_b$; per il rumore si ottiene $N'_0 = \frac{P'_N}{W}$ in cui $P'_N = P_{n_c} = \sigma_{n_c}^2 = \sigma_n^2 = \frac{N_0}{2} 4W$ e quindi $N'_0 = 2N_0$. Pertanto, il valore E'_b/N'_0 su cui si basa ora il decisore è lo stesso E_b/N_0 in ingresso al demodulatore.

e che si riferisce ad un segnale *a coseno rialzato* con $\gamma = 0$ ⁹, ossia a *banda minima* (§ 5.2.2.3), e pertanto le curve di P_e (*bit*) sono quelle di fig. 5.8 a pag. 92, dove si tiene anche conto dell'uso di un codice di Gray (§ 5.2.2.4) per associare i livelli a configurazioni binarie.

Osserviamo esplicitamente che per $L = 2$ la (13.4) si riduce a

$$P_e^{BPSK}(\text{bit}) = \frac{1}{2} \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0}} \right\} \quad (13.5)$$

che rappresenta le prestazioni ottenibili per la modulazione BPSK, graficate in fig. 13.2, ed identica alla (5.16) ottenuta per il caso di banda base, ed alla (9.12) relativa al filtro adattato. Per completare i confronti, osserviamo che ora all'aumentare di L la banda (13.2) (per $\gamma = 0$)

$$B_{L-ASK} = f_s = \frac{f_b}{\log_2 L}$$

si riduce, mentre la P_e (13.4) aumenta: ciò può tornare utile in presenza di canali con limitazioni di banda ma non di potenza, dato che in tal caso la P_e può essere ripristinata aumentando la potenza e quindi E_b/N_0 , in base al cosiddetto *compromesso banda-potenza*. Al § 13.5.1 vedremo come nella tecnica di FSK ortogonale lo stesso compromesso operi in direzione opposta, ovvero riuscendo a migliorare P_e aumentando la banda occupata.

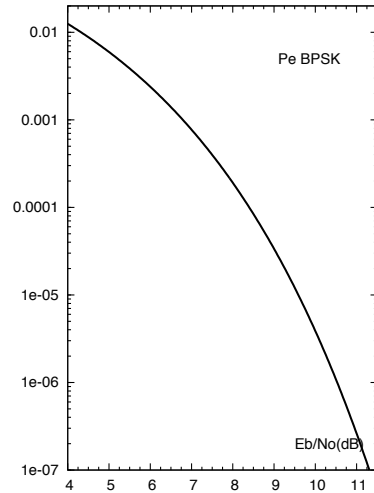


Figura 13.2: Prestazioni BPSK
Al § 13.5.1 vedremo come nella tecnica di FSK ortogonale lo stesso compromesso operi in direzione opposta, ovvero riuscendo a migliorare P_e aumentando la banda occupata.

13.2 Modulazione di fase

Prendiamo ora in esame il caso in cui l'informazione viene rappresentata dalla fase della portante.

13.2.1 Modulazione QPSK ed L-PSK

Questi due acronimi si riferiscono alla possibilità di usare *quattro* oppure $L > 4$ scelte diverse¹⁰ per la fase, dando luogo ad un segnale modulato con espressione

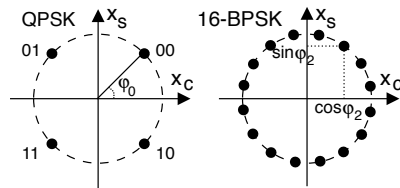
$$x_{L-PSK}(t) = a \cos(\omega_0 t + \varphi(t))$$

e quindi un inviluppo complesso

$$\underline{x}_{L-PSK}(t) = a e^{j\varphi(t)} = a \cos \varphi(t) + ja \sin \varphi(t) \quad (13.6)$$

in cui

$$\varphi(t) = \sum_{k=-\infty}^{\infty} \varphi_k \operatorname{rect}_{T_s}(t - kT_s) \quad \text{e} \quad \varphi_k \in \{\varphi_0, \varphi_1, \dots, \varphi_{L-1}\} \quad (13.7)$$



⁹Se $\gamma \neq 0$, valgono le considerazioni svolte al § 5.3.5.

¹⁰Il caso in $L = 2$ ricade nel BPSK già discusso

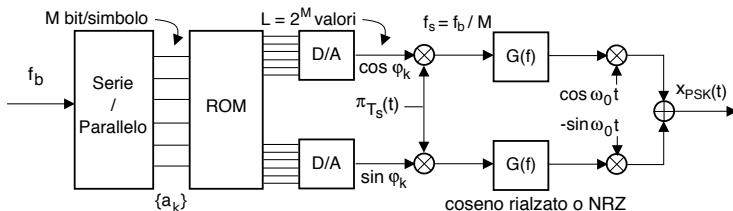


Figura 13.3: Modulatore L-PSK

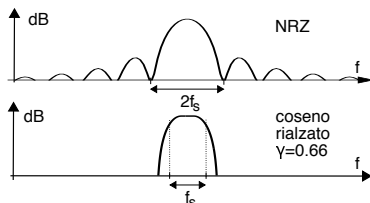
La generica fase $\varphi_i = \frac{\pi}{L} + i \cdot \frac{2\pi}{L}$ con $i = 0, 1, \dots, L - 1$ rappresenta una delle $L = 2^M$ possibili combinazioni di M bit di ingresso, e corrisponde ad uno dei punti mostrati nelle *costellazioni* di figura, che individuano il valore dell'involucro complesso ricevuto in assenza di rumore negli istanti di simbolo $t = kT_s$. Lo stesso valore di fase è quindi mantenuto per tutto il periodo di simbolo se $\varphi(t)$ è realizzata mediante rettangoli come indicato nella (13.7). Ma l'espressione (13.6) di $\underline{x}_{L-PSK}(t)$ in termini di $\{x_c, x_s\}$ evidenzia come il risultato sia ottenibile mediante una modulazione AM in fase e quadratura¹¹, suggerendo l'implementazione del ricevitore secondo lo schema di fig. 13.3, in cui i valori di $\cos \varphi_i$ e $\sin \varphi_i$ per gli L diversi gruppi di M bit sono precalcolati in una memoria a sola lettura, ed usati come ampiezze per realizzare due segnali dati di banda base, usati quindi come c.a di b.f. per il modulatore in fase e quadratura.

Occupazione di banda L'uso di un codice NRZ per $\varphi(t)$, e quindi per x_c ed x_s , produce una occupazione di banda elevata per $\underline{x}_{L-PSK}(t)$, la cui densità di potenza in tal caso acquisisce un andamento $(\frac{\sin x}{x})^2$ centrato in f_0 e con lobo principale di estensione¹² pari ad $2f_s = 2f_b/M$, come rappresentato in figura per una densità in dB.

L'occupazione di banda può essere limitata a $B_{L-PSK} = f_s(1 + \gamma)$ se si realizza $\varphi(t)$ mediante impulsi di Nyquist $g(t)$ a coseno rialzato, potendo così scrivere

$$\underline{x}_{L-PSK}(t) = a \sum_{k=-\infty}^{\infty} e^{j\varphi_k} \cdot g(t - kT_s)$$

Traiettoria di $\underline{x}(t)$ Contrariamente al caso di $g(t)$ rettangolare¹³, adottando una $g(t)$ a coseno rialzato $\underline{x}(t)$ passa dai punti della costellazione solo negli istanti significativi, mentre nell'intervallo tra due istanti segue traiettorie con ampiezza variabile, non rispettando quindi la proprietà di ampiezza costante che una modulazione angolare dovrebbe avere. Pertanto la scelta tra NRZ e coseno rialzato dipende dalla necessità di limitare la dinamica delle ampiezze, oppure l'estensione della banda. In fig. 13.4 si mostrano x_c ed x_s per una modulazione QPSK valutate con $g(t)$ a coseno rialzato ($\gamma = 0.5$) per 10 campioni per simbolo, e mostrate per 10 simboli in coordinate cartesiane



¹¹che non è una modulazione AM-BLU dato che $x_s \neq \hat{x}_c$

¹²Infatti un rettangolo di durata T_s ha trasformata $\text{sinc}(T_s f)$, con il primo zero in $f = 1/T_s = f_s$, e la modulazione AM-BLD produce un raddoppio della banda occupata.

¹³Se $g(t) = \text{rect}_{T_s}(t)$, $|\underline{x}|$ giace su di un cerchio, spostandosi *istantaneamente* da un punto all'altro della costellazione

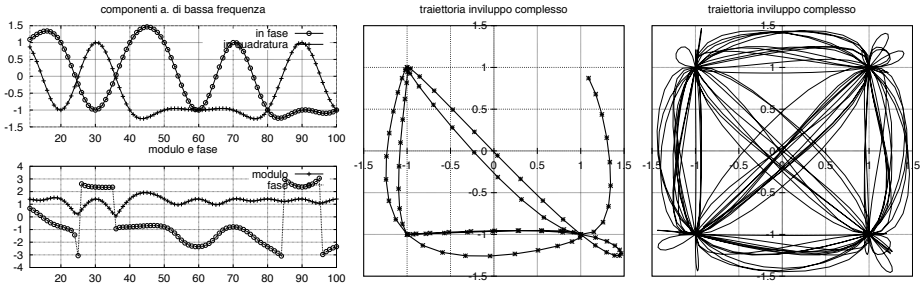


Figura 13.4: x_c ed x_s per QAM in coordinate cartesiane e polari, traiettoria di $\underline{x}(t)$ per 10 e 100 simboli

e polari; al centro è mostrato il corrispondente $\underline{x}(t)$, mentre a destra si mostra la traiettoria di $\underline{x}(t)$ visualizzata per 100 simboli.

Efficienza spettrale Per L-PSK l'efficienza spettrale è identica a quanto ottenuto per l'ASK con ugual numero di livelli, dato che per entrambe la frequenza di simbolo risulta pari a $f_s = \frac{f_b}{\log_2 L}$, e dunque (per coseno rialzato con $\gamma = 0$) si ottiene $\rho_{L-PSK} = \frac{f_b}{B} = \frac{f_s \log_2 L}{f_s} = \log_2 L$.

Dal punto di vista delle prestazioni, occorre distinguere il caso in cui $L = 4$ (indicato come QPSK = *Quadrature Phase Shift Keying*) da quello con L generico, in quanto sussistono due diverse architetture per il demodulatore.

13.2.2 Prestazioni QPSK

In questo caso (PSK con 4 livelli) il demodulatore è costituito da due rami indipendenti in fase e quadratura (vedi fig. 13.5), operanti a frequenza di simbolo f_s metà di quella binaria, in ognuno dei quali si effettua una decisione per uno dei due bit che compongono il simbolo, e quindi le due decisioni vengono re-serializzate. Entrambi i rami si comportano pertanto come un demodulatore L-ASK (§ 13.1.2) con $L=2$, ovvero (a parte una rotazione di fase) un BPSK, e dunque per un segnale a coseno rialzato la probabilità di errore relativa ad ogni singolo ramo è espressa¹⁴ dalla (13.5), fornendo

¹⁴In effetti, dovremmo mostrare che l'attuale valore di E_b/N_0 è lo stesso del caso BPSK: ma mentre per N_0 è evidente che si ha lo stesso valore, sembra che il valore di E_b si dimezzi. Infatti, a parità di potenza ricevuta, i punti di costellazione del BPSK giacciono all'intersezione tra l'asse cartesiano della c.a di b.f. ed il cerchio di raggio pari all'ampiezza a del segnale ricevuto, mentre nel QPSK

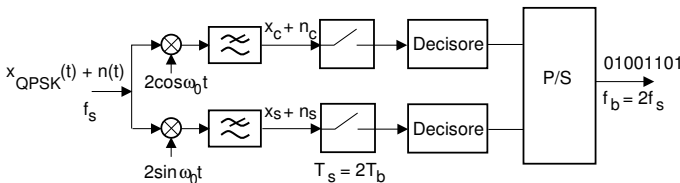


Figura 13.5: Demodulatore QPSK

(per $\gamma = 0$):

$$P_e^c = P_e^s = \frac{1}{2} \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0}} \right\}$$

che rappresenta la probabilità che una componente analogica di b. f. ottenuta demodulando il rumore sovrapposto al segnale, valutata all'istante di decisione kT_s , giaccia nell'area mostrata in figura 13.6. La probabilità di errore in un bit della sequenza re-serializzata risulta quindi¹⁵

$$P_e^{QPSK} (bit) = P_e^c \cdot Pr\{c\} + P_e^s \cdot Pr\{s\} = \frac{1}{2} (P_e^c + P_e^s) = \frac{1}{2} \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0}} \right\}$$

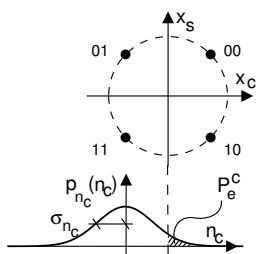


Figura 13.6: Costellazione QPSK e calcolo della P_e^c

in cui $Pr\{c\} = Pr\{s\} = 1/2$ sono le probabilità che il bit ricevuto provenga dal ramo in fase o da quello in quadratura; d'altra parte, la probabilità di errore per simbolo risulta $P_e (simbolo) = P_e^c + P_e^s = \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0}} \right\}$ (di nuovo trascurando la probabilità di un errore contemporaneo su entrambi i rami, vedi nota precedente).

Osserviamo quindi come il QPSK consenta di ottenere le stesse prestazioni del BPSK, graficate in fig. 13.2, utilizzando solo metà banda ($\gamma = 0$):

$$B_{QPSK} = f_s = \frac{f_b}{2}$$

Quest'ultima osservazione permette di scrivere $\rho_{QPSK} = \frac{f_b}{B_{QPSK}} = 2 = 2 \cdot \rho_{BPSK}$, e suggerisce un ulteriore¹⁶ punto di vista rispetto all'invarianza delle prestazioni: infatti, dimezzando la banda si dimezza anche la varianza del rumore in ingresso al demodulatore, e ciò compensa la riduzione di ampiezza delle c. a. di b. f. del segnale per il caso QPSK.

13.2.3 Prestazioni L-PSK

In questo caso il demodulatore ha una differente architettura (vedi fig. 13.7), ed il decisore opera congiuntamente su entrambi i rami, per ottenere la stima del gruppo di $M = \log_2 L$ bit associati ad una delle possibili fasi φ_k . Indicando con $\tilde{x}_{c,s}$ le c.a di b.f. ricevute, la decisione avviene calcolando $\hat{\varphi}_k = \arctan \frac{\sin \hat{\varphi}_k}{\cos \hat{\varphi}_k}$ e stabilendo all'interno di quale regione di decisione $\hat{\varphi}_k$ cada la fase ricevuta $\hat{\varphi}_k$. All'aumentare del numero di livelli L , la potenza di rumore (che concorre alla probabilità di errore) diminuisce con la stessa legge di riduzione della banda, ovvero con il $\log_2 L$. Al contrario, la spaziatura tra le regioni di decisione diminuisce con legge lineare rispetto ad L ; pertanto,

(fig. 13.6) le fasi formano un angolo di 45° rispetto agli assi, riducendone di $\sqrt{2}$ la distanza dallo zero, e riducendo dunque la potenza della c.a di b.f. di un fattore 2, e così pure il valore E_b . In realtà, la durata doppia del periodo di simbolo T_s compensa questa riduzione, e dunque $E_b = P_x T_b$ si mantiene invariato.

¹⁵In effetti all'istante di decisione potrebbe verificarsi errore su entrambi i rami, ma tale evento avviene con probabilità $P_e^{(c)} \cdot P_e^{(s)}$ che si ritiene $\ll P_e^{(c)}$, e quindi trascurabile.

¹⁶Ulteriore rispetto al commento alla nota 14, dove il ragionamento è svolto sull' E_b/N_0 , mentre ora sull' SNR .

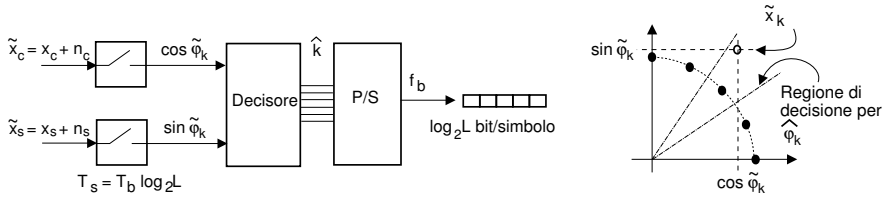


Figura 13.7: Demodulatore L-PSK

l'aumento del numero di livelli produce un peggioramento della P_e . Senza approfondire i relativi conti, forniamo direttamente il risultato (con $\gamma = 0$) della probabilità di errore sul simbolo

$$P_e^{L-PSK}(\text{simbolo}) = \text{erfc} \left\{ \sin \left(\frac{\pi}{L} \right) \sqrt{\frac{E_b}{N_0} \log_2 L} \right\} \quad (13.8)$$

che rappresenta la probabilità di decidere di aver ricevuto un $\hat{\varphi}_k \neq \varphi_k$ (diverso da quello trasmesso) e che, se $P_e \ll 1$, è approssimata con la probabilità di invadere (a causa del rumore) una regione di decisione contigua (vedi fig. 13.7).

Confrontando il risultato con quello (eq. 13.4) per l'ASK, osserviamo che l'assenza del termine $(1 - \frac{1}{L})$ è dovuto alla *circularità* della costellazione, che il termine $\sin(\frac{\pi}{L})$ è un fattore che rappresenta il peggioramento all'aumentare di L , ed il $\log_2 L$ sotto radice è il miglioramento dovuto alla riduzione di banda. Il risultato (13.8) è una approssimazione valida se $P_e \ll 1$, e via via più accurata con L crescente.

Nella tabella a fianco è riportato il risultato del confronto, per uno stesso valore di P_e , dei valori $\frac{E_b}{N_0}$ necessari per L-PSK (13.8), contro quelli necessari (13.4) per L-ASK: si è

eseguito il rapporto tra gli argomenti degli $\text{erfc}\{\}$, si è elevato al quadrato, indicato come Δ , ed il risultato è espresso in dB: quest'ultimo rappresenta (a parte il termine $(1 - \frac{1}{L})$ dell'L-ASK) il miglioramento di prestazioni in dB dell'L-PSK rispetto ad L-ASK, ovvero i dB di potenza risparmiata a parità di probabilità di errore. Tale risultato (4-5 dB di miglioramento) porta a prediligere sempre il PSK rispetto all'ASK.

E' opportuno osservare che, qualora si desideri ottenere un valore di probabilità di errore *per bit*, questo è pari a

$$P_e(\text{bit}) = \frac{P_e(\text{simbolo})}{\log_2 L}$$

a patto di associare, a livelli contigui, gruppi di bit differenti in una sola posizione, come previsto dal codice di Gray¹⁷ (mostrato nella figura 13.8), in modo che l'errore tra due livelli provochi l'errore di un solo bit nel gruppo di $\log_2 L$ bit associati a ciascun livello. Le curve di probabilità di errore per bit, riportate anch'esse in fig. 13.8, sono calcolate in questo modo.

¹⁷vedi il § 5.2.2.4

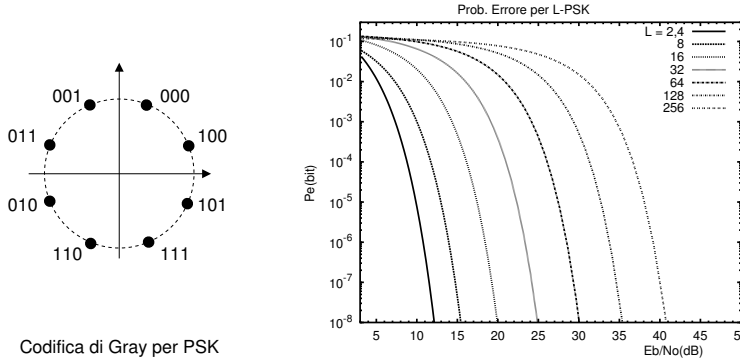


Figura 13.8: Prestazioni L-PSK con codice di Gray

13.3 Modulazione QAM

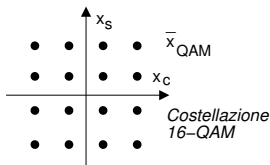
L'acronimo sta per *Quadrature Amplitude Modulation*, ed individua la tecnica di modulazione che utilizza due portanti in quadratura come il PSK

$$x_{QAM}(t) = x_c(t) \cos \omega_0 t - x_s(t) \sin \omega_0 t$$

ma a differenza del PSK, ora le componenti di banda base x_c ed x_s *non* dipendono da una stessa sequenza di fasi, ma sono originate da due flussi di dati distinti.

Con riferimento alla fig. 13.9, osserviamo che sebbene $x_c(t)$ e $x_s(t)$ si ottengano a partire da una medesima sequenza numerica $\{a_k\}$, i bit di quest'ultima sono distribuiti alternativamente sui due rami (sequenze $b_c[n]$ e $b_s[n]$ in fig.) a velocità dimezzata¹⁸, suddividendo un gruppo di M bit in due simboli costituiti da $M' = M/2$ bit. Dalle sequenze b_c e b_s ad M' bit/simbolo si ottengono quindi i valori c_c e c_s con $L' = 2^{M'} = 2^{M/2} = \sqrt{2^M} = \sqrt{L}$ livelli, che attraversando il filtro $G(f)$, danno luogo ai segnali di banda base x_c ed x_s .

La sequenza di operazioni descritte determina una costellazione *quadrata*, composta da $L = (L')^2$ punti, che rappresentano le coordinate (nel piano dell'involuppo complesso) in cui \underline{x} è forzato a transitare in corrispondenza degli istanti di Nyquist multipli del periodo di simbolo T_s , che risulta essere pari a



$$T_s = T_b \cdot M = \frac{1}{f_b} \log_2 L$$

Occupazione di banda Se $G(f)$ è a coseno rialzato con roll off γ , allora la banda a frequenze positive di x_c ed x_s risulta pari a $\frac{f_s}{2} (1 + \gamma) = \frac{f_b}{2 \log_2 L} (1 + \gamma)$, mentre quella di x_{QAM} è pari al doppio, a causa della modulazione AM-BLD-PS operata sui due rami del modulatore, ovvero

$$B_{QAM} = \frac{f_b}{\log_2 L} (1 + \gamma) = f_s (1 + \gamma)$$

¹⁸In pratica, l'indice n si incrementa ogni due incrementi dell'indice k .

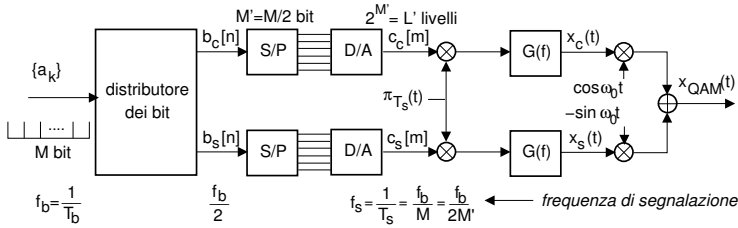


Figura 13.9: Modulatore QAM

e quindi uguale al PSK con uguale numero di livelli (di cui condivide quindi anche l'efficienza spettrale).

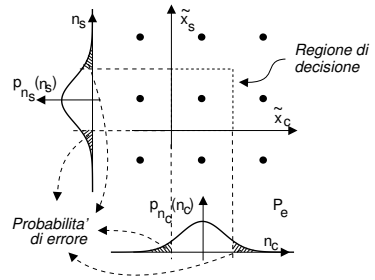
Notiamo che, per come si è impostato lo schema di distribuzione dei bit tra i rami, L deve risultare un quadrato perfetto. Nulla impedisce di elaborare schemi più complessi in cui L' è diverso per i due rami, o più in generale la cui costellazione non sia quadrata. E' appena il caso di notare che l'ampiezza del segnale modulato varia notevolmente tra un simbolo e l'altro.

13.3.1 Prestazioni di QAM

Notiamo subito che la "distanza" tra due punti della costellazione QAM è maggiore (a parità di L) del caso PSK; pertanto, c'è da aspettarsi un miglioramento delle prestazioni (a parità di E_b/N_0) in quanto l'area che individua la probabilità di errore è ridotta. Il valore della probabilità di errore si determina dopo aver osservato che ciascuno dei due rami in fase e quadratura costituisce un segnale ASK multilivello con $L' = \sqrt{L}$.

Abbiamo già calcolato che, per tale segnale, il rapporto E_b/N_0 dopo demodulazione è esattamente pari a quello del segnale modulato; pertanto (considerando un impulso $g(t)$ a banda minima) dalla eq. (13.4) otteniamo

$$\begin{aligned}
 P_\alpha &= P_e^c(\text{simbolo}) = P_e^s(\text{simbolo}) = \\
 &= \left(1 - \frac{1}{L'}\right) \operatorname{erfc} \left\{ \sqrt{3 \frac{E_b}{N_0} \frac{\log_2 L'}{(L')^2 - 1}} \right\}
 \end{aligned}$$



Ricordando ora che $L' = \sqrt{L} = (L)^{1/2}$ e dunque $\log_2 L' = \frac{1}{2} \log_2 L$, si ottiene

$$P_\alpha = \left(1 - \frac{1}{\sqrt{L}}\right) \operatorname{erfc} \left\{ \sqrt{\frac{3 E_b}{2 N_0} \frac{\log_2 L}{L - 1}} \right\}$$

La probabilità di errore (a simbolo) complessiva, cioè la probabilità che il segnale ricevuto $\tilde{x} = \underline{x} + \underline{n}$ cada fuori della regione di decisione relativa all' \underline{x} trasmesso, risulta $P_e(\text{simbolo}) \simeq P_\alpha + P_\alpha = 2P_\alpha$, assumendo trascurabile la probabilità di sbagliare entrambe x_c ed x_s . Questa stessa ipotesi, assieme all'utilizzo di un codice di Gray per codificare i gruppi di bit associati a livelli dei due rami, consente di esprimere la

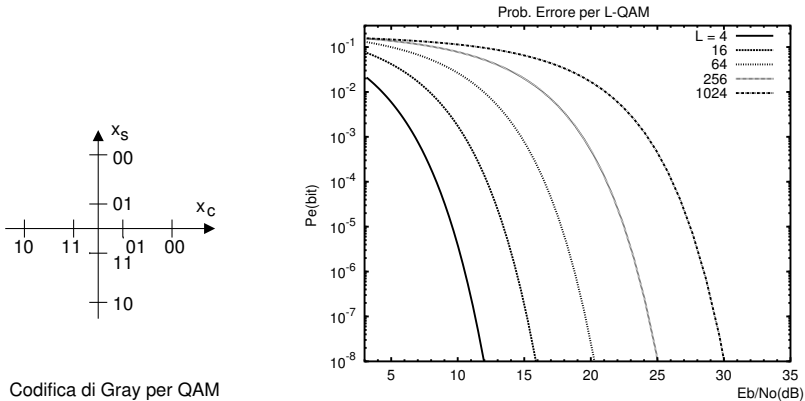


Figura 13.10: Prestazioni L-QAM con codice di Gray

probabilità di errore per bit come

$$\begin{aligned}
 P_e^{QAM}(\text{bit}) &= \frac{P_e(\text{carattere})}{\log_2 L} = \\
 &= \frac{2}{\log_2 L} \left(1 - \frac{1}{\sqrt{L}}\right) \operatorname{erfc} \left\{ \sqrt{\frac{3 E_b \log_2 L}{2 N_0 L - 1}} \right\} \quad (13.9)
 \end{aligned}$$

In figura 13.10 troviamo le curve dei valori di $P_e(\text{bit})$, per diversi valori di L , al variare di $\frac{E_b}{N_0}$ espresso in dB; il confronto con le curve relative al PSK permette di valutare l'entità del miglioramento di prestazioni. Come è evidente, la modulazione QAM offre prestazioni sensibilmente superiori rispetto alla PSK.

Esercizio Consideriamo un sistema di modulazione numerica PSK con 16 fasi, per il quale si riceve una potenza di segnale $\mathcal{P}_x = 10^{-3}$ (Volt)², in presenza di una densità di potenza di rumore $\mathcal{P}_N(f) = 2 \cdot 10^{-11}$ (Volt)²/Hz. Si desidera ricevere un flusso numerico a velocità $f_b = 1$ Mbit/sec e si considerino impulsi a coseno rialzato con $\gamma = 0$.

- 1) Quale è la P_e per bit al ricevitore? E la banda occupata?
- 2) Quale nuovo valore di P_e si ottiene usando invece una modulazione QAM con lo stesso numero di punti di costellazione?
- 3) Nel caso 16-QAM, qualora si desideri ancora la P_e ottenuta al punto 1), quanta potenza è sufficiente ricevere?
- 4) Nel caso QAM con la P_e del punto 1), qualora si desideri dimezzare la banda occupata, che potenza è necessario ricevere?
- 5) Nel caso 16-QAM con la P_e del punto 1) e $\mathcal{P}_x = 10^{-3}$ (Volt)², quale nuova f_b è possibile raggiungere?

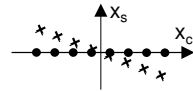
Soluzione

1. Osserviamo che $E_b = \mathcal{P}_x \cdot T_b = \frac{\mathcal{P}_x}{f_b} = \frac{10^{-3}}{10^6} = 10^{-9}$ (Volt)²/Hz, mentre $N_0 = 2\mathcal{P}_N(f) = 4 \cdot 10^{-11}$ (Volt)²/Hz, pertanto $\frac{E_b}{N_0} = 25$ e $\left(\frac{E_b}{N_0}\right)_{dB} = 10 \log_{10} 25 \simeq 14$ dB.

- (a) Dalle curve delle prestazioni per il PSK si trova che con $E_b/N_0 = 14$ dB, si ottiene $P_e = 10^{-3}$ qualora si utilizzino 16 livelli.
- (b) La banda occupata risulta $B = \frac{f_b}{\log_2 L} = \frac{10^6}{4} = 250$ KHz.
2. Le curve delle prestazioni per il QAM mostrano che con $E_b/N_0 = 14$ dB e 16 livelli, si ottiene $P_e \simeq 3 \cdot 10^{-6}$.
3. le stesse curve mostrano che, con il 16-QAM, la $P_e = 10^{-3}$ si ottiene con $E_b/N_0 = 10.5$ dB, ovvero $14 - 10.5 = 3.5$ dB in meno, che corrispondono ad una potenza $\mathcal{P}'_x = \frac{\mathcal{P}_x}{10^{0.35}} = \frac{10^{-3}}{2.24} = 4.47 \cdot 10^{-5}$ (Volt)².
4. Dimezzare la banda equivale a raddoppiare $\log_2 L$, ovvero utilizzare $M = 8$ bit/simbolo, e dunque un numero di livelli $L = 2^M = 256 = (L')^2$. Le curve delle prestazioni per il 256-QAM mostrano che per ottenere $P_e = 10^{-3}$ occorre $E_b/N_0 \simeq 18.3$ dB, pari ad un aumento di $18.3 - 14 = 4.3$ dB, che equivale ad una potenza $\mathcal{P}'_x = 10^{0.43} \mathcal{P}_x \simeq 2.7 \cdot 10^{-3}$ (Volt)².
5. Ci ritroviamo nelle stesse condizioni del punto 3), con un eccesso di 3.5 dB nel valore di E_b/N_0 , che può essere speso riducendo di egual misura T_b , e quindi aumentando f_b . Risulta: $T'_b = \frac{T_b}{10^{0.35}}$ e quindi $f'_b = \frac{1}{T'_b} = \frac{10^{0.35}}{T_b} = 10^{0.35} \cdot f_b = 10^{0.35} \simeq 2.24$ Mb/sec.
- E se $\gamma \neq 0$? La trattazione del caso di banda base, mostra che l'argomento sotto radice della $\text{erfc}\{\}$ subisce un peggioramento di $(1 + \gamma)(1 - \frac{\gamma}{4})$, che (per esempio) con $\gamma = 0.5$ fornisce 1.31, che deve essere compensato da una uguale diminuzione di E_b/N_0 . Nel caso 5), ad esempio, la f_b risulterà quindi limitata a $f''_b = f'_b/1.31 = 1.71$ Mb/sec.

13.4 Codifica differenziale

Qualora la portante con cui si effettua la demodulazione omodina presenti un errore di fase¹⁹, il piano dell'involuppo complesso subisce una rotazione, producendo decisioni sistematicamente errate a causa dello spostamento dei punti di costellazione ottenuti campionando le c.a. di b.f. (croci in figura) rispetto a quelli che si otterrebbero nel caso di demodulazione coerente (pallini). Per rimediare al problema, si può estendere il principio espresso a pag. 80 per segnali di banda base al caso delle modulazioni numeriche, rendendo la decisione su quale punto di costellazione sia stato ricevuto, indipendente dalla fase della portante di demodulazione, ma dipendente invece dalla fase dell'involuppo complesso osservata *per il simbolo precedente*. Ciò si realizza modificando il modo in cui vengono determinati i punti di costellazione da trasmettere, scegliendoli ora in funzione di una coppia di simboli consecutivi, anziché di uno solo.



13.4.1 Modulazione DBPSK

Per illustrare la tecnica, procediamo con un esempio relativo al caso di trasmissione BPSK della sequenza $x_k = 001011011010010$ e mostriamo come la codifica differenziale consenta di neutralizzare un errore di fase di π . La fig. 13.11, simile²⁰ a quella a pag. 441, mostra la sequenza delle operazioni necessarie, e che consistono nel

¹⁹Il cui effetto su $x_c(t)$ è discusso al § 11.2.1.1. Facciamo ricadere in questa casistica l'ambiguità di fase dei sistemi di recupero portante come al § 11.4.2, ma anche la distorsione di fase introdotta dal canale non selettivo in frequenza, vedi pag. 361.

²⁰La similitudine non è per nulla casuale, dato che qualora il predittore mostrato a pag. 441 sia realizzato mediante un elemento di ritardo, i due schemi di elaborazione coincidono.

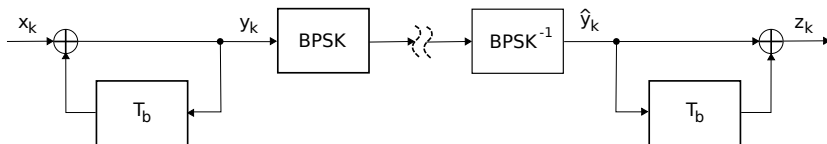


Figura 13.11: Codifica e detezione differenziali

trasformare il messaggio binario x_k in quello y_k in base alla relazione

$$y_k = x_k \oplus y_{k-1} \quad (13.10)$$

(in cui l'operatore \oplus rappresenta un *or esclusivo*), e quindi effettuare la modulazione BPSK di y_k anziché di x_k . Dal lato ricevente, il segnale BPSK viene demodulato ottenendo la sequenza \hat{y}_k , che viene a sua volta trasformata in z_k in base all'espressione

$$z_k = \hat{y}_k \oplus \hat{y}_{k-1} \quad (13.11)$$

che, in assenza di errori (ossia se $\hat{y}_k = y_k$ per tutti i k), permette di ottenere nuovamente i valori del messaggio originario²¹ x_k a partire dalla sequenza z_k , come la figura del riquadro precedente consente di verificare tramite un esempio. Notiamo come i bit della sequenza y_k *cambiano* nel caso in cui il corrispondente bit di x_k è uno, mentre *non cambiano* se è uno zero.

Se assumiamo ora di rappresentare lo zero con una fase nulla, e l'uno con una fase di π , riscriviamo la (13.1) come

$$x_{BPSK}(t) = a \cos(\omega_0 t + \varphi(t)) \quad \text{con} \quad \varphi(t) = \pi \cdot \sum_{k=-\infty}^{\infty} x_k \text{rect}_{T_b}(t - kT_b)$$

e ponendo per semplicità il periodo di bit pari ad un ciclo di portante, otteniamo in fig. 13.12 la forma d'onda BPSK associata alla sequenza originaria x_k , posta a confronto con quella relativa invece a y_k ed indicata come DBPSK, in cui la D sta per *differenziale*. Anche se i simboli del segnale DBPSK sono in corrispondenza univoca con i bit della sequenza y_k , osserviamo che possono anche essere derivati *direttamente* dall'esame visivo della x_k , dato che la fase del DBPSK si inverte per gli $x_k = 1$, e non si inverte per $x_k = 0$. Pertanto, è possibile realizzare un demodulatore che, senza utilizzare la formula che calcola z_k , prende le corrette decisioni a riguardo dei bit di x_k basandosi sui *cambiamenti* di fase anziché sui loro valori assoluti.

D'altra parte però lo schema di fig. 13.11 mostra l'uso di un modem BPSK *convenzionale*, delegando le operazioni di differenziazione e sua inversa a circuiti addizionali. Verifichiamo quindi che anche in tale forma si riesce a recuperare il segnale originario, nel caso in cui ad esempio la portante di demodulazione presenti un errore di fase di π ²², tale da causare l'*inversione* di tutti i bit decodificati, producendo così un messaggio $\hat{y}_k = 110010010011100$. E' facile verificare che anche in questo caso, applicando l'operatore $z_k = \hat{y}_k \oplus \hat{y}_{k-1}$, si ottiene nuovamente la sequenza originaria.

²¹A parte per il primo bit, che ha il solo scopo di stabilire il riferimento di fase per la decodifica del successivo. Da un punto di vista formale, combinando le 13.10 e 13.11 e in assenza di errori (ossia $\hat{y}_k = y_k$) si ottiene $\hat{z}_k = \hat{y}_k \oplus \hat{y}_{k-1} = x_k \oplus y_{k-1} \oplus y_{k-1} = x_k$.

²²E' questo un caso tutt'altro che anomalo, in quanto il segnale BPSK contiene entrambe le fasi in egual misura (se i bit sono equiprobabili) e quindi un circuito come quello di pag. 291 può *agganciarsi* indifferentemente all'una o all'altra.

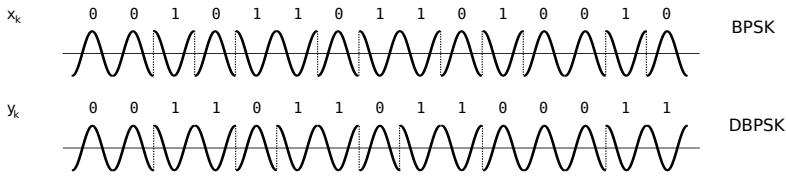


Figura 13.12: Segnale BPSK e BPSK differenziale

Prestazioni DBPSK Esaminiamo ora cosa accade in presenza di errori: supponiamo di ricevere un messaggio $y_k = 000\underline{1}01101100011$, in cui il terzo bit (sottolineato) è errato. Calcolando $z_k = \hat{y}_k \oplus \hat{y}_{k-1}$ questa volta si ottiene $z_k = /00\underline{1}11011010010$ che risulta uguale a x_k tranne che nel terzo e quarto bit. Infatti, dato che z_k dipende dagli indici k e $k - 1$ di y , l'effetto dell'errore non si propaga oltre il bit successivo a quello errato. Dato quindi che ad ogni errore del decisore si ottengono due bit errati anziché uno, a parità di E_b/N_0 il DBPSK è affetto da un tasso di errore circa *doppio* di quello del BPSK.

13.4.2 DQPSK

Il concetto di codifica differenziale può essere facilmente esteso al caso di L-PSK, semplicemente mettendo in corrispondenza le configurazioni di bit previste dal codice di Gray con rotazioni di fase $\Delta\theta$ (tra simboli successivi) contigue, come esemplificato nella tabella a fianco²³ per $L = 4$ (²⁴), ovvero nel caso della modulazione DQPSK. L'involuppo complesso di tale segnale quindi assumerà, negli istanti di simbolo, valori la cui fase dipende dalla fase del simbolo precedente, incrementata del $\Delta\theta$ mostrato in tabella, consentendo la corretta ricezione anche in presenza di una portante di demodulazione affetta da errori di fase multipli di $\frac{\pi}{2}$. Anche qui se (a causa del rumore) si verifica un errore di ricezione, questo si propaga anche al simbolo successivo.

$x_{k-1}x_k$	$\Delta\theta$
00	0
01	$\pi/2$
11	π
10	$-\pi/2$

Nel caso del QAM si può nuovamente applicare una forma di codifica differenziale, ma lo schema di corrispondenza tra gruppi di bit e punti della costellazione è più complesso²⁵, e non viene qui trattato.

13.4.3 OFDM differenziale

Infine, citiamo l'uso che viene fatto della codifica differenziale nei sistemi di trasmissione multiportante come l'OFDM (§ 13.7), in cui il sincronismo di fase viene acquisito in corrispondenza di una sottoportante, ma la risposta in frequenza del canale causa una rotazione del piano dell'involuppo complesso relativo alle altre sottoportanti. In tale circostanza, la codifica differenziale viene applicata ai gruppi di bit associati a portanti contigue, limitando così l'errore di fase introdotto dal canale, a quello che si verifica in un intervallo di frequenza pari alla spaziatura tra sottoportanti.

²³Tratta da *Andrea Goldsmith*, *Wireless Communications*, pag. 151.

²⁴Nel caso di $L > 4$ la tabella si modifica molto semplicemente scrivendo accanto al codice di Gray ai L livelli, la sequenza crescente delle fasi differenziali $\Delta\theta_k = k \frac{2\pi}{L-1}$.

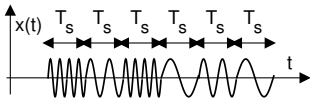
²⁵Vedi ad es. *Krzysztof Wesolowski*, *Introduction to Digital Communication Systems*, Wiley, pag. 328.

13.5 Modulazione di frequenza L-FSK

Qualora si desideri che l'ampiezza del segnale modulato si mantenga strettamente costante si può adottare la tecnica di modulazione FSK (*Frequency Shift Keying*), che associa ad ogni simbolo un valore (o *livello*) di frequenza f_k che si somma a quello della portante, in accordo all'espressione

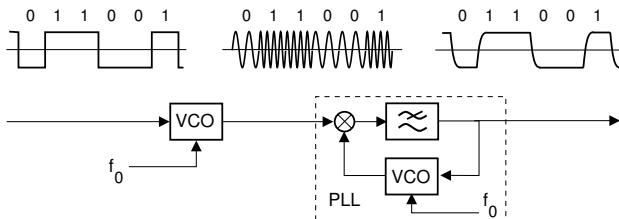
$$x_{FSK}(t) = \cos[2\pi(f_0 + m(t))t] \quad \text{dove} \quad m(t) = \Delta \cdot \sum_{k=-\infty}^{\infty} f_k \cdot \text{rect}_{T_s}(t - kT_s)$$

in cui ogni elemento della sequenza f_k assume uno tra i valori $\{0, 1, 2, \dots, L-1\}$. Si tratta in pratica di una portante la cui frequenza nominale f_0 è alterata di una quantità $\Delta \cdot f_k$ Hz per un intervallo temporale pari al periodo di simbolo T_s , in cui Δ rappresenta ora la spaziatura (uniforme) tra le frequenze associate agli L livelli. Pertanto l'espressione può essere riscritta come



$$x_{FSK}(t) = \sum_{k=-\infty}^{\infty} \cos[2\pi(f_0 + \Delta f_k)t] \cdot \text{rect}_{T_s}(t - kT_s)$$

Il risultato è senza dubbio ad ampiezza costante; se $T_s \gg \frac{1}{f_0}$ si può adottare uno schema di mo-demodulazione basato su di un PLL (vedi § 11.2.1.2 e 11.3.1.1) riportato (per $L = 2$) in figura, in cui all'uscita del passa basso ritroviamo il segnale modulante.



Lo schema è effettivamente utilizzato per modem a bassa velocità e basso costo, ed ha il pregio di funzionare anche in presenza di errori tra l' f_0 usata dal VCO del trasmettitore e quella del ricevitore. Per raggiungere velocità f_b più elevate a parità di L , occorre ridurre T_s , in modo da aumentare $f_s = \frac{f_b}{\log_2 L}$. In tal caso può essere necessario ricorrere ad un demodulatore più complesso, come descritto al § seguente.

13.5.1 Modulazione FSK ortogonale

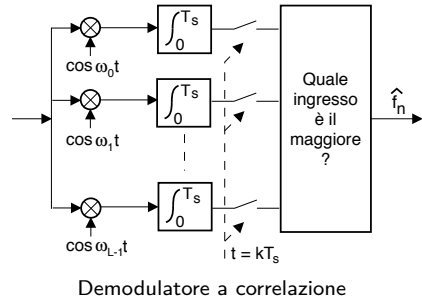
Nel caso in cui si realizzi la condizione $\Delta = \frac{i}{2T_s}$ con i intero, le diverse frequenze $f_0 + \Delta f_k$ sono *ortogonali*²⁶, e può essere adottato un *demodulatore a correlazione*

²⁶La discussione al riguardo è sviluppata al § 13.10.2, definendo anche le condizioni di demodulazione *incoerente* e *coerente*, ovvero se le portanti generate in ricezione $\cos[2\pi(f_0 + \Delta f_k)t + \phi_k]$ presentano o meno una fase ϕ_k casuale rispetto a quella trasmessa. In particolare, la spaziatura Δ multipla di $\frac{1}{2T_s}$ garantisce ortogonalità solo nel caso di modulazione *coerente*, mentre nel caso incoerente occorre una spaziatura *doppia*, ossia Δ multiplo di $\frac{1}{T_s}$.

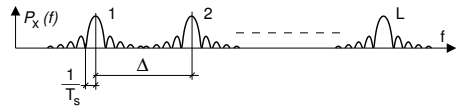
(vedi § 9.5.2), realizzato mediante un banco di correlatori ed un decisione di massimo, in cui l' n -esimo correlatore esegue

$$\int_0^{T_s} \cos [2\pi (f_0 + m\Delta) t] \cos [2\pi (f_0 + n\Delta) t] dt \quad (13.12)$$

dove $f_0 + m\Delta$ rappresenta la frequenza (incognita) in arrivo, mentre $f_0 + n\Delta$ è una delle frequenze possibili, con $n \in \{0, 1, 2, \dots, L-1\}$. Essendo tali frequenze tra loro ortogonali²⁷ entro l'intervallo di integrazione, al termine del calcolo una sola delle uscite sarà diversa da zero. Come discusso al § 9.5, in presenza di rumore l'uscita di ogni correlatore diviene una v. a. con varianza $\frac{N_0}{2} \mathcal{E}_G$, corrompendo l'ortogonalità tra simboli, e dunque si rende necessaria l'operazione di confronto per realizzare una decisione di *massima verosimiglianza*.



Occupazione di banda In generale, se ogni diversa f_k è equiprobabile, l'FSK ha una densità spettrale del tipo mostrato in figura; se $L \gg 1$, l'occupazione di banda complessiva risulta quindi (circa) uguale a $L \cdot \Delta$. Nel caso di *modulazione coerente* (vedi nota 26) la minima spaziatura è di $\Delta = \frac{1}{2T_s} = \frac{f_s}{2}$, e dunque nel caso di L elevato la minima banda occupata può essere approssimata come



$$B_{FSK} \simeq L \cdot \frac{f_s}{2} = \frac{f_b}{2} \cdot \frac{L}{\log_2 L} \quad (13.13)$$

mentre per quanto riguarda l'efficienza spettrale si ottiene

$$\rho_{FSK} = \frac{f_b}{B} = \frac{f_b 2 \log_2 L}{L f_b} = \frac{2}{L} \log_2 L$$

ossia $\frac{L}{2}$ volte peggiore dell'L-ASK (pag. 309). Ma: se l'efficienza spettrale è così bassa, che vantaggi ci sono ad usare l'FSK? ... a sua difesa, portiamo i seguenti argomenti:

Il caso semplice con $T_s \gg \frac{1}{f_0}$ può essere demodulato con lo schema a PLL descritto al § 13.5, di facile realizzazione e poco costoso; ad esempio, veniva usato per salvare su *compact cassette* audio i dati degli *home computer* degli anni '70²⁸

²⁷L'ortogonalità tra le forme d'onda associate ai diversi simboli corrisponde alla intercorrelazione nulla tra le forme d'onda in un periodo (§ 9.1.4), ed infatti l'integrale 13.12 vale $\mathcal{R}_{nm} = \begin{cases} .5 \cdot T_s & \text{se } n = m \\ 0 & \text{altrimenti} \end{cases}$: ricordando che $\cos^2 \alpha = \frac{1}{2} + \frac{1}{2} \cos 2\alpha$, l'uscita del correlatore per $m = n$ vale $\frac{1}{2} \int_0^{T_s} (1 + \cos (4\pi (f_0 + m\Delta) t)) dt$, e scegliendo opportunamente f_0 (vedi § 13.10.2), il coseno che viene integrato descrive un numero intero di periodi all'interno dell'intervallo $(0, T_s)$, fornendo quindi un contributo nullo. Se invece $n \neq m$ la funzione integranda non contiene il termine costante, ma solo termini a media nulla, dato che $\cos \alpha \cos \beta = \frac{1}{2} \cos (\alpha + \beta) \cos (\alpha - \beta)$.

²⁸tipo: Sinclair Spectrum, Commodore Vic20 e 64 ... Come noto, le cassette audio soffrono di variazioni di velocità di trascinamento del nastro (*wow & flutter*), ma il PLL non ne risente.

Nel caso a due livelli l'efficienza spettrale risulta $\rho_{BFSK} = \frac{2}{L} \log_2 L|_{L=2} = 1$, come per il caso del BPSK²⁹. Al contrario, al crescere di L l'efficienza spettrale diviene sempre peggiore.

La probabilità di errore può essere resa *piccola a piacere*, nei limiti della teoria dell'informazione³⁰. La figura a fianco mostra i valori di E_b/N_0 necessari per ottenere le varie P_e con diversi valori di L , e illustra come, all'aumentare di L , sia necessaria sempre minor potenza per ottenere la stessa P_e , a patto che risulti

$$E_b/N_0 > 1/\log_e 2 = 0,69$$

Questo valore è noto come *limite di Shannon-Hartley*³¹ ricavato alla (17.30) a pag. 455.

Il miglioramento di P_e con L è una manifestazione del *compromesso banda-potenza*: osserviamo infatti che anche la banda occupata $B_{FSK} \simeq \frac{f_b}{2} \frac{L}{\log_2 L}$ *aumenta* (a parità di f_b) all'aumentare di L , e dunque a parità di E_b/N_0 l'FSK riesce ad ottenere P_e arbitrariamente piccole, a spese di una occupazione di banda sempre maggiore. L'aumento di L però non può essere illimitato, sia per le limitazioni di banda del canale, che a causa della complessità del ricevitore, per non parlare del ritardo temporale necessario ad accumulare tutti i bit che realizzano un simbolo con un enorme numero L di livelli.

In appendice 13.10.2 è sviluppato un approfondimento di analisi dell'FSK ortogonale.

13.6 Schema riassuntivo delle prestazioni

La tabella 13.1 mette a confronto le prestazioni ottenibili con le tecniche di modulazione fin qui discusse, per un segnale dati a banda minima, e con simboli costituiti da $M = \log_2 L$ bit. Se $\gamma \neq 0$, si deve aggiungere un termine $(1 + \gamma) (1 - \frac{\gamma}{4})$ al denominatore sotto radice, procedendo come indicato al § 5.3.5

13.7 Modulazione OFDM

La sigla sta per ORTHOGONAL FREQUENCY DIVISION MULTIPLEX, ossia *multiplazione a divisione di frequenza ortogonale*. Si tratta della tecnica di modulazione numerica adottata per le trasmissioni ADSL³², DVB-T, e WiFi, ed ha la particolarità di utilizzare in modo *ottimo* la banda del canale, e di ricondurre l'operazione di equalizzazione ad un prodotto tra vettori.

²⁹Tranne che, essendo ora presenti solo 2 frequenze, l'approssimazione (13.13) non è più corretta.

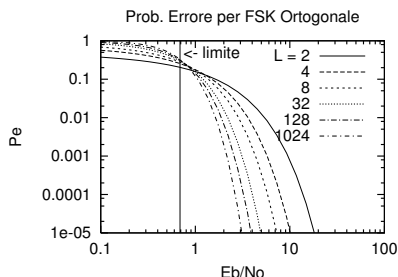
³⁰ Ovvero tenendo conto che (vedi § 17.2.3) f_b non può superare la capacità di canale (eq. (17.29)), che a sua volta non può superare il limite C_∞ espresso dalla (17.31). Ma per l'esattezza, l'espressione teorica della probabilità di errore per simbolo risulta essere

$$P_e^{FSK}(\text{simbolo}) = 1 - \frac{1}{\sqrt{\pi L}} \int_{-\infty}^{\infty} e^{-z^2} \left(\int_{-\infty}^{z + \sqrt{\log_2 L \cdot E_b/N_0}} e^{-y^2} dy \right)^{L-1} dz$$

che deve essere valutata numericamente per ricavare le curve mostrate.

³¹A pagina 347 è esposta una motivazione informale del comportamento descritto.

³²ADSL = *Asymmetric Digital Subscriber Line*, vedi § 6.9.4.



Metodo	M bit/simbolo	P_e (bit) con codifica di Gray ed impulso a banda minima	Banda RF	ρ [bit/sec/Hz]
BPSK	1	$\frac{1}{2} \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0}} \right\}$	$f_b (1 + \gamma)$	$\frac{1}{(1+\gamma)}$
L-ASK	$\log_2 L$	$\frac{1}{M} \left(1 - \frac{1}{L} \right) \operatorname{erfc} \left\{ \sqrt{3 \frac{E_b}{N_0} \frac{M}{(L^2-1)}} \right\}$	$f_b \frac{(1+\gamma)}{M}$	$\frac{M}{(1+\gamma)}$
QPSK	2	$\frac{1}{2} \operatorname{erfc} \left\{ \sqrt{\frac{E_b}{N_0}} \right\}$	$\frac{f_b}{2} (1 + \gamma)$	$\frac{2}{(1+\gamma)}$
L-PSK	$\log_2 L$	$\frac{1}{M} \operatorname{erfc} \left\{ \sin \left(\frac{\pi}{L} \right) \sqrt{\frac{E_b}{N_0} M} \right\}$	$f_b \frac{(1+\gamma)}{M}$	$\frac{M}{(1+\gamma)}$
L-QAM	$\log_2 L$	$\frac{2}{M} \left(1 - \frac{1}{\sqrt{L}} \right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2} \frac{E_b}{N_0} \frac{M}{L-1}} \right\}$	$f_b \frac{(1+\gamma)}{M}$	$\frac{M}{(1+\gamma)}$
L-FSK incoerente	$\log_2 L$		$f_b \frac{L}{M}$	$\frac{M}{L}$
L-FSK coerente	$\log_2 L$	nota 30 a pag. 324	$f_b \frac{L}{2M}$	$\frac{M}{2L}$

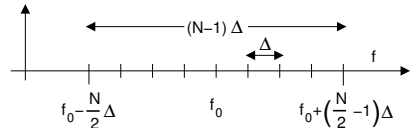
Tabella 13.1: Confronto tra metodi di modulazione numerica

13.7.1 Rappresentazione nel tempo ed in frequenza

La modulazione OFDM suddivide una sequenza binaria su N diversi flussi, trasmessi a divisione di frequenza mediante forme d'onda ortogonali. Concettualmente possiamo pensare l'OFDM come una evoluzione³³ della modulazione FSK, in cui le diverse frequenze spaziate tra loro di Δ Hz in accordo all'espressione

$$f_n = f_0 + \Delta \cdot \left(n - \frac{N}{2} \right) \tag{13.14}$$

con $n = 0, 1, \dots, N-1$ sono utilizzate contemporaneamente, e su ognuna si realizza una modulazione numerica anche a più livelli (es. QPSK o QAM) con impulso NRZ rettangolare di durata T . Indicando ora con $\underline{a}_n^k = a_{n_c}^k + ja_{n_s}^k$ le coordinate nel piano dell'involuppo complesso di un generico punto della costellazione realizzata per la portante f_n all'istante $t = kT$, il segnale OFDM può essere scritto come



$$x_{OFDM}(t) = \sum_k \operatorname{rect}_T(t - kT) \sum_{n=0}^{N-1} \left(a_{n_c}^k \cos \omega_n(t - kT) - a_{n_s}^k \sin \omega_n(t - kT) \right) \tag{13.15}$$

$$= \sum_{k=-\infty}^{\infty} \delta(t - kT) * \left(\operatorname{rect}_T(t) \sum_{n=0}^{N-1} \left(a_{n_c}^k \cos \omega_n t - a_{n_s}^k \sin \omega_n t \right) \right) \tag{13.16}$$

in cui la prima sommatoria (su k) identifica gli istanti di simbolo, e la seconda (su n) le diverse portanti. Tale segnale presenta³⁴ un involuppo complesso rispetto a f_0 pari

³³La trasmissione numerica contemporanea su più portanti è a volte indicata con il nome di *Multi Carrier Modulation* (MCM) o *Discrete Multi Tone* (DMT). La modulazione *FSK* utilizza invece una portante alla volta, in quanto la sua definizione prevede la presenza di un solo oscillatore.

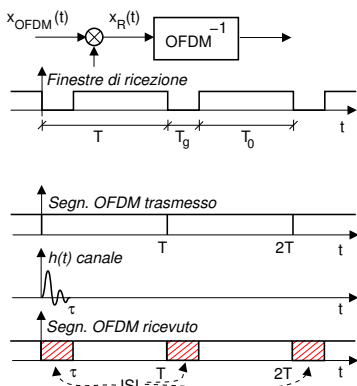
³⁴Osserviamo innanzitutto che per un segnale

$$x(t) = \cos \omega_1 t = \frac{1}{2} \left(e^{j\omega_1 t} + e^{-j\omega_1 t} \right)$$

a

$$\underline{x}_{OFDM}(t) = \sum_{k=-\infty}^{\infty} \delta(t - kT) * \left(\text{rect}_T(t) \sum_{n=0}^{N-1} \underline{a}_n e^{j2\pi[\Delta(n - \frac{N}{2})]t} \right) \quad (13.17)$$

L'espressione (13.16) non vincola la durata T di un simbolo ad un valore particolare; deve però risultare $T \geq T_0 = \frac{1}{\Delta}$, in quanto il ricevitore opera sul segnale $x_R(t)$ ottenuto per moltiplicazione con una finestra temporale di estensione $T_0 = \frac{1}{\Delta}$, allo scopo di rendere ortogonali tra loro³⁵ le frequenze $f_n = f_0 + \Delta \cdot (n - \frac{N}{2})$, e mettere in grado il ricevitore di calcolare i valori \underline{a}_n^k per tutti gli n presenti all'istante $t = kT$, mediante un ricevitore concettualmente simile a quello a correlazione presentato a pag. 322.



L'intervallo T_0 è detto *periodo principale* del simbolo OFDM, mentre la differenza $T_g = T - T_0$ è indicata come *tempo di guardia*, od anche *preambolo*, ed il segnale ricevuto durante T_g non è usato in ricezione. Il motivo di tale "spreco"³⁶ risiede nel fatto che, in presenza di un canale non perfetto, la parte iniziale di ogni simbolo risulta corrotta (vedi figura) da una interferenza intersimbolica (ISI) dovuta al risultato della convoluzione tra la coda del simbolo precedente e l' $h(t)$ del canale. Consideriamo ora un solo simbolo (fissiamo $k = 0$ e consideriamo l'origine dei tempi ritardata di T_g) ricevuto nell'intervallo $T_0 = \frac{1}{\Delta} \leq T$, con inviluppo complesso

$$\underline{x}_{T_0}(t) = \text{rect}_{T_0}(t) \cdot \sum_{n=0}^{N-1} \underline{a}_n e^{j2\pi[\Delta(n - \frac{N}{2})]t} \quad (13.18)$$

e calcoliamone la trasformata per determinare l'occupazione di banda:

$$\underline{X}_{T_0}(f) = T_0 \text{sinc}(fT_0) * \sum_{n=0}^{N-1} \underline{a}_n \delta\left(f - \Delta\left(n - \frac{N}{2}\right)\right) = \quad (13.19)$$

$$= T_0 \sum_{n=0}^{N-1} \underline{a}_n \text{sinc}\left(\left(f - \Delta\left(n - \frac{N}{2}\right)\right)T_0\right) \quad (13.20)$$

risulta $x^+(t) = \frac{1}{2}e^{j\omega_1 t}$, e quindi il suo inviluppo complesso $\underline{x}(t)$ calcolato rispetto ad f_0 vale

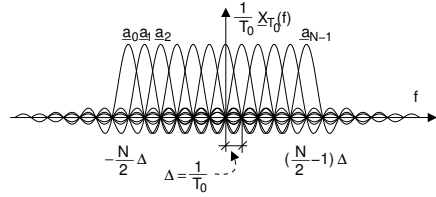
$$\underline{x}(t) = 2x^+(t) e^{-j\omega_0 t} = 2 \frac{1}{2} e^{j\omega_1 t} e^{-j\omega_0 t} = e^{j(\omega_1 - \omega_0)t}$$

Allo stesso modo, si ottiene che per $y(t) = \sin \omega_1 t$ risulta $\underline{y}(t) = \frac{1}{j} e^{j(\omega_1 - \omega_0)t}$. Pertanto, considerando che $\frac{1}{j} = \frac{-j}{j^2} = -j$, ad ogni termine $z_k(t) = a_{n_c}^k \cos \omega_n t - a_{n_s}^k \sin \omega_n t$ corrisponde un $\underline{z}(t) = a_{n_c}^k e^{j(\omega_n - \omega_0)t} + j a_{n_s}^k e^{j(\omega_n - \omega_0)t} = \underline{a}_n^k e^{j2\pi(f_n - f_0)t}$. Applicando ora la (13.14) si ottiene $f_n - f_0 = \Delta \cdot (n - \frac{N}{2})$ e quindi la (13.17).

³⁵Come mostrato per il caso *incoerente* discusso al § 13.10.2

³⁶Infatti la frequenza di simbolo $f_s = \frac{1}{T} = \frac{1}{T_0 + T_g}$ risulta ridotta rispetto al caso in cui T_g sia nullo.

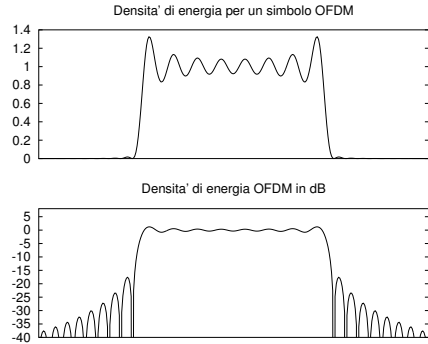
Otteniamo allora la costruzione grafica mostrata alla figura a lato, dove si evidenzia come ogni funzione *sinc* risulti moltiplicata per uno dei coefficienti \underline{a}_n , che potrebbero quindi essere ri-ottenuti in ricezione campionando (in modo complesso) $\underline{X}(f)$ su frequenze spaziate di Δ .



Per quanto riguarda la densità di potenza $\mathcal{P}_{\underline{x}_R}(f)$ dell'involuppo complesso $\underline{x}_R(t)$ ricevuto e finestrato, consideriamo l'espressione (vedi § 9.3.1)

$$\mathcal{P}_{\underline{x}_R}(f) = \frac{1}{T} E \left\{ |\underline{X}_{T_0}(f)|^2 \right\} \quad (13.21)$$

in cui $|\underline{X}_{T_0}(f)|^2$ è la densità di energia di un simbolo OFDM (eq. (13.20)): la figura a lato ne mostra l'andamento (in scala lineare ed in dB) per un simbolo a 32 portanti, di cui 16 (esterne) *spente* (vedi appresso), mentre per le 16 centrali si è posto $\underline{a}_n = 1$. Notiamo come si ottenga una densità spettrale di potenza *quasi rettangolare* pur utilizzando simboli a durata finita.



Potenza complessiva Mostriamo ora come mettere in relazione la potenza ricevuta complessiva \mathcal{P}_{x_R} di $x_R(t)$ e del suo involuppo complesso $\mathcal{P}_{\underline{x}_R}$

$$\mathcal{P}_{x_R} = \int \mathcal{P}_{\underline{x}_R}(f) df \quad (13.22)$$

con la dinamica dei valori \underline{a}_n utilizzati per modulare le singole portanti: nel seguito ci riferiamo a costellazioni L-QAM, indicando con M_n ed $L_n = 2^{M_n}$ rispettivamente il numero di bit e di punti di costellazione per la portante n-esima, ad ognuna delle quali la (13.16) attribuisce una potenza \mathcal{P}_n .

Per calcolare la (13.22) mediante la (13.21) utilizzando l'espressione di $\underline{X}_{T_0}(f)$ fornita dalla (13.20), osserviamo che le funzioni $\text{sinc}(fT_0)$ che vi compaiono *sono ortogonali* se spaziate per un multiplo di $\Delta = \frac{1}{T_0}$, ovvero (vedi § 4.1.3) sussiste la condizione

$$\int \text{sinc}((f - n\Delta)T_0) \text{sinc}((f - m\Delta)T_0) df = \begin{cases} 0 & \text{se } n \neq m \\ \frac{1}{T_0} & \text{se } n = m \end{cases}$$

Pertanto, introducendo una insignificante³⁷ traslazione di f pari a $\frac{N}{2}\Delta$, si ha

$$\begin{aligned} \mathcal{P}_{\underline{x}_R} &= \frac{1}{T} \int E \left\{ |\underline{X}_{T_0}(f)|^2 \right\} df = \\ &= \frac{1}{T} \int T_0^2 \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} E \{ \underline{a}_n \underline{a}_m^* \} \text{sinc}((f - \Delta n)T_0) \text{sinc}((f - \Delta m)T_0) df = \\ &= \frac{T_0^2}{T} \sum_{n=0}^{N-1} E \{ \underline{a}_n^2 \} \int \text{sinc}^2((f - \Delta n)T_0) df = \frac{T_0^2}{T} \sum_{n=0}^{N-1} E \{ \underline{a}_n^2 \} \frac{1}{T_0} = \frac{T_0}{T} \sum_{n=0}^{N-1} \sigma_{\underline{a}_n}^2 \end{aligned}$$

³⁷ Infatti le variazioni di fase non influenzano lo spettro di energia ma solo quello di ampiezza.

in quanto i termini *incrociati* prodotti dalla doppia sommatoria si annullano³⁸.

Scegliendo il lato della costellazione QAM in modo opportuno³⁹, si può ottenere $\sigma_{\underline{a}_n}^2 = E\{\underline{a}_n^2\} = 2\mathcal{P}_n$, in cui \mathcal{P}_n è la potenza per la n -esima portante QAM; considerando infine (vedi eq. (10.7) a pag. 258) che

$$\mathcal{P}_{x_R} = \mathcal{P}_{x_R}^+ + \mathcal{P}_{x_R}^- = \frac{2}{4}\mathcal{P}_{\underline{x}_R} = \frac{1}{2}\mathcal{P}_{\underline{x}_R}$$

possiamo scrivere

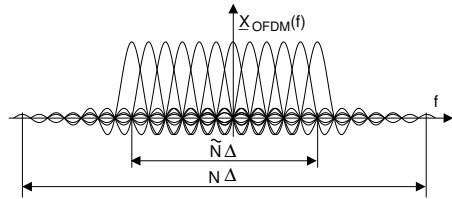
$$\mathcal{P}_{x_R} = \frac{1}{2} \frac{T_0}{T} \sum_{n=0}^{N-1} 2\mathcal{P}_n = \frac{T_0}{T} \sum_{n=0}^{N-1} \mathcal{P}_n$$

in cui è evidenziata la perdita di potenza legata alla presenza del preambolo.

13.7.2 Architettura di modulazione

Una caratteristica fondamentale della modulazione OFDM è quella di essere realizzata senza *oscillatori e integratori*, ma completamente tramite circuiti digitali. Con riferimento alla figura 13.13, il flusso binario a frequenza f_b viene parallelizzato per formare simboli ad $L = 2^M$ livelli a frequenza $f_s = \frac{f_b}{M} = \frac{f_b}{\log_2 L}$. Questi M bit/simbolo sono suddivisi in \tilde{N} gruppi di M_n ($n = 0, 1, \dots, \tilde{N}-1$) bit ciascuno, con $M = \sum_{n=0}^{\tilde{N}-1} M_n$, e ad ogni gruppo di M_n bit corrisponde un punto di costellazione \underline{a}_n , scelto tra $L_n = 2^{M_n}$ punti possibili.

La sequenza $\{\underline{a}_n\}$ viene poi *arricchita* con $N - \tilde{N}$ valori nulli (metà all'inizio e metà alla fine) ottenendo una nuova sequenza $\{\underline{a}_n\}$ di N valori, in modo che la sommatoria in (13.20) dia luogo ad un involuppo complesso praticamente limitato in banda (vedi figura) tra (circa) $\pm \frac{N}{2} \cdot \Delta$ Hz, che può essere pertanto rappresentato dai suoi campioni presi a frequenza $f_c = N \cdot \Delta$ campioni
secondo. Il blocco indicato come FFT⁻¹ svolge proprio questa operazione: esso infatti esegue efficientemente⁴⁰ il



³⁸Vedi nota 12 a pag. 24

³⁹Al § 9.10.4 si è mostrato che se gli a_n sono v.a. indipendenti e distribuite uniformemente su L' livelli tra $\pm A$, si ottiene $\sigma_a^2 = \frac{A^2}{3} \frac{L'+1}{L'-1}$. Nel caso di una costellazione QAM quadrata ad L livelli si ha $L' = \sqrt{L}$, e se le realizzazioni sui rami in fase e quadratura sono indipendenti risulta $\sigma_{\underline{a}_n}^2 = E\{(a_{nc} + ja_{ns})^2\} = 2\sigma_a^2 = \frac{2A^2}{3} \frac{\sqrt{L}+1}{\sqrt{L}-1}$; volendo eguagliare tale valore a $2\mathcal{P}_n$, occorre quindi scegliere $A = \sqrt{3\mathcal{P}_n \frac{\sqrt{L}-1}{\sqrt{L}+1}}$.

⁴⁰La (13.23) è in qualche modo simile alla formula di ricostruzione (2.3) (vedi pag. 17) per il segnale periodico limitato in banda $\pm \frac{N}{2}F$

$$x(t) = \sum_{m=-N/2}^{N/2} X_m e^{j2\pi m F t}$$

che calcolata per $t = hT_c = \frac{h}{NF}$ fornisce $x(hT_c) = \sum_{m=-N/2}^{N/2} X_m e^{j2\pi \frac{m}{N} h}$. Ponendo ora $n = m + \frac{N}{2}$ e $Y_n = X_{n - \frac{N}{2}}$ otteniamo

$$x(hT_c) = \sum_{n=0}^{N-1} Y_n e^{j2\pi \frac{n - \frac{N}{2}}{N} h} = e^{-j\pi h} \sum_{n=0}^{N-1} Y_n e^{j2\pi \frac{n}{N} h}$$

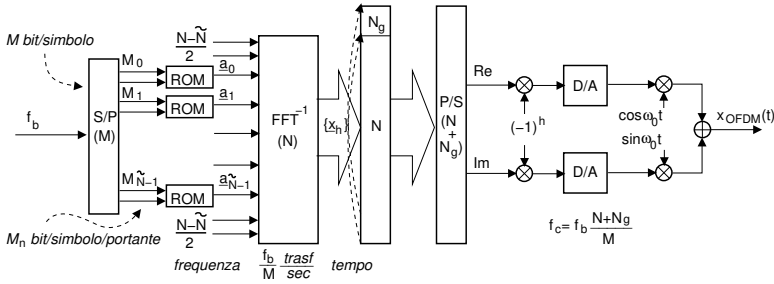


Figura 13.13: Architettura di un modulatore OFDM numerico

calcolo

$$\sum_{n=0}^{N-1} a_n e^{j2\pi \frac{n}{N} h} = \frac{1}{(-1)^h} x_{T_0}(hT_c) \quad (13.23)$$

Il risultato della FFT^{-1} è quindi una sequenza di coefficienti complessi $\{x_h\}$, che a meno di un segno alterno sono uguali ai campioni dell'involucro complesso $x_{T_0}(t)$ fornito dalla (13.18) relativamente ad un simbolo. Il preambolo da trasmettere durante il tempo di guardia T_g si ottiene *aggiungendo* in testa a $\{x_h\}$ un gruppo di campioni prelevati dalla coda⁴¹.

Infine, le parti reale ed immaginaria di $\{x_h\}$ sono inviate ad una coppia di convertitori D/A operanti a $f_c = \frac{N+N_g}{T} = \frac{N}{T_0} = N\Delta$ in modo da ottenere le c.a. di b.f., utilizzate per produrre il segnale $x_{OFDM}(t)$ mediante una coppia di modulatori in fase e quadratura.

13.7.3 Efficienza dell'OFDM

Come vedremo al § 13.7.8, questa è una tra le tecniche di modulazione che meglio approssima i risultati della teoria dell'informazione, tanto più quanto maggiore è la sua efficienza. Quest'ultima si ottiene considerando che solo \tilde{N} portanti su N trasportano informazione, e che solo $f_c \cdot T_0$ campioni su $f_c \cdot T$ sono unici; combinando queste quantità si ottiene

$$\rho = \frac{\tilde{N}}{N} \frac{T_0}{T} = \frac{\tilde{N}}{N} \frac{T - T_g}{T} = \frac{\tilde{N}}{N} \left(1 - \frac{T_g}{T}\right)$$

che misura la frazione di segnale utile rispetto all'occupazione di banda ed al numero di campioni/simbolo presenti in $x_{OFDM}(t)$. La ridondanza introdotta (le portanti

Osservando ora che $e^{-j\pi h} = (-1)^h$ e confrontando con la (13.18) si ottiene la (13.23). La coppia di relazioni

$$X_n = \frac{1}{N} \sum_{h=0}^{N-1} x_h e^{-j2\pi \frac{h}{N} n} \quad \text{e} \quad x_h = \sum_{n=0}^{N-1} X_n e^{j2\pi \frac{n}{N} h}$$

sono chiamate *Discrete Fourier Transform* (DFT) diretta e inversa, in quanto costituiscono la versione discreta della trasformata di Fourier (vedi § 4.4), e consentono il calcolo di una serie di campioni in frequenza a partire da campioni nel tempo e viceversa.

⁴¹In effetti la (13.23) fornisce un risultato periodico rispetto ad h , con periodo N , ossia con periodo $N \cdot T_c = N \frac{1}{f_c} = N \frac{1}{N\Delta} = \frac{1}{\Delta} = T_0$ per la variabile temporale. Per questo motivo il preambolo dell'OFDM è detto anche *estensione ciclica*.

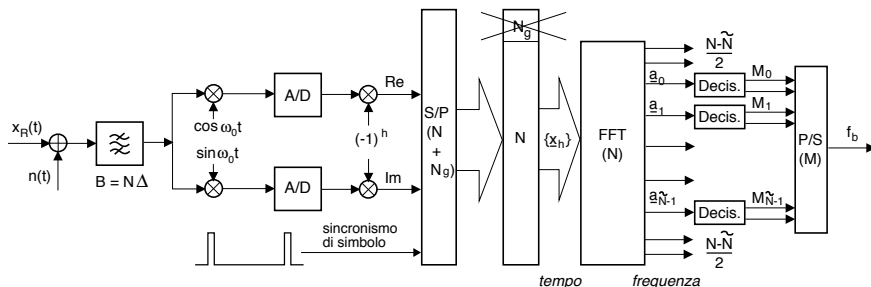


Figura 13.14: Architettura di un demodulatore OFDM numerico

vuote ed il preambolo) è della stessa natura di quella introdotta dal roll-off γ di un impulso a coseno rialzato, in quanto ha lo scopo di evitare che si verifichino fenomeni di interferenza tra simboli. Osserviamo che l'efficienza migliora all'aumentare di T e di N , dato che T_g ed $N - \tilde{N}$ sono fissi.

13.7.4 Architettura di demodulazione

Per ottenere gli elementi della sequenza $\{a_n\}$ e quindi il gruppo di M bit che hanno originato il simbolo, si adotta l'architettura mostrata in figura 13.14 che svolge una azione del tutto inversa a quella del modulatore. Innanzitutto il ricevitore deve acquisire il sincronismo di frequenza (vedi § 13.7.10), in modo che il segnale ricevuto possa essere demodulato in fase e quadratura, e le C.A. di B.F. campionate a frequenza $f_c = \frac{N+N_g}{T}$. Dopo l'inversione di segno ad indici alterni, ed avere acquisito il sincronismo di simbolo, $f_c \cdot T$ campioni complessi sono bufferizzati, quindi gli N_g campioni del preambolo rimossi, e sugli N valori del periodo principale viene calcolata una FFT, ottenendo i valori

$$\frac{1}{N} \sum_{h=0}^{N-1} x_h e^{-j2\pi \frac{h}{N} n} = \underline{X}_{T_0} \left(\left(n - \frac{N}{2} \right) \Delta \right) = a_n \quad (13.24)$$

Solo gli \tilde{N} valori centrali sono avviati verso altrettanti decisori, che determinano il punto di costellazione più vicino all' a_n ricevuto per ogni portante, associandovi il relativo codice di M_n bit, ed il risultato finale è nuovamente serializzato per produrre gli M bit che hanno dato origine al simbolo.

13.7.5 Prestazioni

Il calcolo della P_e per bit si basa su quello relativo alle probabilità di errore P_{e_n} condizionato alle singole portanti. Dato che la portante n -esima trasporta M_n bit/simbolo, la probabilità che un bit generico provenga dalla portante n -esima risulta pari a $Pr(n) = \frac{M_n}{M}$ e quindi la probabilità che sia errato è pari a

$$P_e = \sum_{n=0}^{\tilde{N}-1} Pr(n) P_{e/n} = \frac{1}{M} \sum_{n=0}^{\tilde{N}-1} M_n P_{e_n} \quad (13.25)$$

13.7.5.1 Calcolo della P_e per portante

Per determinare il valore di P_{e_n} per la portante n -esima, si applica il risultato trovato al § 13.3.1 per la modulazione QAM, che esprime P_{e_n} in funzione del numero di livelli $L_n = 2^{M_n}$ e dal rapporto $\left(\frac{E_b}{N_0}\right)_n$ per tale frequenza. Ma l'eq. (13.9) suppone la presenza di un filtro passabasso in ingresso al ricevitore per limitare la densità di potenza del rumore a quella che ricade nella banda di segnale, mentre ora tale filtro non c'è, ed al suo posto deve essere invece tenuto conto l'effetto della FFT sul rumore. Inoltre, vorremmo pervenire ad un risultato valido anche in presenza di rumore non bianco, e/o di una distribuzione di potenza sulle portanti non uniforme. Pertanto, al posto del rapporto E_b/N_0 che compare nella (13.9) utilizziamo ora il rapporto SNR_n tra la potenza di segnale e quella del rumore per l' n -esima portante, ottenendo così

$$P_{e/n} = \frac{2}{\log_2 L_n} P_{\alpha_n} \quad \text{in cui} \quad P_{\alpha_n} = \left(1 - \frac{1}{\sqrt{L_n}}\right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2} SNR_n \frac{1}{L_n - 1}} \right\} \quad (13.26)$$

derivata⁴² dalla (13.9) considerando $\frac{E_{b_n}}{N_0} = \frac{\mathcal{P}_{R_n}}{N_0 f_b} = \frac{SNR_n}{\log_2 L}$, ed in cui P_{α_n} esprime la probabilità di errore su di uno dei rami (in fase od in quadratura) della n -esima costellazione QAM con L_n punti, che rappresentano gruppi di bit secondo la codifica di Gray. Per il calcolo di

$$SNR_n = \frac{\mathcal{P}_{R_n}^c}{\mathcal{P}_{N_n}^c} = \frac{\mathcal{P}_{R_n}^s}{\mathcal{P}_{N_n}^s} = \frac{\frac{1}{2} \mathcal{P}_{R_n}}{\frac{1}{2} \mathcal{P}_{N_n}} = \frac{\mathcal{P}_{R_n}}{\mathcal{P}_{N_n}}$$

osserviamo che la potenza \mathcal{P}_{R_n} dell'involuppo complesso del segnale ricevuto sulla portante n -esima è pari a

$$\mathcal{P}_{R_n} = 2\mathcal{P}_{R_n} = 2 \frac{T_0}{T} \alpha_n \mathcal{P}$$

in cui \mathcal{P} è la potenza totale ricevuta, e $\alpha_n = \frac{\mathcal{P}_n}{\mathcal{P}}$ è la frazione di potenza assegnata alla n -esima portante. Resta quindi da determinare \mathcal{P}_{N_n} .

13.7.5.2 Potenza di rumore per portante

Per quanto riguarda \mathcal{P}_{N_n} , si tratta di applicare la (13.24) alla sequenza $\left\{(-1)^h \underline{n}(hT_c)\right\}$ dei campioni dell'involuppo complesso del rumore, e determinare il valore

$$\mathcal{P}_{N_n} = E \left\{ (\underline{N}_n)^2 \right\} = \sigma_{N_n}^2 \quad \text{in cui} \quad \underline{N}_n = \frac{1}{N} \sum_{h=0}^{N-1} (-1)^h \underline{n}(hT_c) e^{-j2\pi \frac{h}{N} n}$$

in virtù del fatto che i valori $\underline{n}(hT_c)$ sono a media nulla, che (con n fissato) la FFT ne effettua una combinazione lineare con coefficienti $e^{-j2\pi \frac{h}{N} n}$, e che essendo $\underline{n}(t)$ ergodico è possibile scambiare medie temporali e di insieme. Sviluppando

$$(\underline{N}_n)^2 = \underline{N}_n \underline{N}_n^* = \frac{1}{N^2} \sum_{h=0}^{N-1} \sum_{k=0}^{N-1} (-1)^{h-k} \underline{n}(hT_c) \underline{n}^*(kT_c) e^{-j2\pi \frac{h-k}{N} n}$$

⁴²Si tratta di invertire l'eq. 5.12 $SNR = \frac{E_b}{N_0} \frac{2 \log_2 L}{(1+\gamma)}$ considerando $\gamma = 0$ e notando che tale espressione è ottenuta nel caso di banda base, mentre per segnali AM la banda (e la potenza di rumore) raddoppia.

e tenendo conto che $E \left\{ (-1)^{h-k} \underline{n}(hT_c) \underline{n}^*(kT_c) \right\} = e^{j\pi(h-k)} \mathcal{R}_{\underline{N}}((h-k)T_c)$ otteniamo⁴³

$$\begin{aligned} \mathcal{P}_{\underline{N}_n} &= \frac{1}{N^2} \sum_{h=0}^{N-1} \sum_{k=0}^{N-1} \mathcal{R}_{\underline{N}}((h-k)T_c) e^{j\pi(h-k)} e^{-j2\pi \frac{h-k}{N} n} = \\ &= \frac{1}{N} \sum_{m=-(N-1)}^{N-1} \frac{N-|m|}{N} \mathcal{R}_{\underline{N}}(mT_c) e^{j2\pi \frac{mT_c}{2T_c}} e^{-j2\pi \frac{m}{N} n} = \\ &= \frac{1}{N} \sum_{m=-(N-1)}^{N-1} z(m) e^{-j2\pi \frac{m}{N} n} \end{aligned} \quad (13.27)$$

in cui l'ultima riga semplifica l'espressione introducendo la sequenza $\{z(m)\}$ di lunghezza N , che si ottiene campionando

$$z(t) = \left(1 - \frac{|t|}{NT_c}\right) \mathcal{R}_{\underline{N}}(t) e^{j2\pi \frac{t}{2T_c}} \quad (13.28)$$

agli istanti $t = mT_c$ con $T_c = \frac{1}{N\Delta}$. Mostriamo ora come, per N sufficientemente elevato, la (13.27) possa essere calcolata in funzione dei campioni di $Z(f) = \mathcal{F}\{z(t)\}$, ed in particolare di come risulti

$$\mathcal{P}_{\underline{N}_n} \simeq \Delta \cdot Z(f)|_{f=n\Delta} \simeq 4\Delta \cdot \mathcal{P}_N(f_n)$$

Analizzando i termini che compaiono in (13.28), osserviamo che il prodotto $\mathcal{R}_{\underline{N}}(t) e^{j2\pi \frac{t}{2T_c}}$ ha trasformata pari a $\mathcal{P}_{\underline{N}}(f)$, traslata in frequenza di $-\frac{1}{2T_c} = -\frac{N\Delta}{2}$, ovvero

$$\mathcal{F}\left\{\mathcal{R}_{\underline{N}}(t) e^{j2\pi \frac{t}{2T_c}}\right\} = \mathcal{P}_{\underline{N}}\left(f - \frac{N\Delta}{2}\right)$$

mentre il termine $\left(1 - \frac{|t|}{NT_c}\right) = \text{tri}_{2NT_c}(t) = \text{tri}_{\frac{2}{\Delta}}(t)$ possiede come noto trasformata $\mathcal{F}\left\{\text{tri}_{\frac{2}{\Delta}}(t)\right\} = \frac{1}{\Delta} \text{sinc}^2\left(\frac{f}{\Delta}\right)$; pertanto per N elevato il prodotto $z(t) = \mathcal{R}_{\underline{N}}(t) e^{j2\pi \frac{t}{2T_c}} \cdot \text{tri}_{\frac{2}{\Delta}}(t)$ ha trasformata

$$Z(f) = \mathcal{P}_{\underline{N}}\left(f - \frac{N\Delta}{2}\right) * \frac{1}{\Delta} \text{sinc}^2\left(\frac{f}{\Delta}\right) \simeq \mathcal{P}_{\underline{N}}\left(f - \frac{N\Delta}{2}\right)$$

avendo approssimato $\frac{1}{\Delta} \text{sinc}^2\left(\frac{f}{\Delta}\right)$ come un impulso di area unitaria, per $N\Delta$ grande rispetto a Δ .

Dato che $\mathcal{P}_{\underline{N}}(f)$ è limitato in banda tra $\pm \frac{N\Delta}{2}$, allora $Z(f)$ è limitato in una banda compresa tra $f = 0$ ed $f = N\Delta$, e $z(t)$ è perfettamente rappresentato dai suoi campioni

⁴³La riduzione da due ad una sommatoria, si ottiene scrivendo esplicitamente tutti i termini della doppia sommatoria, e notando che si ottiene per N volte lo stesso termine $\mathcal{R}_{\underline{N}}(0)$, $N-1$ volte i termini $\mathcal{R}_{\underline{N}}(T_c) e^{j\pi} e^{-j2\pi \frac{1}{N} n}$ e $\mathcal{R}_{\underline{N}}(-T_c) e^{-j\pi} e^{j2\pi \frac{1}{N} n}$, $N-2$ volte quelli $\mathcal{R}_{\underline{N}}(2T_c) e^{j2\pi} e^{-j2\pi \frac{2}{N} n}$ e $\mathcal{R}_{\underline{N}}(-2T_c) e^{-j2\pi} e^{j2\pi \frac{2}{N} n}$, e così via.

$z(m) = z(mT_c)$ che compaiono nella (13.27); in particolare, per N sufficientemente elevato, si ottiene⁴⁴ che

$$\begin{aligned} \mathcal{P}_{N_n} &= \frac{1}{N} \sum_{m=-(N-1)}^{N-1} z(m) e^{-j2\pi \frac{m}{N} n} \simeq \Delta \cdot Z(f)|_{f=n\Delta} = \\ &= \Delta \cdot \mathcal{P}_N \left(n\Delta - \frac{N\Delta}{2} \right) = \Delta \cdot \mathcal{P}_N \left(\Delta \left(n - \frac{N}{2} \right) \right) = \\ &= 4\Delta \cdot \mathcal{P}_N^+ \left(f_0 + \Delta \left(n - \frac{N}{2} \right) \right) = 4\Delta \cdot \mathcal{P}_N(f_n) = 2\Delta \cdot \mathcal{N}_0(f_n) \end{aligned}$$

in cui si è tenuto conto che $\mathcal{P}_N(f) = 4\mathcal{P}_N^+(f + f_0)$ e si è indicata la densità di potenza in ingresso come $\mathcal{P}_N(f) = \frac{\mathcal{N}_0(f)}{2}$.

13.7.5.3 Prestazioni per portante

Siamo finalmente in grado di scrivere

$$\begin{aligned} SNR_n &= \frac{\mathcal{P}_{R_n}}{\mathcal{P}_{N_n}} = \frac{2\frac{T_0}{T}\alpha_n \mathcal{P}}{2\Delta \mathcal{N}_0(f_n)} = \frac{T_0}{T} \alpha_n \frac{T_0 \mathcal{P}}{\mathcal{N}_0(f_n)} = \frac{T_0}{T} \alpha_n \frac{E_s}{\mathcal{N}_0(f_n)} = \\ &= \frac{T_0}{T} \alpha_n \frac{E_b M}{\mathcal{N}_0(f_n)} = \frac{T_0}{T} \frac{E_{b_n}}{E_b} \frac{E_b M}{\mathcal{N}_0(f_n)} = \frac{T_0}{T} \frac{E_{b_n} M}{\mathcal{N}_0(f_n)} \end{aligned}$$

⁴⁴Se campioniamo $z(t)$ con periodo $T_c = \frac{1}{N\Delta}$, il segnale $Z^*(f) = \sum_{m=-\infty}^{\infty} Z(f - m \cdot N\Delta)$ non presenta aliasing (vedi figura), ed il passaggio di $z^*(t) = \sum_{m=-\infty}^{\infty} z(mT_c) \delta(t - mT_c)$ attraverso un filtro di ricostruzione $H(f) = \frac{1}{N\Delta} \text{rect}_{N\Delta} \left(f - \frac{N\Delta}{2} \right)$ restituisce il segnale originario. Scriviamo pertanto

$$z(t) = z^*(t) * h(t) = \sum_{m=-\infty}^{\infty} z(mT_c) \delta(t - mT_c) * \text{sinc}(N\Delta t) e^{j\pi N\Delta t}$$

ed effettuiamone la trasformata:

$$\begin{aligned} Z(f) &= \mathcal{F} \left\{ \sum_{m=-\infty}^{\infty} z(mT_c) \delta(t - mT_c) \right\} \cdot \frac{1}{N\Delta} \text{rect}_{N\Delta} \left(f - \frac{N\Delta}{2} \right) \\ &= \left[\sum_{m=-\infty}^{\infty} z(mT_c) e^{-j2\pi \frac{m}{N\Delta} f} \right] \cdot \frac{1}{N\Delta} \text{rect}_{N\Delta} \left(f - \frac{N\Delta}{2} \right) \end{aligned}$$

che, calcolata alle frequenze $f = n\Delta$ con $n = 0, 1, \dots, N-1$ fornisce

$$Z(f)|_{f=n\Delta} = \frac{1}{N\Delta} \sum_{m=-\infty}^{\infty} z(mT_c) e^{-j2\pi \frac{m}{N} n}$$

Se ora non disponiamo di tutti i campioni $z(mT_c)$, ma solo degli $2N-1$ valori con $m = -(N-1), \dots, 0, 1, \dots, N-1$, la relazione precedente si applica ad un nuovo segnale $z'(t) = z(t) \cdot \text{rect}_{2NT_c}(t)$, fornendo

$$Z'(f)|_{f=n\Delta} = \frac{1}{N\Delta} \sum_{m=-(N-1)}^{N-1} z(mT_c) e^{-j2\pi \frac{m}{N} n}$$

In virtù delle proprietà delle trasformate, risulta

$$Z'(f) = Z(f) * \mathcal{F}\{\text{rect}_{2NT_c}(t)\} \simeq Z(f) * \delta(f) = Z(f)$$

in cui l'approssimazione è lecita per N elevato.

avendo posto $T_0 \mathcal{P} = E_s = E_b M$ pari all'energia di un simbolo di durata $T_0 = \frac{1}{\Delta}$, ed avendo riscritto $\alpha_n = \frac{\mathcal{P}_n}{\mathcal{P}}$ come $\alpha_n = \frac{E_{b_n}}{E_b}$ in modo da porre in evidenza la E_{b_n} della portante n -esima. La P_e per portante risulta quindi

$$P_{e/n} = \frac{2}{M_n} \left(1 - \frac{1}{\sqrt{L_n}} \right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2} \frac{T_0}{T} \frac{E_{b_n}}{\mathcal{N}_0(f_n)} \frac{M}{L_n - 1}} \right\} \quad (13.29)$$

13.7.5.4 Caso di rumore bianco

Se $\mathcal{P}_N(f)$ non dipende da f , possiamo scrivere

$$\mathcal{P}_N^+(f) = \frac{\mathcal{N}_0}{2} \operatorname{rect}_{N\Delta}(f - f_0)$$

e semplificare la (13.29), sostituendo ad $\mathcal{N}_0(f_n)$ la costante \mathcal{N}_0 . In questo caso, il risultato $\mathcal{P}_{N_n} = 2\Delta \cdot \mathcal{N}_0$ può essere ottenuto direttamente dalla (13.27): infatti, risulta

$$\mathcal{R}_{N_n}(t) = \mathcal{F}^{-1} \{ \mathcal{P}_{N_n}(f) \} = \mathcal{F}^{-1} \{ 4\mathcal{P}_N^+(f + f_0) \} = 2\mathcal{N}_0 N \Delta \operatorname{sinc}(N\Delta t)$$

e dunque $\mathcal{R}_{N_n}(t) = 0$ con $t = mT_c = \frac{m}{N\Delta}$ per $m \neq 0$. Ciò permette di scrivere in definitiva

$$\mathcal{P}_{N_n} = \frac{1}{N} \mathcal{R}_{N_n}(0) = \frac{1}{N} 2\mathcal{N}_0 N \Delta = 2\Delta \cdot \mathcal{N}_0$$

13.7.5.5 Confronto con la portante singola

Proviamo a verificare se la modulazione OFDM è vantaggiosa in termini di prestazioni rispetto ad una QAM monoportante con lo stesso numero di bit/simbolo M , per una medesima occupazione di banda, ed a parità di potenza. Nel caso OFDM, considerando un tempo di guardia $T_g = T - T_0$ nullo, in presenza di rumore bianco, e scegliendo un intervallo di simbolo $T_0 = \frac{1}{\Delta}$ da cui derivare $M^{OFDM} = T_0 \cdot f_b$, si ottiene $\alpha_n = \frac{1}{N}$ e dunque valori $E_{b_n} = \alpha_n E_b = \frac{E_b}{N}$ uguali per le diverse portanti; pertanto la 13.29 diviene

$$P_e^{OFDM} = P_{e/n} = \frac{2\tilde{N}}{M^{OFDM}} \left(1 - \frac{1}{\sqrt{L_n}} \right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2} \frac{E_b}{\mathcal{N}_0} \frac{1}{\tilde{N}} \frac{M^{OFDM}}{L_n - 1}} \right\}$$

Nel caso QAM a portante singola, considerando un impulso a coseno rialzato e roll-off $\gamma = \frac{N}{N} - 1$ si determina una occupazione di banda pari a $B = f_s(1 + \gamma)$ che, se eguagliata a quella del caso OFDM, fornisce $f_s = \tilde{N}\Delta = \frac{\tilde{N}}{T_0}$ e quindi $M^{QAM} = \frac{f_b}{f_s} = \frac{M^{OFDM}}{\tilde{N}}$. Pertanto, visto il risultato del § 13.3.1 si ottiene

$$\begin{aligned} P_e^{QAM} &= \frac{2}{M^{QAM}} \left(1 - \frac{1}{\sqrt{L}} \right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2} \frac{E_b}{\mathcal{N}_0} \frac{M^{QAM}}{L - 1}} \right\} \\ &= \frac{2\tilde{N}}{M^{OFDM}} \left(1 - \frac{1}{\sqrt{L}} \right) \operatorname{erfc} \left\{ \sqrt{\frac{3}{2} \frac{E_b}{\mathcal{N}_0} \frac{1}{\tilde{N}} \frac{M^{OFDM}}{L - 1}} \right\} \end{aligned}$$

che risulta identico a P_e^{OFDM} qualora si noti che $L_n = 2^{M_n} = 2^{\frac{M^{OFDM}}{N}}$ e $L = 2^{M^{QAM}} = 2^{\frac{M^{OFDM}}{N}} = L_n$.

E allora dov'è la convenienza? È il tema delle prossime sottosezioni.

13.7.6 Equalizzazione

Consideriamo il caso in cui la trasmissione attraverso un canale descritto da un involuppo complesso $\underline{H}(f)$ in cui il modulo non è costante e/o la fase non è lineare: in tal caso $\underline{X}_{T_0}(f)$ di (13.20) si altera, ed i valori \underline{a}_n restituiti dalla (13.24) si modificano in $\underline{b}_n = \underline{a}_n \cdot \underline{H}(f - \Delta(n - \frac{N}{2}))$. Come anticipato, l'equalizzazione è pertanto ridotta ad eseguire un semplice prodotto scalare tra il vettore dei valori $\{\underline{b}_n\}$ e quello dei valori $\left\{ \frac{1}{\underline{H}(f - \Delta(n - \frac{N}{2}))} \right\}$, ovviamente purché si conosca $\underline{H}(f)$, od una sua stima.

13.7.6.1 Codifica differenziale

Nel caso in cui la distorsione lineare introdotta dal canale non sia eccessiva, si può evitare del tutto lo stadio di equalizzazione, e ricorrere ad una codifica differenziale (§ 13.4). Infatti, in presenza di distorsione di fase, il piano dell'involuppo complesso subisce, per ogni portante consecutiva, una rotazione pari alla differenza della fase di $\underline{H}(f)$ in corrispondenza delle due frequenze contigue. Se questa rotazione non è eccessiva, si può prendere come riferimento di fase il risultato della demodulazione della portante precedente.

13.7.7 Sensibilità alla temporizzazione

Nel caso in cui il ricevitore non acquisisca una perfetta sincronizzazione di simbolo, il calcolo della FFT su di un gruppo di campioni presi a partire dalla coda del preambolo non altera per nulla il risultato⁴⁵, in virtù della caratteristica di periodicità dello stesso.

13.7.8 Ottimalità

Come stiamo per mostrare, questa proprietà è intimamente legata alla possibilità dell'OFDM di assegnare valori di potenza differenti alle diverse portanti.

La trasmissione numerica con una f_b elevata, eseguita utilizzando una sola portante, deve necessariamente occupare una banda molto ampia; nel caso in cui $H(f)$ presenti una elevata distorsione di ampiezza, l'equalizzazione della stessa causa una colorazione del rumore in ingresso al demodulatore, ed un peggioramento delle prestazioni. Un problema analogo nasce nel caso in cui il rumore non sia bianco, ad esempio perché derivante da un segnale interferente.

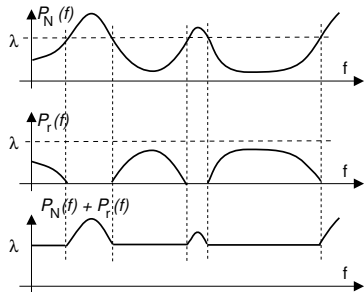
In entrambi i casi, la teoria di Shannon che definisce⁴⁶ una *capacità di canale* pari a $C = W \log_2 \left(1 + \frac{\mathcal{P}_r}{W N_0} \right)$ in presenza di un rumore bianco, con $\mathcal{P}_N(f) = \frac{N_0}{2}$ e con una potenza ricevuta \mathcal{P}_r in una banda (positiva) W , si modifica nel seguente modo, per tenere conto dell'andamento incostante di $\mathcal{P}_r(f)$ e $\mathcal{P}_N(f)$. Considerando il canale scomposto in infinite sottobande entro le quali le densità di potenza possono ritenersi costanti, l'espressione della capacità diviene ora:

$$C = \sup_{\mathcal{P}_r(f)} \int_{f \in I_f} \log_2 \left(1 + \frac{\mathcal{P}_r(f)}{\mathcal{P}_N(f)} \right) df \quad (13.30)$$

⁴⁵ Infatti *non siamo* nelle condizioni di demodulazione coerente dell'FSK, e le portanti del simbolo OFDM ricevuto mantengono ortogonalità purché finestrate su di un periodo $T_0 = \frac{1}{\Delta}$.

⁴⁶ Come discusso ai § 17.2.3 e 17.2.4, il risultato della teoria di Shannon asserisce che $f_b = C$ è la massima velocità di trasmissione per cui si può (teoricamente) conseguire una probabilità di errore nulla, ma anche che il canale consegue capacità C massima a seguito di una oculata scelta di come trasmettere il messaggio.

in cui $\mathcal{P}_r(f)$ viene fatto variare in tutti i modi possibili e tali che $\int_{f \in I_f} \mathcal{P}_r(f) df = \mathcal{P}_r$, con $\mathcal{P}_r(f) \geq 0$, ed I_f rappresenta l'insieme delle frequenze in cui è presente il segnale: $I_f = \{f : \mathcal{P}_r(f) > 0\}$. La (13.30) asserisce quindi che, nel caso in cui $\mathcal{P}_N(f)$ in ingresso al canale non sia bianco, le migliori prestazioni (f_b che tende a C) si ottengono solamente sagomando la densità di potenza del segnale ricevuto in modo opportuno.



Per determinare l'andamento ottimo di $\mathcal{P}_r(f)$ si ricorre allora al *calcolo delle variazioni* basato sui *moltiplicatori di Lagrange*⁴⁷, che in questa sede non affrontiamo, e che fornisce la condizione

$$\mathcal{P}_r(f) + \mathcal{P}_N(f) = \begin{cases} \lambda & \text{se } \mathcal{P}_N(f) < \lambda \\ \mathcal{P}_N(f) & \text{se } \mathcal{P}_N(f) \geq \lambda \end{cases} \quad (13.31)$$

detta anche del *riempimento d'acqua* perché asserisce che (vedi figura) il segnale debba essere

presente in misura maggiore nelle regioni di frequenza dove il rumore è minore. La costante λ è scelta in modo tale da ottenere $\int \mathcal{P}_r(f) df = \mathcal{P}_r$.

In un sistema di modulazione numerica a singola portante, $\mathcal{P}_r(f)$ non può essere modificato a piacere, in quanto il suo andamento è vincolato dalla particolare caratteristica di Nyquist $G(f)$ scelta per ottenere una ricezione priva di ISI. Nel caso dell'OFDM invece, la potenza assegnata a ciascuna portante può essere variata liberamente, e se la $\mathcal{P}_r(f)$ che realizza le condizioni (13.31) può essere comunicata al modulatore, è possibile avvicinarsi alla velocità massima permessa dalla (13.30).

In particolare, si ottiene che la massima velocità f_b è conseguibile attribuendo a tutte le portanti la medesima probabilità di errore, e quindi in definitiva determinando dei valori $\left(\frac{E_b}{N_0}\right)_n$ per ogni portante $n = 0, 1, \dots, \tilde{N} - 1$ tali da rendere le $P_{e/n} = P_e$. Questo risultato può essere ottenuto scegliendo le potenze \mathcal{P}_n in accordo alla (13.31), e quindi trasmettere più bit M_n sulle portanti n per le quali \mathcal{P}_n è maggiore.

13.7.9 Codifica

Abbiamo appena mostrato come, conoscendo la $H(f)$ e la $\mathcal{P}_N(f)$ del canale, sia possibile equalizzare $\mathcal{P}_x(f) = \frac{\mathcal{P}_r(f)}{|H(f)|^2}$ e al contempo soddisfare (13.31) e rendere massima la f_b . Nel caso di collegamenti tempo-varianti però, la $H(f)$ non è nota, ed anche se lo fosse non esiste garanzia che rimanga costante. In tal caso allora non ha senso determinare una distribuzione ottima della potenza e dei bit sulle portanti, mentre invece occorre aggiungere della ridondanza al segnale trasmesso mediante un codice di canale, allo scopo di correggere i bit errati.

Osserviamo ora che, nel caso di una modulazione a portante singola, in presenza di una $H(f)$ tempo-variante, il processo di equalizzazione è particolarmente complesso in quanto deve *inseguire* le variazioni di $H(f)$. Se l'equalizzazione non è perfetta, insorge ISI e la trasmissione diviene rapidamente così piena di errori da renderne impossibile la correzione anche adottando codici di canale.

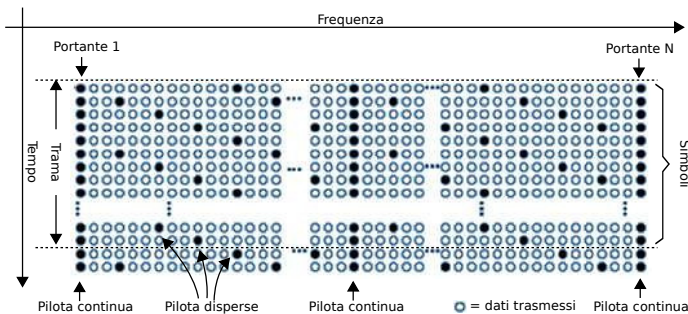
⁴⁷Vedi ad es. http://it.wikipedia.org/wiki/Metodo_dei_moltiplicatori_di_Lagrange

Nel caso dell'OFDM, al contrario, l'andamento di $H(f)$ determina un peggioramento di prestazioni solamente per quelle portanti per le quali $H(f)$ si è ridotto⁴⁸. Pertanto, l'applicazione di un codice di canale al blocco di M bit che costituisce un simbolo, seguito da una operazione di scrambling, consente al lato ricevente di recuperare l'informazione trasmessa anche nel caso in cui per alcune portanti si determini un elevato tasso di errore.

La trasmissione OFDM in cui è presente una codifica di canale prende il nome di trasmissione COFDM (*Coded OFDM*).

13.7.10 Portanti pilota

Fin qui abbiamo assunto che il ricevitore OFDM mostrato in fig. 13.14 operi in condizioni di sincronismo sia per quanto riguarda la portante di demodulazione, che per quanto gli intervalli di simbolo. A questo scopo alcune delle sottoportanti - dette *pilota* - non sono usate per trasmettere dati, ma sono mantenute costantemente attive, con potenza di poco superiore, allo scopo di facilitare la sincronizzazione in frequenza. In figura è rappresentato il caso per il DVB-T, in cui ogni riga rappresenta le portanti di un simbolo, e quelle pilota si trovano in posizione fissa; sono inoltre mostrate delle *portanti disperse* (SCATTERED) le cui posizioni evolvono ciclicamente di simbolo in simbolo, e consentono di acquisire un sincronismo sia di simbolo che di trama, oltre che eventualmente a permettere una stima della $H(f)$ del canale attraversato.



13.8 Sistemi a spettro espanso

In questa tecnica di modulazione la stessa banda di frequenze è utilizzata contemporaneamente da più trasmissioni differenti, che non interferiscono tra loro grazie all'uso di forme d'onda mutuamente ortogonali; ciò avviene adottando una *manipolazione* del messaggio da trasmettere, in modo che questo occupi una banda molto maggiore di quella originaria, e sulla manipolazione inversa in ricezione: tale caratteristica è quindi indicata con il termine di *Spread Spectrum*⁴⁹, e la tecnica di trasmissione risultante prende anche il nome di *Moltiplicazione a Divisione di Codice*.

Sebbene la doppia operazione di *spreading/despreading* non produca nessun vantaggio effettivo nei riguardi delle prestazioni ottenibili nel caso in cui la ricezione sia

⁴⁸Si consideri ad esempio il caso in cui $H(f)$ ha origine da un fenomeno di cammini multipli, che determina un andamento di $H(f)$ oscillante in frequenza.

⁴⁹To spread = spalmare, vedi ad es. lo spread butter.

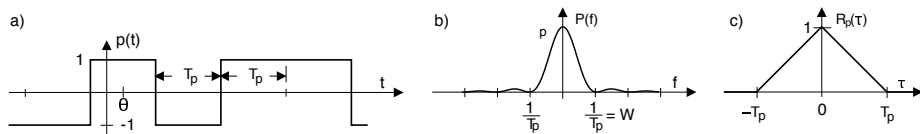


Figura 13.15: a) - sequenza pseudonoise; b) densità di potenza; c) autocorrelazione

disturbata dalla sola presenza di rumore additivo gaussiano, si ottengono invece i seguenti altri benefici:

- la presenza di eventuali altre trasmissioni e/o disturbi nella stessa banda del segnale espanso causa una potenza interferente ridotta;
- la densità spettrale del segnale trasmesso può confondersi con quella del rumore, rendendo la trasmissione stessa poco rilevabile da parte di soggetti ostili;
- l'ignoranza della effettiva manipolazione adottata rende la trasmissione indecifrabile da parte di soggetti ostili.

13.8.1 Sequenze pseudo-casuali

La manipolazione che produce l'espansione spettrale si basa a sua volta sull'utilizzo di una sequenza cosiddetta *pseudo-noise* (PN)⁵⁰, le cui caratteristiche statistiche *si avvicinano* a quelle di un rumore stazionario bianco (ovvero a valori incorrelati), tranne per il fatto che tali valori non sono casuali ma deterministici, in modo che la loro ripetizione ciclica rende la sequenza PN riproducibile dal lato ricevente. La fig. 13.15a mostra una parte di un possibile segnale dati $p(t)$ pseudo casuale, bipolare, di durata $L \cdot T_p$

$$p(t) = \sum_{k=0}^{L-1} a_k g(t - kT_p - \theta) \tag{13.32}$$

realizzato mediante impulsi rettangolari $g(t) = \text{rect}_{T_p}(t)$ di durata T_p chiamati *chip*, con polarità stabilita dagli L valori $a_k = \pm 1$ scelti⁵¹ in modo da produrre media nulla e varianza unitaria ($\overline{a_k} = 0, \overline{a_k^2} = 1$), mentre la correlazione $\mathcal{R}_a(n)$ tende a zero con $n \neq 0$ (⁵²), *imitando* così la proprietà di indipendenza statistica. Al § 9.10.3 abbiamo mostrato che un segnale del genere presenta uno spettro di densità di potenza

$$\mathcal{P}_p(f) = \sigma \frac{|\mathcal{E}_G(f)|^2}{T_p} = T_p \text{sinc}^2(fT_p) \tag{13.33}$$

rappresentato in fig. 13.15b, e per il quale la frequenza $W_p = \frac{1}{T_p}$ ne approssima l'occupazione di banda. Dalla (13.33) consegue che l'autocorrelazione di $p(t)$ si esprime

⁵⁰http://en.wikipedia.org/wiki/Pseudorandom_binary_sequence

⁵¹Un modo di generare la sequenza PN è descritto presso http://en.wikipedia.org/wiki/Maximum_length_sequence

⁵²Data la sequenza deterministica $a_k = \{a_0, a_1, \dots, a_{L-1}\}$ di lunghezza $L - 1$, la correlazione tra coppie di elementi a distanza n è definita come

$$\mathcal{R}_a(n) = \frac{1}{L-n} \sum_{k=0}^{L-n-1} a_k a_{k+n}$$



Considerando invece la sequenza periodica ottenuta ripetendo gli a_k , possiamo definire la stessa grandezza come $\mathcal{R}_a(n) = \frac{1}{L} \sum_{k=0}^{L-1} a_k a_{(k+n) \bmod L}$

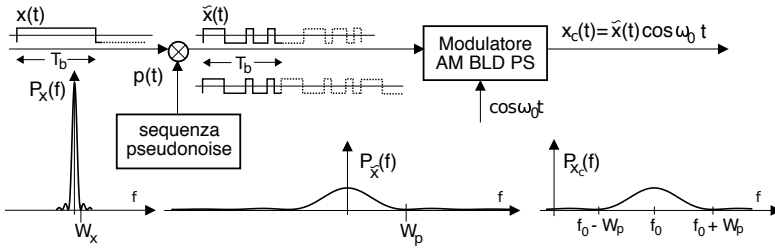


Figura 13.16: Generazione di un segnale modulato DSSS

come⁵³

$$\mathcal{R}_p(\tau) = \mathcal{F}^{-1} \{ \mathcal{P}_p(f) \} = \text{tri}_{2T_p}(\tau)$$

mostrata in fig. 13.15c, e che appunto si azzera per $\tau > T_p$. Le sequenze pseudo-noise utilizzate realmente non soddisfano in pieno tutti questi requisiti, ma vi si avvicinano molto⁵⁴.

13.8.2 Sequenza diretta

La moltiplicazione del segnale PN (13.32) per un segnale di banda base $x(t)$ realizza la tecnica di espansione spettrale nota come *Direct Sequence Spread Spectrum* (o DSSS) il cui effetto, sebbene valido per $x(t)$ qualsiasi, viene ora discusso con riferimento ad un segnale numerico⁵⁵ binario NRZ polare, con frequenza binaria che determina una densità di potenza $\mathcal{P}_x(f)$ del tipo di (13.33) ma con banda $W_x \ll W_p$; il risultato del prodotto $x(t)p(t)$ quindi modula AM-BLD-PS una portante a frequenza f_0 . La fig. 13.16 indica con $\tilde{x}(t) = x(t)p(t)$ il segnale *allargato*, la cui potenza è la stessa⁵⁶ \mathcal{P}_x di $x(t)$, che ora risulta però *spalmata* sulla banda W_p di $p(t)$. Infatti, la densità di potenza $\mathcal{P}_{\tilde{x}}(f)$ è il risultato della convoluzione in frequenza⁵⁷

$$\mathcal{P}_{\tilde{x}}(f) = \mathcal{P}_x(f) * \mathcal{P}_p(f) \simeq \int_{-W_x}^{W_x} \mathcal{P}_x(\lambda) \mathcal{P}_p(f - \lambda) d\lambda$$

in cui la definizione degli estremi di integrazione tiene conto del fatto che $\mathcal{P}_x(f) \approx 0$ per $|f| > W_x$. Considerando ora che $W_p \gg W_x$, notiamo che per $|\lambda| \leq W_x$ si ha $\mathcal{P}_p(f - \lambda) \simeq \mathcal{P}_p(f)$, e quindi

$$\mathcal{P}_{\tilde{x}}(f) \approx \left[\int_{-W_x}^{W_x} \mathcal{P}_x(\lambda) d\lambda \right] \mathcal{P}_p(f) = \mathcal{P}_x \mathcal{P}_p(f)$$

⁵³Si veda l'esempio a pag. 43

⁵⁴Ad esempio per i collegamenti WiFi si utilizza http://en.wikipedia.org/wiki/Barker_code

⁵⁵Quando il messaggio $x(t)$ è di natura numerica, esso è rappresentato da un flusso binario a velocità $f_b = 1/T_b$, e l'operazione di *spreading* si ottiene scegliendo $T_p = T_b/L \ll T_b$. Moltiplicando modulo due ogni bit del messaggio per la sequenza di chip della PN, si ottiene di fatto una *sequenza di sequenze* PN, ognuna con segno invertito o meno a seconda del valore dei singoli bit del messaggio. Il segnale dati viene poi realizzato a partire dalla sequenza risultante a velocità $f_b \cdot W_p$.

⁵⁶Considerando $x(t)$ realizzazione di un processo ergodico indipendente da $p(t)$, la potenza di $\tilde{x}(t)$ risulta $\overline{\tilde{x}^2} = E \{ x^2(t) p^2(t) \} = \overline{x^2} = \mathcal{P}_x$

⁵⁷L'autocorrelazione del prodotto di processi indipendenti è pari al prodotto delle autocorrelazioni (§ 9.7.1), ed a questo si applica teorema di Wiener (§ 9.2.1).

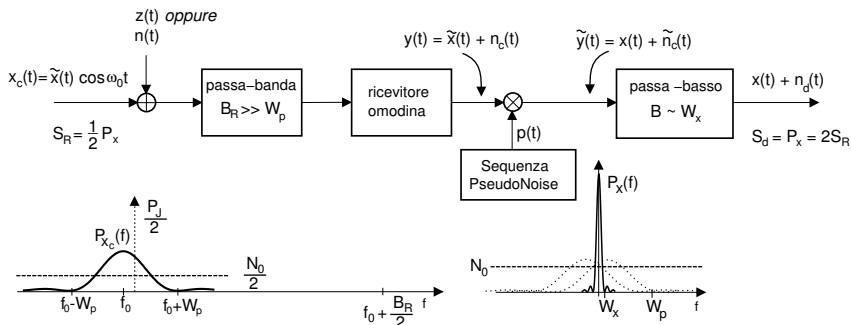


Figura 13.17: Ricevitore DSSS con rumore additivo $n(t)$ o interferenza $z(t)$

13.8.2.1 Guadagno di processo

Il rapporto di espansione spettrale $\frac{W_p}{W_x}$ tra la banda del segnale allargato e quella del segnale di partenza varia tipicamente tra 10 e 10000 volte, ossia tra 10 e 40 dB, e viene indicato anche come *guadagno di processo* (o *processing gain*), in quanto come vedremo rappresenta una misura del miglioramento dell'SNR nel caso di presenza di segnali interferenti.

13.8.2.2 Despreading

Per proseguire nell'analisi, consideriamo lo schema di ricevitore mostrato in fig. 13.17, nella cui parte sinistra è mostrato il segnale modulato ricevuto $x_c(t) = \tilde{x}(t) \cos \omega_0 t$ con potenza⁵⁸ $S_R = \frac{1}{2} \mathcal{P}_x$, a cui si sovrappone un disturbo $n(t)$, ed insieme attraversano il filtro passabanda di ricezione caratterizzato da una banda di rumore $B_R \gg W_p \gg W_x$, dato che deve lasciar passare l'intero spettro *allargato*, compresi i suoi lobi laterali. Il fatto che la potenza S_R sia spalmata su tutta la banda B_R rende il segnale utile poco distinguibile dal livello di rumore, e dunque difficilmente intercettabile.

Dopo demodulazione omodina si ottiene il nuovo segnale di banda base $y(t) = \tilde{x}(t) + n_c(t)$ in cui $n_c(t)$ è la componente in fase del rumore. A questo punto avviene l'operazione di *despreading* che si avvale della possibilità per il ricevitore di generare la stessa sequenza PN usata in trasmissione, in forma temporalmente sincrona, in modo da poter scrivere

$$\tilde{y}(t) = [\tilde{x}(t) + n_c(t)]p(t) = x(t)p^2(t) + n_c(t)p(t) = x(t) + \tilde{n}_c(t)$$

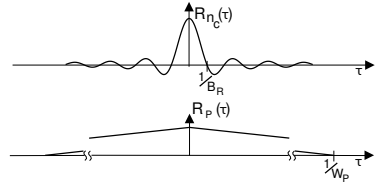
Pertanto, mentre il messaggio $x(t)$ è di nuovo quello precedente all'allargamento, il rumore ha subito a sua volta un effetto di allargamento. Un successivo filtraggio passa-basso con banda W_x pari a quella di segnale produce infine il risultato $y_d(t) = x(t) + n_d(t)$, in cui il segnale utile ha potenza $S_d = \mathcal{P}_x = 2S_R$, mentre al termine di disturbo additivo $n_d(t)$ è stata rimossa la potenza che cade al di fuori della banda di segnale.

⁵⁸ $\tilde{x}(t) \cos(\omega_0 t + \varphi)$ con φ v.a. a d.d.p. uniforme può essere considerato come il prodotto di due processi statisticamente indipendenti, la cui potenza è il prodotto delle potenze.

13.8.2.3 Prestazioni in presenza di rumore

Nel caso in cui $n(t)$ sia rumore bianco con densità di potenza $\mathcal{P}_n(f) = N_0/2$, la sua componente in fase dopo demodulazione omodina ha densità⁵⁹ $\mathcal{P}_{n_c}(f) = N_0 \text{rect}_{B_R}(f)$ e dunque autocorrelazione

$$\mathcal{R}_{n_c}(\tau) = N_0 B_R \text{sinc}(B_R \tau)$$



Allo scopo di valutare la densità di potenza $\mathcal{P}_{\tilde{n}_c}(f)$ del rumore dopo despreading, con l'aiuto della figura a lato osserviamo che l'autocorrelazione di $\tilde{n}_c(t)$ è pari a $\mathcal{R}_{\tilde{n}_c}(\tau) = \mathcal{R}_{n_c}(\tau) \mathcal{R}_p(\tau)$, e che $\mathcal{R}_{n_c}(\tau) \simeq 0$ con $|\tau| \gg \frac{1}{B_R} \ll \frac{1}{W_p}$, mentre $\mathcal{R}_p(\tau) \simeq 1$ con $|\tau| \ll T_p = \frac{1}{W_p}$: pertanto possiamo scrivere $\mathcal{R}_{\tilde{n}_c}(\tau) \simeq \mathcal{R}_{n_c}(\tau)$ e quindi

$$\mathcal{P}_{\tilde{n}_c}(f) \simeq \mathcal{P}_{n_c}(f) = N_0 \text{rect}_{B_R}(f)$$

La componente di rumore $n_d(t)$ in uscita dall'ultimo passa basso ha quindi una potenza $N_d = 2N_0 W_x$, permettendo di valutare il rapporto segnale-rumore dopo demodulazione come

$$\left(\frac{S}{N}\right)_d = \frac{2S_R}{2N_0 W_x} = \frac{S_R}{N_0 W_x}$$

ossia proprio pari all'*SNR di riferimento* (pag. 300), mostrando come la concatenazione delle operazioni di spreading e despreading *non alteri* le prestazioni del processo di modulazione nei confronti del rumore bianco.

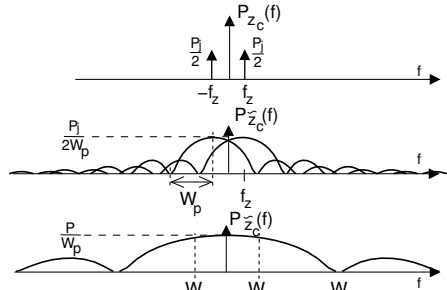
13.8.2.4 Prestazioni in presenza di un tono interferente

Mostriamo che se il termine di disturbo additivo $z(t)$ occupa una banda relativamente stretta in rapporto a B_R , allora la sua potenza dopo demodulazione risulterà *ridotta* di un fattore pari al guadagno di processo W_p/W_x . Ad esempio, consideriamo il caso di un tono interferente (*jammer*), in cui

$$z(t) = \sqrt{2\mathcal{P}_j} \cos(\omega_0 + \omega_z) t$$

con potenza $\overline{z^2} = \mathcal{P}_j$ alla frequenza $f_0 + f_z$. In tal caso $z_c(t) = \sqrt{2\mathcal{P}_j} \cos \omega_z t e$

$$\mathcal{P}_{z_c}(f) = \frac{\mathcal{P}_j}{2} [\delta(f - f_z) + \delta(f + f_z)] \tag{13.34}$$



Una volta che il tono interferente viene demodulato e moltiplicato per $p(t)$, esso produce $\mathcal{P}_{\tilde{z}_c}(f) = \mathcal{P}_{z_c}(f) * \mathcal{P}_p(f)$, mostrato alla riga centrale della figura a lato⁶⁰. Notiamo ora che la massima interferenza si ottiene quando $|f_z| \ll W_p$, al limite pari a zero, come mostrato all'ultima riga della figura in scala espansa per il

⁵⁹vedi § 12.1.2

⁶⁰Il risultato si ottiene tenendo conto delle eq. (13.33) e (13.34)

caso limite di $f_z = 0$. Pertanto, il limite superiore della potenza interferente uscente dal passa basso con banda W_x è

$$\overline{z_d^2} = \int_{-W_x}^{W_x} \mathcal{P}_{z_c}(f) df \leq 2W_x \frac{\mathcal{P}_j}{W_p}$$

e dunque il rapporto segnale-interferente diviene

$$\left(\frac{S}{z_d^2} \right)_d \geq 2S_R \frac{W_p}{2W_x \mathcal{P}_j} = \frac{S_R}{\mathcal{P}_j} \frac{W_p}{W_x}$$

mostrando quindi un miglioramento pari proprio al guadagno di processo.

13.8.2.5 Accesso multiplo

Una frequente applicazione delle tecniche spread spectrum consiste nel permettere la comunicazione *contemporanea* di una pluralità di soggetti, possibile qualora ognuno di essi adotti una diversa sequenza PN: la tecnica prende allora il nome di CDMA (*Code Division Multiple Access*). Mostriamo che in tal caso per ogni comunicazione l'effetto delle altre si riduce ad un modesto innalzamento del rumore di fondo, tanto più piccolo quanto minore è il valore della *intercorrelazione* tra i codici PN utilizzati.

Dopo la demodulazione, il termine interferente $z(t)$ causato da N diversi utenti, ognuno con un diverso codice $p_n(t)$ e segnale dati $x_n(t)$, può essere scritto come

$$z(t) = \sum_{n=1}^N A_n x_n(t - \tau_n) p_n(t - \tau_n) \cos \theta_n$$

in cui A_n , τ_n e $\cos \theta_n$ sono rispettivamente ampiezza, ritardo di simbolo e fase della portante relativi all' n -esimo utente. Assumendo ora eguali tra loro le ampiezze del segnale utile $x(t)$ e degli interferenti, dopo il despreading otteniamo

$$\tilde{y}(t) = x(t) + \left[\sum_{n=1}^N x_n(t - \tau_n) p_n(t - \tau_n) \cos \theta_n \right] p(t)$$

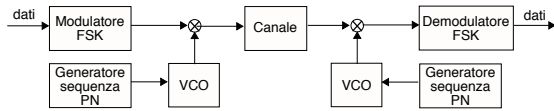
Se rappresentiamo ora il filtro passa basso di fig. 13.17 come un integratore esteso ad un periodo di bit, ovvero un filtro adattato al segnale NRZ, il valore della sua uscita campionata al termine della durata del k -esimo periodo di bit risulta

$$\begin{aligned} \tilde{y}(t_k) &= x(t_k) + \sum_{n=1}^N \left[\cos \theta_n \int_{kT_b}^{(k+1)T_b} x_n(t - \tau_n) p_n(t - \tau_n) p(t) dt \right] \\ &= x(t_k) + z(t_k) \end{aligned}$$

in cui $z(t_k)$ rappresenta il termine di interferenza complessiva da parte di tutti gli altri N utenti; dato che i valori di x_n possono essere ± 1 , l'integrale calcola in effetti l'intercorrelazione $\mathcal{R}_{pp_n}(\tau_n)$ (§ 9.1.4) tra $p(t)$ e le altre $p_n(t)$, calcolata per un ritardo τ_n . Pertanto, scegliendo la famiglia di sequenze pseudo-noise in modo che esibiscano una intercorrelazione molto ridotta (in teoria nulla, se le PN fossero *ortogonali*), è possibile rendere trascurabile l'effetto degli interferenti.

13.8.3 Frequency Hopping

Si tratta di una diversa tecnica *spread spectrum*, in cui la sequenza PN è di tipo *multivello*, ed è utilizzata in uno schema L-FSK incoerente (§ 13.5) per cambiare in continuazione la frequenza portante a cui avviene la trasmissione, tipicamente FSK anch'essa (vedi figura), da cui il nome di *Saltando di Frequenza* (traduzione di FREQUENCY HOPPING). Per una corretta ricezione, è necessaria una accurata sincronizzazione temporale tra la PN usata in trasmissione e quella in ricezione.



Dunque l'espansione spettrale si verifica ancora, ma stavolta non tutta la banda è occupata in modo *permanente* come nel DSSS, ma anzi durante ogni *salto* si occupa solo la banda necessaria alla modulazione *non allargata*. In questo caso un disturbo a banda stretta provoca interferenza solo durante il salto che occupa la sua stessa frequenza, e dunque può essere facilmente contrastata adottando una codifica di canale (§ 17.3). Inoltre, la tecnica FHSS è proficuamente impiegata in sistemi di accesso multiplo CDMA, dato che possono avvenire contemporaneamente più trasmissioni FHSS utilizzando per esse differenti sequenze PN a bassa intercorrelazione.

Se il periodo di chip (ovvero il tempo per cui il VCO permane sulla stessa frequenza) è più breve del periodo di simbolo, il sistema è detto *fast frequency hopping* o FFHSS, mentre se è maggiore è detto *slow FH* o SFHSS.

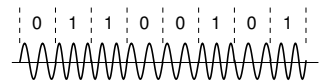
13.8.4 Time Hopping

In questo caso la trasmissione avviene su intervalli temporali molto ridotti, e dunque con una occupazione di banda molto elevata (a volte indicata come *ultra wide band* o UWB⁶¹); l'altro aspetto in comune con le tecniche a spettro espanso è il posizionamento pseudo-casuale degli impulsi nell'ambito di una trama temporale, in base ad una sequenza PN⁶².

13.9 Altre possibilità

Diamo un accenno ad altre tecniche di modulazione numerica, che non vengono sviluppate in questo capitolo.

MSK (*Minimum Shift Keying*) Simile all'FSK, ma con la variante di mantenere una *continuità di fase* tra simboli contigui, come accade per la famiglia di tecniche dette *Continuous phase modulation* o CPM⁶³. Questa caratteristica consente una riduzione della banda occupata, in virtù dell'assenza di brusche variazioni di ampiezza.



⁶¹<http://en.wikipedia.org/wiki/Ultra-wideband>

⁶²<http://en.wikipedia.org/wiki/Time-hopping>

⁶³http://en.wikipedia.org/wiki/Continuous_phase_modulation

Offset keying⁶⁴ Una variante del QAM e QPSK, in cui i periodi di simbolo per i 2 rami sono sfasati del 50%. La capacità di sincronizzazione del ricevitore risulta migliore, e la banda ridotta.

Partial response QAM⁶⁵ Il segnale modulato è filtrato, e si introduce deliberatamente una ISI in modo controllato. Migliora l'efficienza spettrale.

TCM (*Trellis Coded Modulation*)⁶⁶ TRELLIS significa *traliccio*, e TCM rappresenta un modo di realizzare una codifica di canale che impone vincoli alle possibili sequenze. Il numero di livelli è artificialmente aumentato, ma i punti della costellazione risultante non sono tutti possibili, anzi solo un ristretto numero lo è, in funzione dei simboli precedenti. Il risultato è un miglioramento della P_e (od una riduzione di E_b necessaria) a spese di una maggiore occupazione di banda.

Sistemi MIMO⁶⁷ Acronimo di *Multiple Input Multiple Output*, è realizzato utilizzando più di una antenna sia in trasmissione che in ricezione, attuando così una *trasmissione in diversità* in grado di aumentare la capacità di un collegamento radio grazie allo sfruttamento del fenomeno dei *cammini multipli*.

Radio Cognitiva⁶⁸ Analizza come trasmettere *intelligentemente* in modo da usare i migliori canali radio a disposizione: analizzando con continuità lo spettro radio, modifica conseguentemente i parametri di trasmissione e ricezione per permettere un utilizzo più efficiente dello spettro radio nella propria regione di spazio.

13.10 Appendici

13.10.1 Sincronizzazione

Nelle trasmissioni numeriche occorre ottenere allo stesso tempo sia la sincronizzazione della portante di demodulazione, nei limiti delle ambiguità di fase residue, sia la corretta temporizzazione di simbolo, per campionare le c.a. di b.f. ricevute al centro del periodo di simbolo⁶⁹, ed evitare l'insorgenza di ISI. Le due problematiche possono essere affrontate l'una di seguito all'altra, adottando le soluzioni già esposte⁷⁰. D'altra parte sono ora possibili varianti, come ad esempio il *Costas loop*⁷¹ che, utilizzando entrambe le c.a. di b.f., realizza l'aggancio di frequenza anche per tecniche di modulazione a portante soppressa, oppure procedure che tentano di acquisire per primo il sincronismo di simbolo, e quindi usano i valori delle c.a. di b.f. ricevute per effettuare correzioni alla fase dell'oscillatore di demodulazione⁷².

⁶⁴http://en.wikipedia.org/wiki/Phase-shift_keying#Offset_QPSK

⁶⁵<http://complextoreal.com/wp-content/uploads/2013/01/qpr.pdf>

⁶⁶http://en.wikipedia.org/wiki/Trellis_modulation

⁶⁷<http://en.wikipedia.org/wiki/MIMO>, <http://complextoreal.com/wp-content/uploads/2013/01/mimo.pdf>

⁶⁸http://en.wikipedia.org/wiki/Cognitive_radio

⁶⁹Od alla fine, come nel caso di un ricevitore a correlazione.

⁷⁰Per il recupero della portante si possono usare circuiti del tipo di § 11.4.2, mentre l'uso del PLL (§ 11.2.1.2) non è possibile a causa della assenza di residui di portante. Una volta acquisto il sincronismo di frequenza, quello di simbolo può essere ottenuto mediante schemi operanti in banda base, come quelli a pag. 5.6.2.

⁷¹http://en.wikipedia.org/wiki/Costas_loop

⁷²Vedi as es. http://en.wikipedia.org/wiki/Carrier_recovery#Decision-directed

Qualora la portante di demodulazione presenti una ambiguità di fase residua si può applicare la tecnica della codifica differenziale esposta al § 13.4, oppure inserire una sequenza di simboli noti (o *flag*) all'inizio della trama trasmissiva, in modo che il confronto tra i valori attesi e quelli ricevuti permetta di correggere tale ambiguità. Da notare che i *flag* o *trailer* ad inizio trama possono essere vantaggiosamente usati anche da schemi di recupero del clock del tipo di quelli al § 5.6.2.

13.10.2 Ortogonalità tra simboli sinusoidali

Al § 13.5.1 si è introdotta la modulazione FSK ortogonale, e nelle note è iniziata la discussione relativa alla condizione di ortogonalità tra la forma d'onda sinusoidale di durata T_s ricevuta, e quella prodotta al ricevitore come ingresso ai correlatori di un banco. Prendiamo pertanto ora in considerazione segnali del tipo $\cos[2\pi(f_0 + \Delta f_k)t + \phi_k]$, in cui è inclusa una differenza (o errore) di fase aleatorio tra le forme d'onda, in modo da esaminare le differenze tra il caso di modulazione coerente ed incoerente.

Iniziamo dunque con lo sviluppare l'espressione dell'integrale di intercorrelazione $\rho = \int_0^{T_s} \cos[2\pi(f_0 + m\Delta)t] \cos[2\pi(f_0 + n\Delta)t + \phi] dt$ facendo uso della relazione $\cos \alpha \cos \beta = \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)]$ e riferendoci per semplicità al caso di due frequenze contigue (ponendo $m = 0$ ed $n = 1$):

$$\rho = \frac{1}{2} \int_0^{T_s} \{ \cos[2\pi(2f_0 + \Delta)t + \phi] + \cos[2\pi\Delta t - \phi] \} dt = \quad (13.35)$$

$$= \frac{1}{2} \int_0^{T_s} \cos[2\pi(2f_0 + \Delta)t + \phi] dt + \frac{1}{2} \int_0^{T_s} \cos[2\pi\Delta t - \phi] dt \quad (13.36)$$

Per quanto riguarda il primo integrale, esso assume un valore nullo se $2f_0 + \Delta = \frac{k}{T_s}$, perché in tal caso in un intervallo T_s entrano un numero intero di periodi, ed il coseno ha valor medio nullo. Concentriamoci allora sul valore di Δ che annulla anche il secondo integrale, che riscriviamo facendo uso della relazione $\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta$:

$$\begin{aligned} & \int_0^{T_s} \cos(2\pi\Delta t - \phi) dt = \\ &= \int_0^{T_s} [\cos(2\pi\Delta t) \cos \phi + \sin(2\pi\Delta t) \sin \phi] dt = \\ &= \frac{\sin(2\pi\Delta t)}{2\pi\Delta} \Big|_0^{T_s} \cdot \cos \phi - \frac{\cos(2\pi\Delta t)}{2\pi\Delta} \Big|_0^{T_s} \cdot \sin \phi = \\ &= T_s \left[\frac{\sin(2\pi\Delta T_s)}{2\pi\Delta T_s} \cdot \cos \phi + \frac{1 - \cos(2\pi\Delta T_s)}{2\pi\Delta T_s} \cdot \sin \phi \right] \end{aligned} \quad (13.37)$$

Osserviamo ora che, nel caso in cui $\phi = 0$, il secondo termine della (13.37) si annulla per qualunque Δ . Esaminiamo quindi ora solamente il primo termine, individuando così il risultato relativo al caso di

Modulazione coerente Il termine $\frac{\sin(2\pi\Delta T_s)}{2\pi\Delta T_s} = \text{sinc}(2\Delta T_s)$ si annulla per $\Delta = \frac{k}{2T_s}$, e quindi la minima spaziatura tra portanti risulta $\Delta = \frac{1}{2T_s} = \frac{f_s}{2}$; pertanto, le frequenze utilizzate dovranno essere del tipo $f_0 + k \frac{f_s}{2}$.

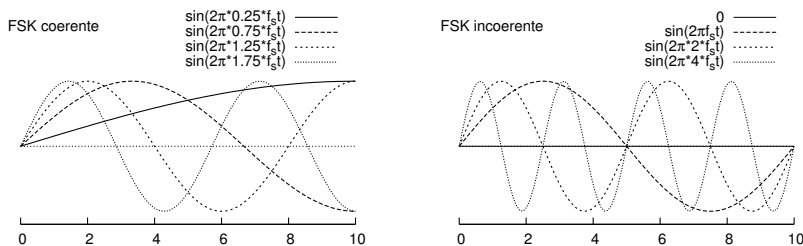


Figura 13.18: Forme d'onda ortogonali nei casi di modulazione *coerente* ed *incoerente*

Per fare in modo che il primo termine della (13.36) si annulli, deve sussistere la relazione $2f_0 + \Delta = 2f_0 + \frac{f_s}{2} = \frac{k}{T_s} = kf_s$, che fornisce la condizione $f_0 = f_s \frac{2k-1}{4}$, ossia f_0 deve essere scelta come uno tra i valori $\frac{1}{4}f_s, \frac{3}{4}f_s, \frac{5}{4}f_s, \frac{7}{4}f_s, \dots$. Notiamo come la spaziatura $\frac{f_s}{2}$ ora individuata tra i possibili valori per la portante, coincide con quella Δ tra le frequenze di segnalazione. Pertanto la parte sinistra della figura 13.18 rappresenta, disegnate in un intervallo pari a T_s , sia le portanti che possono essere usate, sia le prime frequenze che è possibile adottare per un modulazione FSK *coerente* basata sul valore minimo di f_0 pari a $\frac{1}{4}f_s$ ⁷³.

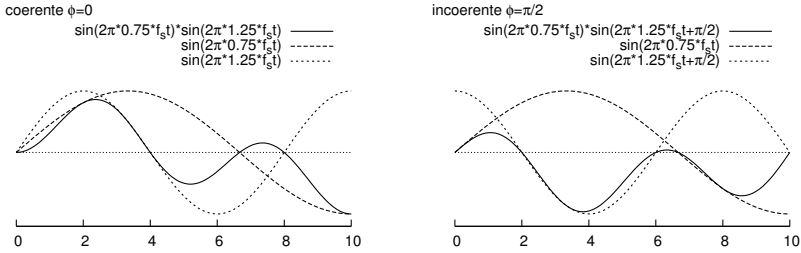
Nel caso in cui f_0 non assuma uno dei valori individuati, il primo termine di (13.36) non si annulla, ma se $f_0 \gg \frac{1}{T_s}$, risulta trascurabile rispetto al secondo. Pertanto, nel caso di trasmissioni su canali di tipo passa-banda, la scelta di f_0 non è più determinante.

Modulazione incoerente In questo caso si ha $\phi \neq 0$. In generale la (13.37) presenta entrambi i termini; mentre il primo (come già osservato) si annulla per $\Delta = \frac{k}{2T_s}$, il secondo invece è nullo solo se $\Delta = \frac{k}{T_s}$. Questa circostanza determina il risultato che occorre ora adottare una spaziatura tra portanti doppia della precedente, e pari cioè a $\Delta = f_s$.

Tornando ad esaminare la (13.36): ora il suo primo termine si annulla se $2f_0 + \Delta = 2f_0 + f_s = kf_s$, che determina la condizione $f_0 = f_s \frac{k-1}{2}$, ossia $f_0 = 0, \frac{1}{2}f_s, f_s, \frac{3}{2}f_s, \dots$. Notiamo come la spaziatura $\frac{f_s}{2}$ tra i possibili valori per la portante sia identica al caso precedente, mentre la spaziatura necessaria alle frequenze di segnalazione è raddoppiata. La circostanza che ora sia ammessa anche una “portante a frequenza nulla” consente quindi di tracciare la parte destra della figura 13.18, che mostra le prime frequenze di segnalazione che è possibile adottare per una modulazione FSK *incoerente* basata sul valore minimo di $f_0 = 0$.

Verifica grafica La figura che segue mostra il risultato del prodotto di due frequenze ortogonali distanti $\frac{f_s}{2}$ e calcolate in assenza di errore di fase (a sinistra) e con un errore di fase pari a $\phi = \frac{\pi}{2}$. Si può notare come in questo secondo caso si perda l’ortogonalità tra i segnali, essendo il risultato prevalentemente negativo.

⁷³Possiamo notare come la spaziatura tra le frequenze di segnalazione di $\frac{f_s}{2}$ fa sì che due forme d’onda con una differenza di frequenza $n\Delta = n\frac{f_s}{2}$ accumulino in un intervallo T_s una differenza di fase di $n\pi$, ovvero un numero intero di semiperiodi.



Discussione sull'ottimalità per $L \rightarrow \infty$ Osserviamo innanzitutto che il ricevitore a correlazione commette errore nel caso in cui il rumore sovrapposto al segnale di ingresso sia casualmente “simile” ad una delle cosinusoidi utilizzate per la trasmissione. In tal caso, l'uscita dell'integratore relativo alla frequenza “simile” può superare quella relativa alla frequenza trasmessa, e corrotta dal medesimo rumore.

All'aumentare di L (per f_b fisso) aumenta il periodo di simbolo $T_s = \frac{\log_2 L}{f_b}$ e quindi diventa sempre più “difficile” per il rumore emulare “bene” una delle frequenze di segnalazione, e quindi si riduce la probabilità di errore.

Chiaramente, all'aumentare di L aumenta proporzionalmente la complessità del ricevitore, che deve disporre di un numero di correlatori crescente. Pertanto, le prestazioni ideali per L che tende ad infinito rivestono solamente un interesse teorico.

Capitolo 14

Transito nei circuiti e distorsioni

Trattiamo qui della descrizione dei circuiti elettrici in termini di sistemi fisici, e delle modifiche indesiderate o *distorsioni* che questi generano nei segnali in transito, nei termini dell'effetto prodotto sulle caratteristiche spettrali e temporali per segnali di banda base o modulati. Dopo aver determinato le condizioni circuitali necessarie al massimo trasferimento di potenza, viene illustrata la contromisura alla distorsione *lineare* costituita dall'*equalizzazione*, nelle sue varie declinazioni. Senza farci mancare una doverosa riflessione sulla rappresentazione delle grandezze in *decibel*. Ma prendiamo la cosa per gradi.

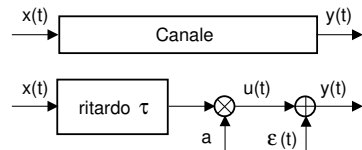
Elaborazione e Distorsione Quando un segnale viene “manipolato” di proposito, si dice che questo è *elaborato*. Se viceversa il segnale si altera per una causa indipendente dalla volontà, allora il segnale subisce una *distorsione*¹.

Canale perfetto Per valutare l'entità della distorsione, stabiliamo un criterio con cui distinguere la componente di segnale *utile* dal *disturbo*.

La ricezione di un segnale identico a quello trasmesso, tranne che per un fattore di scala ed un ritardo temporale, non altera la sostanza del messaggio: pertanto, un canale che presenti una risposta impulsiva

$$h(t) = a\delta(t - \tau)$$

viene indicato come *canale perfetto*, ed il segnale ricevuto $y(t) = u(t) = ax(t - \tau)$ è *tutto utile*. Se invece non vale quest'uguaglianza, viene definito *disturbo additivo* la differenza $\varepsilon(t) = y(t) - ax(t - \tau) = y(t) - u(t)$ (²). Infine, il rap-



¹ *Elaborazione di segnale* è la traduzione di SIGNAL PROCESSING, e così il risultato viene detto anche *processato*.

² Nella pratica, i valori a e τ non si conoscono, mentre invece possiamo disporre di coppie di segnali $(x(t), y(t))$. I valori vengono dunque definiti come quelli che *rendono SNR massimo* ovvero \mathcal{P}_ε minimo. Considerando segnali di potenza, ossia processi stazionari ergodici, si ha

$$\begin{aligned} \mathcal{P}_\varepsilon(a, \tau) &= E \left\{ (y(t) - ax(t - \tau))^2 \right\} = E \left\{ y^2(t) \right\} + a^2 E \left\{ x^2(t) \right\} - 2aE \left\{ y(t)x(t - \tau) \right\} = \\ &= \mathcal{P}_y + a^2 \mathcal{P}_x - 2a\mathcal{R}_{xy}(\tau) \end{aligned}$$

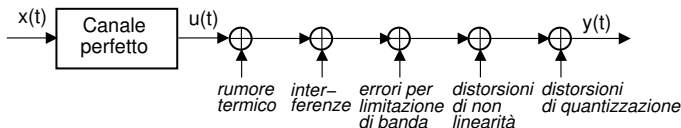
in cui si è operata la sostituzione $E \{ y(t)x(t - \tau) \} = \mathcal{R}_{yx}(-\tau) = \mathcal{R}_{xy}^*(\tau) = \mathcal{R}_{xy}(\tau)$.

porto $\frac{\mathcal{P}_u}{\mathcal{P}_\varepsilon}$ tra la potenza del segnale utile e quella del disturbo, prende il nome di *rapporto segnale rumore* (SNR).

Potenza di segnale e grandezze elettriche La caratterizzazione energetica dei segnali è stata finora svolta *a prescindere* dalla natura fisica degli stessi: ovvero, non si è mai specificato se si trattasse di tensioni o correnti, né si sono indicate le impedenze in gioco. Trattando ora di grandezze elettriche, le potenze di segnale, di tensione o di corrente, saranno misurate in $(Volt)^2$ o in $(Ampere)^2$ rispettivamente.

Esempio Sia $x(t)$ un segnale di tensione. La sua potenza \mathcal{P}_x ha unità di misura $[V^2]$, mentre la sua densità di potenza $\mathcal{P}_x(f)$ si esprime in $\left[\frac{V^2}{Hz}\right]$.

Disturbi additivi Per quanto riguarda \mathcal{P}_u , questa viene determinata a partire dalla conoscenza di \mathcal{P}_x applicando i risultati relativi ai trasferimenti energetici mostrati in questo capitolo alle condizioni reali determinate dai mezzi trasmissivi descritti al capitolo 15. D'altra parte, il termine di errore $\varepsilon(t)$ può essere dovuto al rumore additivo di natura termica (Cap. 16), a quello di natura interferente (ad es., vedi pag. 382), o di quantizzazione (§ 4.2.1.1), così come può insorgere a causa delle distorsioni lineari (§ 14.5) e/o di non linearità (§ 14.7) introdotte da un canale non perfetto.

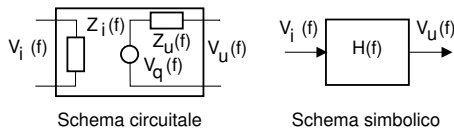


Spesso però queste diverse fonti di disturbo sono analizzate in forma separata e indipendente, ottenendo in definitiva un modello di ricezione in cui per ognuna di esse si è in grado di ottenere un valore di SNR_i dovuto a quella sola fonte. In tal caso, l' SNR_T complessivo viene calcolato come $SNR_T = \frac{1}{\sum_{i=1}^n \frac{1}{SNR_i}}$, vedi § 14.8.1.

14.1 Caratterizzazione dei circuiti

Numero di porte Le coppie di morsetti a cui applicare o da cui prelevare un segnale sono anche denominate *porte*. In questo senso i *generatori* ed le *impedenze* costituiscono reti ad *una* porta, mentre un sistema fisico dotato di relazione ingresso-uscita è una rete *due* porte.

Modello di rappresentazione Un circuito può essere rappresentato mediante il suo *modello circuitale*, in cui sono evidenziati generatori, resistenze, impedenze, generatori controllati..., oppure



Il valore di a che rende minimo $\mathcal{P}_\varepsilon(a, \tau)$ si ottiene eguagliando a zero la derivata: $\frac{\partial}{\partial a} \mathcal{P}_\varepsilon(a, \tau) = 2a\mathcal{P}_x - 2\mathcal{R}_{xy}(\tau) = 0 \Rightarrow a_{opt} = \frac{\mathcal{R}_{xy}(\tau)}{\mathcal{P}_x}$, che sostituita nell'espressione di \mathcal{P}_ε fornisce

$$\mathcal{P}_\varepsilon(\tau) = \mathcal{P}_y + \left(\frac{\mathcal{R}_{xy}(\tau)}{\mathcal{P}_x}\right)^2 \mathcal{P}_x - 2\frac{\mathcal{R}_{xy}(\tau)}{\mathcal{P}_x} \mathcal{R}_{xy}(\tau) = \mathcal{P}_y - \frac{(\mathcal{R}_{xy}(\tau))^2}{\mathcal{P}_x} = \mathcal{P}_y \left(1 - \frac{(\mathcal{R}_{xy}(\tau))^2}{\mathcal{P}_x \mathcal{P}_y}\right)$$

Il valore di \mathcal{P}_ε evidentemente è minimo per quel valore di $\tau = \tau_{opt}$ che rende massima $(\mathcal{R}_{xy}(\tau))^2$, ovvero per quella traslazione temporale che rende "più simili" i segnali di ingresso ed uscita.

il suo *schema simbolico*, in cui sono solo mostrate le relazioni funzionali tra i segnali in transito.

Proprietà delle reti due porte Le proprietà di *linearità, permanenza, realizzabilità ideale e fisica, stabilità*, già definite al § 9.6 per i sistemi fisici, possono essere verificate o meno nelle reti due porte.

14.2 Bipoli

Passivi Non contengono generatori, e sono caratterizzati dalle relazioni esistenti tra la tensione ai loro capi e la corrente che vi scorre (entrante). Il legame tra le due grandezze è una *convoluzione*

$$v(t) = i(t) * z(t)$$

in cui si suppone $i(t)$ la causa e $v(t)$ l'effetto. La trasformata di Fourier fornisce $V(f) = I(f) \cdot Z(f)$ in cui $Z(f)$ prende il nome di *impedenza*, e può scriversi nei termini di parte reale ed immaginaria:

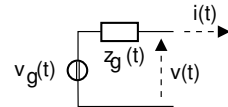
$$Z(f) = R(f) + jX(f)$$

in cui $R(f)$ (*resistenza*) è una funzione *pari* di f (e sempre positiva), mentre $X(f)$ (*reattanza*) è *dispari*: pertanto, $Z(f) = Z^*(-f)$ e quindi $z(t)$ è reale. Allo stesso tempo, è definita *l'ammettenza*

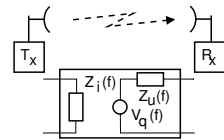
$$Y(f) = \frac{1}{Z(f)} = \frac{R(f) - jX(f)}{|Z(f)|^2}$$

e la corrispondente $y(t) = \mathcal{F}^{-1}\{Y(f)\}$, che permette di scrivere $i(t) = v(t) * y(t)$.

Attivi Sono bipoli al cui interno è presente un generatore. Per il teorema di THEVENIN,³ qualunque circuito può essere ridotto ad un generatore di tensione con in serie una impedenza (vedi figura), in cui $V_g(f)$ rappresenta la tensione a vuoto, ossia quando $I(f) = 0$ (considerata uscente nei bipoli attivi).

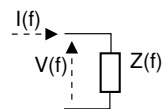


Esempio Una antenna trasmittente è schematizzabile come un bipolo passivo, di impedenza pari all'impedenza di ingresso dell'antenna, che assorbe la potenza erogata dal trasmettitore. Una antenna ricevente è schematizzabile come un generatore di tensione con in serie la propria impedenza di uscita, e trasferisce allo stadio di ingresso del ricevitore la potenza ricevuta per via elettromagnetica.



14.2.1 Potenza assorbita da un bipolo

Se ad un bipolo passivo di impedenza $Z(f)$ è applicato un segnale di tensione con spettro di densità di potenza $\mathcal{P}_v(f)$, la potenza *dissipata* sul bipolo, indicata come $\mathcal{W}_z(f)$ per distinguerla da quella di segnale, ha densità



$$\mathcal{W}_z(f) = \mathcal{P}_v(f) \cdot \Re\{Y(f)\} = \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} \quad \left[\frac{V^2}{\Omega \cdot Hz} \right] = \left[\frac{Watt}{Hz} \right] \quad (14.1)$$

³Vedi ad es. http://it.wikipedia.org/wiki/Teorema_di_Thévenin

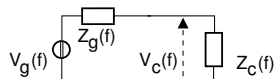
La dimostrazione della relazione illustrata è fornita in appendice 14.8.2. La dipendenza di $Y(f)$ dalla frequenza svolge pertanto *una azione filtrante*, e la potenza totale assorbita (o dissipata) su $Z(f)$ vale

$$\mathcal{W}_z = \int_{-\infty}^{\infty} \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} df \quad [Watt]$$

14.2.2 Connessione tra generatore e carico

La tensione ai capi del carico è valutabile applicando la *regola del partitore*⁴:

$$V_c(f) = V_g(f) \frac{Z_c(f)}{Z_c(f) + Z_g(f)}$$



ossia $V_c(f) = V_g(f) H(f)$ con $H(f) = \frac{Z_c(f)}{Z_c(f) + Z_g(f)}$. La potenza di segnale ai capi del carico vale $\mathcal{P}_{v_c}(f) = \mathcal{P}_{v_g}(f) |H(f)|^2$, e la potenza dissipata su $Z_c(f)$ risulta

$$\begin{aligned} \mathcal{W}_{z_c}(f) &= \mathcal{P}_{v_c}(f) \frac{R_c(f)}{|Z_c(f)|^2} = \mathcal{P}_{v_g}(f) \left| \frac{Z_c(f)}{Z_c(f) + Z_g(f)} \right|^2 \frac{R_c(f)}{|Z_c(f)|^2} = \\ &= \mathcal{P}_{v_g}(f) \frac{R_c(f)}{|Z_c(f) + Z_g(f)|^2} \end{aligned} \quad (14.2)$$

Osserviamo dunque che la potenza dissipata dal carico dipende da $Z_c(f)$, che compare sia a denominatore, che a numeratore con $R_c(f)$. Ci chiediamo allora quale sia il valore di Z_c che realizza il *massimo trasferimento di potenza* tra generatore e carico, sfruttando così appieno la *potenzialità* del generatore detta *potenza disponibile*.

14.2.2.1 Potenza disponibile e massimo trasferimento di potenza

La $\mathcal{W}_{z_c}(f)$ espressa da (14.2) risulta massimizzata qualora il suo denominatore viene reso minimo, e in appendice 14.8.3 si mostra che ciò avviene, per qualunque valore di $R_c(f)$, se si pone $X_c = -X_g$. Pertanto, per $Z_g(f)$ fissato, occorre porre anche $R_c = R_g$, ottenendo così il risultato cercato:

$$\begin{aligned} \text{se} \quad Z_c(f) &= Z_g^*(f) \\ \text{allora} \quad \mathcal{W}_{z_c}(f) &= \max_{Z_c(f)} \{\mathcal{W}_{z_c}(f)\} = \frac{\mathcal{P}_{v_g}(f)}{4R_g(f)} = \mathcal{W}_{d_g}(f) \end{aligned} \quad (14.3)$$

Il valore $\mathcal{W}_{d_g}(f) = \frac{\mathcal{P}_{v_g}(f)}{4R_g(f)}$ prende il nome di spettro di potenza *disponibile* del generatore, dipende solo dai suoi parametri $\mathcal{P}_{v_g}(f)$ e $R_g(f)$, e rappresenta la *massima potenza ceduta* ad un carico che è *adattato* per il *massimo trasferimento di potenza*⁵.

La potenza disponibile $\mathcal{W}_{d_g}(f)$ è pertanto *una grandezza caratteristica* del generatore; la potenza effettivamente ceduta ad un carico generico $Z_c(f) \neq Z_g^*(f)$, risulta inferiore a $\mathcal{W}_{d_g}(f)$ di una quantità

$$\alpha(f) = \frac{4R_g(f) R_c(f)}{|Z_g(f) + Z_c(f)|^2}$$

⁴http://it.wikipedia.org/wiki/Partitore_di_tensione

⁵E' bene notare esplicitamente che questo massimo è valido solo nel caso in cui non sia possibile modificare la $Z_g(f)$. Altrimenti, per un qualunque valore fissato di $Z_c(f)$, il massimo di $\mathcal{W}_{z_c}(f)$ si ottiene quando $Z_g(f) \rightarrow 0$.

(vedi appendice 14.8.4) e quindi in generale si ha $\mathcal{W}_{z_c}(f) = \alpha(f) \mathcal{W}_{d_g}(f)$.

14.2.2.2 Assenza di distorsioni lineari

Abbiamo già osservato come la tensione ai capi del carico abbia valore $V_c(f) = V_g(f) \cdot \frac{Z_c(f)}{Z_c(f)+Z_g(f)} = V_g(f) H(f)$. Ci chiediamo ora quali condizioni debbano sussistere affinché $H(f)$ si comporti come un *canale perfetto*, ovvero risulti $|H(f)| = \text{cost}$ e $\arg\{H(f)\} = 2\pi f\tau$: tali condizioni sono anche indicate come *assenza di distorsioni lineari*. Il risultato cercato si ottiene qualora si ponga

$$Z_c(f) = \alpha Z_g(f) \quad \text{con } \alpha \text{ reale}$$

infatti in tal caso risulta $H(f) = \frac{\alpha Z_g(f)}{(1+\alpha)Z_g(f)} = \frac{\alpha}{1+\alpha}$, ossia $H(f)$ costante. La condizione $Z_c(f) = \alpha Z_g(f)$ prende il nome di *adattamento di impedenza*, a volte ristretta al caso in cui $\alpha = 1$.

14.2.2.3 $Z_g(f)$ reale

Notiamo che massimo trasferimento di potenza ed assenza di distorsioni lineari possono sussistere *congiuntamente*, a patto che $Z_g(f) = R_g$, ovvero che sia il generatore che il carico siano caratterizzati da una impedenza reale.

14.3 Reti due porte

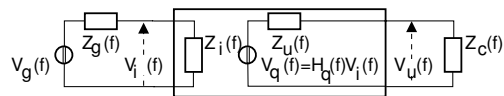
Come anticipato, un circuito accessibile mediante due coppie di morsetti è detto rete due porte, e può essere rappresentata secondo almeno due diversi formalismi: il *modello circuitale* e lo *schema simbolico*.

14.3.1 Modello circuitale

In figura è mostrato un possibile modello circuitale⁶ per una rete due porte, caratterizzata in termini di impedenza di ingresso $Z_i(f)$, di uscita $Z_u(f)$, e di un generatore controllato con tensione a vuoto $V_q(f) = H_q(f) V_i(f)$; le condizioni di chiusura sono quelle di un generatore $V_g(f)$ con impedenza $Z_g(f)$ in ingresso, e di una impedenza di carico $Z_c(f)$ in uscita.

La tensione $V_i(f)$ all'ingresso della rete

$$V_i(f) = V_g(f) H_i(f)$$



dipende da quella del generatore $V_g(f)$ mediante il rapporto di partizione $H_i(f) = \frac{Z_i(f)}{Z_g(f)+Z_i(f)}$, così come la tensione in uscita

$$V_u(f) = V_q(f) H_u(f)$$

⁶Sono chiaramente possibili modelli diversi, basati su topologie e relazioni differenti. Esistono infatti circuiti a T, ad L, a scala, a traliccio, a pigreco; le relazioni tra le grandezze di ingresso ed uscita possono essere espresse mediante modelli definiti in termini di impedenze, ammettenze, e parametri ibridi.

Il caso qui trattato è quello di un modello ibrido, con la particolarità di non presentare influenze esplicite dell'uscita sull'ingresso. Qualora il circuito che si descrive presenti una dipendenza, ad esempio di Z_i da Z_c , o Z_u da Z_g , questo deve risultare nell'espressione della grandezza dipendente. Viceversa, qualora il circuito presenti in ingresso un generatore controllato da una grandezza di uscita, il modello non è più applicabile.

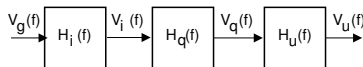
dipende da quella del generatore controllato $V_q(f)$ mediante il rapporto di partizione $H_u(f) = \frac{Z_c(f)}{Z_u(f) + Z_c(f)}$. Combinando queste relazioni, si ottiene che la funzione di trasferimento complessiva $H(f)$ risulta:

$$V_u(f) = V_g(f) H_i(f) H_q(f) H_u(f) = V_g(f) H(f) \quad (14.4)$$

La relazione mostra come $H(f)$ dipenda, oltre che dalla risposta in frequenza intrinseca della rete $H_q(f)$, anche dalle condizioni di adattamento che si realizzano in ingresso ed in uscita.

14.3.2 Schema simbolico

Lo stesso modello circuitale descritto può essere rappresentato equivalentemente mediante lo schema simbolico rappresentato a lato, in cui sono evidenziate le tre funzioni di trasferimento sopra ricavate, e che operano sui segnali indicati. Lo schema simbolico ha il vantaggio di trascendere dal modello circuitale adottato, e di rendere del tutto evidente come la funzione di trasferimento complessiva abbia origine dal prodotto di tre termini di cui solo uno ($H_q(f)$) rappresenta strettamente la rete.



14.3.3 Trasferimento energetico

Applicando ora la (14.1) alla potenza ceduta al carico $Z_c(f)$ dal generatore controllato $V_q(f)$, e tenendo conto della (14.4), si ottiene:

$$W_c(f) = P_{v_u}(f) \frac{R_c(f)}{|Z_c(f)|^2} = P_{v_g}(f) |H(f)|^2 \frac{R_c(f)}{|Z_c(f)|^2}$$

Proseguiamo ora l'analisi cercando di individuare una relazione di trasferimento energetico che possa rappresentare caratteristiche esclusive della rete.

Guadagno di tensione E' definito come il rapporto tra tensione di uscita e di ingresso:

$$G_v(f) = \frac{V_u(f)}{V_i(f)} = H_q(f) H_u(f)$$

Evidentemente, dipende dalle condizioni di chiusura della rete.

Guadagno di potenza E' il rapporto tra la potenza ceduta al carico e quella assorbita all'ingresso della rete:

$$\begin{aligned} G_W(f) &= \frac{W_c(f)}{W_i(f)} = P_{v_g}(f) |H(f)|^2 \frac{R_c(f)}{|Z_c(f)|^2} \cdot \frac{1}{P_{v_g}(f)} \frac{|Z_g(f) + Z_i(f)|^2}{R_i(f)} = \\ &= |H(f)|^2 \frac{R_c(f)}{R_i(f)} \cdot \left| \frac{Z_g(f) + Z_i(f)}{Z_c(f)} \right|^2 = \\ &= |H_q(f)|^2 \cdot \frac{R_c(f)}{R_i(f)} \cdot \left| \frac{Z_i(f)}{Z_u(f) + Z_c(f)} \right|^2 \end{aligned}$$

ed evidentemente è ancora funzione di $Z_c(f)$ ⁽⁷⁾. Notiamo ora che, qualora il carico sia adattato per il massimo trasferimento di potenza ($Z_c(f) = Z_u^*(f)$), la potenza ceduta a $Z_c(f)$ (e quindi $G_{\mathcal{W}}(f)$) è massima, e la dipendenza di $G_{\mathcal{W}}(f)$ da $Z_c(f)$ decade, risultando

$$G_{\mathcal{W}_{Max}}(f) = |H_q(f)|^2 \cdot \frac{|Z_i(f)|^2}{4R_u(f)R_u(f)} \quad (14.5)$$

Guadagno disponibile Il rapporto tra la potenza disponibile di uscita, e quella disponibile del generatore posto in ingresso della rete (indipendentemente dal fatto che l'ingresso della rete presenti o meno le condizioni per il massimo trasferimento di potenza) è detto *guadagno disponibile*, e risulta:

$$\begin{aligned} G_d(f) &= \frac{\mathcal{W}_{d_u}(f)}{\mathcal{W}_{d_g}(f)} = \frac{\mathcal{P}_{v_q}(f)}{4R_u(f)} \cdot \frac{4R_g(f)}{\mathcal{P}_{v_g}(f)} = \\ &= \mathcal{P}_{v_g}(f) |H_i(f)|^2 |H_q(f)|^2 \cdot \frac{R_g(f)}{R_u(f)} \cdot \frac{1}{\mathcal{P}_{v_g}(f)} = \\ &= |H_i(f)|^2 |H_q(f)|^2 \cdot \frac{R_g(f)}{R_u(f)} \end{aligned} \quad (14.6)$$

La relazione trovata mostra la dipendenza di $G_d(f)$ dalle condizioni di chiusura in ingresso; se l'impedenza $Z_g(f)$ del generatore è scelta in modo da conseguire il massimo trasferimento di potenza $Z_g(f) = Z_i^*(f)$, la dipendenza decade ed $|H_i(f)|^2 = \left| \frac{Z_i(f)}{Z_i(f) + Z_i^*(f)} \right|^2 = \frac{|Z_i(f)|^2}{4R_i^2(f)}$; considerando inoltre che $R_g(f) = R_i(f)$, la (14.6) diviene:

$$G_{d_{Max}}(f) = |H_q(f)|^2 \cdot \frac{|Z_i(f)|^2}{4R_u(f)R_i(f)} \quad (14.7)$$

Quest'ultima quantità è chiamata *guadagno disponibile DELLA RETE DUE PORTE* ed è quella che appunto dipende solo dai parametri della rete stessa. Confrontando (14.7) con (14.5) notiamo che $G_{d_{Max}}(f)$ coincide con $G_{\mathcal{W}_{Max}}(f)$. Confrontando (14.7) con (14.6), troviamo che $G_d(f) = |H_i(f)|^2 G_{d_{Max}}(f) \frac{4R_g(f)R_i(f)}{|Z_i(f)|^2}$. Considerando ora che $|H_i(f)|^2 \frac{1}{|Z_i(f)|^2} = \left| \frac{Z_i(f)}{Z_i(f) + Z_g(f)} \right|^2 \frac{1}{|Z_i(f)|^2} = \frac{1}{|Z_i(f) + Z_g(f)|^2}$, otteniamo

$$G_d(f) = \frac{4R_g(f)R_i(f)}{|Z_g(f) + Z_i(f)|^2} \cdot G_{d_{Max}}(f)$$

che ci consente di valutare $G_d(f)$ nelle reali condizioni di chiusura in ingresso, a partire da $G_{d_{Max}}(f) = G_{\mathcal{W}_{Max}}(f)$ che dipende solo dalla rete.

Collegamento generatore-carico mediante rete due porte

- Considerando generatore e porta di ingresso della rete adattati per il massimo trasferimento di potenza, la densità di potenza disponibile in uscita risulta

$$\mathcal{W}_{d_u}(f) = G_{d_{Max}}(f) \mathcal{W}_{d_g}(f)$$

⁷L'ultimo passaggio tiene conto che (omettendo la dipendenza da f):

$$|H|^2 \cdot \left| \frac{Z_g + Z_i}{Z_c} \right|^2 = \left| \frac{Z_i}{Z_i + Z_g} H_q \frac{Z_c}{Z_c + Z_u} \right|^2 \left| \frac{Z_g + Z_i}{Z_c} \right|^2 = |H_q|^2 \cdot \left| \frac{Z_i}{Z_u + Z_c} \right|^2$$

e dunque l'uscita della rete due porte si comporta come un generatore equivalente, caratterizzato da una nuova $\mathcal{W}_{du}(f)$ ed una diversa impedenza interna $Z_u(f)$.

- Nel caso in cui in ingresso non si verifichi il massimo trasferimento di potenza, $G_d(f)$ si riduce di un fattore $\beta(f) = \frac{4R_g(f)R_i(f)}{|Z_g(f)+Z_i(f)|^2}$, e dunque la nuova potenza disponibile di uscita risulta

$$\mathcal{W}_{du}(f) = \beta(f) \cdot G_{d_{Max}}(f) \mathcal{W}_{dg}(f) = \frac{4R_g(f)R_i(f)}{|Z_g(f)+Z_i(f)|^2} \cdot G_{d_{Max}}(f) \mathcal{W}_{dg}(f)$$

- Nel caso infine in cui il carico $Z_c(f)$ in uscita alla rete non sia adattato, quest'ultimo assorbe una potenza inferiore a $\mathcal{W}_{du}(f)$ e pari a (vedi Appendice 14.8.4)

$$\mathcal{W}_c(f) = \alpha(f) \cdot \mathcal{W}_{du}(f) = \frac{4R_u(f)R_c(f)}{|Z_u(f)+Z_c(f)|^2} \cdot \mathcal{W}_{du}(f)$$

Reti passive Se una rete non contiene elementi attivi, allora $G_{d_{Max}}(f) \leq 1$ per qualunque f . In questo caso si parla più propriamente di *attenuazione disponibile* $A_d(f) = \frac{1}{G_d(f)}$ ovvero $A_d(f)$ [dB] = $-G_d(f)$ [dB].

Reti in cascata Se più reti sono connesse tra loro l'una di seguito all'altra, e si verificano per ciascuna coppia le condizioni di massimo trasferimento di potenza tra lo stadio di uscita di una e quello di ingresso della successiva, il guadagno complessivo è il prodotto dei singoli guadagni disponibili: $G_{d_{Tot}} = G_{d1} \cdot G_{d2} \cdot \dots \cdot G_{dN}$.

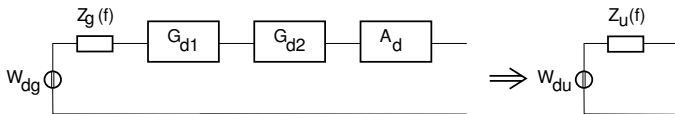
Esempio Alla figura seguente è mostrato un generatore con potenza disponibile \mathcal{W}_{dg} collegato ad una serie di tre reti due porte; l'effetto complessivo è quello di un nuovo generatore di uscita con potenza disponibile \mathcal{W}_{du} pari al prodotto di quella del generatore originario, moltiplicata per i guadagni disponibili delle reti attraversate, tenendo anche eventualmente conto delle attenuazioni supplementari:

$$\mathcal{W}_{du} = \mathcal{W}_{dg} \cdot G_{d1} \cdot G_{d2} \cdot \frac{1}{A_d} \cdot \frac{1}{A_s}$$

che può essere egualmente valutato operando in decibel, come

$$\mathcal{W}_{du} [dBW] = \mathcal{W}_{dg} [dBW] + G_{d1} [dB] + G_{d2} [dB] - A_d [dB] - A_s [dB]$$

in cui ovviamente, qualora \mathcal{W}_{dg} fosse espresso in *dBm* anziché *dBW*, lo stesso accadrebbe per \mathcal{W}_{du} .



Collegamento radio Con riferimento al circuito equivalente per una coppia di antenne di pag. 351, puntualizziamo che la potenza trasmessa è quella assorbita dall'impedenza di ingresso dell'antenna trasmittente, mentre quella ricevuta è quella caduta dal generatore equivalente dell'antenna ricevente, all'impedenza di ingresso del ricevitore.

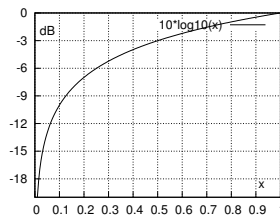
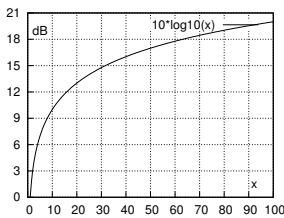
14.4 Misure di potenza in decibel

Ci sono almeno due buoni motivi matematici per misurare le grandezze in unità logaritmiche: la prima è che in tal modo si rappresentano in modo uniforme anche grandezze dalla dinamica molto elevata, e la seconda è che prodotti e rapporti, si trasformano in somme e sottrazioni. Inoltre, c'è almeno un buon motivo fisiologico, e cioè che la sensibilità dei nostri sensi segue naturalmente una legge logaritmica, ossia è necessario uno stimolo che aumenta in progressione geometrica, per produrre una sensazione che aumenta linearmente. Ciò posto, va anche detto che l'esperienza di insegnamento mostra come, anche se le misure in dB sono qui per aiutarci nei calcoli, esse sono anche uno degli argomenti in cui lo studente medio tende più facilmente a perdersi. Proviamo quindi a fare un pò di ordine!

La misura logaritmica Data una qualsiasi grandezza α , la sua misura in decibel⁸ è definita come

$$\alpha_{dB} = 10 \cdot \log_{10} \alpha \tag{14.8}$$

e descrive le relazioni mostrate nella figura sottostante, a sinistra per valori $\alpha > 1$, ed a destra per $\alpha < 1$, a cui corrispondono rispettivamente valori in decibel positivi e negativi. Inoltre, è mostrata una tabella con alcune corrispondenze che possono comunemente ricorrere: ad esempio, dato che $\log_{10} 2 = 0.30102\dots$, un valore α pari a 2 equivale a circa 3 dB.



α	α_{dB}
0	$-\infty$
10^{-3}	-30
1	0
2	~ 3
5	~ 7
10	10
10^n	$n \cdot 10$

Nota una grandezza espressa in dB, si può risalire al suo valore naturale, mediante l'ovvia relazione inversa

$$\alpha = 10^{\frac{\alpha_{dB}}{10}}$$

Misura relativa dei rapporti Per esprimere un rapporto $R = \frac{\alpha}{\beta}$ molto grande o molto piccolo, si ricorre spesso alla *scala logaritmica* definita dai dB, calcolando direttamente il rapporto in tali termini, ovvero eseguendo la differenza tra le grandezze α e β espresse in dB, in quanto

$$R_{dB} = 10 \cdot \log_{10} \frac{\alpha}{\beta} = 10 \cdot \log_{10} \alpha - 10 \cdot \log_{10} \beta = \alpha_{dB} - \beta_{dB} \tag{14.9}$$

Se le due grandezze α e β sono omogenee, come ad esempio due potenze di segnale \mathcal{P}_x e \mathcal{P}_y espresse in V^2 , o due potenze \mathcal{W}_x e \mathcal{W}_y espresse in *Watt*, allora il loro rapporto è un *numero puro*, e le sua misura in dB esprime *di quanti dB* il numeratore è maggiore

⁸Un decibel, per come è definito, è la decima parte del Bel. Chissà, forse dopo che definirono il Bel, si accorsero che era troppo grande ? :-)

(o minore) del denominatore. Conoscendo una delle due grandezze, ed il valore del loro rapporto, si può ovviamente risalire al valore dell'altra, ovvero ad esempio

$$\alpha_{dB} = R_{dB} + \beta_{dB} \quad (14.10)$$

ma, perché questa ovvia relazione possa avere una utilità pratica, occorre sapere cosa rappresenta β , dopodiché potremo concludere che α rappresenta la stessa cosa, ma R_{dB} decibel più grande. Per questo, si definisce la

Misura assoluta delle grandezze La (14.8) può essere usata per esprimere il *valore assoluto* di una grandezza, assieme alla sua unità di misura, se viene pensata come una applicazione della (14.9), ponendo il denominatore pari *all'unità di misura stessa*. Così ad esempio, la potenza \mathcal{W}_x di α *Watt* viene espressa come

$$\mathcal{W}_{x\,dBW} = 10 \cdot \log_{10} \frac{\alpha_{Watt}}{1\,Watt} \, dBW \quad (14.11)$$

ovvero, misurandola in dB *sopra il Watt*. Quindi, una potenza misurata in *dBW* può ricondursi alla corrispondente potenza in *Watt*, calcolando

$$\beta_{Watt} = 10^{\frac{\beta_{dBW}}{10}}$$

ed allo stesso modo, si può finalmente applicare la (14.10) per ottenere una grandezza effettiva:

$$\alpha_{dBW} = R_{dB} + \beta_{dBW}$$

Ovviamente, se qui β fosse stato *riferito al milliWatt* (e quindi misurato in dBm), anche per α si sarebbe ottenuta la medesima unità di misura.

Misura delle densità Spesso non si ha a che fare con una potenza complessiva, bensì con una *densità* espressa in $\frac{V^2}{Hz}$ o $\frac{W}{Hz}$ (a seconda se si tratti di potenza di segnale o fisica), ovvero con i suoi multipli e sottomultipli (*MHz*, *mWatt*...). Anche in questo caso, è possibile applicare la (14.8) per esprimere in unità logaritmiche la grandezza, purché intesa nel senso della (14.11), ossia indicando l'unità di misura di partenza, individuando così delle grandezze *assolute* misurate in $\frac{dBV^2}{Hz}$, $\frac{dBW}{Hz}$, $\frac{dBW}{MHz}$, $\frac{dBm}{MHz}$...

Quando poi si tratta di applicare una formula come quelle di progetto per i collegamenti, come ad es. la (15.12), occorre prestare attenzione a mantenere congruità dimensionale tra le grandezze usate, eventualmente convertendo dall'una e all'altra.

Corrispondenze tra grandezze Passare da una unità di misura in dB all'altra è molto semplice, basta infatti ricordare le equivalenti relazioni nelle unità lineari rispettive, ed aver presente la misura in dB dei rapporti più comuni. Per questo, ad esempio

- zero *dBW* equivalgono a 30 *dBm*, dato che 1 Watt = 1000 mWatt;
- -60 $\frac{dBW}{Hz}$ equivalgono a -30 $\frac{dBm}{Hz}$, per lo stesso motivo;
- 20 $\frac{dBW}{MHz}$ equivalgono a -40 $\frac{dBW}{Hz}$, dato che 1 MHz = 10^6 Hz

14.5 Distorsione lineare

E' dovuta al passaggio del segnale $x(t)$ attraverso un canale che non è perfetto, ossia un sistema fisico la cui risposta in frequenza $H(f) = H(f)e^{j\varphi_h(f)}$ non ha modulo costante e/o fase lineare. Questo determina una *modifica* della forma d'onda a causa della dipendenza delle componenti frequenziali del segnale di uscita $Y(f) = |Y(f)|e^{j\varphi_y(f)}$ oltre che da quelle in ingresso, anche dai valori di $H(f)$ come descritto a pag. 235. Queste distorsioni sono dette *lineari* in quanto risultato di una operazione *lineare* come è la convoluzione, verificando così il principio di sovrapposizione degli effetti, e rendendo la distorsione invertibile mediante un dispositivo di equalizzazione (§ 14.6). D'altra parte, risultando $|Y(f)| = |X(f)| \cdot |H(f)|$ e $\varphi_y(f) = \varphi_x(f) + \varphi_h(f)$, è possibile descrivere l'effetto della distorsione lineare considerando separatamente la risposta in frequenza $|H(f)|$ e quella in fase $\varphi_h(f)$, anche adottando per queste ultime rappresentazioni particolarmente utili a descrivere in modo complessivo l'entità delle distorsioni stesse.

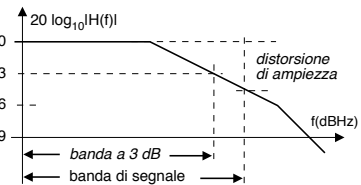
14.5.1 Guadagno di potenza in dB

E' definito dall'espressione

$$G_{dB}(f) = 10 \log_{10} |H(f)|^2 = 20 \log_{10} |H(f)|$$

Qualora l'occupazione di banda del segnale in transito si estenda su di una regione entro la quale $G_{dB}(f)$ non è costante, l'escursione di $G_{dB}(f)$ in tale banda è indicata come *distorsione lineare di ampiezza*, e quantificata appunto in dB.

Al contrario, qualora si specifichi quale debba essere la massima distorsione lineare in dB, mediante il grafico si individua la banda entro la quale $G_{dB}(f)$ si mantiene all'interno della fascia consentita, ottenendo così il valore della banda per distorsione assegnata (ad esempio, la *banda a 3 dB*, corrispondente alla frequenza di taglio, è definita per questa via).



14.5.2 Tempo di ritardo di gruppo

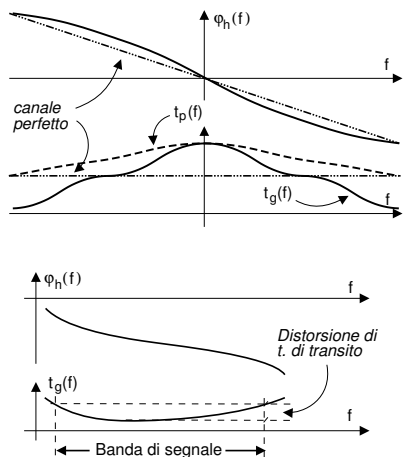
Le conseguenze dello scostamento della risposta in fase $\varphi_h(f)$ rispetto alle condizioni di canale perfetto possono essere sinteticamente espresse mediante la quantità

$$t_g(f) = -\frac{1}{2\pi} \frac{d}{df} \varphi_h(f) \quad (14.12)$$

che rappresenta il ritardo subito da *un gruppo di frequenze vicine a f*. Per chiarire il senso di questa definizione, consideriamo un segnale AM-BLD-PS $x(t) = a(t) \cos(2\pi f_0 t)$ la cui densità di potenza è come noto concentrata attorno a f_0 : in appendice 14.8.5 viene mostrato che, in prima approssimazione, questo si presenta in uscita da $H(f)$ con espressione

$$y(t) = a(t - t_g(f_0)) \cos(2\pi f_0(t - t_p(f_0))) \quad (14.13)$$

dove $t_p(f) = -\frac{\varphi_h(f)}{2\pi f}$ rappresenta il ritardo della portante, mentre $t_g(f)$ è fornito dalla (14.12). Nel caso di canale perfetto si avrebbe $\varphi_h(f) = -2\pi f\tau$, e quindi $t_g = t_p = \tau$ per qualunque frequenza; in caso contrario i due valori possono differire, come mostrato



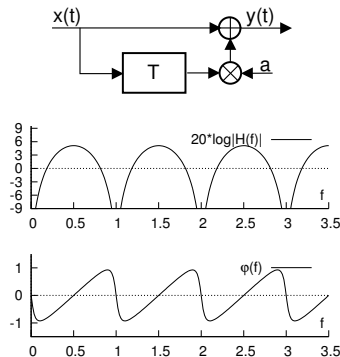
nella figura a lato, in cui notiamo che $t_g(f) \simeq \tau$ alle frequenze per cui $\varphi_h(f)$ viaggia parallela alla risposta in fase del canale perfetto, mentre risulta $t_p(f) = \tau$ quando $\varphi_h(f)$ la interseca.

La rappresentazione di $\varphi_h(f)$ nei termini della sua derivata normalizzata $t_g(f) = -\frac{1}{2\pi} \frac{d}{df} \varphi(f)$, misurabile strumentalmente, è utilizzata per valutare in modo grossolano l'entità della *distorsione di fase*, per questo detta anche *distorsione di tempo di transito*, espressa nei termini del massimo scarto tra i valori di $t_g(f)$ nell'ambito della banda di segnale, come mostrato in figura. Ovviamente, minore è questa differenza, e minore risulta l'entità della distorsione subita.

Esempio Il filtro trasversale in figura rappresenta un collegamento radio in cui si verifica una eco dovuta a riflessione. Risulta⁹:

$$|H(f)|^2 = 1 + a^2 + 2a \cos 2\pi fT$$

e dunque è presente sia distorsione lineare di ampiezza che di fase (mostrate in figura per $\alpha = .8$ e $T = 1$). In particolare, $|H(f)|^2$ è periodica di periodo $f = \frac{1}{T}$ e dunque può produrre una forte attenuazione (per $a \simeq 1$) anche a frequenze elevate. Se poi T cambia (perché si sposta il corpo riflettente, oppure si spostano trasmettitore o ricevitore) allora si ha a che fare con un canale tempo-variante, il cui studio non è per ora affrontato.



14.5.3 Effetto della distorsione lineare sui segnali

14.5.3.1 Segnali di banda base

E' noto che l'orecchio umano non è sensibile alle spettro di fase¹⁰ e dunque le distorsioni di fase non modificano la qualità del segnale audio - mentre alterazioni di modulo sì.

Se il segnale è numerico, una risposta di fase *lineare* è importante, perché altrimenti i diversi ritardi di transito alterano l'arrivo delle componenti spettrali dell'impulso elementare, che si deforma e perde la caratteristica di Nyquist. In conseguenza, insorge il fenomeno di interferenza intersimbolica (ISI) e aumenta la probabilità di errore.

⁹L'espressione di $|H(f)|^2$ è stata ricavata al § 9.8.2. Per la fase (mostrata in figura), osservando che $H(f) = 1 + ae^{-j2\pi fT}$ e che $\varphi(f) = \arctan \frac{\Im}{\Re}$, si ottiene $\varphi(f) = \arctan \frac{a \sin 2\pi fT}{1 + \cos 2\pi fT}$.

¹⁰Al contrario, è sensibile alle sue variazioni: queste ultime sono infatti elaborate dal cervello per estrarne informazioni relative al "movimento" dei suoni percepiti. Confrontando i ritardi differenti e variabili dei segnali pervenuti alle orecchie, si può comprendere se la sorgente degli stessi è in movimento.

14.5.3.2 Segnali modulati

Abbiamo mostrato al § 10.3.2 che l'involuppo complesso $\underline{y}(t) = \frac{1}{2}\underline{x}(t) * \underline{h}(t)$ presenta C.A di B.F. pari a

$$\begin{cases} y_c(t) = \frac{1}{2} [x_c(t) * h_c(t) - x_s(t) * h_s(t)] \\ y_s(t) = \frac{1}{2} [x_s(t) * h_c(t) + x_c(t) * h_s(t)] \end{cases}$$

Si è anche osservato come, nel caso in cui $H(f)$ risulti a simmetria coniugata rispetto ad f_0 , si ottiene $h_s(t) = 0$. In tal caso, l'effetto si riduce a quello che si avrebbe in banda base, filtrando il messaggio modulante mediante $h_c(t)$. Pertanto, l'effetto può essere rimosso *equalizzando* i segnali di banda base $y_c(t)$ e $y_s(t)$ con un filtro $H_e(f) = \frac{ae^{j2\pi f\tau}}{H_c(f)}$ (vedi § 10.3.3).

Segnali a banda stretta Nel caso in cui un segnale modulato $x(t)$ presenti una occupazione di frequenza B molto piccola rispetto alla frequenza portante f_0 , si assume spesso che $H(f)$ nella banda di segnale non vari di molto: ad esempio dal punto di vista del modello circuitale, ciò corrisponde a realizzare le condizioni di adattamento di impedenza in forma approssimata ponendo $Z_g(f) = Z_i(f_0)$ e $Z_c(f) = Z_u(f_0)$, dato che per frequenze $|f - f_0| < \frac{B}{2}$ con $B \ll f_0$, le impedenze $Z_i(f)$ e $Z_u(f)$ non variano di molto. Se invece si preferisce optare per le condizioni di massimo trasferimento di potenza, o più in generale per ciò che riguarda $H_q(f)$, la condizione $B \ll f_0$ spesso permette comunque di approssimare sia modulo che fase come pressoché costanti e pari al valore assunto per $f = f_0$, ossia $H(f) \simeq H(f_0) = G_0 e^{j\phi_0}$. Mostriamo che questa approssimazione permette di trascurare l'effetto delle distorsioni lineari, che nel caso attuale equivalgono ad una semplice rotazione degli assi dell'involuppo complesso, reversibile scegliendo opportunamente la fase della portante di demodulazione.

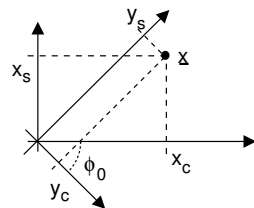
Nelle ipotesi poste, risulta infatti $\underline{H}(f) = 2H^+(f + f_0) = 2G_0 e^{j\phi_0}$ e quindi $\underline{h}(t) = \mathcal{F}^{-1}\{\underline{H}(f)\} = 2G_0 e^{j\phi_0} \delta(t)$; pertanto l'involuppo complesso di $h(t)$ è un impulso di area complessa $2G_0 e^{j\phi_0} = 2G_0 (\cos \phi_0 + j \sin \phi_0)$. All'uscita del canale $H(f)$ troviamo quindi

$$\begin{aligned} \underline{y}(t) &= \frac{1}{2} \cdot \underline{x}(t) * \underline{h}(t) = (x_c(t) + jx_s(t)) * G_0 (\cos \phi_0 + j \sin \phi_0) \delta(t) = \\ &= G_0 [(x_c(t) \cos \phi_0 - x_s(t) \sin \phi_0) + j(x_c(t) \sin \phi_0 + x_s(t) \cos \phi_0)] \end{aligned}$$

che identifica la trasformazione subita come una rotazione

$$\begin{bmatrix} y_c(t) \\ y_s(t) \end{bmatrix} = \begin{bmatrix} \cos \phi_0 & -\sin \phi_0 \\ \sin \phi_0 & \cos \phi_0 \end{bmatrix} \begin{bmatrix} x_c(t) \\ x_s(t) \end{bmatrix}$$

con una matrice dei coefficienti costante. Il risultato della rotazione è esemplificato in figura.



Modulazione di ampiezza

BLD-PS In questo caso $x_s(t) = 0$ ed allora $\begin{cases} y_c(t) = \frac{1}{2} [x_c(t) * h_c(t)] \\ y_s(t) = \frac{1}{2} [x_c(t) * h_s(t)] \end{cases}$. Adottando, ad esempio, una demodulazione omodina, si ottiene un segnale demodulato pari a $d(t) = x_c(t) * h_c(t)$, equivalente al caso di distorsione lineare di banda base.

BLD-PI Il problema maggiore con la portante intera può verificarsi se $H(f_0)$ è molto ridotto (ad esempio a causa di una attenuazione selettiva, esemplificata all'esempio di pag. 360), perché in tal caso il demodulatore involuppo non funziona più.

BLU In questo caso il segnale modulato contiene ambedue le C.A. di B.F., e dunque la presenza di distorsioni lineari provoca il fenomeno noto come *intermodulazione tra componenti analogiche di bassa frequenza*, in quanto in entrambe ($y_c(t)$, $y_s(t)$), si trovano entrambe ($x_c(t)$, $x_s(t)$), mescolate tra loro tramite ($h_c(t)$, $h_s(t)$).

Modulazione angolare Qualora un segnale modulato angolarmente attraversi un canale che presenta distorsioni lineari (di modulo, di fase, od entrambe), si verificano due fenomeni indicati come conversione PM-AM e PM-PM. Si manifesta infatti una modulazione di ampiezza sovrapposta, ed anche la modulazione di fase presenta delle alterazioni. Mentre la modulazione AM “parassita” può essere rimossa da un limitatore in ricezione, quella di fase no; inoltre quest’ultima presenta anche termini non-lineari e dunque non eliminabili mediante equalizzazione.

14.5.4 Calcolo dell’SNR

Se il segnale ricevuto ha subito distorsioni di ampiezza, lo spettro di densità di potenza si è deformato, ed il calcolo dell’SNR al ricevitore deve tenere conto dell’effetto filtrante introdotto dal canale attraversato, ossia:

$$SNR = \frac{\mathcal{P}_y}{\mathcal{P}_N} = \frac{\int_B \mathcal{P}_y(f) df}{\int_B \mathcal{P}_N(f) df} = \frac{\int_B \mathcal{P}_x(f) |H(f)|^2 df}{N_0 B}$$

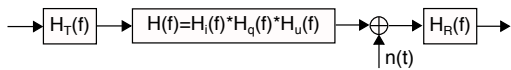
in cui si è indicata con B la banda a frequenze positive occupata dal segnale.

14.6 Equalizzazione

Come anticipato, si tratta dell’operazione necessaria qualora la funzione di trasferimento complessiva del mezzo trasmissivo e delle sue chiusure $H(f) = H_i(f) H_q(f) H_u(f)$ non soddisfi la condizione di canale perfetto, e può essere intrapresa inserendo elementi filtranti presso il trasmettitore ($H_T(f)$) ed il ricevitore ($H_R(f)$), realizzati in modo da ottenere $H_T(f) H(f) H_R(f) = ae^{-j2\pi f\tau}$. Nel seguito ci riferiamo a segnali di banda base, ovvero alle componenti analogiche di b.f. di segnali modulati.

Se non si utilizzano entrambi i filtri, ma uno solo di essi, si può delegare il compito dell’equalizzazione al solo $H_T(f)$, riservando ad $H_R(f)$ il ruolo di filtrare il rumore esterno alla banda di segnale, come fatto ai cap. 12 e 15. Nel caso quindi in cui la $H(f)$ del canale sia nota a priori, la sua inversione può essere approssimata sintetizzando $H_T(f)$ mediante un metodo di progetto filtri analogici¹¹.

D’altra parte, se l’equalizzazione è svolta solo in ricezione, si possono applicare tecniche volte a *stimare* la $H(f)$ non nota¹², e quindi sintetizzare una realizzazione



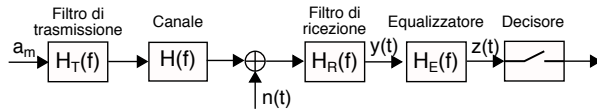
¹¹http://en.wikipedia.org/wiki/Analogue_filter

¹²Operazione necessaria se il canale *cambia* da un collegamento all’altro, come nel transito in una rete commutata, o nel caso di una comunicazione radiomobile.

numerica di $H_R(f)$ tale da invertire $H(f)$. In questo caso però si può determinare una “colorazione” del rumore in ingresso al ricevitore, che può essere esaltato anche di molto nelle bande per le quali $|H(f)|$ è praticamente nullo, come per un canale con echi di ampiezza simile, vedi pag. 360.

14.6.1 Equalizzazione numerica

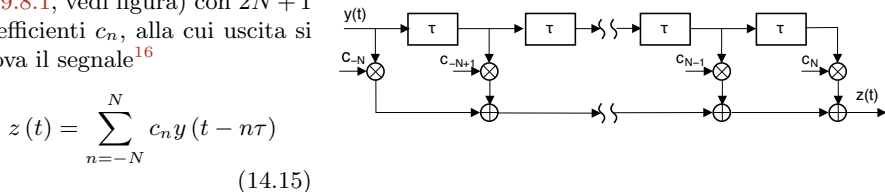
Per le trasmissioni numeriche è già prevista la presenza di un filtro $H_T(f)$ con funzione di sagomatore di impulso (come componente del codificatore di linea (§ 5.1.2), mentre la teoria del filtro adattato (§ 9.5) mostra che realizzando¹³ $H_R(f) = H_T^*(f)$ si ottiene il minimo SNR all’istante di decisione; pertanto ora lo schema di elaborazione è del tipo



in cui $H_R(f)H_T^*(f) = G(f)$ assieme realizzano un filtro di Nyquist¹⁴, e l’equalizzazione di $H(f)$ è demandata per intero al ricevitore, per mezzo di $H_E(f)$, al cui ingresso si presenta il segnale

$$y(t) = \sum_{m=-\infty}^{\infty} a_m x(t - mT_s) + \nu(t) \tag{14.14}$$

che dipende oltre che dai simboli a_m della sequenza informativa, dalla risposta impulsiva complessiva $x(t) = h_T(t) * h(t) * h_R(t)$, mentre $\nu(t)$ tiene conto dell’effetto di $H_R(f)$ su $n(t)$. Il segnale $y(t)$ viene quindi campionato con periodo di campionamento $\tau \leq T_s$ ¹⁵, ed elaborato realizzando $H_E(f)$ in forma numerica come un filtro FIR (§ 9.8.1, vedi figura) con $2N + 1$ coefficienti c_n , alla cui uscita si trova il segnale¹⁶



$$z(t) = \sum_{n=-N}^N c_n y(t - n\tau) \tag{14.15}$$

Osserviamo ora che qualora $H_E(f)$ riesca nel suo intento l’ISI introdotta dal canale è completamente rimossa, e la sua uscita in corrispondenza degli istanti di simbolo $z(mT_s)$ dipende dal valore di un solo a_m , oltre che dal rumore filtrato via

¹³Si veda l’eq. (9.8) a pag. 230, tralasciando il termine di fase lineare necessario a rendere $h_R(t)$ causale.

¹⁴Sia che si abbia $H_R(f) = H_T(f) = \sqrt{G(f)}$ come per il caso del ricevitore ottimo § 5.7, sia nel caso in cui $H_T(f) = G(f)$ e $H_R(f)$ filtra il rumore.

¹⁵Dato che per evitare aliasing (§ 4.1.1) la frequenza di campionamento $f_c = \frac{1}{\tau}$ deve risultare maggiore del doppio della massima frequenza presente in $y(t)$, scegliamo $\frac{1}{\tau} = 2f_s$, ovvero adottiamo una frequenza di campionamento pari al doppio di quella di simbolo. Infatti, un segnale dati di banda base a coseno rialzato occupa una banda a frequenze positive pari a $B = \frac{f_s}{2}(1 + \gamma)$, e dunque scegliendo $f_c = 2f_s$ ci cauteliamo anche nel caso di $\gamma = 1$. La scelta di porre $\tau < T_s$ viene indicata anche con il termine di *fractionally spaced equalizer* o FSE.

¹⁶La (14.15) rappresenta l’uscita di un filtro FIR *anti-causale*, dato che $z(t)$ dipende da valori futuri di ingresso, fino a $y(t + N\tau)$, mentre l’espressione *causale* che rispecchia la notazione usata nella figura dovrebbe essere $z(t) = \sum_{n=-N}^N c_n y(t - n\tau - N\tau)$, che corrisponde ad un semplice ritardo dell’ingresso di $N\tau$, trascurato per semplicità di notazione nel seguito.

$H_R(f)H_E(f)$; pertanto, i coefficienti c_n possono essere scelti direttamente con l'obiettivo di annullare l'ISI, anziché con quello di invertire $H(f)$; in particolare, la (14.15) è dello stesso tipo della (5.20) a pag. 115, e difatti l'approccio dello *zero forcing* ivi esposto è tuttora applicabile, tranne per il fatto che in quel caso è stata completamente ignorata la presenza del rumore; illustriamo dunque una tecnica che invece ne tiene conto.

14.6.1.1 Equalizzatore MMSE

Questa impostazione individua i coefficienti del filtro FIR $H_E(f)$ come quelli di un *filtro di WIENER*¹⁷, ovvero in modo da minimizzare la potenza dell'errore commesso alla sua uscita, cioè tali da rendere *minimo l'errore quadratico medio* \bar{e}^2 (o MEAN SQUARE ERROR) tra i campioni $z(mT_s)$ della (14.15) in uscita da $H_E(f)$, ed i corrispondenti valori a_m della sequenza informativa:

$$\bar{e}^2 = E \{ (z(mT_s) - a_m)^2 \} = E \left\{ \left(\sum_{n=-N}^N c_n y(mT_s - n\tau) - a_m \right)^2 \right\} \quad (14.16)$$

La minimizzazione di \bar{e}^2 si ottiene calcolandone le derivate rispetto ai c_k ed eguagliando il risultato a zero, ovvero

$$\begin{aligned} \frac{\partial}{\partial c_k} \bar{e}^2 &= 2E \left\{ \left(\sum_{n=-N}^N c_n y(mT_s - n\tau) - a_m \right) y(mT_s - k\tau) \right\} = \\ &= 2 \left(\sum_{n=-N}^N c_n E \{ y(mT_s - n\tau) y(mT_s - k\tau) \} - E \{ a_m y(mT_s - k\tau) \} \right) \end{aligned} \quad (14.17)$$

da cui si ottengono le condizioni

$$\sum_{n=-N}^N c_n \mathcal{R}_Y(n-k) = \mathcal{R}_{YA}(k) \quad k = 0, \pm 1, \pm 2, \dots, \pm N \quad (14.18)$$

che rappresentano un sistema di $2N+1$ equazioni nelle incognite c_n , note come equazioni di *Wiener-Hopf*, e la cui matrice dei coefficienti e vettore dei termini noti sono costituiti rispettivamente dall'autocorrelazione $\mathcal{R}_Y(n-k) = E \{ y(mT_s - n\tau) y(mT_s - k\tau) \}$ tra campioni del segnale ricevuto, e dall'intercorrelazione $\mathcal{R}_{YA}(k) = E \{ y(mT_s - k\tau) a_m \}$ tra i campioni di $y(t)$ e la sequenza dei simboli trasmessi, dove il valore atteso è calcolato rispetto alla variabilità della sequenza a_m e del rumore.

Soluzione delle equazioni di Wiener-Hopf In effetti, $\mathcal{R}_Y(n)$ e $\mathcal{R}_{YA}(n)$ *non sono note* al ricevitore, ma nell'ipotesi di stazionarietà ed ergodicità per $y(t)$ ed a_n , queste possono essere *stimate* a partire da una fase iniziale di *apprendimento* durante la quale viene trasmesso un segnale di test, per il quale il ricevitore calcola le medie temporali $\hat{\mathcal{R}}_Y(n) = \frac{1}{K} \sum_{k=1}^K y(kT_s - n\tau) y(kT_s)$ e $\hat{\mathcal{R}}_{YA}(n) = \frac{1}{K} \sum_{k=1}^K y(kT_s - n\tau) a_k$, che quindi utilizza al posto delle medie di insieme nella (14.18).

Adottando quindi per la (14.18) una notazione matriciale del tipo

$$\mathbf{Bc} = \mathbf{d} \quad (14.19)$$

¹⁷http://en.wikipedia.org/wiki/Wiener_filter

dove $\mathbf{B} = [b_{nk}] = [\mathcal{R}_Y(n-k)]$ è una matrice $(2N+1) \times (2N+1)$, $\mathbf{c} = [c_n]$ sono i coefficienti del filtro e $\mathbf{d} = [d_n] = [\mathcal{R}_{YA}(n)]$, i coefficienti ottimi possono ottenersi come $\mathbf{c}_{opt} = \mathbf{B}^{-1}\mathbf{d}$; la proprietà di simmetria di $\mathcal{R}_Y(n)$ rende poi \mathbf{B} una matrice di *Toeplitz*¹⁸, per la quale esistono metodi che semplificano in parte la soluzione diretta del sistema di equazioni, vedi anche la nota 14 a pag. 477.

Metodo del gradiente La (14.16) può essere minimizzata anche adottando metodi *iterativi* che evitano completamente il problema di risolvere la (14.19) in modo diretto. In particolare, può essere usata la tecnica nota come *discesa del gradiente*¹⁹, in quanto \bar{e}^2 può essere interpretato come una funzione $\bar{e}^2(\mathbf{c})$ ⁽²⁰⁾ dei coefficienti \mathbf{c} , ed il suo *gradiente* \mathbf{g}_i alla iterazione i -esima è definito come un *vettore* i cui elementi sono le derivate parziali di \bar{e}^2 rispetto agli elementi di \mathbf{c} , calcolate per un determinato valore \mathbf{c}_i del vettore dei coefficienti del filtro²¹; pertanto il suo orientamento indica la direzione verso la quale $\bar{e}^2(\mathbf{c})\big|_{\mathbf{c}=\mathbf{c}_i}$ cresce di più.

Scegliendo casualmente un vettore di coefficienti iniziale \mathbf{c}_0 (eventualmente nullo), ad ogni iterazione sono calcolati nuovi valori dei coefficienti

$$\mathbf{c}_{i+1} = \mathbf{c}_i - \Delta \mathbf{g}_i$$

spostando cioè ogni coefficiente c_n in direzione *opposta* alla relativa componente di \mathbf{g} di una quantità proporzionale alla derivata parziale corrispondente, ed al valore del parametro Δ (un piccolo numero positivo) che determina la velocità di convergenza dell'algoritmo²². Quindi, viene calcolato un nuovo valore $\mathbf{g}_{i+1}(\mathbf{c}_{i+1})$, ed il procedimento iterato fino a quando si verifica $|\mathbf{c}_{i+1} - \mathbf{c}_i| < \varepsilon$, nel cui caso si ritiene raggiunto l'ottimo \mathbf{c}_{opt} .

Equalizzazione adattiva LMS Il metodo esposto si basa sul calcolo ripetuto di $\mathbf{g}_i = 2(\mathbf{B}\mathbf{c}_i - \mathbf{d})$ (nota 21), ma... se la $H(f)$ del canale varia nel tempo, lo stesso avviene per \mathbf{B} e per \mathbf{d} , così come per il valore dei coefficienti \mathbf{c}_{opt} ! In tal caso il metodo del gradiente si modifica, utilizzando ad ogni iterazione una *stima* dello stesso ed eguagliando la frequenza delle iterazioni con quella di simbolo: osserviamo infatti che la (14.17) permette di esprimere (a parte un fattore due) il gradiente anche come

$$\mathbf{g}_m = E\{e_m \mathbf{y}_m\} \quad (14.20)$$

in cui $e_m = z(mT_s) - a_m$ è l'errore commesso dall'equalizzatore rispetto al simbolo trasmesso (e , nella fase di apprendimento, noto in ricezione²³) ed \mathbf{y}_m è il vettore dei

¹⁸http://en.wikipedia.org/wiki/Toeplitz_matrix

¹⁹http://it.wikipedia.org/wiki/Discesa_del_gradiente

²⁰ Infatti, sviluppando la (14.16) si ottiene $\bar{e}^2(\mathbf{c}) = \sum_{n=-N}^N \sum_{k=-N}^N c_n c_k \mathcal{R}_Y(n-k) - 2 \sum_{k=-N}^N c_k \mathcal{R}_{YA}(k) + E\{a_m^2\}$, che è una forma quadratica (o paraboloidale) e pertanto possiede un solo minimo globale.

²¹ Dalla (14.17) osserviamo che $\mathbf{g}(\mathbf{c}_k)$ è valutabile come $\mathbf{g}_k = 2(\mathbf{B}\mathbf{c}_k - \mathbf{d})$.

²² Una regola pratica suggerisce di porre $\Delta = \frac{1}{5(2N+1)\mathcal{P}_Y}$ in cui \mathcal{P}_Y è la potenza ricevuta di segnale più rumore. Valori maggiori possono accelerare la convergenza, ma anche dar luogo ad instabilità della soluzione, mentre valori più piccoli rallentano la convergenza, ma possono produrre errori finali minori.

²³ La sequenza dei simboli trasmessi è nota all'inizio della trasmissione, e spesso generata nella forma di una sequenza *pseudo-noise*. Dopo questa fase iniziale di *apprendimento*, l'equalizzatore commuta su di una modalità *dipendente dalle decisioni*, i cui sono invece usati i valori \tilde{a}_k emessi dal decisore: se la probabilità di errore per simbolo non è troppo elevata, la maggior parte delle volte la decisione è esatta, ed il metodo continua a funzionare correttamente.

$2N + 1$ campioni memorizzati nello stato del filtro al momento di produrre l'uscita $z(mT_s)$. La (14.20) viene quindi approssimata tralasciando di effettuare la media di insieme, e valutando la stima del gradiente come $\hat{\mathbf{g}}_m = e_m \mathbf{y}_m$, da usare poi nella formula di aggiornamento dei coefficienti c_n ²⁴:

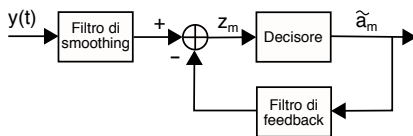
$$\mathbf{c}_{m+1} = \mathbf{c}_m - \Delta e_m \mathbf{y}_m \quad (14.21)$$

Il metodo ora esposto prende il nome di algoritmo del *gradiente stocastico* o LMS²⁵ (LEAST MEAN SQUARE).

14.6.1.2 Equalizzatore decision feedback

Nel caso in cui l'ISI introdotta dal canale sia rilevante, i metodi descritti non sono più sufficienti, e l'equalizzatore deve far uso anche dell'informazione relativa agli errori commessi per i simboli già ricevuti in precedenza, prendendo il nome di *decision feedback equalizer* o DFE.

Ciò è realizzato mediante lo schema mostrato a lato, in cui il filtro di *smoothing* opera come discusso finora, mentre quello di *feedback* è anch'esso di tipo FIR, ma opera alla frequenza di simbolo f_s e viene alimentato dalle decisioni precedenti \tilde{a}_m ²⁶, la cui combinazione lineare è usata per *correggere* i valori emessi dal primo filtro: il valore in ingresso al decisore può quindi essere ora espresso come



$$z_m = \sum_{n=0}^{N_1} c_n y(mT_s - n\tau) - \sum_{n=1}^{N_2} d_n \tilde{a}_{m-n}$$

I coefficienti d_n del filtro di feedback possono essere anch'essi stimati in maniera adattiva mediante il metodo del gradiente stocastico, dando così luogo alla architettura computazionale di figura 14.1, in cui i cerchi con i simboli di sommatoria rappresentano l'aggiornamento dei coefficienti in accordo alla (14.21); e che permette di verificare il funzionamento *real-time* dell'algoritmo.

14.6.1.3 Equalizzatore di massima verosimiglianza

Le tecniche illustrate finora effettuano la decisione simbolo per simbolo, mentre il metodo che analizziamo ora produce una stima di quanto ricevuto solo dopo aver preso in esame una intera sequenza di N simboli, operando una *maximum likelihood sequence detection* o MLSD. L'esposizione inizia con il riconsiderare i campioni $y_m = y(mT_s)$ del segnale ricevuto (14.14) come

$$y_m = \sum_{n=0}^N a_n h(mT_s - nT_s) + \nu_m = h_0 a_m + \sum_{n=0, n \neq m}^N a_n h_{m-n} + \nu_m \quad (14.22)$$

²⁴Da un punto di vista intuitivo, la (14.21) si commenta notando che il generico coefficiente c_n viene *aumentato* nel caso in cui e_m e y_n sono di segno opposto, ovvero se l'uscita è troppo piccola ($z(mT_s) < a_m$) ed il contributo dell' n -esima memoria (a cui corrisponde un c_n positivo) va fatto pesare di più, oppure se l'uscita è troppo grande ($z(mT_s) > a_m$) ed il contributo dell' n -esima memoria (a cui corrisponde un c_n negativo) deve essere (in modulo) ridotto. Considerazioni opposte, nel caso e_m e y_n abbiano lo stesso segno.

²⁵http://en.wikipedia.org/wiki/Least_mean_squares_filter

²⁶Nella fase di apprendimento sono usati i valori a_m noti a priori, mentre in quella successiva si usano quelli \tilde{a}_m emessi dal decisore, supposti per la maggior parte esatti.

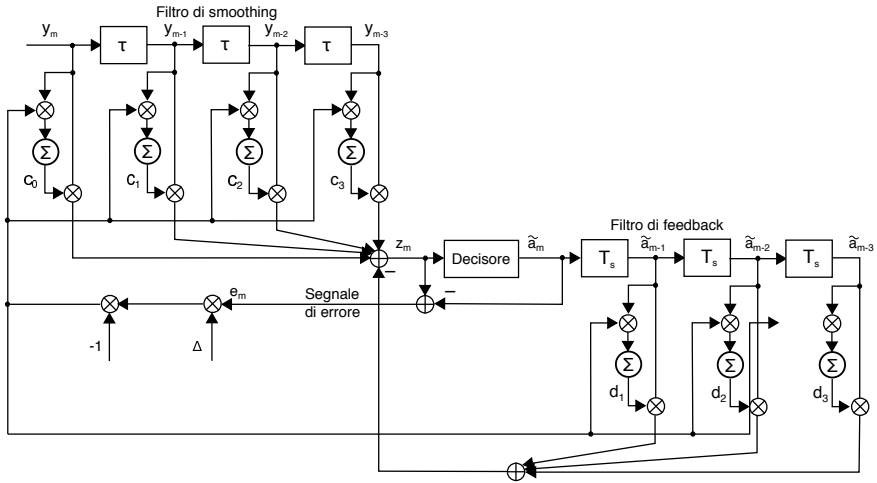


Figura 14.1: Architettura di un equalizzatore DFE basato sul gradiente stocastico

in cui $h(t) = h_T(t) * h_{canale}(t) * h_R(t)$, evidenziando le componenti di errore dovuta all'ISI (la $\sum_{n \neq m} a_n h_{m-n}$) ed al rumore ν_m ; impostando il problema come quello di stimare la sequenza $\mathbf{a} = \{a_m\}$ conoscendo solo quella di osservazione $\mathbf{y} = \{y_m\}$, la soluzione di massima verosimiglianza individua la sequenza $\hat{\mathbf{a}}$ in modo che $p(\mathbf{y}/\mathbf{a})|_{\mathbf{a}=\hat{\mathbf{a}}}$ sia massima.

La (14.22) evidenzia inoltre come la componente aleatoria di \mathbf{y} sia tutta riconducibile a quella della sequenza di rumore \mathbf{v} , dato che l'effetto del canale si riduce ad una convoluzione discreta e dunque è deterministico; pertanto y_m è una v.a. con media $\bar{y}_m(\mathbf{a}) = \sum_{n=0}^N a_n h_{m-n}$ e d.d.p. gaussiana $p_\nu(y_m/\mathbf{a})$ condizionata ad \mathbf{a} , che ne determina la sequenza dei valori medi. Se le realizzazioni ν_m sono staticamente indipendenti²⁷, la $p(\mathbf{y}/\mathbf{a})$ è il prodotto dei singoli contributi $p_\nu(y_m/\mathbf{a})$, da valutare per tutti gli m e per tutti i modi di scegliere \mathbf{a} , e quindi trovare $\hat{\mathbf{a}} = \arg \max \prod_{m=0}^M p_\nu(y_m/\mathbf{a})$ o, equivalentemente

$$\hat{\mathbf{a}} = \arg \min \sum_{m=0}^M -\ln(p_\nu(y_m/\mathbf{a})) \quad (14.23)$$

Quest'ultimo passaggio trova motivazione nella espressione della d.d.p. gaussiana $p_\nu(y) = \frac{1}{\sqrt{2\pi}\sigma_\nu} e^{-\frac{(y-\bar{y})^2}{2\sigma_\nu^2}}$ per cui $\ln(p_\nu(y)) = -\ln(\sqrt{2\pi}\sigma_\nu) - \frac{(y-\bar{y})^2}{2\sigma_\nu^2}$, e dato che σ_ν è la stessa per tutti gli istanti m , il suo valore non contribuisce alla minimizzazione (14.23), che diviene pertanto

$$\hat{\mathbf{a}} = \arg \min \sum_{m=0}^M (y_m - \bar{y}_m(\mathbf{a}))^2$$

La sequenza ottima $\hat{\mathbf{a}}$ è quindi quella che minimizza la somma degli scarti quadratici tra le osservazioni y_m ed il corrispondente valore atteso, e viene trovata utilizzando l'algoritmo di Viterbi (pag. 466) riconducendo il problema a quello di individuare un

²⁷A questo proposito, è opportuno che $H_R(f)$ sia realizzato come un passa basso e non come un filtro ottimo, oppure che si effettui una operazione di sbiancamento di $y(t)$.

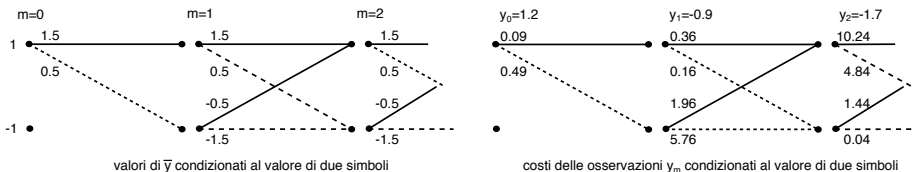


Figura 14.2: Traliccio per la decodifica di Viterbi e corrispondenti costi di verosimiglianza

percorso di minimo costo nell'ambito del traliccio che rappresenta tutte le possibili sequenze **a**.

Esempio Ipotizziamo una $h_n = \begin{cases} 1 & \text{con } n = 0 \\ 0.5 & \text{con } n = 1 \end{cases}$ ed una trasmissione antipodale binaria con $a = \{1, -1\}$; la (14.22) si scrive quindi come

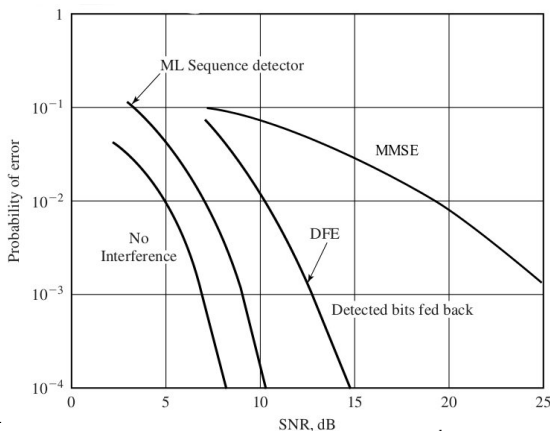
$$y_m = a_m h_0 + a_{m-1} h_1 + \nu_m$$

che rappresenta un filtro FIR di primo ordine, la cui uscita in assenza di rumore assume uno dei 4 possibili valori mostrati in tabella in funzione di $a_m a_{m-1}$. Il traliccio di ricerca può quindi essere tracciato come nella parte sinistra di fig. 14.2, identificando i valori di a_{m-1} con le righe +1 e -1, e quelli di a_m con il tratteggio (-1) o la linea continua (1) degli archi che uniscono le colonne, considerando $a_{-1} = a_{-2} = 1$; sopra ogni arco, è quindi scritto il valore di \bar{y}_m corrispondente. Nella parte destra di fig. 14.2 gli archi sono invece etichettati con i costi $(y_m - \bar{y}_m(\mathbf{a}))^2$ associati all'osservazione della sequenza $\mathbf{y} = 1.2, -0.9, -1.7$.

Il metodo con cui l'algoritmo di Viterbi esplora il traliccio per individuare il percorso di minimo costo, e l'associata sequenza $\hat{\mathbf{a}}$, è illustrato a pag. 466 e non viene qui ripetuto.

14.6.1.4 Prestazioni

La figura a lato²⁸ confronta P_e vs. SNR per una modulazione binaria antipodale nei casi di assenza di ISI oppure di un canale con $h_0 h_1 h_2 = 0.4, 0.8, 0.4$, ed eseguendo l'equalizzazione con le tecniche illustrate. Come si vede, il MLSD si comporta meglio di tutti, e determina il valore delle prestazioni ottime. Al contrario, le prestazioni di MMSE appaiono come le peggiori, essenzialmente a causa della amplificazione del rumore nelle



²⁸Tratta da J.G. Proakis, M. Salehi, *Communication systems engineering*, 2nd Ed., 2002 Prentice-Hall

bande in cui il canale ha guadagno nullo; d'altro canto, il DFE

ne migliora sensibilmente le prestazioni, pur non raggiungendo quelle di MLSD. Ciò che rende poco diffuso l'MLSD però è l'elevata complessità computazionale, oltre che il ritardo da attendere per una decisione su intere sequenze, nonché le difficoltà a rendere adattativo l'algoritmo nel caso di condizioni non stazionarie.

14.7 Distorsione di non linearità

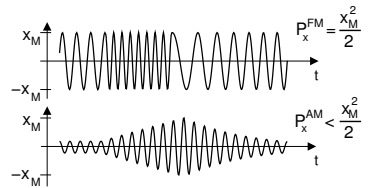
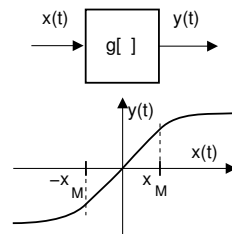
Descrive il deterioramento subito dal segnale nel transito attraverso un dispositivo dal comportamento non-lineare, per il quale cioè la relazione ingresso-uscita è del generico tipo $y(t) = g[x(t)]$; espandendo tale relazione in serie di potenze, ed arrestando lo sviluppo al terzo ordine, otteniamo:

$$y(t) = G [x(t) + \alpha x^2(t) + \beta x^3(t)] \tag{14.24}$$

Un caso *tipico* di questo fenomeno si osserva allo stadio finale di un amplificatore di potenza²⁹ che, per ampiezze del segnale di ingresso maggiori di $\pm x_M$, presenta fenomeni di *saturazione* dei valori di uscita (vedi figura a lato). Notiamo subito come il coefficiente β , possa rappresentare questo comportamento.

Nel caso in cui il dispositivo che introduce saturazione sia attraversato da segnali con modulazione AM, il fenomeno è particolarmente grave, dato che questi presentano valori di ampiezza direttamente dipendenti da quelli del segnale modulante. Per evitare di operare in regione non lineare, la potenza del segnale in ingresso all'elemento non lineare deve quindi essere ridotta (questa operazione è chiamata *back-off*), e conseguentemente la trasmissione avviene ad un livello di potenza \mathcal{P}_x^{AM} inferiore a quello consentito ($\frac{x_M^2}{2}$) dall'amplificatore.

Nel caso invece di trasmissione FM, abbiamo visto che il segnale modulato mantiene sempre la stessa ampiezza, e dunque si può effettuare la trasmissione a piena potenza; in altre parole, una volta fissato il livello di trasmissione, non occorre ricorrere ad un amplificatore sovradimensionato. Esaminiamo ora come quantificare l'effetto delle distorsioni non lineari.



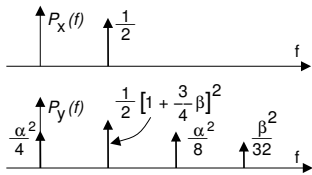
14.7.1 Ingresso sinusoidale

Ponendo $x(t) = A \cos \omega_0 t$ la (14.24) si riscrive come³⁰

$$\begin{aligned} y(t) &= G [A \cos \omega_0 t + \alpha A^2 \cos^2 \omega_0 t + \beta A^3 \cos^3 \omega_0 t] = \\ &= GA \left[\frac{\alpha A}{2} + \left(1 + \frac{3}{4} \beta A^2 \right) \cos \omega_0 t + \frac{\alpha A}{2} \cos 2\omega_0 t + \frac{\beta A^2}{4} \cos 3\omega_0 t \right] \end{aligned}$$

²⁹ come ad esempio è il caso dei TWTA introdotti a pag. 414

³⁰ Si fa uso delle relazioni $\cos^2 \alpha = \frac{1}{2} + \frac{1}{2} \cos 2\alpha$ e $\cos^3 \alpha = \frac{3}{4} \cos \alpha + \frac{1}{4} \cos 3\alpha$.



che corrisponde allo spettro di densità di potenza disegnato a lato (che è uno spettro *unilatero*, e calcolato per $A = G = 1$), dove si osserva la comparsa di termini a frequenza multipla di quella di ingresso, oltre che di una componente continua.

Nella pratica, i valori di α e β non sono noti, e le relazioni ottenute sono usate per derivarli, ponendo in ingresso una sinusoidale di potenza nota, ed osservando la potenza delle sue armoniche presenti in uscita.

Fattori di intermodulazione Le caratteristiche tecniche che accompagnano gli amplificatori riportano, invece di α e β , i valori dei *fattori di intermodulazione* μ_2 e μ_3 (detti di seconda e di terza armonica), ottenuti utilizzando appunto un ingresso sinusoidale e misurando le potenze \mathcal{P}_I , \mathcal{P}_{II} e \mathcal{P}_{III} alla frequenza in ingresso ed alla sua seconda e terza armonica, e derivando³¹ da queste le quantità

$$\mu_2 \doteq \frac{\mathcal{P}_{II}}{\mathcal{P}_I^2} \quad \text{e} \quad \mu_3 \doteq \frac{\mathcal{P}_{III}}{\mathcal{P}_I^3}$$

da cui si ottengono i coefficienti α e β mediante le relazioni $\alpha \simeq 0.7 \cdot \mu_2 \cdot G$, $\beta = 2 \cdot \mu_3 \cdot G^2$ come mostrato alla nota³².

Scrivendo $\mathcal{P}_{II} = \mu_2^2 \mathcal{P}_I^2$ e $\mathcal{P}_{III} = \mu_3^2 \mathcal{P}_I^3$, osserviamo che per piccoli valori di \mathcal{P}_I , la distorsione prodotta sia da \mathcal{P}_{II} che da \mathcal{P}_{III} è trascurabile; all'aumentare di \mathcal{P}_I , \mathcal{P}_{II} cresce con il quadrato, mentre \mathcal{P}_{III} con il cubo, e pertanto è quest'ultima componente che poi predomina.

14.7.2 Ingresso aleatorio

Nel caso in cui l'ingresso dell'elemento non lineare sia un processo gaussiano, la densità spettrale in uscita può ottenersi come \mathcal{F} -trasformata della funzione di autocorrelazione dell'uscita: in virtù di alcune proprietà³³ dei momenti di variabili aleatorie gaussiane,

³¹Le relazioni mostrate si ottengono scrivendo

$$\begin{aligned} \mathcal{P}_I &= \frac{G^2 A^2}{2} \left(1 + \frac{3}{4} \beta A^2 \right)^2 \simeq \frac{G^2 A^2}{2} \quad \left(\text{se } \beta \ll \frac{4}{3A^2} \right) \\ \mathcal{P}_{II} &= \frac{G^2 A^4 \alpha^2}{8} = \frac{G^4 A^4}{4} \frac{1}{G^2} \frac{\alpha^2}{2} = \mathcal{P}_I^2 \mu_2^2 \\ \mathcal{P}_{III} &= \frac{G^2 A^6 \beta^2}{32} = \frac{G^6 A^6}{8} \frac{1}{G^4} \frac{\beta^2}{4} = \mathcal{P}_I^3 \mu_3^2 \end{aligned}$$

³²Scrivendo $\mu_2 \doteq \frac{\mathcal{P}_{II}}{\mathcal{P}_I^2} = \frac{G^2 A^4 \alpha^2}{8} \cdot \frac{4}{2G^2}$ si ottiene $\mu_2 = \frac{\alpha}{\sqrt{2}G}$ e quindi $\alpha \simeq 0.7 \cdot \mu_2 \cdot G$; allo stesso modo da $\mu_3 \doteq \frac{\mathcal{P}_{III}}{\mathcal{P}_I^3} = \frac{G^2 A^6 \beta^2}{32} \cdot \frac{8}{G^6 A^6} = \frac{\beta^2}{4G^4}$ si ha $\mu_3 = \frac{\beta}{2G^2}$ e dunque $\beta = 2 \cdot \mu_3 \cdot G^2$.

³³Dato che l'uscita ha espressione $y(t) = G [x(t) + \alpha x^2(t) + \beta x^3(t)]$, il calcolo di $\mathcal{R}_y(\tau) = E \{ y(t) y(t + \tau) \}$ si sviluppa calcolando i momenti misti $m_x^{(i,j)}(\tau) = E \{ x^i(t) x^j(t + \tau) \}$. Se $x(t)$ è un processo gaussiano a media nulla, accade che $m_x^{(i,j)}(\tau) = 0$ se $i + j$ è dispari, mentre in caso contrario si applica il risultato per il valore atteso del prodotto di più v.a. estratte in tempi diversi:

$$E \{ x_1 \cdot x_2 \cdot \dots \cdot x_n \} = \sum \left(E \{ x_{p_1} \cdot x_{p_2} \} \cdot E \{ x_{p_3} \cdot x_{p_4} \} \cdot \dots \cdot E \{ x_{p_{n-1}} \cdot x_{p_n} \} \right)$$

in cui la somma è estesa a tutte le possibili permutazioni non equivalenti di $(1, 2, \dots, n)$ (sono equivalenti se accoppiano con ordine diverso o in posizione diversa le stesse v.a.). Ad esempio, per quattro

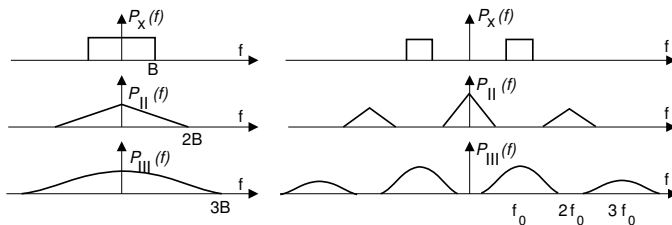


Figura 14.3: Densità spettrale affette da distorsioni non lineari; a sinistra per banda base, a destra per segnale modulato

si ottiene in questo caso:

$$\mathcal{P}_{II}(f) = G^2 2\alpha^2 \mathcal{P}_x(f) * \mathcal{P}_x(f); \quad \mathcal{P}_{III}(f) = G^2 6\beta^2 \mathcal{P}_x(f) * \mathcal{P}_x(f) * \mathcal{P}_x(f)$$

ovvero compaiono termini di distorsione di “2^a e 3^a armonica” che hanno origine dalla convoluzione della densità di potenza del segnale utile con se stesso.

Nel caso in cui il processo $x(t)$ sia limitato in banda contigua all’origine, i termini $\mathcal{P}_{II}(f)$ e $\mathcal{P}_{III}(f)$ hanno una banda rispettivamente doppia e tripla di quella di $\mathcal{P}_x(f)$ (vedi fig. 14.3). Nel caso di segnali modulati, oltre ad un allargamento di banda, avviene un fatto diverso e degno di commento: $\mathcal{P}_{II}(f)$ giace in bande diverse da quelle di $\mathcal{P}_x(f)$, pertanto può essere non considerato fonte di disturbo. $\mathcal{P}_{III}(f)$ invece ha una componente anch’essa centrata su f_0 , e dunque è solo questa la fonte disturbo. La potenza totale delle due componenti di disturbo nella banda di $\mathcal{P}_x(f)$ risulta inoltre pari a

$$\mathcal{P}_{II} = 4\mu_2^2 \mathcal{P}_I^2 \quad \text{e} \quad \mathcal{P}_{III} = 32\mu_3^2 \mathcal{P}_I^3$$

In definitiva, vi sono almeno tre buone ragioni per tenere d’occhio il valore di β , che è causa delle distorsioni di terza armonica:

- è il coefficiente che tiene conto dei fenomeni di saturazione;
- produce interferenza “in banda” per i segnali modulati;
- produce interferenza “fuori banda” che danneggia le trasmissioni a frequenza tripla.

14.7.3 Effetto sulla modulazione FM

Se consideriamo un segnale $x(t) = \cos[\omega_0 t + \varphi(t)]$, l’effetto della non linearità produce il segnale

$$y(t) = G \left[\frac{\alpha}{2} + \left(1 + \frac{3}{4}\beta \right) \cos[\omega_0 t + \varphi(t)] + \frac{\alpha}{2} \cos[2\omega_0 t + 2\varphi(t)] + \frac{\beta}{4} \cos[3\omega_0 t + 3\varphi(t)] \right]$$

v.a. si ha:

$$E\{x_1 \cdot x_2 \cdot \dots \cdot x_n\} = E\{x_1 \cdot x_2\} \cdot E\{x_3 \cdot x_4\} + E\{x_1 \cdot x_3\} \cdot E\{x_2 \cdot x_4\} + E\{x_1 \cdot x_4\} \cdot E\{x_2 \cdot x_3\}$$

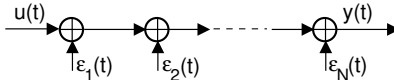
Osserviamo che i termini a frequenza $2\omega_0$ e $3\omega_0$, nonché il livello in continua, possono essere eliminati mediante un filtro passa-banda centrato in $f = f_0 = \frac{\omega}{2\pi}$; dopo tale operazione, la modulazione di fase $\varphi(t)$ è proprio quella impressa dal modulatore, e pertanto i fenomeni non lineari *non hanno conseguenze sulla FM* (tranne che per le interferenze causate ai canali vicini !)

Il risultato appena illustrato è stato sfruttato nei ponti radio progettati per trasmettere un segnale FDM in FM. Si usa un basso indice di modulazione (risparmiando banda) e si trasmette a piena potenza (senza backoff). La potenza del segnale modulato *non dipende* dal numero di canali contemporaneamente attivi.

14.8 Appendici

14.8.1 Valutazione dell' SNR dovuto a diverse fonti di disturbo

Consideriamo la situazione alla figura seguente, in cui il segnale utile $u(t)$ è affetto da diverse cause di disturbo $\varepsilon_i(t)$ indipendenti tra loro, per ognuna delle quali è noto il relativo $SNR_i = \mathcal{P}_u/\mathcal{P}_{\varepsilon_i}$. In tal caso l'effetto di tutte le cause contemporaneamente attive determina un valore di SNR complessivo definito come $SNR_T = \mathcal{P}_u/\sum_{i=1}^N \mathcal{P}_{\varepsilon_i}$; dato che $\mathcal{P}_{\varepsilon_i} = \mathcal{P}_u/SNR_i$ si ottiene



$$SNR_T = \frac{\mathcal{P}_u}{\mathcal{P}_u \sum_{i=1}^N \frac{1}{SNR_i}} = \frac{1}{\sum_{i=1}^N \frac{1}{SNR_i}}$$

Questo risultato ricorda quello della impedenza equivalente a più impedenze poste in parallelo, il che porta a descrivere l' SNR complessivo come *il parallelo* degli SNR . Una applicazione di questo risultato verrà esposta al § 16.4.

14.8.2 Potenza assorbita da un bipolo

La dimostrazione inizia definendo una potenza *istantanea* assorbita dal bipolo come $w(t) = v(t)i(t) = v(t) \cdot (v(t) * y(t))$. La potenza *media* (nel tempo) allora risulta

$$\begin{aligned} \mathcal{W}_z &= \lim_{\Delta t \rightarrow \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} w(t) dt = \\ &= \lim_{\Delta t \rightarrow \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} v(t) \left[\int_{-\infty}^{\infty} v(t-\tau) y(\tau) d\tau \right] dt = \\ &= \int_{-\infty}^{\infty} y(\tau) \left[\lim_{\Delta t \rightarrow \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} v(t) v(t-\tau) dt \right] d\tau \\ &= \int_{-\infty}^{\infty} y(\tau) \mathcal{R}_v(-\tau) d\tau = \int_{-\infty}^{\infty} Y(f) \mathcal{P}_v(f) df = \\ &= \int_{-\infty}^{\infty} [\Re\{Y(f)\} + j\Im\{Y(f)\}] \mathcal{P}_v(f) df = \\ &= \int_{-\infty}^{\infty} \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} df \end{aligned}$$

Nel terzultimo passaggio si è fatto uso del teorema di Parseval, e del fatto che $\mathcal{R}_v(\tau)$ è pari; nell'ultimo, si è tenuto conto che $\mathcal{P}_v(f)$, $R(f)$ e $|Z(f)|^2 = R^2(f) + X^2(f)$ sono

funzioni pari di f , mentre $X(f)$ è dispari: pertanto il termine $\int_{-\infty}^{\infty} \Im \{Y(f)\} \mathcal{P}_v(f) df = \int_{-\infty}^{\infty} \mathcal{P}_v(f) \frac{X(f)}{|Z(f)|^2} df$ è nullo. Notiamo che quest'ultimo termine rappresenta la *potenza reattiva*, che non è trasformata in altre forme di energia, e viene accumulata e restituita dalla componente reattiva del carico. Al contrario, il termine relativo a $\Re \{Y(f)\}$ rappresenta la potenza assorbita dalla componente resistiva, nota come *potenza attiva*, che viene completamente dissipata.

Avendo espresso la potenza assorbita \mathcal{W}_z nella forma di un integrale in f , la funzione integranda è intuitivamente associabile allo spettro di densità di potenza: $\mathcal{W}_z(f) = \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2}$. Lo stesso risultato può essere confermato svolgendo il seguente calcolo più diretto, pensando al bipolo come ad un filtro la cui grandezza di ingresso è $v(t)$ e quella di uscita $i(t)$.

La definizione di potenza media $\mathcal{W}_z = \lim_{\Delta t \rightarrow \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} w(t) dt$, in cui $w(t) = v(t)i(t)$, mostra come questa sia equivalente alla funzione di intercorrelazione tra i e v calcolata in $\tau = 0$: $\mathcal{W}_z = \mathcal{R}_{vi}(0)$. Allora, è ragionevole assumere che $\mathcal{W}_z(f) = \mathcal{F}\{\mathcal{R}_{vi}(\tau)\}$. Indicando infatti con \otimes l'integrale di intercorrelazione, e ricordando che gli operatori di convoluzione e correlazione godono della proprietà commutativa, possiamo scrivere

$$\mathcal{R}_{vi}(\tau) = v(t) \otimes i(t) = v(t) \otimes (v(t) * y(t)) = (v(t) \otimes v(t)) * y(t) = \mathcal{R}_v(\tau) * y(t)$$

quindi, risulta che

$$\mathcal{W}_z(f) = \mathcal{F}\{\mathcal{R}_{vi}(\tau)\} = \mathcal{P}_v(f) Y(f) = \mathcal{P}_v(f) \frac{R(f) - jX(f)}{|Z(f)|^2}$$

In base alle stesse considerazioni già svolte, si verifica come il termine immaginario non contribuisce alla potenza media assorbita, e quindi può essere omissa dalla definizione di *potenza attiva* $\mathcal{W}_z(f)$.

14.8.3 Condizioni per il massimo trasferimento di potenza

Svolgiamo per intero la dimostrazione delle (14.3). Verifichiamo subito che, mantenendo $Z_g(f)$ fisso e per qualunque valore di $R_c(f)$, la potenza ceduta ad un carico espressa dalla (14.2): $\mathcal{W}_{z_c}(f) = \mathcal{P}_{v_g}(f) \frac{R_c(f)}{|Z_c(f) + Z_g(f)|^2}$ risulta massima se il suo denominatore è il più piccolo possibile, e ciò si verifica quando $X_c(f) = -X_g(f)$, ed in tal caso risulta

$$\mathcal{W}_{z_c}(f) = \mathcal{P}_{v_g}(f) \frac{R_c(f)}{(R_c(f) + R_g(f))^2} = \frac{\mathcal{P}_{v_g}(f)}{R_c(f) + 2R_g(f) + (R_g(f))^2/R_c(f)} \quad (14.25)$$

Per individuare ora la condizione su $R_c(f)$ che rende minimo il denominatore (e dunque $\mathcal{W}_{z_c}(f)$ massima), eseguiamone la derivata rispetto ad R_c (omettendo per brevità la dipendenza da f) ed eguagliamola a zero:

$$\frac{d}{dR_c} \left(R_c + 2R_g + \frac{R_g^2}{R_c} \right) = 1 - \left(\frac{R_g}{R_c} \right)^2 = 0 \quad (14.26)$$

che fornisce la condizione $R_c(f) = \pm R_g(f)$ in cui il valore negativo viene scartato mentre quello positivo, assieme alla condizione $X_c(f) = -X_g(f)$ determina la condizione $Z_c(f) = Z_g^*(f)$ espressa alla (14.3). Volendo verificare che la (14.26) individui

effettivamente un minimo e non un massimo del denominatore di (14.25), se ne può eseguire la derivata seconda, ottenendo

$$\frac{d^2}{dR_c^2} \left(R_c + 2R_g + \frac{R_g^2}{R_c} \right) = \frac{d}{dR_c} \left[1 - \left(\frac{R_g}{R_c} \right)^2 \right] = 2 \frac{R_g^2}{R_c^3}$$

che verifichiamo immediatamente essere sempre positiva.

14.8.4 Potenza ceduta ad un carico $Z_c(f) \neq Z_g^*(f)$

Avendo a disposizione un generatore di segnale di potenza disponibile $\mathcal{W}_d(f)$ ed impedenza interna $Z_g(f)$ assegnate, la tensione a vuoto ai suoi capi ha densità di potenza (di segnale) pari a $\mathcal{P}_v(f) = \mathcal{W}_d(f) 4R_g(f)$. Collegando al generatore un carico generico $Z_c(f)$, la potenza dissipata da quest'ultimo risulta pari a $\mathcal{W}_{z_c}(f) = \mathcal{P}_v(f) \frac{R_c(f)}{|Z_g(f) + Z_c(f)|^2}$. Il rapporto tra la densità di potenza effettivamente ceduta a $Z_c(f)$, e quella che le sarebbe ceduta se questa fosse adattata per il massimo trasferimento di potenza, fornisce la perdita di potenza subita:

$$\frac{\mathcal{W}_{z_c}(f)}{\mathcal{W}_d(f)} = \mathcal{W}_d(f) 4R_g \frac{R_c(f)}{|Z_g(f) + Z_c(f)|^2} \cdot \frac{1}{\mathcal{W}_d(f)} = \frac{4R_g(f) R_c(f)}{|Z_g(f) + Z_c(f)|^2} = \alpha(f)$$

Pertanto, se $Z_c(f) \neq Z_g^*(f)$, su $Z_c(f)$ si dissipa una potenza pari a $\mathcal{W}_{z_c}(f) = \alpha(f) \mathcal{W}_d(f)$. Il medesimo risultato è valido anche per l'analisi dell'accoppiamento tra il generatore equivalente di uscita di una rete due porte ed un carico.

Esempio Consideriamo un generatore con $Z_g(f)$ resistiva e pari a 50Ω , e con densità di potenza disponibile (a frequenze positive)

$$\mathcal{W}_d^+(f) = \frac{\mathcal{W}_d}{4W} \text{rect}_{2W}(f - f_0)$$

in cui $\mathcal{W}_d = 1$ Watt è la potenza disponibile totale, distribuita uniformemente in una banda $2W = 10$ KHz centrata a frequenza $f_0 = 1$ MHz. Il generatore è collegato ad un carico

$$Z_c(f) = R_c(f) + jX_c(f)$$

con $R_c(f) = 50 \Omega$ ed $X_c(f) = 2\pi fL = 50 \Omega$ per $f = f_0$ (da cui $L = \frac{50}{2\pi 10^6} = 7.96 \mu H$).

Essendo la banda di segnale $2W \ll f_0$, approssimiamo la dipendenza da f di $X_c(f)$ come una costante. In queste ipotesi, la potenza effettivamente ceduta al carico risulta $\mathcal{W}_{z_c} = \alpha \mathcal{W}_d$, con

$$\alpha = \frac{4R_g R_c}{|Z_g + Z_c|^2} = \frac{4 \cdot 50 \cdot 50}{|50 + 50 + j50|^2} = \frac{10000}{12500} = 0.8$$

e quindi $\mathcal{W}_{z_c} = 0.8$ Watt ovvero, in dBm: $10 \log_{10} 0.8 = -0.97$ dBW = 29.03 dBm.

Il valore $\alpha_{dB} = 10 \log_{10} \alpha = 0.97$ dB rappresenta il valore della perdita di potenza causata dal mancato verificarsi delle condizioni di massimo trasferimento di potenza, e può essere tenuto in conto come una attenuazione supplementare al collegamento, in fase di valutazione del *link budget* (vedi capitolo seguente).

14.8.5 Derivazione del tempo di ritardo di gruppo

Dimostriamo qui il risultato (14.13). Consideriamo di operare nello spazio dell'involuppo complesso, e di disporre di un canale affetto dalla sola distorsione di fase, ovvero con risposta in frequenza $H(f) = 1 \cdot e^{j\varphi(f)}$; la \mathcal{F} -trasformata di $\underline{y}(t)$ si scrive allora come $\underline{Y}(f) = \frac{1}{2} \underline{X}(f) \underline{H}(f)$, dove per $x(t) = a(t) \cos(2\pi f_0 t)$ si ha $\underline{X}(f) = X_c(f) = A(f)$, mentre $\underline{H}(f) = 2H^+(f + f_0) = 2e^{j\varphi(f+f_0)} = 2e^{j\varphi_{pb}(f)}$ dove il pedice $_{pb}$ simboleggia la caratteristica *passa-basso* dell'involuppo complesso. Sviluppando $\varphi_{pb}(f)$ in serie di Maclaurin e troncandola al primo termine si ottiene³⁴ per f prossimo a zero

$$\varphi_{pb}(f) \simeq -2\pi(f_0 t_p(f_0) + f t_g(f_0))$$

e quindi

$$\begin{aligned} \underline{Y}(f) &= A(f) e^{j\varphi_{pb}(f)} = A(f) e^{-j2\pi(f_0 t_p(f_0) + f t_g(f_0))} \\ &= e^{-j2\pi f_0 t_p(f_0)} \cdot A(f) e^{-j2\pi f t_g(f_0)} \end{aligned}$$

da cui, ricordando la proprietà di traslazione temporale, si ottiene l'antitrasformata

$$\underline{y}(t) = e^{-j2\pi f_0 t_p(f_0)} \cdot a(t - t_g(f_0))$$

a cui corrisponde³⁵ il segnale modulato espresso dalla (14.13).

³⁴Infatti sussistono i seguenti passaggi

$$\begin{aligned} \varphi_{pb}(f) &\simeq \varphi_{pb}(0) + f \cdot \left. \frac{d\varphi_{pb}(f)}{df} \right|_{f=0} = \varphi(f_0) + f \cdot \left. \frac{d\varphi(f)}{df} \right|_{f=f_0} \\ &= 2\pi \left(f_0 \frac{\varphi(f_0)}{2\pi f_0} + f \frac{1}{2\pi} \left. \frac{d\varphi(f)}{df} \right|_{f=f_0} \right) = -2\pi(f_0 t_p(f_0) + f t_g(f_0)) \end{aligned}$$

³⁵Infatti al termine $\underline{z}(t) = e^{-j2\pi f_0 t_p}$ corrispondono le componenti a frequenza positiva e negativa

$$z^\pm(t) = \frac{1}{2} z(t) e^{\pm j2\pi f_0 t} = \frac{1}{2} e^{-j2\pi f_0 t_p} e^{\pm j2\pi f_0 t} = \frac{1}{2} e^{\pm j2\pi f_0 (t - t_p)}$$

da cui otteniamo in modo semplice il corrispondente segnale

$$z(t) = z^+(t) + z^-(t) = \frac{1}{2} \left(e^{j2\pi f_0 (t - t_p)} + e^{-j2\pi f_0 (t - t_p)} \right) = \cos(2\pi f_0 (t - t_p))$$

Capitolo 15

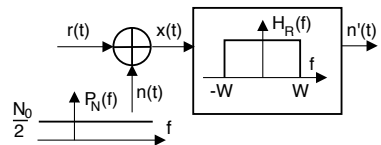
Collegamenti e mezzi trasmissivi

Sono qui analizzati i mezzi trasmissivi cavo, radio e fibra ottica, e derivate le relazioni tra i parametri fisici che li caratterizzano, i fenomeni che possono manifestarsi, e la descrizione dei mezzi nei termini della rappresentazione tempo-frequenza di un canale di comunicazione. Le particolarità dei diversi mezzi intervengono quindi nella analisi del *bilancio di collegamento*, che mira a quantificare la distanza che è possibile raggiungere (o *portata*), e la potenza che è necessario trasmettere.

15.1 Bilancio di collegamento

Introduciamo subito alcune definizioni e concetti, che aiutano a meglio analizzare questo problema. I parametri fondamentali *del collegamento* sono espressi dalla potenza disponibile del trasmettitore W_{dT} , dalla minima potenza che occorre ricevere W_{RMin} (spesso indicata come *sensibilità* del ricevitore) e dall'attenuazione disponibile A_d del mezzo di trasmissione che si intende utilizzare.

D'altra parte, il parametro che più caratterizza il segnale trasmesso ai fini dimensionamento è la *banda occupata* W , in quanto determina la potenza di rumore in ingresso al ricevitore, come illustrato ai § 5.3.1 e 12.1.



Determinazione della sensibilità Si ottiene in base alla conoscenza del livello di rumore $\frac{N_0}{2}$ in ingresso al ricevitore (vedi cap. 16) e dell' SNR (cap. 12) o della P_e (vedi eq. (5.17) a pag. 92 e cap. 13) che si desidera ottenere. Nel caso analogico, se si richiede un valore $SNR = \alpha SNR_0 = \alpha \frac{W_R}{N_0 W}$, si ottiene¹

$$W_{RMin} = N_0 W \cdot \frac{SNR}{\alpha}$$

mentre in quello numerico, la necessità di conseguire un valore di P_e^{bit} consente di determinare il valore di $\frac{E_b}{N_0} = \frac{W_R}{f_b N_0}$, e quindi

$$W_{RMin} = N_0 f_b \cdot \frac{E_b}{N_0}$$

¹ Come definito al § 12.2.1.1, SNR_0 dipende solo dalle caratteristiche del collegamento, mentre il coefficiente α rappresenta la dipendenza dal tipo di modulazione adottata, e differisce da uno nei casi di modulazione FM, AM-PI e AM-PPS.

Benché la valutazione delle prestazioni svolta ai precedenti capitoli consideri potenze *di segnale*, lo stesso valore SNR esprime anche un rapporto tra potenze *disponibili*, dato che sia segnale che rumore hanno origine da generatori che condividono la stessa impedenza interna (vedi § 16.2). Infatti

$$\frac{W_{dR}}{W_{dN}} = \frac{P_R}{4R_g} \frac{4R_g}{P_N} = \frac{P_R}{P_N}$$

così come l' SNR non varia se, anziché le potenze *disponibili*, si considerano quelle assorbite da un carico (lo stadio di ingresso del ricevitore), dato che segnale e rumore subiscono il medesimo rapporto di partizione (vedi § 14.2.2).

Guadagno di sistema Il rapporto

$$G_s = \frac{W_{dT}}{W_{RMin}}$$

è detto *guadagno di sistema*² e rappresenta il massimo valore di attenuazione *disponibile* A_d che è possibile superare. La differenza in decibel $G_{s_{dB}} = W_{dT} [dBW] - W_{RMin} [dBW]$ rappresenta la stessa quantità, in una forma che rende più intuitivo il suo utilizzo nel determinare un limite alla massima attenuazione disponibile: deve infatti risultare

$$A_{dB} \leq G_{s_{dB}}$$

Margine di sistema La differenza tra $G_{s_{dB}}$ ed A_{dB} , che per quanto appena detto deve risultare ≥ 0 , prende il nome di *Margine di sistema*, e rappresenta l'eccesso di potenza (in dB) che viene trasmessa, rispetto alla minima indispensabile:

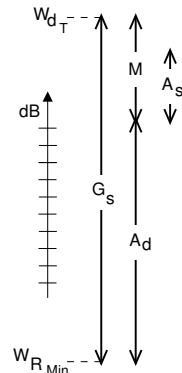
$$M_{dB} = G_{s_{dB}} - A_{dB}$$

Attenuazione supplementare L'eccesso di potenza M_{dB} deve comunque risultare maggiore della somma (in dB) di tutte le possibili ulteriori cause di attenuazione del segnale, indicate collettivamente come *attenuazioni supplementari*:

$$\sum A_{s_{dB}} \leq M_{dB}$$

In questa categoria rientrano tutte le cause di attenuazione non previste nella situazione ideale e che possono, ad esempio, avere origine dal fallimento delle condizioni per il massimo trasferimento di potenza, oppure essere causate da un fenomeno piovoso in un collegamento radio, o dipendere dalla perdita di segnale dovuta alla giunzione tra tratte in fibra ottica....

²Notiamo che G_s è definito come ingresso/uscita, contrariamente agli altri guadagni. Infatti, non è una *grandezza* del collegamento, bensì una *potenzialità* dello stesso.



Grado di servizio Nel capitolo 8 il concetto di grado di servizio è stato associato al valore di probabilità con cui può verificarsi un fenomeno di blocco in un elemento di commutazione. Un concetto del tutto analogo sussiste, qualora le attenuazioni supplementari siano grandezze aleatorie, e la loro somma possa superare il valore del margine a disposizione: in tal caso, la potenza ricevuta si riduce sotto la minima $W_{R_{Min}}$, ed il collegamento “*va fuori specifiche*”. Pertanto, in sede di dimensionamento di un collegamento, indichiamo con grado di servizio la percentuale di tempo per la quale si mantiene $W_R > W_{R_{Min}}$, ovvero la probabilità che le attenuazioni supplementari *non* superino il margine, ossia

$$\text{Grado di Servizio} = Pr \left\{ \sum A_{s_{dB}} < M_{dB} \right\}$$

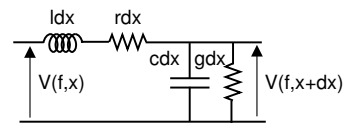
Esempio Un grado di servizio del 99.99 % equivale a poco meno di 1 ora l'anno di fuori servizio, e corrisponde a richiedere che $Pr \{ \sum A_{s_{dB}} > M_{dB} \} = 10^{-4}$.

15.2 Collegamenti in cavo

La descrizione completa delle caratteristiche e delle prestazioni dei cavi in rame è una materia molto vasta, di cui forniamo di seguito solo alcuni risultati, strettamente legati agli aspetti di telecomunicazione, il più rilevante dei quali è senz'altro il manifestarsi dell'effetto pelle, che determina (per $f > 100$ KHz) una attenuazione in dB proporzionale a \sqrt{f} . La sezione è completata da una breve catalogazione dei cavi usati per telecomunicazioni.

15.2.1 Costanti distribuite, grandezze derivate, e condizioni generali

Un conduttore elettrico uniforme e di lunghezza infinita, è descritto in base ad un modello a costanti distribuite, espresso in termini delle *costanti primarie* costituite dalla resistenza r , la conduttanza g , la capacità c e l'induttanza l per unità di lunghezza. La teoria delle linee uniformi definisce quindi due grandezze derivate dalle costanti primarie: l'*impedenza caratteristica* $Z_0(f)$ e la *costante di propagazione* $\gamma(f)$.



Impedenza caratteristica E' definita come

$$Z_0(f) = R_0(f) + jX_0(f) = \sqrt{\frac{r + j2\pi fl}{g + j2\pi fc}} \quad (15.1)$$

e rappresenta il rapporto tra $V(f)$ ed $I(f)$ in un generico punto del cavo, permettendo di scrivere

$$I(f) = \frac{V(f)}{Z_0(f)}$$

Costante di propagazione E' definita come

$$\gamma(f) = \alpha(f) + j\beta(f) = \sqrt{(r + j2\pi fl)(g + j2\pi fc)} \quad (15.2)$$

mentre la grandezza $e^{-\gamma(f)d}$ rappresenta il rapporto dei valori di tensione presenti tra due punti di un cavo di lunghezza infinita, distanti d , permettendo di scrivere

$$V(f, x + d) = e^{-\gamma(f)d} V(f, x)$$

Condizioni di chiusura Qualora il cavo di lunghezza d sia chiuso ai suoi estremi su di un generatore con impedenza $Z_g(f)$ e su di un carico $Z_c(f)$, risultano definiti i *coefficienti di riflessione* del generatore e del carico:

$$r_g(f) = \frac{Z_g(f) - Z_0(f)}{Z_g(f) + Z_0(f)} \quad e \quad r_c(f) = \frac{Z_c(f) - Z_0(f)}{Z_c(f) + Z_0(f)} \quad (15.3)$$

Osserviamo subito che nel caso in cui $Z_g(f) = Z_c(f) = Z_0(f)$, risulta $r_g(f) = r_c(f) = 0$.

Quadripolo equivalente L'impedenza vista dai morsetti di *ingresso* e di *uscita* di un cavo, interposto tra generatore e carico, vale rispettivamente

$$Z_i(f) = Z_0(f) \frac{1 + r_c(f) \cdot e^{-2d\gamma(f)}}{1 - r_c(f) \cdot e^{-2d\gamma(f)}} \quad e \quad Z_u(f) = Z_0(f) \frac{1 + r_g(f) \cdot e^{-2d\gamma(f)}}{1 - r_g(f) \cdot e^{-2d\gamma(f)}} \quad (15.4)$$

Allo stesso tempo, la funzione di trasferimento *intrinseca* risulta

$$H_q(f) = 2 \frac{e^{-d\gamma(f)}}{1 - r_g(f) \cdot r_c(f) \cdot e^{-2d\gamma(f)}} \quad (15.5)$$

Condizioni di adattamento Nel caso in cui $Z_g(f) = Z_c(f) = Z_0(f)$, come sappiamo, il quadripolo si comporta in modo perfetto. In tal caso, risultando $r_g(f) = r_c(f) = 0$, si ottiene che $Z_i(f) = Z_u(f) = Z_0(f)$ e $H_q(f) = \frac{V_q(f)}{V_i(f)} = 2e^{-d\gamma(f)}$: il cavo si comporta allora come se avesse lunghezza infinita. In tal caso, inoltre, risulta che $H_i(f) = \frac{1}{2}$ ed $R_g(f) = R_u(f)$; pertanto il guadagno disponibile risulta

$$G_d(f) = |H_i(f)|^2 |H_q(f)|^2 \frac{R_g(f)}{R_u(f)} = \frac{1}{4} \left| 2e^{-d[\alpha(f) + j\beta(f)]} \right|^2 = e^{-2d\alpha(f)}$$

Condizione di Heaviside Nel caso in cui i valori delle costanti primarie siano tali da risultare $r \cdot c = l \cdot g$, relazione nota come *condizione di Heaviside*, le (15.1) e (15.2) si modificano, e si ottiene

$$\gamma(f) = \sqrt{rg} + j2\pi f\sqrt{lc} \quad e \quad Z_0(f) = \sqrt{\frac{r}{g}} = \sqrt{\frac{l}{c}} = R_0$$

e pertanto, risultando $\alpha(f)$ costante e $\beta(f)$ linearmente crescente con la frequenza, si realizzano le condizioni di un canale perfetto; dato inoltre che l'impedenza caratteristica $Z_0(f) = R_0$ è solo resistiva ed indipendente dalla frequenza, diviene semplice realizzare la condizione di adattamento $Z_g(f) = Z_c(f) = R_0$, il che determina al contempo anche il massimo trasferimento di potenza, e implica che $r_g(f) = r_c(f) = 0$, e quindi

$$H_q(f) = 2e^{-d\alpha(f)} e^{-jd\beta(f)} = 2e^{-d\sqrt{rg}} e^{-jd2\pi f\sqrt{lc}}$$

In definitiva, la funzione di trasferimento complessiva in questo caso vale

$$H(f) = H_i(f) H_q(f) H_u(f) = \frac{1}{2} 2e^{-d\sqrt{rg}} e^{-jd2\pi f\sqrt{lc}} \frac{1}{2} = \frac{1}{2} e^{-d\sqrt{rg}} e^{-jd2\pi f\sqrt{lc}}$$

equivalente quindi ad un canale perfetto con guadagno $G = \frac{1}{2} e^{-d\sqrt{rg}}$ e ritardo $t_R = d\sqrt{lc}$; al contempo, l'attenuazione disponibile risulta indipendente da f , e pari a

$$A_d(f) = 1/G_d(f) = e^{2d\sqrt{rg}}$$

15.2.2 Trasmissione in cavo

In generale, le costanti primarie del cavo non soddisfano le condizioni di Heaviside, e le impedenze di chiusura non sono adattate. In tal caso si ha $r_g(f) \neq 0$ e/o $r_c(f) \neq 0$, e devono essere applicate le (15.4) e (15.5).

Cavo molto lungo Se il cavo è sufficientemente lungo da poter porre $e^{-2d\gamma(f)} \ll 1$, ossia $|e^{-2d\gamma(f)}| = e^{-2d\alpha(f)} \ll 1$, le (15.4) divengono $Z_i(f) = Z_u(f) \simeq Z_0(f)$, mentre la (15.5) si semplifica in $H_q(f) = 2e^{-d\gamma(f)}$; nel caso generale risulta pertanto

$$G_d(f) = |H_q(f)|^2 \cdot |H_i(f)|^2 \cdot \frac{R_g(f)}{R_u(f)} = 4 \cdot e^{-2d\alpha(f)} \cdot |H_i(f)|^2 \cdot \frac{R_g(f)}{R_u(f)}$$

che evidenzia due cause di distorsione lineare, di cui la prima dipende dal disadattamento di impedenze in ingresso ed uscita: qualora invece si realizzi la condizione $Z_g(f) = Z_c(f) = Z_0(f)$, si ottiene

$$A_d(f) = \frac{1}{G_d(f)} = e^{2d\alpha(f)}$$

La seconda causa di distorsione dipende dal comportamento non perfetto di $H_q(f) = 2e^{-d\gamma(f)}$, che secondo la teoria può essere completamente corretto, solo nel caso in cui le costanti primarie soddisfino le condizioni di Heaviside. In pratica, però, il risultato è diverso, perché... le "costanti primarie" *non sono costanti !!!*

Effetto pelle Si tratta di un fenomeno legato all'addensamento del moto degli elettroni verso la superficie del cavo, al crescere della frequenza. Per questo motivo, si riduce la superficie del conduttore realmente attraversata da corrente elettrica, a cui corrisponde un aumento della resistenza per unità di lunghezza r . Si può mostrare che, per frequenze maggiori di 50-100 KHz, la resistenza per unità di lunghezza r aumenta proporzionalmente a \sqrt{f} , e quindi si può scrivere $\alpha(f) = \alpha_0\sqrt{f}$, in cui la costante α_0 dipende dal tipo di cavo.

In tali condizioni, l'attenuazione disponibile risulta $A_d(f) = e^{2d\alpha(f)} = e^{2d\alpha_0\sqrt{f}}$, a cui corrisponde un valore in dB pari a

$$A_d(f)|_{dB} = 10 \log_{10} e^{2d\alpha_0\sqrt{f}} = d\alpha_0\sqrt{f} \cdot 10 \log_{10} e^2 = A_0 \cdot d \cdot \sqrt{f}$$

Il valore A_0 riassume in se tutte le costanti coinvolte, prende il nome di *attenuazione chilometrica*, ed è espresso in dB/Km; il suo valore dipende dal tipo di cavo, ed è fornito con riferimento ad una determinata frequenza f_R (ad es. 1 MHz), permettendo di scrivere

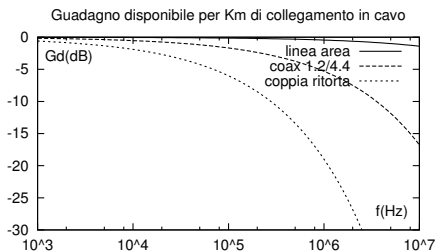
$$A_d(f)|_{dB} = A_0(f_R) \cdot d_{Km} \cdot \sqrt{\frac{f}{f_R}} \quad (15.6)$$

in cui f_R rappresenta appunto la frequenza per la quale è disponibile il valore di A_0 , ed il valore della f per cui si calcola A_d va espresso nella stessa unità di misura di f_R . Questo risultato può essere usato come formula di progetto, e mette in evidenza come l'attenuazione in dB dei cavi sia linearmente proporzionale alla lunghezza³.

³Questa circostanza è comune con le trasmissioni in fibra ottica (vedi fig. 15.6 a pag. 407), ed è legato alla presenza nel mezzo di una componente dissipativa, in questo caso la resistenza.

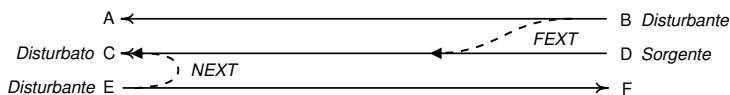
La figura a lato mostra l'andamento di $G_d(f)|_{dB} = -A_d(f)|_{dB}$ adottando i valori di A_0 e f_R riportati al § 15.2.3 per alcune tipologie di cavo.

Equalizzazione In presenza di effetto pelle, la funzione di trasferimento intrinseca $H_q(f) = 2e^{-d\gamma(f)}$ presenta una dipendenza da f tutt'altro che perfetta, causando potenzialmente distorsioni lineari sui segnali in transito. Un problema analogo insorge anche in assenza di effetto pelle, qualora si manifesti un disadattamento di impedenze ed il cavo non sia sufficientemente lungo (vedi appresso).



Se la banda di segnale è sufficientemente estesa da causare una distorsione lineare non trascurabile, o se la particolare natura del segnale (ad es. numerico) richiede la presenza di un ritardo strettamente costante con f , è necessario prevedere uno stadio di equalizzazione.

Diafonia La diafonia, indicata in inglese con il termine di *crossstalk*, consiste nei fenomeni di interferenza tra i messaggi trasportati su cavi disposti in prossimità reciproca, e dovuti a fenomeni di induzione elettromagnetica ed accoppiamenti elettrostatici. Il fenomeno è particolarmente rilevante in tutti i casi in cui molti cavi giacciono affasciati in una medesima canalizzazione, condividendo un lunghezza significativa di percorso. Nel caso di telefonia analogica, la diafonia può causare l'ascolto indesiderato di altre comunicazioni⁴; nel caso di trasmissioni numeriche o di segnali modulati, la diafonia produce un disturbo additivo supplementare, che peggiora le prestazioni espresse in termini di probabilità di errore o di SNR.



Con riferimento allo schema della figura soprastante, consideriamo un collegamento D-C su cui gravano due cause di interferenza di diafonia: il collegamento da E ad F produce il fenomeno di *paradiafonia* (in inglese NEXT, *near end crosstalk*), mentre il collegamento da B ad A produce il fenomeno di *telediafonia* (FEXT, *far end crosstalk*). Nel primo caso, il segnale disturbante ha origine in prossimità del punto di prelievo del segnale disturbato, mentre nel secondo ha origine in prossimità del punto di immissione.

L'entità del disturbo è quantificata mediante un valore di attenuazione di diafonia tra le sorgenti disturbanti e l'estremo disturbato. La circostanza che, nei rispettivi punti di immissione, i segnali disturbanti hanno la stessa potenza della sorgente che emette il segnale disturbato, permette di definire lo *scarto di paradiafonia*

$$\Delta A_{EC}|_{dB} = A_{EC}|_{dB} - A_{DC}|_{dB}$$

come la differenza in dB tra l'*attenuazione di paradiafonia* $A_{EC}|_{dB}$ e l'*attenuazione del collegamento* $A_{DC}|_{dB}$. Il livello di potenza del segnale disturbante proveniente da

⁴... le famose *interferenze telefoniche*, praticamente scomparse con l'avvento della telefonia numerica (PCM), da non confondere con ... *le intercettazioni*.

E ed osservato al punto C risulta quindi pari a⁵ $W_E^{next} = W_E - A_{EC} = W_D - A_{EC} = W_C + A_{DC} - A_{EC} = W_C - \Delta A_{EC}$, ossia di ΔA_{EC} dB inferiore al segnale utile. Una definizione del tutto analoga risulta per la telediafonia (FEXT), per la quale il livello di potenza del segnale disturbante proveniente da B ed osservato al punto C risulta $W_B^{fext} = W_C - \Delta A_{BC}$ in cui lo scarto di telediafonia ha il valore

$$\Delta A_{BC}|_{dB} = A_{BC}|_{dB} - A_{DC}|_{dB}$$

15.2.2.1 Casi limite

Cavo a basse perdite E' un modello applicabile per tutte quelle frequenze per cui risulti $r \ll 2\pi fl$ e $g \ll 2\pi fc$. In tal caso le (15.1) e (15.2) forniscono

$$Z_0(f) = R_0 = \sqrt{\frac{l}{c}} \text{ reale} \quad e \quad \gamma(f) = j2\pi f\sqrt{lc}$$

Di conseguenza, è facile realizzare $Z_g = Z_c = R_0$, che determina

$$H_q(f) = 2e^{-jd2\pi f\sqrt{lc}}$$

quindi il cavo non presenta distorsioni di ampiezza, ha una attenuazione trascurabile, e manifesta una distorsione di fase lineare in f , realizzando quindi le condizioni di canale perfetto.

Cavo corto E' il caso di collegamenti interni agli apparati, o tra un trasmettitore-ricevitore e la relativa antenna. La ridotta lunghezza del cavo permette di scrivere

$$e^{-d\gamma(f)} = e^{-d\alpha(f)}e^{-jd\beta(f)} \simeq e^{-jd\beta(f)}$$

in quanto $e^{-d\alpha(f)} \simeq 1$.

Qualora si verifichi un disadattamento di impedenze, i coefficienti di riflessione $r_g(f)$ e $r_c(f)$ risultano diversi da zero, rendendo

$$H_q(f) = 2 \frac{e^{-jd\beta(f)}}{1 - r_g(f) \cdot r_c(f) \cdot e^{-j2d\beta(f)}}$$

periodica con d e con f (quest'ultimo in assenza di effetto pelle). In particolare, se il carico viene sconnesso, o l'uscita del cavo posta in corto circuito, l'eq. (15.3) mostra come risulti $r_c(f) = \pm 1$ rispettivamente, e la prima delle (15.4) diviene

$$Z_i(f) = Z_0(f) \frac{1 \pm e^{-j2d\beta(f)}}{1 \mp e^{-j2d\beta(f)}}$$

e si vede che per quei valori (ricorrenti) di frequenza f che rendono $e^{-jd\beta(f)} = \pm 1$ (6), l'impedenza di ingresso del cavo può risultare infinita o nulla.

Evidentemente, le distorsioni lineari prodotte in questo caso hanno un andamento del tutto dipendente dalle particolari condizioni operative, e dunque la loro equalizzazione deve prevedere componenti in grado di adattarsi alla $H_q(f)$ del caso⁷. D'altra

⁵Omettiamo di indicare di operare in dB per chiarezza di notazione.

⁶Ovvero, tali che $|\cos 2d\beta(f) - j \sin 2d\beta(f)| = 1$, e quindi $d\beta(f) = k\frac{\pi}{2}$ con $k = 0, 1, 2, \dots$

⁷Può ad esempio rendersi necessario "tarare" un trasmettitore radio, la prima volta che lo si collega all'antenna.

parte, una volta equalizzato il cavo, non sono necessari ulteriori aggiustamenti, a parte problemi di deriva termica. Diverso è il caso dal punto di vista di un terminale di rete, per il quale il cavo effettivamente utilizzato può essere diverso da collegamento a collegamento, e pertanto i dispositivi modem a velocità più elevate devono disporre di un componente di equalizzazione adattiva, da regolare ogni volta ad inizio del collegamento⁸.

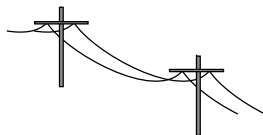
15.2.3 Tipologie di cavi per le telecomunicazioni

Descriviamo i principali tipi di cavo utilizzati, per i quali forniamo in tabella i valori tipici delle grandezze essenziali, nelle condizioni illustrate nel testo che segue.

Tipo di cavo	A_0 [dB/Km]	Z_0 [Ω]	r, g, l, c per 1 Km
Linee aeree	0.036 ad 1 KHz 0.14 a 100 KHz	600	$5, 10^{-6}, 2 \cdot 10^{-3}, 5 \cdot 10^{-9}$
Coppie ritorte	1.2 ad 1 KHz 6 a 100 KHz 20 a 1 MHz	$600e^{-j\frac{\pi}{4}}$	$100, 5 \cdot 10^{-5}, 10^{-3}, 5 \cdot 10^{-8}$
Coax 1.2/4.4 mm	5.3 ad 1 MHz	75 con politene	$89, 1.88 \cdot 10^{-7}, .26 \cdot 10^{-6}, 10^{-10}$
" 2.6/9.5 mm	2.3 ad 1 MHz	50 con aria	41, " , " , "
" 8.4/38 mm	.88 ad 1 MHz	$\frac{138}{\sqrt{\epsilon_r}} \log_{10} \frac{D}{d}$	1.45, " , " , "

15.2.3.1 Coppie simmetriche

Linea aerea E' costituita da una coppia di conduttori nudi, di bronzo od acciaio rivestito in rame, con diametro ϕ da 2 a 4 mm, sostenuti da una palificazione che li mantiene a distanza di 15 - 30 cm. L'uso delle linee aeree è andato estinguendosi con il tempo, ma rimane largamente diffuso nei paesi più poveri.



I valori riportati in tabella sono riferiti a conduttori con $\phi = 3$ mm, a frequenza di 1 KHz; la r già a 100 KHz cresce al valore di 20 Ω /Km, mentre la conduttanza g a 100 KHz e con tempo molto umido, può crescere fino a decine di volte il suo valore nominale ad 1 KHz. I valori riportati mostrano come le condizioni di Heaviside non siano rispettate, in quanto $rc \gg lg$, anche se lo scarto è inferiore rispetto al caso delle coppie ritorte.

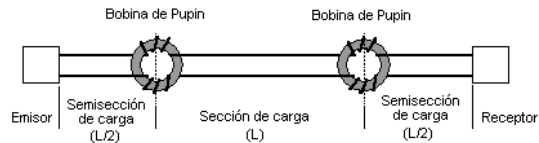
L'impedenza caratteristica riportata in tabella, di circa 600 Ω , è ottenuta applicando il modello a basse perdite, con le costanti primarie indicate.

Coppia ritorta E' costituita da una coppia di conduttori in rame, con ϕ da 0.4 ad 1.3 mm, rivestiti di materiale isolante, ed avvolti tra loro secondo eliche con passo grande rispetto al diametro. Un numero variabile di tali coppie (tra qualche decina e qualche centinaio) sono poi raggruppate assieme, e rivestite con guaine protettive isolanti o metalliche; il risultato dell'operazione è interrato o sospeso mediante una fune in acciaio. L'uso delle coppie ritorte, nato allo scopo di realizzare il collegamento tra utente e centrale telefonica, si è esteso al cablaggio di reti locali (LAN) con topologia a stella (IEEE 802.3); in tale contesto, i cavi sono indicati come UTP (*unshielded twisted pair*).

⁸E' questa la fase in cui il modem che usiamo per collegarci al provider internet emette una serie di orribili suoni....

I valori riportati in tabella sono riferiti a conduttori con $\phi = .7$ mm, a frequenza di 1 KHz; la r a 100 KHz è circa doppia. La g dipende sostanzialmente dall'isolante utilizzato, mentre l'aumento di c è evidentemente legato alla vicinanza dei conduttori. Anche in questo caso, risulta $rc \gg lg$, e dunque le condizioni di Heaviside non sono verificate. Nel passato, si è fatto largo uso dell'espedito di innalzare artificialmente l , collocando ad intervalli regolari una induttanza "concentrata" (le cosiddette bobine *Pupin*), realizzando così nella banda del canale telefonico un comportamento approssimativamente perfetto.

Al crescere della frequenza, però, le bobine Pupin producono un effetto passa basso, aumentando di molto il valore di attenuazione; attualmente, se le stesse coppie ritorte sono utilizzate per la trasmissione di segnali numerici PCM, le bobine Pupin sono state rimosse, ed al loro posto inseriti ripetitori rigenerativi.



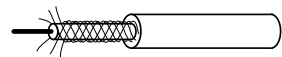
L'impedenza caratteristica riportata in tabella, di circa 600 Ω , è valida a frequenze audio, con cavi $\phi = .7$ mm. Prevalendo l'aspetto capacitivo, al crescere della frequenza Z_0 si riduce a 100-200 Ω , con fase di -10 gradi. L'attenuazione chilometrica riportata, è sempre relativa al caso $\phi = .7$ mm; per diametri di 1.3 mm si ottengono valori circa dimezzati, mentre con $\phi = .4$ mm il valore di A_0 risulta maggiore.

Come ultima osservazione, illustriamo come l'avvolgimento della coppia su se stessa ha lo scopo di ridurre i disturbi di diafonia. Infatti, se il passo dell'elica è diverso tra le coppie affasciate in unico cavo, le tensioni e correnti indotte da una coppia su di un'altra non interessano sempre lo stesso conduttore, ma entrambi in modo alternato. L'avvolgimento della coppia disturbante, inoltre, produce una alternanza dei conduttori in vicinanza della coppia disturbata, aggiungendo una ulteriore alternanza del verso del fenomeno di disturbo. Con questi accorgimenti, si trovano attenuazioni di diafonia a frequenze vocali, dell'ordine di 80-90 dB su 6 Km. All'aumentare della frequenza, e della lunghezza del percorso comune, l'attenuazione di diafonia diminuisce (e quindi l'interferenza aumenta), fino a mostrare valori di 60-70 dB a 750 KHz su 1.6 Km.

15.2.3.2 Cavo coassiale

Un conduttore centrale è ricoperto di dielettrico, su cui è avvolto il secondo conduttore, intrecciato a formare una sorta di calza, e racchiuso a sua volta in una guaina isolante. La particolare conformazione del cavo lo rende molto più resistente ai fenomeni di interferenza; indicando con ϕ il diametro del conduttore interno e con D quello esterno, la teoria mostra che si determina un minimo di attenuazione se $D/\phi = 3.6$; per questo sono stati normalizzati i diametri mostrati nella tabella a pag. 384. Il tipo con $\phi/D = 8.4/38$ mm è sottomarino, e presenta la minima attenuazione chilometrica; A_0 aumenta al diminuire della sezione del cavo.

Finché $D/\phi = 3.6$, l'impedenza caratteristica dipende solo dal dielettrico, con l'espressione generale fornita in tabella, ottenendo i valori di 50 e 75 Ω con dielettrico aria e polietilene rispettivamente.



I valori delle costanti primarie riportati in tabella sono ottenuti facendo uso delle seguenti relazioni: $r = 8.4 \cdot 10^{-8} \sqrt{f} \left(\frac{1}{D} + \frac{1}{\phi} \right) \Omega/m$; $l = 0.46 \log_{10} \frac{D}{\phi} \mu H/m$; $g = 152 \cdot 10^{-12} \frac{f \varepsilon_r \tan \delta}{\log_{10} \frac{D}{\phi}} S/m$; $c = \frac{24.2 \cdot \varepsilon_r}{\log_{10} \frac{D}{\phi}} pF/m$; in cui si è posto f (in Hz nelle formule) pari a 1 MHz, D e d sono espressi in metri, ε_r è la costante dielettrica, e $\tan \delta$ è l'angolo di perdita del dielettrico; nel caso del polietilene, risulta $\varepsilon_r = 2.3$, $\tan \delta = 3 \cdot 10^{-4}$.

Esercizio Si desidera effettuare una trasmissione FDM di 120 canali telefonici, ognuno modulato AM-BLU, su di un cavo coassiale, nella banda di frequenze $1 \div 1.48$ MHz. Desiderando una potenza ricevuta per ogni canale di almeno 1 mW, e disponendo di un trasmettitore in grado di erogare 10 W, determinare la massima lunghezza del collegamento, supponendo verificate le condizioni di adattamento agli estremi del cavo, con impedenza caratteristica resistiva, ed attenuazione chilometrica $A_0 = 5.3$ dB/Km ad 1MHz. Di quanto dovrebbe aumentare la potenza trasmessa W_{dT} per raddoppiare la lunghezza?

Soluzione Supponendo tutti i canali contemporaneamente attivi, la potenza trasmessa per ciascuno di essi risulta pari a

$$W_{dT}^{(n)} = \frac{10}{120} = 83.3 \text{ mW}, \text{ con } n = 1, 2, \dots, 120.$$

Tra tutti i canali, quello che subisce la massima attenuazione chilometrica è quello con portante più elevata, per il quale

$$A_d^{(120)} \text{ (dB/Km)} = A_0 \sqrt{1.48} = 5.3 \cdot 1.22 = 6.46 \text{ dB/Km}.$$

Per questo canale, il *guadagno di sistema* risulta pari a

$$G_s^{(120)} \Big|_{dB} = 10 \log_{10} \frac{W_{dT}^{(120)}}{W_{RMin}^{(120)}} = 10 \log_{10} \frac{83.3}{1} = 19.2 \text{ dB},$$

essendo $W_{RMin} = 1$ mW come indicato nel testo. Come noto, G_s corrisponde alla massima attenuazione A_{dTot} che può essere accettata, e pertanto

$$A_{dTot}^{(120)} \Big|_{dB} = A_d^{(120)} \text{ (dB/Km)} \cdot L_{Km} = 19.2 \text{ dB},$$

da cui si ricava per la massima lunghezza

$$L_{Km} = \frac{A_{dTot}^{(120)} \Big|_{dB}}{A_d^{(120)} \text{ (dB/Km)}} = \frac{19.2}{6.46} = 2.97 \text{ Km},$$

che come vediamo è imposta dal canale più attenuato.

Per il primo canale, invece, si ha $A_d^{(1)} \text{ (dB/Km)} = A_0$, e dunque

$$A_{dTot}^{(1)} \Big|_{dB} = A_0 \text{ (dB/Km)} \cdot L_{Km} = 5.3 \cdot 2.97 = 15.74 \text{ dB}.$$

La differenza tra $G_s \Big|_{dB}$ (uguale per tutti i canali) e $A_{dTot}^{(1)} \Big|_{dB}$ rappresenta il margine di sistema per il primo canale, pari a

$$M = G_s - A_d = 19.2 - 15.74 = 3.46 \text{ dB}.$$

La stessa quantità, è anche uguale alla *distorsione lineare di ampiezza* del cavo nella banda del segnale.

Nel caso in cui si voglia superare una lunghezza doppia, anche $A_{dTot}^{(120)} \Big|_{dB}$ raddoppia, e per mantenere $W_{RMin} = 1$ mW, deve raddoppiare anche $G_s^{(120)} \Big|_{dB}$. Pertanto la nuova potenza/canale risulta

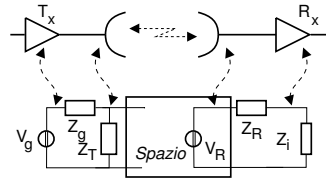
$$W'_{dT} \text{ (dBm)} = W_{Rmin} \text{ (dBm)} + G'_s \text{ (dB)} = 0 + 2G_s \text{ (dB)}; \text{ quindi}$$

$$W'_{dT} \text{ (mW)} = 10^{\frac{W'_{dT} \text{ (dBm)}}{10}} = 10^{\frac{2G_s \text{ (dB)}}{10}} = 10^{\frac{2 \cdot 19.2}{10}} = 10^{3.84} = 6918.3$$

milliWatt, cioè 6.91 Watt/canale, e quindi $6.91 \cdot 120 = 830$ Watt complessivi !

15.3 Collegamenti radio

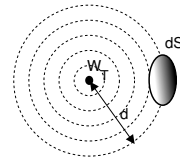
Questa sezione inizia dalla descrizione delle antenne e dei fenomeni di propagazione nei termini di potenza complessiva del segnale modulato, ponendosi nella semplice ipotesi di occupare una banda molto stretta attorno alla portante f_0 , da poterlo quasi assimilare ad una singola sinusoide. Pertanto, le condizioni di massimo trasferimento di potenza tra amplificatore finale e antenna trasmittente ($Z_g = Z_T^*$) e tra antenna ricevente e stadio di ingresso al ricevitore ($Z_R = Z_i^*$) danno luogo, nella banda di segnale, ad una risposta in frequenza $H(f)$ che non dipende dalla frequenza (modulo e fase costanti), e questo corrisponde (a parte una rotazione di fase) all'assenza di distorsione lineare, vedi pag. 361. Tutta la potenza disponibile fornita dall'amplificatore finale $W_{dT} = \frac{V_{Teff}^2}{4R_g}$ viene ceduta all'antenna, e da questa allo spazio. In realtà Z_T dipende dalla geometria dello spazio circostante; perciò l'amplificatore del trasmettitore Tx va *accordato* dopo aver posizionato l'antenna.



15.3.1 Trasduzione elettromagnetica

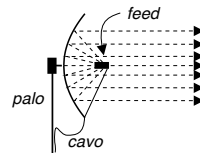
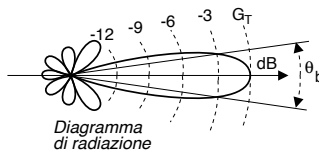
Antenna isotropa Se l'antenna trasmittente irradia allo stesso modo in tutte le direzioni, W_{dT} si distribuisce su di una sfera; dunque una superficie dS , posta a distanza d , è attraversata da una potenza pari a

$$dW = W_{dT} \frac{dS}{4\pi d^2} \quad [\text{Watt}] \quad (15.7)$$



Si noti che il denominatore rappresenta la superficie di una sfera di raggio d .

Antenna direttiva Sono antenne che hanno direzioni privilegiate di emissione. Ad esempio, le antenne paraboliche dispongono di un *illuminatore* o FEED⁹ posto in corrispondenza del *fuoco* della parabola stessa, la cui superficie riflette le onde elettromagnetiche in modo che si propagano in forma pressoché parallela¹⁰.



W_{dT} si distribuisce quindi *non* sfericamente, e la direzione di propagazione massima esibisce un guadagno G_T rispetto all'antenna isotropa, mentre l'intensità di campo irradiato spazialmente è descritta da un *diagramma di radiazione*. Il valore di G_T dipende dal rapporto tra le dimensioni dell'antenna e quelle della lunghezza d'onda λ secondo la relazione

$$G_T = 4\pi \frac{A}{\lambda^2} \quad (15.8)$$

⁹Dall'inglese *to feed* = alimentare.

¹⁰Il processo di focalizzazione parabolica, comunemente usato ad esempio nei *fanali* degli autoveicoli, era ben noto ad un certo siracusano...

avendo indicato con A l'area dell'antenna.

Può essere definita una *larghezza del fascio* (BEAM WIDTH), che misura l'angolo θ_b entro cui la potenza irradiata è superiore alla metà della massima potenza presente nella direzione privilegiata¹¹. Ovviamente minore è θ_b , e maggiore è G_T .

Antenna ricevente Se una antenna identica a quella trasmittente viene usata (dall'altro lato del collegamento) per ricevere, questa mantiene lo stesso guadagno $G_R = G_T$ e lo stesso θ_b . Si definisce allora la sua *area efficace* come il valore

$$A_e = G_R \frac{\lambda^2}{4\pi} \quad (15.9)$$

legato alla forma e dimensione dell'antenna, a meno di un fattore di efficienza ρ (¹²). Perciò una stessa antenna (A_e fisso) aumenta il suo guadagno (e stringe il *beam*) all'aumentare della frequenza, ovvero al diminuire di $\lambda = \frac{c}{f}$ (¹³).

15.3.2 Trasferimento energetico

Potenza ricevuta Usando l'area efficace dell'antenna ricevente (15.9) per intercettare parte della potenza irradiata (15.7), si ottiene

$$W_R = W_{dT} G_T \frac{A_e}{4\pi d^2} = W_{dT} G_T G_R \left(\frac{\lambda}{4\pi d} \right)^2 \text{ [Watt]}$$

Ovviamente, anche il ricevitore ha la propria $Z_i = Z_R^*$ accordata per il massimo trasferimento di potenza, e la banda di segnale è sempre stretta a sufficienza da garantire l'assenza di distorsioni lineari. Quindi la $W_R = W_{dR}$ è proprio la potenza ricevuta.

Attenuazione di spazio libero Il termine

$$A_{sl} = \left(\frac{4\pi d}{\lambda} \right)^2 = \left(\frac{4\pi df}{c} \right)^2 \quad (15.10)$$

è chiamato *attenuazione di spazio libero*, che dipende da f^2 . In realtà ai fini del bilancio energetico, la dipendenza dalla frequenza si elide con quella relativa al guadagno delle antenne: $G_T = A_e \frac{4\pi}{\lambda^2} = A_e \frac{4\pi f^2}{c^2}$ (¹⁴).

¹¹Si tratta di un concetto del tutto analogo alla "frequenza di taglio a 3 dB", ma applicata ad un dominio spaziale con geometria radiale.

¹²Indicando con A_r l'area *reale* (fisica) dell'antenna, risulta $A_e = \rho A_r$, con $\rho < 1$. La disegualianza tiene conto delle perdite dell'antenna, come ad esempio le irregolarità nella superficie della parabola, o l'ombra prodotta dalle strutture di sostegno. Ovviamente, anche l'antenna trasmittente presenta perdite, ed il valore G_T *misurato* è inferiore a quello fornito dalla (15.8), a meno di non usare appunto il valore di area efficace.

¹³La costante $c = 3 \cdot 10^8$ metri/secondo rappresenta la velocità della luce nel vuoto, ossia la velocità di propagazione dell'onda elettromagnetica nello spazio.

¹⁴Mantenendo fissa la dimensione delle antenne, si ottiene il risultato che trasmissioni operanti a frequenze più elevate permettono di risparmiare potenza. Purtroppo però, guadagni superiori a 30-40 dB (corrispondenti a piccoli valori di θ_b) sono controproducenti, per i motivi esposti al §15.3.3.1.

Attenuazione disponibile Il rapporto

$$A_d = \frac{W_{dT}}{W_{dR}} = \left(\frac{4\pi df}{c} \right)^2 \frac{1}{G_T G_R} \quad (15.11)$$

è chiamato *attenuazione disponibile*, ed indica di quanto si riduce la potenza trasmessa. Il suo valore espresso in decibel, tenendo conto delle costanti che vi compaiono, ed usando le unità di misura più idonee, risulta essere

$$A_d \text{ (dB)} = 32.4 + 20 \log_{10} f \text{ (MHz)} + 20 \log_{10} d \text{ (Km)} - G_T \text{ (dB)} - G_R \text{ (dB)} \quad (15.12)$$

nota come *equazione di Friis*. Osserviamo che, a differenza della trasmissione in cavo, l'attenuazione cresce con il quadrato della distanza, e quindi con il suo logaritmo quando espressa in decibel. Infatti ora l'attenuazione è dovuta esclusivamente all'aumentare della superficie su cui si distribuisce la potenza irradiata, e non a fenomeni dissipativi, come accade invece per cavo (eq. (15.6)) e fibra ottica. Per un esempio di applicazione della (15.12), si veda l'esercizio a pag. 425.

Il sistema di telecomunicazione che meglio rappresenta le condizioni di spazio libero è quello tra terra ed satellite, per il semplice fatto che non vi sono fraposti ostacoli, e che approfondiamo brevemente alla appendice 15.5.1. D'altra parte, i collegamenti radio terrestri, casalinghi, e mobili, sono affetti da una serie di ulteriori fenomeni, mentre la (15.12) si limita a considerare un solo aspetto del problema; di seguito ne citiamo un altro paio, ed ai prossimi § approfondiamo il tema:

- *perdite di accoppiamento*: sono dovute al mancato verificarsi delle condizioni di massimo trasferimento di potenza, ed ammontano a qualche dB;
- *assorbimento terrestre*: quando l'antenna è distante dal suolo meno di qualche lunghezza d'onda, l'energia si propaga anche per *onda superficiale*¹⁵, ovvero la crosta terrestre fa da conduttore. Ciò permette la ricezione anche in assenza di visibilità tra antenne, subendo però una attenuazione che *aumenta con la frequenza*, tanto che già a 3 MHz raggiunge i 25 dB¹⁶ ogni 10 Km. Le *onde medie* (0,3-3 MHz) sono meno attenuate, ed ancora meno le *onde lunghe* (10-300 KHz) che viaggiano appunto via terra.

15.3.3 Condizioni di propagazione e attenuazioni supplementari

Sono ora descritti i fenomeni legati alla geometria del territorio ed alle condizioni atmosferiche, che determinano l'insorgenza di attenuazioni *supplementari* A_s ovvero *in più*, da sommare al valore A_d (dB) fornito dalla (15.12) per ricavare la potenza realmente ricevuta.

15.3.3.1 Condizioni di visibilità

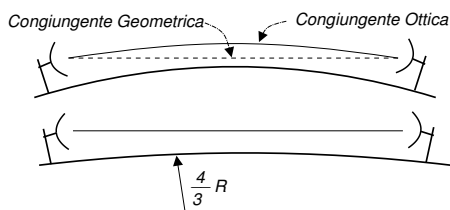
Come ricavabile anche dall'espressione dell'area efficace, all'aumentare della frequenza si possono ottenere antenne di dimensioni ridotte e contemporaneamente di elevato guadagno. Allo stesso tempo, per evitare l'assorbimento terrestre, occorre posizionare l'antenna in alto (in cima ad una torre), e trasmettere per *onda diretta*, condizione nota anche come LOS o *line of sight*.

¹⁵http://it.wikipedia.org/wiki/Onda_superficiale

¹⁶equivalente ad una riduzione di potenza di $10^{2.5} = 316$ volte

A causa della curvatura terrestre, esiste una altezza minima da rispettare: ad esempio con torri da 60 metri si raggiungono distanze (in visibilità) di 50 Km. Ovviamente, il problema si presenta in pianura. Tratte più lunghe richiedono torri più alte, ma anche guadagni di antenna maggiori (e quindi antenne più grandi e più direttive). Questa non è però una soluzione molto praticabile, in quanto in presenza di vento forte le antenne “grandi” possono spostarsi e perdere il puntamento; inoltre, il costo delle torri aumenta esponenzialmente con l’altezza.

Orizzonte radio Nel calcolare l’altezza delle torri (ed il puntamento delle antenne) si deve considerare anche il fenomeno legato al fatto che l’onda elettromagnetica, propagandosi, *si piega* verso gli strati dell’atmosfera con indici di rifrazione maggiori (ossia verso terra). Pertanto, i calcoli vengono effettuati supponendo che il raggio terrestre sia

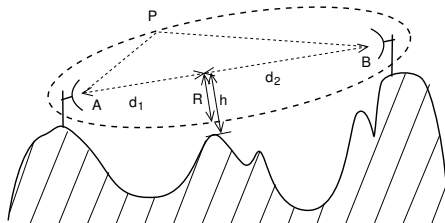


4/3 volte quello reale. Inoltre, l’indice di rifrazione (che aumenta verso il basso) può variare con l’ora e con le condizioni climatiche: pertanto, anche in questo caso, le antenne con guadagni elevati (e molto direttive) possono andare fuori puntamento.

Ellissoidi di Fresnel Anche se le antenne si trovano in condizioni di visibilità, occorre tenere conto dei fenomeni di *diffrazione*¹⁷, che *deviano* le onde radio che transitano in prossimità di ostacoli nella zona in *ombra*¹⁸. Pertanto, la determinazione dell’orizzonte radio deve prevedere un *margin*e di distanza h tra la congiungente delle antenne ed il suolo, od un eventuale ostacolo. La distanza h deve essere almeno pari al raggio del primo ellissoide di Fresnel, che è un solido di rotazione definito come il luogo dei punti P per i quali la somma delle distanze $d(A, P) + d(P, B)$ è pari a $d(A, B) + \frac{\lambda}{2}$, in cui $\lambda = \frac{c}{f}$ è la lunghezza d’onda della trasmissione a frequenza f .

Suddividendo la distanza $d(A, B)$ tra i due fuochi A e B in due segmenti d_1 e d_2 , individuati dalla posizione dell’ostacolo, si trova che il raggio dell’ellissoide è pari a

$$R = \sqrt{\frac{\lambda}{\frac{1}{d_1} + \frac{1}{d_2}}}$$



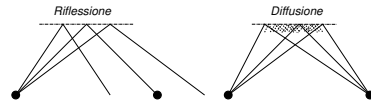
che, nel caso $d_1 = d_2 = \frac{d(A,B)}{2}$, assume il valore massimo $R_M = \frac{1}{2}\sqrt{\lambda d}$. Qualora si determini la condizione $h < R$, il collegamento subisce una attenuazione supplementare che aumenta al diminuire di $\frac{h}{R}$, ed è maggiore per gli *spigoli vivi*, fino ad arrivare ad una decina di dB.

15.3.3.2 Diffusione e riflessione atmosferica

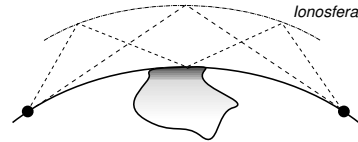
¹⁷<http://it.wikipedia.org/wiki/Diffrazione>

¹⁸Lo stesso fenomeno di diffrazione è egualmente valido per l’energia luminosa, e può essere sperimentato illuminando una fessura, ed osservando le variazioni di luminosità dall’altro lato.

Tra 0,1 e 10 GHz si può verificare il fenomeno della *diffusione troposferica*¹⁹ (lo strato dell'atmosfera fino a 20 Km di altezza), causata da turbolenze e particelle sospese, e che comportano un numero *infinito* di cammini multipli.



Tra qualche MHz e 30 MHz, intervengono fenomeni di radiodiffusione *ionosferica*²⁰ (la fascia oltre gli 80 Km), dove strati ionizzati causano *riflessioni* del segnale, e consentono la trasmissione anche tra luoghi non in visibilità²¹, ma con il rischio di cammini multipli. E' questo il caso tipico delle *onde corte*.



Per frequenze sotto il MHz la propagazione è per *onda di terra*, e l'assorbimento terrestre impedisce di coprire grandi distanze (tranne che per le *onde lunghe*, meno attenuate). Anche in questo caso può verificarsi la diffusione troposferica, specie di *notte*.

15.3.3.3 Assorbimento atmosferico

Per lunghezze d'onda di dimensione comparabile a quella delle molecole di ossigeno, si produce un fenomeno dissipativo *di assorbimento*; le frequenze interessate sono quelle superiori a 30 GHz, con un massimo di 20 dB/Km a 60 GHz⁽²²⁾. Inoltre, il vapor d'acqua (con molecole di dimensioni maggiori) produce una attenuazione supplementare di 1-2 dB/Km (al massimo) a 22 GHz²³. Sotto i 10 GHz non si verifica né assorbimento da ossigeno, né da vapore.

15.3.3.4 Dimensionamento di un collegamento soggetto a pioggia

In caso di pioggia, si manifesta una ulteriore causa di assorbimento atmosferico, detto appunto *da pioggia*, che costituisce la principale fonte di attenuazione supplementare per frequenze superiori a 10 GHz. L'attenuazione supplementare da pioggia aumenta con la frequenza portante, con l'intensità di precipitazione e con l'estensione della zona piovosa lungo il tragitto radio; questi ultimi due fattori sono evidentemente elementi aleatori, e per questo il dimensionamento mira a stabilire quale sia il margine necessario a garantire un grado di servizio prefissato. Il margine necessario, è pertanto pari al valore di attenuazione supplementare che viene superata con una probabilità minore o uguale al grado di servizio.

Una formula sperimentale che consente di determinare il valore in dB dell'attenuazione supplementare che viene superata con probabilità p è:

$$A_s(r_0, d, p) = K \cdot r_0^\alpha \cdot d \cdot \beta(d) \cdot \gamma(p) \quad [\text{dB}]$$

¹⁹http://en.wikipedia.org/wiki/Tropospheric_scatter

²⁰<http://en.wikipedia.org/wiki/Skywave>

²¹Anche, ma non solo, in concorso con la riflessione operata da masse d'acqua, come mostrato in figura.

²²L'elevata attenuazione chilometrica presente a 60 GHz può essere sfruttata nei sistemi di comunicazione cellulare, allo scopo di riusare una stessa banda di frequenze anche a breve distanza.

²³L'assorbimento di potenza da parte delle molecole d'acqua per onde elettromagnetiche a 22 GHz è il principio su cui si basa il forno a microonde.

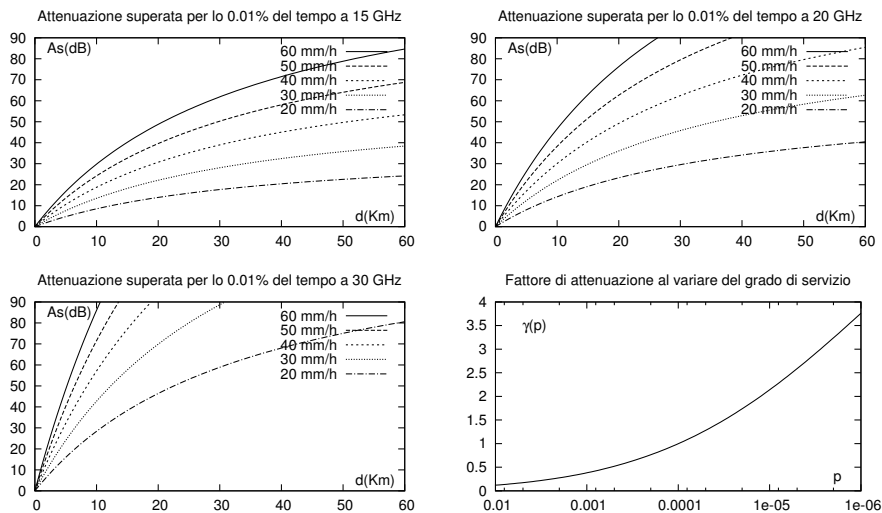


Figura 15.1: Curve di attenuazione supplementare per pioggia

in cui r_0 è l'intensità di precipitazione (in mm/h) che viene superata per lo 0.01 % del tempo, d è la lunghezza del collegamento, e K ed α sono costanti che caratterizzano l'entità dell'interazione dell'onda radio con la pioggia, in funzione della frequenza portante e di altre condizioni climatiche ed ambientali, i cui valori medi sono riportati nella tabella che segue.

f_0 (GHz)	10	15	20	25	30	35
α	1.27	1.14	1.08	1.05	1.01	.97
K	.01	.036	.072	.12	.177	.248

Il valore di r_0 per l'Italia è compreso tra 20 e 60 mm/h, mentre il termine $\gamma(p) = 6.534 \cdot 10^{-3} \cdot p^{-(.718 + .043 \cdot \log_{10} p)}$, che vale 1 per $p = 10^{-4}$, permette di tener conto del grado di servizio che si vuole ottenere. Infine, $\beta(d) = 1/(1 + .0286 \cdot d)$ è un fattore correttivo che tiene conto del fatto che *non piove lungo tutto* il collegamento. I grafici in fig. 15.1 mostrano l'andamento del termine $K \cdot r_0^\alpha \cdot d \cdot \beta(d)$ per diversi valori di f_0 ed r_0 , in funzione dell'estensione del collegamento; infine, è riportato il grafico della funzione $\gamma(p)$ per diversi valori di p .

Dimensionare un collegamento imponendo un margine elevato può dar luogo a problemi dal lato del ricevitore, che potrebbe trovarsi ad operare in regione non lineare a causa dell'eccesso di potenza ricevuta, qualora non siano presenti le attenuazioni supplementari: può essere allora utilizzato un canale di ritorno nell'altra direzione, in modo da regolare la potenza del trasmettitore.

15.3.3.5 Cammini multipli

Oltre i 30 MHz, nonostante la direttività delle antenne, alcuni raggi obliqui possono incontrare superfici riflettenti come laghi o masse d'acqua, essere riflessi dagli strati

atmosferici, o percorrere notevoli distanze nei condotti atmosferici²⁴ per poi tornare al suolo, e causare la ricezione di una (o più) eco ripetuta dello stesso segnale. In questi casi il collegamento si dice affetto da *multipath*, e può essere caratterizzato mediante una risposta impulsiva del tipo

$$h(t) = \sum_{n=1}^N a_n \delta(t - \tau_n) \tag{15.13}$$

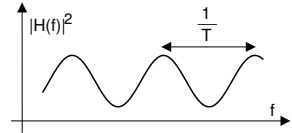
in cui i valori τ_n sono i ritardi con cui si presentano le diverse eco, ognuna caratterizzata da una ampiezza a_n , in accordo allo schema di filtro trasversale presentato al § 9.8.

La corrispondente risposta in frequenza

$$H(f) = \sum_{n=1}^N a_n e^{-j2\pi f \tau_n}$$

può produrre distorsione lineare; ricordiamo infatti (vedi § 9.8.2) che il modello di cammino multiplo per una sola eco con ritardo T , presenta una

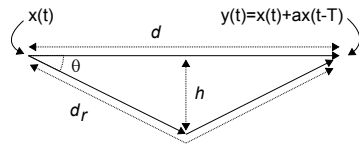
$$|H(f)|^2 = 1 + a^2 + 2a \cos 2\pi f T$$



periodica in frequenza con periodo $f = \frac{1}{T}$, ed osserviamo che all'aumentare di T , le oscillazioni di $|H(f)|^2$ si infittiscono²⁵, e dunque aumenta la possibilità che $|H(f)|^2$ vari di molto nella banda del segnale, causando una distorsione lineare che deve poi essere equalizzata.

Esempio Consideriamo la geometria descritta in figura,

in cui un collegamento di portata d subisce un fenomeno di riflessione a metà della sua lunghezza, da parte di una superficie riflettente che dista h dalla congiungente, e ricaviamo l'espressione del ritardo T . Ricordando che $\text{tempo} = \frac{\text{spazio}}{\text{velocità}}$ e



indicando con d_r la distanza percorsa dall'onda riflessa, otteniamo che la differenza tra i tempi di arrivo dell'onda diretta e riflessa vale $T = \frac{1}{c}(d_r - d)$; inoltre, dalla trigonometria risulta che $\frac{d}{2} = \frac{d_r}{2} \cos \theta$. Combinando le due relazioni, otteniamo che $T = \frac{d}{c} \left(\frac{1}{\cos \theta} - 1 \right)$, in cui $\theta = \arctan \frac{h}{d/2} = \arctan 2\frac{h}{d}$. Attualizzando il risultato ad uno scenario in cui $d = 1$ Km ed $h = 100$ metri, si ottiene $\theta = 11^\circ 31'$, $\cos \theta = 0.98$, e $T = 0,066 \mu\text{secondi}$. Pertanto $|H(f)|^2$ presenta un periodo (in frequenza) di $\frac{1}{T} = 15,15 \cdot 10^6 = 15,15$ MHz.

Qualora la banda del segnale sia invece sufficientemente piccola rispetto a $\frac{1}{T}$, e si possa considerare $|H(f)|^2$ costante (pag. 361), la presenza di cammini multipli può comunque dar luogo ad attenuazione, che prende il nome di *flat fading*²⁶ (§ 15.3.4.5).

²⁴Nel caso in cui una massa d'aria calda ne sovrasti una più fredda, si verifica una *inversione* dell'indice di rifrazione, e l'onda elettromagnetica di propaga come in una guida d'onda, vedi anche http://en.wikipedia.org/wiki/Tropospheric_propagation.

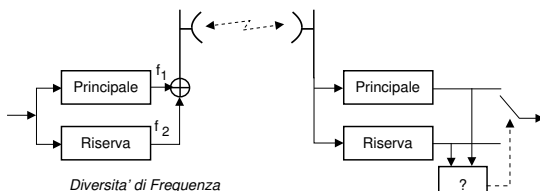
²⁵Ad esempio, desiderando $\frac{1}{T} > 1$ MHz, si ottiene $T_{Max} = 1 \mu\text{sec}$; se l'onda radio si propaga alla velocità $c = 3 \cdot 10^8$ m/sec, la massima differenza di percorso vale $\Delta_{max} = c \cdot T_{Max} = 3 \cdot 10^8 \cdot 10^{-6} = 300$ metri.

²⁶Il termine *fading* si traduce come *affievolimento* o *evanescenza*, ma è spesso usato in inglese, cosicché l'assenza di distorsione lineare per segnali a banda stretta è anche detta condizione di *fading piatto*, sottintendendo *in frequenza*.

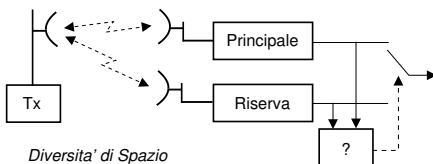
15.3.3.6 Collegamento in diversità

Il *fading* prodotto da cammini multipli viene detto *selettivo in frequenza* quando $|H(f)|^2$ varia in modo rilevante, e ciò accade specialmente quando due repliche del segnale giungono al ricevitore con ampiezze molto simili. Il problema può essere risolto prevedendo una *ridondanza* degli apparati.

Diversità di frequenza La stessa trasmissione è effettuata mediante due diverse portanti: nel caso in cui una delle due subisca attenuazione, la trasmissione che utilizza l'altra portante ne è probabilmente esente (o viceversa). Se la banda è particolarmente affollata, la stessa configurazione può essere adottata per fornire una ridondanza $N : 1$. Ad esempio, in una trasmissione multiplata FDM, la portante di riserva viene assegnata al canale del banco FDM che presenta la maggiore attenuazione.

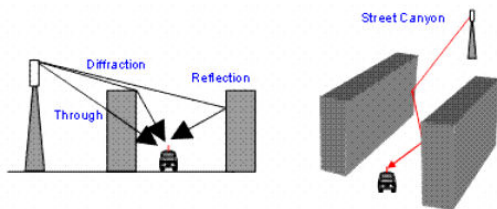


Diversità di spazio Adottando due diverse antenne riceventi in posizioni differenti, la differenza di percorso T tra cammini multipli è differente per le due antenne, e dunque la risposta in frequenza $|H(f)|^2 = 1 + a^2 + 2a \cos 2\pi fT$ ha una diversa periodicità nei due casi. Pertanto, anche se un ricevitore subisce una attenuazione selettiva, l'altro ricevitore ne è esente.



15.3.4 Collegamenti radiomobili

Le condizioni di propagazione per comunicazioni radiomobili, come nel caso della telefonia cellulare, presentano diversi aspetti particolari. Innanzitutto, l'antenna del terminale mobile è molto vicina al suolo, e ciò comporta la presenza di una eco fissa da terra, quasi sempre il mancato rispetto delle condizioni di Fresnel²⁷, ed una attenuazione supplementare da assorbimento terrestre. Inoltre, specialmente in ambito urbano, si



verifica un elevato numero di cammini multipli e diffrazioni, che perdipiù variano nel tempo in conseguenza dello spostamento del terminale. Infine, l'uso condiviso di una stessa banda di frequenze radio da parte di una moltitudine di terminali, determina la necessità di riusare le stesse

frequenze in regioni differenti²⁸, e l'attuazione di meccanismi di codifica di canale per

²⁷Alla frequenza di 1 GHz si ha $\lambda = 30$ cm e per una distanza di 100 metri dal trasmettitore si ottiene un raggio massimo dell'ellissoide pari a $\frac{1}{2}\sqrt{3} \cdot 100 = \frac{1}{2}\sqrt{30} \simeq 2.7$ metri.

²⁸Vedi ad es. il § 10.1.1.2.

ridurre gli effetti delle interferenze e del fading variabile²⁹. Analizziamo nel seguito i primi due fenomeni, fornendo modelli che descrivono le attenuazioni supplementari ed i fenomeni di multipath variabile, rimandando la discussione sulle tecniche di accesso multiplo ad una prossima edizione.

15.3.4.1 Le componenti della attenuazione supplementare

Occorre considerare una componente di attenuazione *nominale* indicata come *path loss* (o attenuazione di percorso), e due componenti aleatorie di attenuazione supplementare legate a posizione e movimento, indicate rispettivamente come attenuazione (o *fading*³⁰) su *larga scala* o *shadowing* (ombreggiatura) e su *piccola scala*.

Il valore del *path loss* A_{pl} risulta maggiore di quello di A_{sl} (eq. 15.10) a causa delle condizioni di propagazione non ideali, causando una A_d più elevata, come analizzato al § 15.3.4.2.

L'attenuazione supplementare su *larga scala* a_s^{ls} tiene conto dei fenomeni lentamente variabili nel tempo, come la frapposizione di rilievi, edifici, ed alberi: essa non varia di molto con il movimento del ricevitore, ed al § 15.3.4.3 si mostra come il suo valore sia una v.a. gaussiana in dB a media nulla e varianza σ_{ls}^2 , consentendo di determinare il *margin* su *larga scala* M_{dB}^{ls} come quel valore di a_s^{ls} che viene superato con probabilità sufficientemente bassa.

La variabilità su *piccola scala* o *fading* tiene invece conto degli innumerevoli cammini multipli presenti in ambito urbano ed *indoor*, che possono produrre una attenuazione supplementare a_s^{ps} maggiore del caso precedente, una $H(f)$ selettiva in frequenza, e se è presente movimento del ricevitore e/o delle superfici riflettenti, la variabilità temporale di a_s^{ps} ; a seconda se la rapidità di variazione sia maggiore o minore del periodo di simbolo, si distingue ulteriormente in *fast* e *slow fading*. Questi effetti sono analizzati al § 15.3.4.4, dove si determina il margine M_{dB}^{ps} tale da rendere trascurabile la probabilità che $a_s^{ps} > M_{dB}^{ps}$; mentre ai § 15.3.4.5 e 15.3.4.6 si illustrano rispettivamente i fenomeni di variabilità in frequenza e nel tempo.

La fig. 15.2 mostra come queste tre componenti di attenuazione si sommano al fine di determinare la potenza che occorre trasmettere

$$W_T = W_{R_{min}} + A_d + M_{dB}^{ls} + M_{dB}^{ps}$$

mentre quella a fianco tenta di rappresentare come varia la somma dei tre contributi di attenuazione con la posizione del ricevitore.

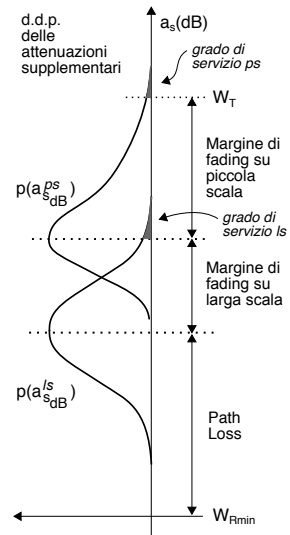
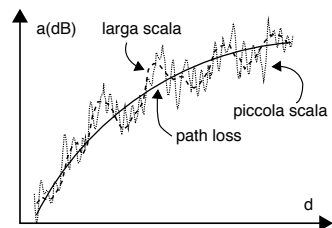


Figura 15.2: Determinazione del margine radiomobile



²⁹Mentre il fading produce una attenuazione variabile al segnale, la stessa variabilità delle condizioni di propagazione può portare a livelli interferenza variabili, causati da altre trasmissioni nella stessa banda. La variabilità temporale della qualità del segnale ricevuto, in particolare quella *veloce* (vedi § 15.3.4.6), produce errori a *burst*, che possono essere corretti mediante codifica di canale ed interleaving (vedi § 5.4.3.3).

³⁰Il termine *fading* si traduce come affievolimento o evanescenza.

15.3.4.2 Path loss

La dipendenza della attenuazione dal quadrato della distanza espressa dalla (15.10) si riferisce al caso ideale di spazio libero; misurazioni *reali* mostrano che invece l'esponente di d aumenta fino alla quarta potenza, a seconda del tipo di ambiente (urbano, rurale) e dell'altezza dell'antenna ricevente³¹. Pertanto, il termine $20 \log_{10} d (Km)$ che compare in (15.12) deve essere sostituito con $A_{pl} = n \cdot 10 \log_{10} d (Km) + \alpha$, e quindi l'espressione per l'attenuazione da usare in questo caso è

$$A_d (dB) = 32.4 + 20 \log_{10} f (\text{MHz}) + n \cdot 10 \log_{10} d (Km) + \alpha - G_T (dB) - G_R (dB)$$

in cui n ed α sono determinati in base a *campagne di misura*, e tengono conto delle condizioni operative. Il valore di n varia da 4 a 3 con $d < 100$ metri, all'aumentare dell'altezza dell'antenna fissa, mentre il termine α può variare da 7 a 15 dB con antenna fissa alta 30 e 10 metri rispettivamente, e subire un incremento di quasi 30 dB passando da un ambiente aperto ad un ambito urbano.

15.3.4.3 Fading su larga scala e shadowing

La stima delle grandezze n ed α ora introdotte è svolta *mediando* i risultati di diverse misure condotte nel territorio che si intende caratterizzare: infatti per territori diversi, si riscontrano valori di attenuazione complessiva diversi, anche per uguali valori di d . Anche questo fenomeno è a volte indicato come *slow fading* oltre che *su larga scala*, poiché non si presenta muovendosi di poco in una stessa zona, dipendendo dalla orografia del territorio e dalla natura degli oggetti limitrofi. Non conoscendo a priori in che zona ci si trovi, l'effetto del *fading su larga scala* (LS) si manifesta come una attenuazione supplementare a_s aleatoria, che risulta avere un andamento gaussiano in dB³² (per questo detto *lognormale*) e cioè del tipo

$$p_{A_s}(a_s(dB)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a_s(dB))^2}{2\sigma^2}}$$

³¹Per una rassegna dei diversi modelli di propagazione, si veda ad es. <http://www.slideshare.net/deepakecrbs/propagation-model>.

³²La d.d.p. gaussiana discende dall'ipotesi che uno dei cammini multipli pervenga al ricevitore con una potenza nettamente predominante rispetto agli altri. In questo caso l'involuppo complesso \underline{x} del segnale ricevuto è adeguatamente rappresentato da una v.a. di Rice (vedi pag. 164) $\underline{x} = a + \underline{r}$, in cui \underline{r} ha d.d.p. di Rayleigh e rappresenta l'effetto di molte cause indipendenti, relative ai cammini multipli, ed a è l'ampiezza della eco di segnale ricevuta con la maggiore ampiezza. Se $a \gg |\underline{r}|$ possiamo scrivere

$$\begin{aligned} a_s(dB) &= 10 \log_{10} \frac{1}{|a + \underline{r}|^2} = -10 \log_{10} \left((a + r_c)^2 + r_s^2 \right) = \\ &= -10 \log_{10} \frac{a^2}{a^2} \left(a^2 + 2ar_c + r_c^2 + r_s^2 \right) = \\ &= 10 \left(\log_{10} a^2 + \log_{10} \left(1 + \frac{2r_c}{a} + \frac{|\underline{r}|^2}{a^2} \right) \right) = \\ &\simeq 10 \left(\log_{10} a^2 + \frac{2r_c}{a} \right) = 10 \log_{10} a^2 + 20 \frac{r_c}{a} \end{aligned}$$

in quanto $\log(1 + \alpha) \simeq \alpha$ con $\alpha \ll 1$, e quindi $a_s(dB)$ ha media $10 \log_{10} a^2$ (compresa nel *path loss*) ed esibisce una d.d.p. gaussiana, la stessa di r_c .

dove σ_{ls}^2 varia tra 6 ed 8 dB per una altezza dell'antenna tra 5 e 15 metri³³. Per velocità del mobile non superiori ai 15 Km/h, si può assumere a_s costante in frequenza per qualche MHz, e nel tempo per poche centinaia di millisecondi.

Esempio Un collegamento radio tra antenne omnidirezionali poste a $d = 20$ KM e con portante $f_0 = 27$ MHz è usato da una trasmissione per cui occorre ricevere una potenza di almeno $W_R = -50$ dBm. Determinare la potenza W_T^{slib} che occorre trasmettere in condizioni di spazio libero, e la nuova potenza W_T^{sfad} necessaria a garantire una probabilità di fuori servizio pari al 5%, in presenza di un fading *su larga scala* caratterizzato da $\sigma_{ls}^2 = 8$ dB. Utilizziamo la (15.12) per calcolare

$$\begin{aligned} A_d \text{ (dB)} &= 32.4 + 20 \log_{10} f \text{ (MHz)} + 20 \log_{10} d \text{ (Km)} - G_T \text{ (dB)} - G_R \text{ (dB)} = \\ &= 32.4 + 20 \log_{10} 27 + 20 \log_{10} 20 = 32.4 + 28.6 + 26 = 87 \text{ dB} \end{aligned}$$

da cui si ottiene

$$W_T^{slib} \text{ (dBm)} = W_R \text{ (dBm)} + A_d \text{ (dB)} = -50 + 87 = 37 \text{ dBm}$$

pari a 7 dBW ovvero $10^{0.7} = 5$ Watt. Il fading su larga scala produce una attenuazione supplementare aleatoria con d.d.p. gaussiana in dB, e la probabilità di fuori servizio del 5% corrisponde al punto della curva di pag. 155 per cui $0.05 = \frac{1}{2} \operatorname{erfc} \left(\frac{a_s \text{ (dB)}}{\sqrt{2} \sigma_{SF}} \right)$, e quindi graficamente si ottiene $\frac{a_s \text{ (dB)}}{\sqrt{2} \sigma_{SF}} = 1.5$, da cui $a_s \text{ (dB)} = 1.5 \cdot \sqrt{2} \cdot \sqrt{8} = 1.5 \cdot 1.41 \cdot 2.82 = 6$ dB, che rappresenta il margine cercato, e che ci consente di calcolare la nuova W_T^{sfad} come $W_T^{sfad} \text{ (dBW)} = W_T^{slib} \text{ (dBW)} + 6 \text{ dB} = 13 \text{ dBW}$, ovvero $10^{1.3} = 20$ Watt.

15.3.4.4 Fading su piccola scala

Consiste nelle fluttuazioni di livello del segnale radio osservate durante *il movimento*, e causate dalle variazioni dei ritardi di fase con cui i cammini multipli giungono al ricevitore: spostandosi infatti di $\frac{\lambda}{2}$ ⁽³⁴⁾ si può passare da una situazione di somma coerente ad una completa opposizione di fase. Analizziamo ora la situazione in dettaglio, distinguendo ulteriormente diversi casi-tipo.

Fading piatto L'involuppo complesso del segnale ricevuto $\underline{y}(t)$ in presenza di cammini multipli può essere espresso in funzione di quello trasmesso $\underline{x}(t)$ sviluppando la (15.13) in modo da ottenere³⁵

$$\underline{y}(t) = \sum_{n=1}^N a_n \underline{x}(t - \tau_n) e^{-j2\pi f_0 \tau_n} \quad (15.14)$$

in cui τ_n è il ritardo dell' n -esimo cammino, ed a_n il rispettivo guadagno. Mostriamo ora che se durante il tempo che intercorre tra l'arrivo della prima replica (ritardata

³³All'aumentare dell'altezza dell'antenna, si estende l'area di copertura della stessa, ma in ambito urbano questo corrisponde ad una maggiore variabilità delle effettive condizioni operative.

³⁴A frequenza di 1 Ghz, si ha $\lambda \simeq 30$ cm. Questo fenomeno può essere facilmente sperimentato quando, durante una sosta al semaforo, si perde la sintonia di una radio FM, riacquistandola per piccoli spostamenti dell'auto; un altro esempio può essere la *ricerca del campo* per poter telefonare.

³⁵La (15.14) si ottiene considerando un generico segnale modulato in ampiezza $x(t) = a(t) \cos 2\pi f_0 t$ ed il suo involuppo complesso $\underline{x}(t) = a(t)$: per ognuna delle repliche

$$x_n(t) = x(t - \tau_n) = a(t - \tau_n) \cos 2\pi f_0 (t - \tau_n) = a(t - \tau_n) \cos (2\pi f_0 t - 2\pi f_0 \tau_n)$$

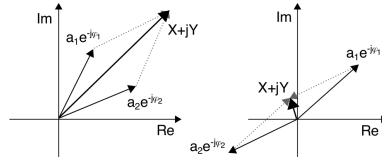
l'involuppo complesso rispetto ad f_0 può essere scritto come

$$\underline{x}_n(t) = a(t - \tau_n) e^{-j2\pi f_0 \tau_n} = \underline{x}(t - \tau_n) e^{-j2\pi f_0 \tau_n}$$

di τ_{min}) e l'arrivo dell'ultima (ritardata di τ_{max}) il segnale $\underline{x}(t)$ non varia di molto (e cioè $\underline{x}(t - \tau_{min}) \simeq \underline{x}(t - \tau_n) \simeq \underline{x}(t - \tau_{max})$)³⁶, l'effetto complessivo equivale ad una moltiplicazione di $\underline{x}(t)$ per un numero complesso, senza quindi produrre distorsione lineare (vedi pag. 361). Infatti, in tal caso la (15.14) può essere riscritta come

$$\begin{aligned} \underline{y}(t) &\simeq \underline{x}(t) \sum_{n=1}^N a_n e^{-j2\pi f_0 \tau_n} = \underline{x}(t) \sum_{n=1}^N a_n (\cos \varphi_n - j \sin \varphi_n) \\ &= \underline{x}(t) \cdot (X + jY) \end{aligned} \quad (15.15)$$

in cui il valore complesso $X + jY = \sum_{n=1}^N a_n \cos \varphi_n - j \sum_{n=1}^N a_n \sin \varphi_n$ riassume l'effetto delle diverse repliche, e rappresenta un v.a. gaussiana complessa, in quanto a partire da valori della portante f_0 dell'ordine dell'inverso di $\frac{1}{\tau_n}$, e tanto più per f_0 più elevate³⁷, bastano piccole variazioni di ritardo τ_n per produrre una fase $\varphi_n = 2\pi f_0 \tau_n$ (nota 35) uniformemente distribuita tra 0 e 2π e del tutto indipendente per le diverse repliche. D'altra parte, se anche i valori a_n sono realizzazioni di v.a. indipendenti ed equidistribuite, e se i cammini multipli sono in numero elevato, si applica il teorema centrale del limite (§ 7.2.4), e quindi i valori di X ed Y nella (15.15) possono considerarsi realizzazioni di v.a. indipendenti, gaussiane, a media nulla ed uguale varianza σ^2 .



Fading di Rayleigh Per ciò che riguarda l'ampiezza del segnale ricevuto, risulta $|\underline{y}(t)| = \rho \cdot |\underline{x}(t)|$, e nelle condizioni descritte $\rho = \sqrt{X^2 + Y^2}$ è una v.a. di RAYLEIGH (pag. 163), la cui d.d.p. ha espressione

$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

con $\rho \geq 0$; pertanto il valore della potenza ricevuta, legata a $|\underline{y}(t)|^2 = \rho^2 |\underline{x}(t)|^2$, risulta variata di una quantità pari a ρ^2 , che è una v.a. esponenziale negativa³⁸, descritta dalla d.d.p. (vedi § 8.2.1)

$$p_E(\rho^2) = \frac{1}{2\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \quad (15.16)$$

Questo risultato ci consente di determinare il margine M_{dB}^f necessario a contrastare un fading di Rayleigh, qualora si desideri una probabilità di fuori servizio pari ad α ³⁹:

$$M_{dB}^f = -10 \log_{10}(-\ln(1 - \alpha)) \quad (15.17)$$

³⁶Si consideri che il risultato dell'esempio di pag. 393 valuta i ritardi in gioco dell'ordine di grandezza delle decine di nanosecondi.

³⁷Se ad esempio i ritardi τ_n sono dell'ordine di 10^{-8} , l'ipotesi è valida per $f_0 > 100$ MHz.

³⁸Impostando il cambiamento di variabile $s = \rho^2$, si possono applicare le regole viste al § 7.4, individuando la funzione inversa come $\rho = \sqrt{s}$, la cui $\frac{d}{ds} \rho(s)$ fornisce $\frac{1}{2\sqrt{s}}$. Pertanto, la d.d.p. della nuova v.a. s vale:

$$p_S(s) = p_P(\sqrt{s}) \cdot \frac{d}{ds} \rho(s) = \frac{\sqrt{s}}{\sigma^2} \exp\left(-\frac{(\sqrt{s})^2}{2\sigma^2}\right) \cdot \frac{1}{2\sqrt{s}} = \frac{1}{2\sigma^2} \exp\left(-\frac{s}{2\sigma^2}\right)$$

³⁹Infatti, osserviamo che il collegamento va fuori servizio quando la potenza ricevuta è inferiore alla sensibilità del ricevitore $W_{R_{min}}$, e dunque (vedi eq. (8.2) a pag. 175) la probabilità di questo

La figura 15.3 mostra come varia in questo caso la potenza in dB del segnale ricevuto, relativamente alle condizioni di ricezione *medie* (ovvero su larga scala), per posizioni del ricevitore via via più distanti: si nota chiaramente come in alcuni luoghi questa possa diminuire anche di molto (caso dei *deep fade*).

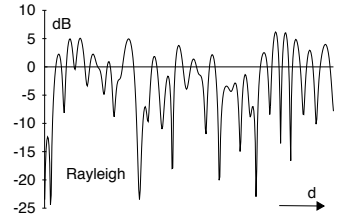
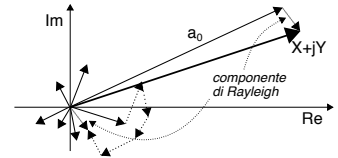
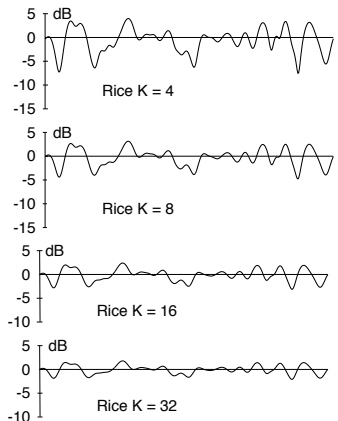


Figura 15.3: Intensità del segnale in presenza di fading di Rayleigh

Fading di Rice Si verifica se i guadagni a_n dei diversi percorsi che compaiono nella (15.14) non sono identicamente distribuiti, ma ne esiste uno (a_0 in figura) che *prevale* su tutti gli altri, come quando l'antenna trasmittente si trova in visibilità (anche parziale) del ricevitore.



In questo caso il canale produce un guadagno aleatorio ρ caratterizzato da una d.d.p. di Rice, espressa dalla eq. (7.11) a pag. 164, essendo la risultante $X + Y$ tipicamente ora *vicina* al cammino prevalente $a_0 e^{-j2\pi f_0 \tau_0}$. In particolare, il rapporto $K = \frac{a_0^2}{2\sigma^2}$ tra la potenza $\frac{a_0^2}{2}$ dell'onda diretta e quella σ^2 della componente dovuta al multipath prende il nome di *fattore di Rice*, ed alla figura a lato si mostra come, in presenza di una forte componente diretta, la profondità del fading si riduca sensibilmente.



Viceversa, qualora la ricezione avvenga principalmente in *assenza di visibilità*, i valori del modulo dell'involuppo complesso del segnale $\rho(t) = |\underline{x}(t)|$ sono soggetti al fading di Rayleigh precedentemente esposto.

15.3.4.5 Fading selettivo in frequenza

Rimuoviamo ora l'ipotesi fatta a pag. 397 che la velocità di variazione del segnale sia minore dell'intervallo temporale $\Delta\tau = \tau_{max} - \tau_{min}$ tra l'arrivo della prima e dell'ultima replica, detto anche *dispersione temporale*. Per analizzare cosa succede⁴⁰, partiamo dalla (15.14) per scrivere l'espressione dell'involuppo complesso della risposta impulsiva del canale come

$$\underline{h}(t) = \sum_{n=1}^N a_n \delta(t - \tau_n) e^{-j2\pi f \tau_n} \tag{15.18}$$

e facciamo l'ipotesi semplificatrice che i diversi ritardi τ_n siano multipli di un comune intervallo T , cioè $\tau_n = nT$, considerando eventualmente nullo qualche valore a_n .

evento si esprime come $\alpha = Pr(\rho^2 < W_{R_{min}}) = 1 - \exp\left(-\frac{W_{R_{min}}}{m \rho^2}\right)$, dato che il valor medio della v.a. esponenziale ρ^2 è $m \rho^2 = 2\sigma^2$. Esprimendo ora il margine M come il rapporto tra la potenza media ricevuta e la sensibilità del ricevitore $M = \frac{m \rho^2}{W_{R_{min}}}$, si ottiene $\alpha = 1 - e^{-\frac{1}{M}}$, e dunque $-\frac{1}{40} = \ln(1 - \alpha)$ e, passando ai decibel, $-10 \log_{10} M = 10 \log_{10} (-\ln(1 - \alpha))$, da cui la 15.17.

⁴⁰ Questa e la prossima analisi sono tratte da http://www.dii.unisi.it/~abrardo/comunicazioni_radiomobili.pdf

Introduciamo quindi la notazione $Z_n = a_n e^{-j2\pi f_0 \tau_n}$, in modo che la (15.18) possa essere assimilata all'espressione di un segnale campionato $\underline{h}(t) = \sum_{n=1}^N Z_n \delta(t - \tau_n)$, interpretando dunque i coefficienti complessi Z_n come campioni di un processo $Z(t)$ ⁴¹, ovvero $Z_n = Z(nT) = a_n e^{-j2\pi f_0 nT}$. Ciò consente di porre la risposta in frequenza dell'involuppo complesso del canale nella forma

$$\underline{H}(f) = \sum_{n=1}^N Z_n e^{-j2\pi f_0 nT}$$

Notiamo ora che i valori di $\underline{H}(f)$ in funzione di f sono variabili aleatorie, in quanto dipendono dalle caratteristiche statistiche dei termini Z_n , che per i motivi illustrati a pag. 398 nel caso di fading di Rayleigh sono da considerare v.a. gaussiane complesse, indipendenti ed a valor medio nullo, e quindi

$$E\{Z_n^* Z_{n+m}\} = \begin{cases} 0 & \text{se } m \neq 0 \\ \sigma_{a_n}^2 & \text{altrimenti} \end{cases} \quad (15.19)$$

in cui $\sigma_{a_n}^2 = E\{a_n^2\}$ è detto *profilo di intensità del multipath* e rappresenta la distribuzione temporale di energia delle repliche del segnale. Quel che ora ci interessa capire, è per quale intervallo Δf si ottengono valori di risposta in frequenza ($\underline{H}(f)$, $\underline{H}(f + \Delta f)$) che iniziano a divenire *incorrelati*, dato che un segnale che occupa una banda comparabile a Δf in tal caso è soggetto a distorsione lineare. A tal fine, calcoliamo la funzione di autocorrelazione *in frequenza*:

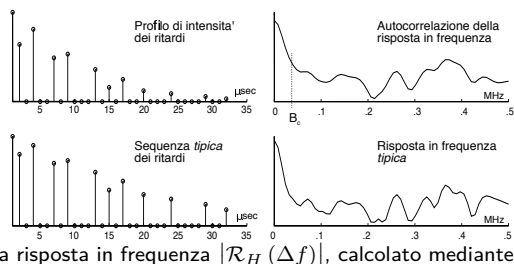
$$\begin{aligned} \mathcal{R}_{\underline{H}}(\Delta f) &= E\{\underline{H}^*(f) \underline{H}(f + \Delta f)\} = \sum_{n=1}^N \sum_{m=1}^N E\{Z_n^* Z_m\} e^{j2\pi f nT} e^{-j2\pi(f + \Delta f)mT} \\ &= \sum_{n=1}^N \sigma_{a_n}^2 e^{-j2\pi \Delta f nT} \end{aligned}$$

in cui all'ultimo passaggio si è applicata la (15.19): $\mathcal{R}_{\underline{H}}(\Delta f)$ è dunque pari alla trasformata di Fourier del *profilo di intensità* $\sigma_{a_n}^2$, che è misurabile.

Banda di coerenza E' definita come l'intervallo di frequenza B_c entro il quale $\underline{H}(f)$ si mantiene correlata, che può essere fatto corrispondere con la *larghezza di banda* di $\mathcal{R}_{\underline{H}}(\Delta f)$. Pertanto, quanto più la dispersione temporale $\Delta\tau$ risulta elevata, tanto minore sarà il valore di B_c ; come misura approssimata⁴², si può convenzionalmente porre

$$B_c = \frac{1}{\Delta\tau} \quad (15.20)$$

Esempio Consideriamo un canale caratterizzato da un profilo di intensità dei ritardi $\sigma_{a_n}^2$ i cui valori sono mostrati in figura, ed a cui corrisponde un delay spread $\Delta\tau$ pari a 32 μsec . A destra di $\sigma_{a_n}^2$ è mostrato il modulo della autocorrelazione della risposta in frequenza $|\mathcal{R}_{\underline{H}}(\Delta f)|$, calcolato mediante



⁴¹Si sottintende pertanto che T sia minore di metà della banda di $Z(t)$.

⁴²Una approssimazione migliore valuta $\Delta\tau$ come la deviazione standard dei ritardi, vedi http://en.wikipedia.org/wiki/Delay_spread.

una FFT; considerando per esso una larghezza di banda pari alla frequenza per cui il suo valore si dimezza, osserviamo che ciò avviene per $\Delta f \simeq 30$ KHz, che è quindi il valore della banda di coerenza B_c , in buon accordo con la (15.20), che fornisce $B_c = 1/32 \cdot 10^{-6} = 31,25$ KHz. Nella seconda riga della figura è quindi mostrata una sequenza tipica dei ritardi, ognuno posto pari alla deviazione standard σ_{a_n} , da cui sempre tramite FFT si ottiene la risposta in frequenza $|\underline{H}(f)|$, mostrata in basso a destra.

Dimensione di cella e velocità trasmissiva Ricapitolando, se la banda W del segnale risulta abbastanza minore di B_c siamo nelle condizioni di fading *piatto*, mentre se $W \gtrsim B_c$ le componenti spettrali di $x(t)$ subiscono alterazioni statisticamente indipendenti, i cammini multipli causano un effetto filtrante, si manifesta ISI, ed il canale corrispondente viene detto *selettivo in frequenza*. Approssimando l'occupazione di banda di un segnale numerico modulato come il reciproco del periodo di simbolo $W \simeq \frac{1}{T_s}$, osserviamo che la condizione di fading piatto $W < B_c$ implica anche che $T_s \simeq \frac{1}{W} > \frac{1}{B_c} = \Delta\tau$, ossia che la dispersione temporale è inferiore al periodo di simbolo, limitando gli effetti dell'ISI. Dato che la correzione degli effetti di distorsione lineare e ISI richiede al ricevitore complesse operazioni di equalizzazione (§ 14.6), si tenta di operare per quanto possibile in condizioni di fading piatto, occupare una banda $W < B_c$, e limitare di conseguenza la velocità di segnalazione f_s .

Per celle molto grandi, la differenza di percorso tra cammini multipli può essere notevole, determinando una banda di coerenza ridotta, e quindi una bassa velocità di trasmissione. Riducendo la dimensione di cella, è possibile invece aumentare la velocità, dato che le differenze di ritardo si riducono. Per questo motivo, se celle con raggio di chilometri e dispersione di ritardo di oltre 10 μsec necessitano di equalizzazione anche per trasmissioni a 64 kbps, al contrario comunicazioni *indoor* con dispersione di ritardo inferiore ad 1 μsec possono presentare *flat fading* per velocità superiori al Mbps. Celle di dimensione minima, dette anche *picocelle*, presentano una dispersione temporale di solo qualche decina di picosecondi, permettendo di operare a molti Mbps anche senza equalizzazione.

15.3.4.6 Dispersione spettrale e variabilità temporale

Finché il ricevitore e gli oggetti riflettenti sono fermi, la distribuzione dei ritardi τ_n non varia nel tempo, e la componente di attenuazione supplementare su piccola scala mantiene uno stesso (casuale) valore; in tal caso pertanto il fading non è né lento né veloce, ma *costante*. Viceversa, nel caso in cui ci sia movimento⁴³ l'inviluppo complesso ricevuto (15.14) si riscrive come

$$\underline{y}(t) = \sum_{n=1}^N a_n(t) \underline{x}(t - \tau_n(t)) e^{-j2\pi f_0 \tau_n(t)}$$

evidenziando come ora sia le ampiezze a_n che i ritardi τ_n dipendono dal tempo.

L'analisi delle conseguenze di questa *non stazionarietà* inizia considerando una portante non modulata $x(t) = \cos 2\pi f_0 t$, con inviluppo complesso $\underline{x}(t) = 1$ ⁴⁴, e che

⁴³Del ricevitore, del trasmettitore, o degli oggetti riflettenti.

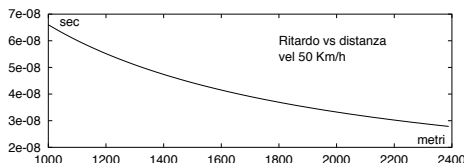
⁴⁴Dall'espressione $x(t) = \cos 2\pi f_0 t = \frac{e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}}{2}$, si riconosce che $x^+(t) = \frac{1}{2} e^{j2\pi f_0 t}$ dato che $X(f) = \frac{1}{2} [\delta(f - f_0) + \delta(f + f_0)]$; ma dato che $\underline{x}(t) = 2x^+(t) e^{-j2\pi f_0 t}$, si ottiene appunto $\underline{x}(t) = 2 \frac{1}{2} e^{j2\pi f_0 t} e^{-j2\pi f_0 t} = 1$.

produce la ricezione di

$$\underline{y}(t) = \sum_{n=1}^N a_n(t) e^{-j2\pi f_0 \tau_n(t)} = \sum_{n=1}^N a_n(t) e^{-j\varphi_n(t)} \quad (15.21)$$

Si ottengono quindi N diversi segnali *modulati* sia in ampiezza che angolarmente, anche se è stata trasmessa una sola frequenza. In generale le ampiezze $a_n(t)$ non variano di molto con il movimento, mentre come già osservato, sono sufficienti piccole variazioni di $\tau_n(t)$ per causarne grandi di $\varphi_n(t) = 2\pi f_0 \tau_n(t)$: ad esempio, con una $f_0 = 1$ GHz, basta una variazione di τ pari ad 1 nsec per produrre una rotazione di 2π .

Esempio Riprendiamo i dati ed il modello usati a pag. 393 per ottenere il risultato in figura, relativo ad un mobile che viaggia a 50 Km/h, e che in 100 sec percorre 1.4 Km a partire da una distanza di 1 Km dal trasmettitore, in presenza di una superficie riflettente posta a 100 metri da metà percorso. Il ritardo del cammino riflesso varia da 66 a 27 nsec, con la legge mostrata in figura, dovuta al variare nel tempo dell'angolo di riflessione.



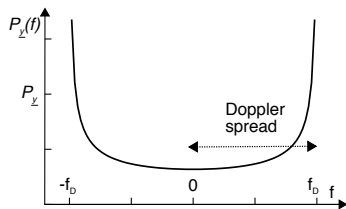
L'effetto del movimento sui diversi percorsi può essere descritto considerando che per ognuno di essi si verifica l'effetto Doppler⁴⁵, che consiste nella ricezione di una frequenza $f_n = f_0 \pm f_d^n$, ossia aumentata (o diminuita) rispetto alla portante f_0 di una quantità detta *frequenza Doppler*

$$f_d^n = f_0 \frac{v}{c} \cos \vartheta_n$$

in cui c è la velocità di propagazione elettromagnetica (pari a quella della luce $3 \cdot 10^8$ m/s), v è la velocità del mobile⁴⁶ e ϑ_n è l'angolo tra la direzione del moto e la posizione del riflettore: con i dati dell'esempio precedente, ponendo $f_0 = 1$ GHz si ottiene una f_d^n massima di 46.3 Hz, relativa al caso di $\vartheta_n = 0$ ⁴⁷; indichiamo con $f_D = \max_n \{f_d^n\}$ tale valore.

Dispersione Doppler Dato che ogni diverso percorso è caratterizzato da una f_d^n compresa tra zero e f_D , il segnale ricevuto contiene frequenze che si discostano da f_0 in più o in meno, entro una deviazione massima pari ad f_D , per questo indicata come *dispersione* (o *spread*) Doppler, ed il canale è detto *dispersivo in frequenza*. Considerando il mobile raggiunto da infiniti percorsi con direzione di arrivo distribuita uniformemente (condizione di *scattering isotropo*), si può arrivare a mostrare che la densità spettrale ricevuta assume una espressione del tipo

$$\mathcal{P}_y(f) = \begin{cases} \frac{\mathcal{P}_y}{\pi f_D \sqrt{1-(f/f_D)^2}} & \text{con } |f| \leq f_D \\ 0 & \text{altrove} \end{cases} \quad (15.22)$$



⁴⁵Lo stesso che produce la variazione del suono di una sirena, vedi http://it.wikipedia.org/wiki/Effetto_Doppler

⁴⁶La stessa analisi è valida anche nel caso di un ricevitore *fermo* ma con i riflettori in movimento, come per la *riflessione ionosferica*: in tal caso l'espressione si scrive come $f_d^n = f_0 \frac{v_n}{c} \cos \vartheta_n$, considerando cioè la possibilità che i riflettori abbiano velocità diverse tra loro.

⁴⁷Notiamo che se $\vartheta_n = 0$ non stiamo trattando di un raggio riflesso, ma di quello diretto.

del tutto simile a quella mostrata a pag. 160. La dispersione Doppler f_D costituisce nella pratica una misura della *velocità di variazione* del canale⁴⁸. Consideriamo infatti che $\underline{y}(t)$ descritto dalla (15.21) è il risultato della somma vettoriale nel piano complesso dei termini $a_n(t) e^{-j2\pi f_0 \tau_n(t)}$, che in virtù dei diversi scostamenti Doppler sono ognuno in rotazione⁴⁹ ad una diversa velocità angolare $2\pi f_d^n$, tanto maggiore quanto più è grande f_D , che quindi determina la rapidità con cui il risultato $\underline{y}(t)$ varia nel tempo.

Tempo di coerenza L'antitrasformata di $\mathcal{P}_y(f)$ è per definizione l'autocorrelazione di $\underline{y}(t)$, e per la 15.22 si ottiene $\mathcal{R}_{yy}(\tau) = J_0(2\pi f_D \tau)$ in cui J_0 è la funzione di Bessel del primo tipo di ordine zero graficata a pag. 286, ed il cui primo passaggio per zero avviene con $\tau \simeq \frac{0.4}{f_D}$, che corrisponde al minimo intervallo di tempo necessario ad osservare valori di $\underline{y}(t)$ *incorrelati*; viceversa, un intervallo sufficientemente più piccolo trova il canale in condizioni pressoché immutate, come ad esempio

$$T_c = \frac{0.1}{f_D}$$

che viene indicato come *tempo di coerenza*. Una trasmissione con periodo di simbolo $T_s \gtrsim T_c$ subisce molto probabilmente condizioni del canale differenti nell'arco di tempo di un simbolo, ostacolando la sincronizzazione di fase⁵⁰. Se invece $T_s \ll T_c$ il canale si mantiene in condizioni pressoché stazionarie per tutto il periodo di simbolo, il fading è lento (*slow*), ed il movimento non produce conseguenze sensibili. Utilizzando di nuovo i dati dell'ultimo esempio, ad un doppler spread $f_D = 46.3$ Hz corrisponde un tempo di coerenza $T_c = 21.6$ msec.

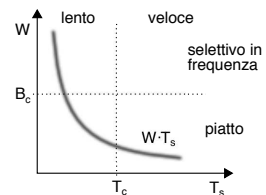
Considerazioni conclusive Notiamo che il verificarsi contemporaneo della assenza di distorsione lineare in quanto $W \ll B_c$, e della stazionarietà del canale in quanto $T_s \ll T_c$, ovvero di *fading lento e piatto* (*slow and flat*), equivale in pratica al verificarsi delle condizioni di canale perfetto. Perché ciò possa accadere è necessario che si verifichi⁵¹

$$B_d \cdot \Delta\tau \ll 1$$

detta condizione di *sottodispersione (underspread)*. Nella pratica, i valori di B_d e $\Delta\tau$ per i canali in uso nelle telecomunicazioni soddisfano tale condizione; la figura a lato suddivide il piano⁵² T_s, W nelle regioni per le quali si verificano i diversi tipi di fading, in funzione dai valori B_c e T_c .

$$\begin{cases} W \ll B_c = \frac{1}{\Delta\tau} & \text{no dist lin} \\ T_s \ll T_c = \frac{1}{B_d} & \text{stazionario} \end{cases}$$

Condizioni di canale perfetto



Tipo di fading in relazione al canale ed al segnale

⁴⁸In questo modo si ottiene una trattazione unificata sia per il caso di un singolo mobile in un contesto statico, sia per quello di un ricevitore fermo con riflettori in movimento. In entrambi i casi il *Doppler spread* f_D può essere effettivamente *misurato* al ricevitore, in presenza di una portante non modulata.

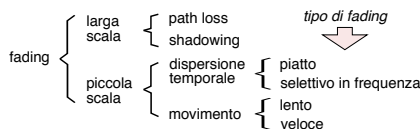
⁴⁹Si veda l'interessante animazione java presso <http://www.wirelesscommunication.nl/reference/chaptr03/rayjav>

⁵⁰Ciò avviene perché in pratica è come se due simboli consecutivi pervenissero attraverso due differenti canali, e dunque non è possibile eseguire operazioni di media.

⁵¹Infatti le due condizioni possono essere riscritte come $B_d T_s \ll 1$ e $W \Delta\tau \ll 1$; moltiplicandole tra loro si ha $B_b T_s W \Delta\tau \ll 1$; ma se poniamo $W \simeq \frac{1}{T_s}$ si ottiene la condizione indicata.

⁵²Il ramo di iperbole che individua il luogo dei punti $W \cdot T_s$ è tracciato in base all'osservazione della nota precedente, ossia $W \simeq \frac{1}{T_s}$, da cui $W \cdot T_s \simeq 1$.

Infine, la figura a lato rappresenta in forma gerarchica i fenomeni di attenuazione supplementare fin qui discussi.



15.4 Collegamenti in fibra ottica

Una fibra ottica è realizzata in vetro o silicio fuso, ovvero qualunque materiale dielettrico trasparente alla luce, tanto che può essere realizzata anche in plastica. Il suo utilizzo è quello di trasportare energia luminosa in modo guidato. Una caratteristica che deriva direttamente dalla sua natura, è l'immunità della fibra ottica ai disturbi di natura elettromagnetica; tale proprietà impedisce fenomeni di interferenza (diafonia), così come non permette di prelevare segnale dall'esterno (intercettazione).

Il segnale luminoso Le lunghezze d'onda delle radiazioni elettromagnetiche nel campo del visibile sono comprese tra circa 100 μm dell'infrarosso (1 nm=10⁻⁹ metri) e 50 nm dell'ultravioletto, che corrispondono a frequenze (ricordando ancora che $f = \frac{c}{\lambda}$) che vanno da 3 · 10¹² fino a 6 · 10¹⁵ Hz. Questi valori individuano una banda

Infrarosso	→	Ultravioletto	λ [metri] f [Hz]
10 ⁻⁴	→	50 · 10 ⁻⁹	
3 · 10 ¹²	→	6 · 10 ¹⁵	

passante veramente notevole se comparata ad altri mezzi trasmissivi: supponiamo infatti di effettuare una modulazione che occupi una banda pari allo

0.1% della frequenza portante. Se $f_0 = 1$ GHz, si ha 1 MHz di banda; ma se $f_0 = 10^{13}$, si ha una banda di 10 GHz!

15.4.1 Trasmissione ottica

Anche se sono teoricamente possibili schemi di modulazione analogici, le fibre ottiche sono usate prevalentemente per trasportare informazione di natura *numerica* secondo lo schema di fig. 15.4, in cui la luce emessa da una sorgente è accesa o spenta (ovvero modulata in ampiezza con uno schema ON/OFF). All'altro estremo della fibra, un fotorivelatore effettua una rivelazione incoerente dell'energia luminosa, che viene nuovamente convertita in un segnale elettrico. Le prime fibre ottiche risalgono al 1970, e fornivano attenuazioni dell'ordine di 20 dB/Km. Attualmente si sono raggiunti valori di attenuazione di 0.2 dB/Km, pari ad un quarto di quella dei migliori cavi coassiali. D'altra parte, a differenza del rame, il materiale utilizzato per le fibre (vetro o silicio) è largamente disponibile in natura. Inoltre, a parità di diametro, una fibra ottica trasporta un numero anche 1000 volte maggiore di comunicazioni rispetto ad un cavo coassiale, fornendo quindi anche un risparmio di spazio.

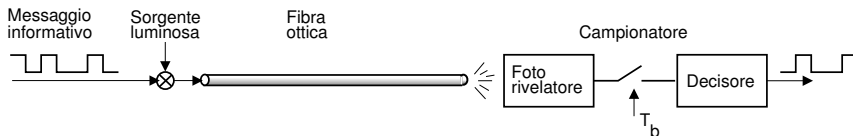


Figura 15.4: Schema di trasmissione in fibra ottica

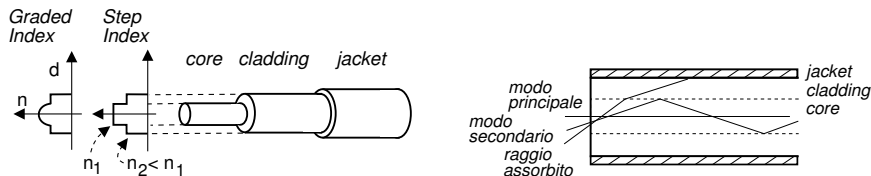
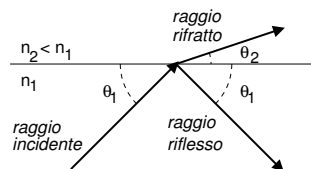


Figura 15.5: Tecnologia della fibra ottica e modi di propagazione

Propagazione luminosa e indice di rifrazione Lo spazio libero è il mezzo di propagazione in cui la luce viaggia più velocemente. Il rapporto tra $c = 3 \cdot 10^8$ m/sec, e la velocità di propagazione v in un mezzo trasparente, è l'*indice di rifrazione*⁵³ n del mezzo stesso: $n = c/v$, risultando $n \geq 1$. Ad esempio, se $n = 2$ allora la velocità è la metà.

Quando un raggio luminoso incontra un mezzo con diverso indice n (ad esempio, da n_1 ad $n_2 < n_1$) una parte di energia si riflette con angolo θ_1 uguale a quello incidente, e la restante parte continua nell'altro mezzo, ma con diverso angolo $\theta_2 < \theta_1$. La relazione tra gli angoli è nota come *legge di Snell*

$$\frac{\cos \theta_1}{\cos \theta_2} = \frac{n_2}{n_1}$$



e dunque il raggio rifratto è più inclinato nel mezzo con n inferiore (dove viaggia più veloce). Esiste un valore $\theta_c = \arccos \frac{n_2}{n_1}$ sotto il quale non si ha rifrazione, ma tutto il raggio viene riflesso ed è proprio su questo fenomeno che si basa l'attitudine delle fibre ottiche di trasportare energia luminosa. La fibra ottica è infatti costituita da un nucleo (*core*) centrale con indice di rifrazione n_1 , circondato da un rivestimento (*cladding*) con indice $n_2 < n_1$; entrambi racchiusi in una guaina (*jacket*) di materiale opaco (vedi fig. 15.5).

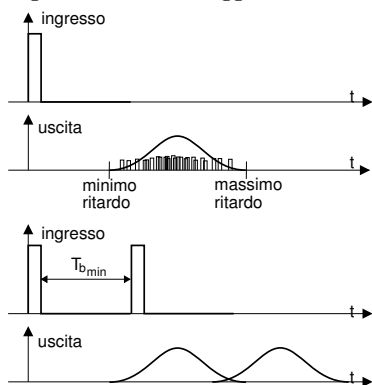
Quando una sorgente luminosa è posta davanti alla fibra, l'energia si propaga mediante diversi *modi di propagazione*, definiti nel contesto della meccanica quantistica, e identificabili in chiave di ottica geometrica come i diversi angoli di incidenza. Il *modo principale* è quello che si propaga lungo l'asse rettilineo, mentre i *modi secondari* sono quelli con angolo $< \theta_c$, che si riflettono completamente al confine tra core e cladding. I modi associati ad angoli più elevati di θ_c vengono progressivamente assorbiti dalla guaina, e dunque non si propagano.

Il valore $\Delta = \sqrt{n_1^2 - n_2^2} = n_1 \sin \theta_c$ prende il nome di *apertura numerica*, e permette di risalire al massimo angolo di incidenza mediante la relazione $\theta_c = \arcsin \frac{\Delta}{n_1}$. Come si vede, Δ è tanto più piccolo quanto più n_1 ed n_2 sono simili; al diminuire di Δ , si riduce anche la potenza luminosa che viene immessa nella fibra ottica, ma si ottiene il beneficio illustrato di seguito.

⁵³Vedi http://it.wikipedia.org/wiki/Indice_di_rifrazione, ma anche il video <https://www.youtube.com/watch?v=k7ohfaMmTKg>.

Dispersione modale Questo fenomeno è dovuto al fatto che i modi di propagazione relativi agli angoli di incidenza più elevati percorrono di fatto *più strada*, e dunque impiegano più tempo per giungere a destinazione.

Pertanto, ogni singolo impulso luminoso presente in ingresso produce in uscita più impulsi distanziati nel tempo, uno per ogni modo di propagazione. Dato che inoltre avviene un continuo scambio di energia tra i diversi modi, si ottiene che l'uscita sarà un segnale con una maggiore estensione temporale (vedi figura seguente). L'entità della



dispersione temporale (differenza tra ritardo max e min) sarà tanto maggiore quanto più il collegamento è lungo, e quanti più modi partecipano alla propagazione. L'effetto più appariscente del fenomeno descritto consiste nella limitazione della massima frequenza con cui gli impulsi luminosi possono essere posti in ingresso alla fibra; impulsi troppo vicini risulterebbero infatti affetti da interferenza intersimbolica (ISI) ed in pratica sarebbero indistinguibili in uscita. Pertanto la massima frequenza di segnalazione in una fibra ottica, dipende dalla *lunghezza* della fibra stessa.

Si chiamano *fibre multimodo* le fibre ottiche in cui sono presenti più modi di propagazione. Queste sono del tipo STEP INDEX se n cambia in modo brusco, o GRADED INDEX se il *core* ha un indice graduato. Nel secondo caso la dispersione temporale è ridotta; infatti quando i modi secondari attraversano la sezione periferica del core, incontrano un indice di rifrazione n ridotto, e quindi viaggiano più veloci. Una diversa (e drastica) soluzione al problema della dispersione temporale, è fornita dalle fibre ottiche *monomodo*: queste sono realizzate con un core di diametro così piccolo⁵⁴, da consentire alla sorgente luminosa di immettere luce nella fibra solo con angolo di incidenza nullo, e quindi di permettere la propagazione del solo modo principale.

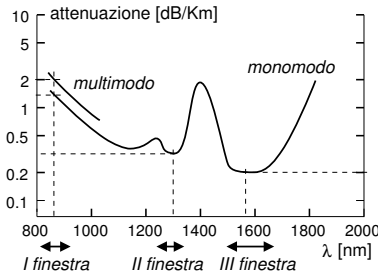
Ovviamente le ultime due soluzioni (graded index e fibra monomodo) si sono rese possibili grazie ai progressi nei processi di fabbricazione.

Attenuazione In modo simile ai cavi elettrici, anche le fibre ottiche sono mezzi dissipativi, in quanto parte dell'energia in transito viene assorbita dalla fibra stessa e trasformata in calore. I fenomeni di assorbimento sono legati alla presenza di impurità chimiche, che possono ridurre la trasparenza oppure avere dimensioni (a livello molecolare) comparabili con le lunghezze d'onda in gioco.

Per questi motivi, la caratteristica di attenuazione chilometrica ha un andamento (vedi fig. 15.6) fortemente dipendente da λ , e sono stati individuati 3 intervalli di lunghezze d'onda (detti *finestre*) per i quali l'assorbimento è ridotto, ed in cui sono effettuate le trasmissioni ottiche.

La prima finestra (con attenuazione maggiore) è stata l'unica disponibile agli inizi, a causa dell'assenza di trasduttori affidabili a frequenze inferiori, ed è tuttora usata per collegamenti economici e scarsamente critici. La seconda finestra ha iniziato ad essere usata assieme alle fibre monomodo, grazie all'evoluzione tecnologica dei trasduttori, mentre l'uso della III finestra si è reso possibile dopo essere riusciti a limitare la *dispersione cromatica* delle fibre (vedi appresso).

⁵⁴Si passa dai 50 μm per le fibre multimodo, a circa 8 μm nel caso monomodo.



Finestra	λ [μm]	A_d [dB/Km]
I	0.8 ÷ 0.9	1.2 (monomodo)
		2 (multimodo)
II	1.2 ÷ 1.3	0.35
III	1.5 ÷ 1.7	0.2

Figura 15.6: Dipendenza della attenuazione chilometrica dalla lunghezza d'onda

Un'altra fonte di attenuazione può avere origine dalle *giunzioni* tra tratte di fibre ottiche: l'uso di connettori produce una perdita di $0.4 \div 1$ dB, ed i giunti meccanici $\simeq 0.2$ dB oppure anche 0.05 dB se ottimizzati per via strumentale. Si possono infine *fondere* tra loro le fibre, con perdite tra 0.01 e 0.1 dB.

15.4.2 Bilancio di collegamento

Dispersione cromatica Dopo aver ridotto od eliminato il fenomeno di dispersione modale, è emersa una ulteriore causa di dispersione temporale dell'energia immessa nella fibra ottica: il problema si verifica se il segnale di ingresso non è perfettamente monocromatico, ovvero sono presenti diverse lunghezze d'onda. Dato che il valore dell'indice di rifrazione dipende dalla lunghezza d'onda, λ diverse si propagano con velocità differenti e raggiungono l'altro estremo della fibra in tempi successivi. La dispersione cromatica *nominale* D_0 della fibra si misura in $[\frac{psec}{Km \cdot nm}]$, e dà luogo ad una effettiva dispersione temporale $D = D_0 \cdot L \cdot \Delta\lambda$, che è direttamente proporzionale alla lunghezza L della fibra ed alla estensione della *gamma cromatica* $\Delta\lambda$ della sorgente⁵⁵. Per ridurre il fenomeno è possibile:

- utilizzare una lunghezza d'onda λ per la quale la dispersione cromatica è ridotta. Ad esempio, una fibra di silicio *normale* produce una dispersione cromatica 15 volte inferiore a $1.3 \mu\text{m}$ che non a $1.5 \mu\text{m}$;
- scegliere una sorgente con la minima estensione cromatica $\Delta\lambda$ possibile.

Trasduttori elettro-ottici Quelli usati per primi sono stati gli economici LED (*Light Emitting Diode*), che richiedono una circuiteria di interfaccia semplice, sono poco sensibili alle condizioni ambientali, e quindi risultano affidabili. D'altra parte, i LED emettono luce su più lunghezze d'onda, mentre per limitare la dispersione cromatica (e quindi raggiungere frequenze di segnalazione più elevate) occorre ricorrere ai *Diodi Laser* (LD).

I LASER forniscono anche una maggiore potenza, e quindi divengono indispensabili per coprire distanze maggiori⁵⁶. D'altra parte sono più costosi, hanno vita media

⁵⁵Il fenomeno della dispersione cromatica è l'equivalente ottico della distorsione di fase (o distorsione di ritardo) introdotta al § 14.5 per i segnali elettrici.

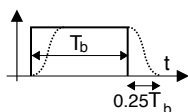
⁵⁶La potenza emessa da un LASER non può aumentare a piacimento: oltre un certo valore intervengono infatti fenomeni *non lineari*, e la luce non è più monocromatica, causando pertanto un aumento della dispersione cromatica.

Sorgente	λ (nm)	W_{dT} (dBm)	$\Delta\lambda$ (nm)
Si LED	850	-16	50
Ge LED	1300	-19	70
InGaAsP LED	1300	-10	120
DFB LASER	1300	-5	1
DFB LASER	1550	-5	0.4
IL/DFB LASER	1550	+2	0.8

ridotta rispetto ai LED, e richiedono condizioni di lavoro più controllate. Notiamo inoltre che una fibra ottica posta inizialmente in opera mediante sorgenti LED, può essere potenziata (in termini di banda) sostituendo il LED con il LASER.

L'uso di sorgenti che operano in III finestra, che (presentando una attenuazione ridotta) permette di operare con tratte più lunghe, obbligherebbe però a ridurre la frequenza di segnalazione, a causa della maggiore dispersione cromatica. Ma questa limitazione è stata superata da un particolare tipo di fibra, detta *dispersion shifted*, che presenta un minimo di dispersione cromatica nominale D_0 in III finestra anziché in II, e che raggiunge valori migliori di 3.5 psec/Km.nm.

Prodotto banda-lunghezza e codici di linea Come anticipato, la dispersione cromatica D risulta proporzionale alla lunghezza del collegamento L ed all'estensione cromatica $\Delta\lambda$ della sorgente. Se pensiamo di effettuare una trasmissione con codici



NRZ e periodo $T_b = \frac{1}{f_b}$, ed imponiamo che la dispersione temporale sia non maggiore di $\frac{1}{4}T_b$, si ottiene

$$D_0 \cdot L \cdot \Delta\lambda \leq 0.25 \cdot T_b \quad (15.23)$$

in cui D_0 è la dispersione cromatica *nominale* [psec/Km.nm], L è la lunghezza [Km], $\Delta\lambda$ è l'estensione cromatica della sorgente [nm], e T_b è la durata di un bit [psec]. Associando ora il concetto di *banda* B alla frequenza di segnalazione $f_b = \frac{1}{T_b}$, la relazione (15.23) può essere riscritta in modo da evidenziare il *prodotto della banda per la lunghezza PBL*, che è pari al valore

$$PBL_{NRZ} = f_b \cdot L = \frac{.25}{D_0 \cdot \Delta\lambda} \quad [Tbps \cdot Km] \quad (15.24)$$

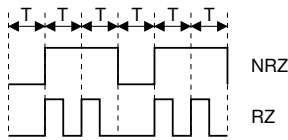
che è una grandezza dipendente dalla coppia fibra-sorgente⁵⁷, e che rappresenta la relazione tra f_b ed L necessaria ad ottenere $D = \frac{1}{4}T_b$. Inserendo dunque i valori di $\Delta\lambda$ (della sorgente) e D_0 (della fibra) nella (15.24), si ottiene una *costante* da usare per calcolare la banda (frequenza) massima trasmissibile per una data lunghezza (o viceversa). Qualora si usi un codice RZ, i cui simboli hanno durata metà del periodo di bit T_b , la dispersione temporale tollerabile può essere elevata al 50% di T_b , e quindi in questo caso il prodotto banda-lunghezza risulta doppio⁵⁸ rispetto al caso precedente:

$$PBL_{RZ} = \frac{.5}{D \cdot \Delta\lambda} = 2 \cdot PBL_{NRZ}$$

Qui sotto è mostrato il confronto tra i codici RZ e NRZ, e la tabella dei valori di *PBL* associati alle coppie fibra-sorgente indicate.

⁵⁷In questo senso, il prodotto *banda-lunghezza* costituisce un parametro di sistema che tiene conto di un concorso di cause. Un pò come il concetto di *tenuta di strada* di una autovettura, che dipende da svariati fattori, come il peso, i pneumatici, la trazione, il fondo stradale....

⁵⁸Tuttavia, il dimezzamento della durata di un bit causa una perdita di potenza di 3 dB, in base alle considerazioni riportate a pag. 410.



Fibra	Sorgente	λ [nm]	PBL [Gbps-Km]
Multimodo	LED	820	0.013
Graded Index	LED	820	2
Monomodo	LASER	1300	250
Monomodo	LASER	1550	36.8
Monomodo DS	LASER	1550	179

Esercizio Determinare la lunghezza massima di un collegamento in fibra ottica monomodo, operante con $\lambda=1.3 \mu\text{m}$, e che garantisca una velocità $f_b=417$ Mbps, assumendo un guadagno di sistema di 42 dB (ovvero disponendo di una potenza di trasmissione 42 dB maggiore della minima potenza necessaria in ricezione).

Soluzione Dal grafico di fig. 15.6 si trova che per $\lambda=1300$ nm, l'attenuazione chilometrica è di 0,35 dB/Km, che determina una $A_d = 0,35 \cdot L_{Km}$ dB. Imponendo ora $G_s \geq A_d$ si ottiene una lunghezza non superiore a $\frac{42}{0,35}=120$ Km, che identifica il *Limite di Attenuazione* del collegamento. Verifichiamo quindi che non intervenga un limite più stringente a causa della dispersione cromatica. Supponendo di utilizzare la sorgente laser in grado di conseguire un PBL di 250 Gbps-Km, si ottiene una lunghezza massima pari a $\frac{250.000}{417}=600$ Km, che costituisce il *Limite di Dispersione*.

Massima lunghezza di tratta L'esercizio svolto ha lo scopo di mostrare la metodologia di progetto per un collegamento in fibra ottica, in cui vengono calcolati entrambi i limiti di *Attenuazione* e di *Dispersione*, e la massima lunghezza del collegamento è determinata dal vincolo più stringente. Nel caso dell'esercizio, la lunghezza è determinata dal limite di attenuazione, ed il progetto può essere rivisto utilizzando una sorgente *meno pura* per risparmiare, oppure una sorgente *più potente* per aumentare il guadagno di sistema e conseguentemente migliorare il limite di attenuazione. In questo secondo caso, può essere opportuno prestare attenzione al fatto che, aumentando la potenza di emissione, la purezza cromatica della sorgente può degradare (in quanto si verifica un aumento di $\Delta\lambda$ dovuto a fenomeni non lineari) con un conseguente peggioramento del limite di dispersione; è pertanto possibile ricercare la soluzione di migliore compromesso tra potenza di emissione e purezza spettrale.

Qualora non si riesca a rientrare nelle specifiche di progetto con una unica tratta, occorrerà suddividere il collegamento in più tratte, collegate da *ripetitori rigenerativi*, oppure ripartire la banda su più fibre poste in parallelo, ovvero bilanciarsi tra queste due soluzioni.

Trasduttori ottico-elettrici La conversione del segnale uscente dalla fibra ottica meriterebbe una ampia trattazione approfondita, ma qui ci limitiamo a riferire esclusivamente poche cose fondamentali.

Il trasduttore utilizzato fin dall'inizio, economico ed affidabile, è il diodo P-I-N⁵⁹, che però non è adatto all'impiego con λ più elevate. Un secondo tipo di trasduttore molto usato è il diodo APD⁶⁰ (*Avalanche Photo Detector*), caratterizzato da un "effetto valanga" che lo rende più sensibile di 10-15 dB rispetto ai P-I-N; d'altra parte però gli APD sono più delicati, e più sensibili alla temperatura.

⁵⁹http://it.wikipedia.org/wiki/Diodo_PIN

⁶⁰http://it.wikipedia.org/wiki/Fotodiodo_a_valanga

Fotorivelatore	λ [nm]	Sensibilità [dBm]	f_b [Mbps]
Si P-I-N	850	-48	50
Si APD	850	-58	50
InGaAs P-I-N	1310	-35	420
InGaAs APD	1310	-43	420
InGaAs APD	1550	-37.5	678

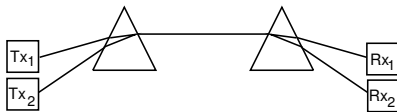
La tabella a lato riporta i valori di sensibilità W_R (ossia la minima potenza che è necessario ricevere) di diversi fotorecettori, necessaria a conseguire⁶¹ una probabilità di errore per bit $P_e = 10^{-11}$.

Dipendenza della sensibilità dalla durata del simbolo Nella tabella è riportato anche il valore della frequenza di segnalazione f_b a cui si riferisce la sensibilità, in quanto le prestazioni conseguite dal decisore che si trova a valle del trasduttore dipendono (pag. 377) da $\frac{E_b}{N_0}$, in cui E_b è l'energia per bit che vale $E_b = W_R \cdot T_b = \frac{W_R}{f_b}$. Pertanto, i trasduttori dimezzano la sensibilità (che aumenta di 3 dB) se la velocità f_b raddoppia, in quanto si dimezza l'energia per bit E_b . La sensibilità a frequenze diverse da quelle in tabella può quindi essere calcolata come⁶²

$$W_R(f'_b) \text{ [dBm]} = W_R(f_b) \text{ [dBm]} + 10 \log \frac{f'_b}{f_b}$$

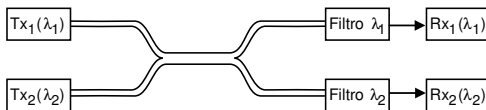
15.4.3 Multiplicazione a divisione di lunghezza d'onda - WDM

Nel campo delle trasmissioni ottiche, per le quali è spesso sottinteso che si sta trasportando un segnale già multiplexato TDM, si aggiunge una nuova possibilità di multiplexazione, la WDM (*Wavelength Division Multiplex*), in cui si usano diverse lunghezze d'onda (nella stessa finestra) per diverse comunicazioni contemporanee.



Il modo più semplice ed intuitivo di realizzare la moltiplicazione di lunghezza d'onda è di adottare dei *rifrattori prismatici*, realizzando un circuito ottico del tipo illustrato in figura⁶³. I dispositivi di moltiplicazione di forma

di questo tipo vengono detti *passivi e reversibili*, in quanto non necessitano di alimentazione, ed uno stesso apparato può indifferentemente svolgere una funzione e la sua inversa. La passività del WDM rende questa tecnica attraente, qualora si pensi di distribuire fibre ottiche di casa in casa⁶⁴: ognuno avrebbe una sua lunghezza d'onda λ_i , e la fibra sarebbe una per tutto il condominio.



Se le λ_i sono troppo vicine tra loro (con una spaziatura dell'ordine di 0.015 nm), allora i prismi non riescono più nel compito di separazione

⁶¹La consuetudine del dimensionamento dei collegamenti in fibra ottica porta a considerare ogni bit in transito *nella sua purezza*, senza cioè confidare (o meno) nella presenza di elaborazioni terminali come la codifica di canale, e/o il numero di bit/simbolo. In tale prospettiva, si ritiene che un valore di $P_e = 10^{-11}$ sia più che sufficiente a qualunque tipo di trasmissione: un errore ogni 100.000 miliardi di bit!

⁶²Questo metodo di calcolo della sensibilità ad una f_b diversa da quella delle tabelle è approssimato, in quanto nei trasduttori avvengono fenomeni non-lineari che legano il livello di potenza del rumore alla potenza di segnale ricevuta. Trascurando questo effetto, si può applicare l'espressione sopra riportata.

⁶³Si sfrutta il principio "dell'arcobaleno" (o dei *Pink Floyd* ?), in quanto uno stesso materiale (il prisma) presenta indici di rifrazione differenti per lunghezze d'onda diverse, e quindi è in grado di focalizzare più sorgenti di diverso colore in un unico raggio.

⁶⁴Soluzione nota come *Fiber to the home* o FTTH: <http://it.wikipedia.org/wiki/Fttx>.

geometrica, e conviene ricorrere ad una separazione della potenza (si fa uscire parte di segnale luminoso da *jacket*) ed un filtraggio (realizzato otticamente mediante gelatine) di ognuna delle λ_i . Così facendo però si perde molta potenza.

15.4.4 Ridondanza e pericoli naturali

Le fibre vengono normalmente interrato, e per questo sono esposte ai pericoli di essere attaccate da talpe e topi, o di essere interrotte a causa di lavori stradali od agricoli. Quelle sottomarine sono a rischio per via di squali e reti a strascico. E' più che opportuno prevedere una adeguata ridondanza (vedi § 6.6.3), in modo che in caso di interruzione di un collegamento sia possibile deviarne il traffico su di un altro.

15.4.5 Sonet e SDH

SONET è l'acronimo di *Synchronous Optical Network*, lo standard americano definito allo scopo di permettere l'interconnessione diretta tra reti in fibra ottica, perfettamente interoperabile con l'equivalente europeo *Synchronous Digital Hierarchy* (SDH); l'unità di moltiplicazione fondamentale per i due casi è pari a 51.84 Mbps per Sonet, ovvero 3 volte tanto (155,52 Mbps) per l'Europa. Per approfondimenti, vedi § 6.4.

Data Rate [Mbps]	Sigla ITU-T
51.84	
155.52	STM-1
466.56	STM-3
622.08	STM-4
933.12	STM-6
1244.16	STM-8
1866.24	STM-12
2488.32	STM-16

15.4.6 Dalle fibre ottiche alle reti ottiche

Lo sviluppo tecnologico permette attualmente⁶⁵ di realizzare dispositivi in grado di operare direttamente sul segnale ottico WDM, e cioè

- amplificatori ottici - questa funzione realizzata direttamente nel dominio ottico abbatte la complessità ed i ritardi legati alla doppia conversione altrimenti necessaria;
- moltiplicatori add and drop - per inserire o estrarre una specifica λ ;
- convertitori di lunghezza d'onda - per convertire una λ in un'altra;
- commutatori di lunghezza d'onda - per estrarre una λ da un segnale WDM ed inserirla in un altro.

Questi dispositivi permettono la realizzazione di trasmissioni DWDM (*Dense Wavelength Division Multiplex*), che ospitano quasi un centinaio di diverse λ nella stessa fibra, e che hanno trovato impiego dapprima nei collegamenti a lunga distanza, favoriti dagli amplificatori ottici, e quindi nelle sezioni di rete via via più periferiche, sfruttando i dispositivi di commutazione ed accesso, ed infine hanno trovato un ruolo anche nei meccanismi di protezione dai guasti.

Un ulteriore importante risultato della trasmissione DWDM è che, ospitando differenti tributari ad alta velocità su diverse λ , decadono quelle esigenze di sincronizzazione tipiche dei sistemi TDM, e si realizza una sorta di *trasparenza* in quanto scompaiono i dispositivi strettamente legati al tipo di segnale trasportato.

⁶⁵ A titolo integrativo, si propongono i riferimenti a
http://it.wikipedia.org/wiki/Amplificatore_ottico,
http://en.wikipedia.org/wiki/Wavelength-division_multiplexing#Dense_WDM,
http://www.fiber-optics.info/articles/dense_wavelength-division_multiplexing

15.5 Appendici

15.5.1 Collegamenti satellitari

Tutti i satelliti artificiali hanno, ovviamente, l'esigenza di mantenere un collegamento radio con il centro di controllo orbitale terrestre; in tutti i modi, un buon numero di satelliti è stato lanciato per svolgere un ruolo nell'ambito dei sistemi di comunicazione e telerilevamento, come ad esempio nei casi dei satelliti meteorologici, di radiolocalizzazione (il GPS, ma non solo), per ponti radio televisivi, di telefonia, di broadcast televisivo. Senza molto togliere alla generalità dell'esposizione, questa procede illustrando l'ultimo caso citato, detto DVB (*Digital Video Broadcast*), in cui il satellite semplicemente ritrasmette verso una estesa area geografica i segnali ricevuti da terra, come mostrato in figura 15.7, assieme all'*ipsogramma*⁶⁶ relativo.

Studio di produzione Non volendo assolutamente entrare qui negli innumerevoli dettagli che andrebbero illustrati, limitiamoci a descrivere i passi necessari a generare il segnale inviato al satellite:

- si effettua la codifica digitale MPEG2 (§ 18.3.1.4) del segnale televisivo, ottenendo un flusso numerico chiamato PS (*Program Stream*);
- più PS sono pacchettizzati e multiplati (§ 18.3.2.1) in un nuovo flusso chiamato MPEG-TS (*Transport Stream*), assegnando loro un identificativo noto come PID (*Packet Identifier* o *Program ID*);
- alcuni PID sono riservati per indicare l'inserimento all'interno del TS di informazioni di controllo (*o tabelle*) note come PAT (*Program Association Table*), PMT (*Program Map Table*), CAT (*Conditional Access Table*), NIT (*Network Information Table*), etc;
- il TS è sottoposto ad un processo di *scrambling* basato su di un generatore binario pseudocasuale, in modo da renderne la densità spettrale più uniforme possibile;
- il risultato è sottoposto ad una codifica di canale FEC (vedi pag. 93) a tre stadi, in cui è prima applicato un codice di *Reed-Solomon* (§ 17.3.1.2), poi un *interleaver*, (§ 5.4.3.3) e quindi un codificatore *convoluzionale* (§ 17.3.2), rendendo il segnale particolarmente robusto agli errori di trasmissione sia singoli che a burst;
- il nuovo flusso numerico è modulato QPSK (a due bit per simbolo) con codifica *di Gray*, sagomando i simboli con un filtro a coseno rialzato con $\gamma = 0.35$ ripartito tra trasmettitore e ricevitore finale, ossia adottando un formatore di impulsi a *radice di coseno rialzato* (vedi § 5.2.2.3).

Uplink Il *collegamento in salita* (UPLINK) è quello mediante il quale lo studio di produzione invia al satellite l'MPEG-TS che deve essere re-distribuito. Il segnale sopra descritto è quindi amplificato a potenza W_{dT} , parte della quale si perde nel cavo che collega l'antenna trasmittente di guadagno G_T^e . L'EIRP^e (*Equivalent Isotropically Radiated Power*) rappresenta la potenza effettivamente irradiata⁶⁷, che si riduce notevolmente nella trasmissione da terra a satellite. Considerando una portante di 2 GHz e

⁶⁶Dal greco *hypso* che significa *altezza*. Mentre l'*ipsografia* è un diagramma che individua il rilievo altimetrico terrestre, il termine *ipsogramma* è a volte usato nelle telecomunicazioni per descrivere l'andamento del livello di potenza in un collegamento.

⁶⁷Più precisamente, l'EIRP è la potenza che erogherebbe una antenna isotropa, per generare lo stesso campo elettrico prodotto dalla antenna direttiva nella direzione di massimo guadagno.

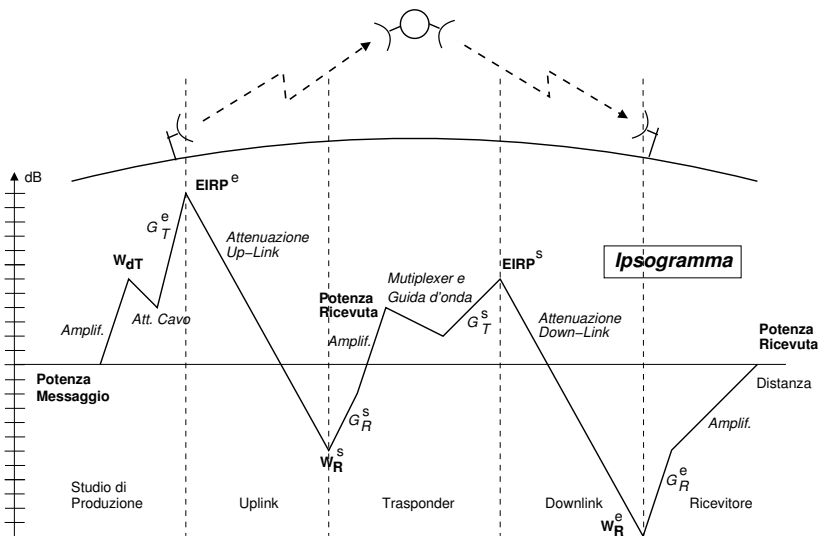


Figura 15.7: Andamento del livello di potenza in dB per un collegamento satellitare

la quota di un satellite in orbita geostazionaria⁶⁸ (36.000 Km da terra), l'attenuazione di spazio libero dell'UP-LINK (eguale a quella del DOWN-LINK da satellite a terra) è di circa 190 dB.

Transponder Il segnale ricevuto, di potenza W_R^s , è captato dall'antenna ricevente del satellite di guadagno G_R^s , e l'amplificatore seguente eleva ulteriormente il livello del segnale ricevuto, che poi subisce alcune perdite nel collegamento con l'antenna trasmittente del satellite di guadagno G_T^s , determinando così il valore della $EIRP^s$ all'uscita del transponder satellitare. Questo termine descrive la circostanza che il satellite non si limita ad amplificare il segnale in transito, ma *traspone* anche la banda di frequenze occupata dalla trasmissione. Infatti, essendo la differenza tra $EIRP^s$ e W_R^s molto elevata, se la frequenza portante utilizzata nell'uplink fosse uguale a quella del down-link il segnale trasmesso costituirebbe un rilevante termine di *interferenza* per il lato ricevente del satellite, nonostante l'elevata direttività delle antenne, dando così luogo ad un fenomeno di *diafonia*⁶⁹. La Fig. 15.8 mostra come il segnale a banda larga (che trasporta molteplici canali) ricevuto da terra viene prima filtrato alla banda del segnale utile, quindi amplificato una prima volta, poi miscelato con un oscillatore

⁶⁸Un satellite in orbita geostazionaria è visto da terra sempre nella stessa posizione (e ciò consente di puntare l'antenna in modo permanente) in quanto la sua orbita giace sul piano definito dall'equatore, ed il suo periodo di rivoluzione attorno all'asse terrestre coincide con quello di rotazione della terra (pari ad un giorno). Il moto orbitale è causa di una forza centrifuga, che è bilanciata da quella centripeta prodotta dall'attrazione terrestre. Dato che all'aumentare della distanza dalla terra, la prima aumenta (con orbite più grandi, deve aumentare la velocità tangenziale) e la seconda diminuisce, la quota di 36.000 Km costituisce un punto di equilibrio, al disotto del quale il satellite precipiterebbe al suolo, ed al disopra del quale si perderebbe nello spazio.

⁶⁹Le considerazioni sulla diafonia si applicano altrettanto bene anche al caso di ripetitori terrestri.

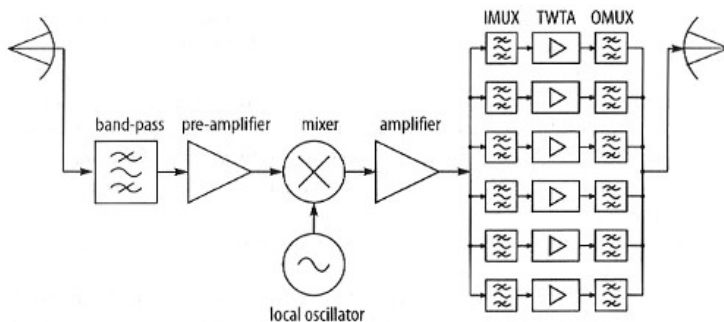
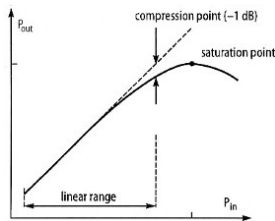


Figura 15.8: Elaborazione di bordo per un trasponder DVB satellitare

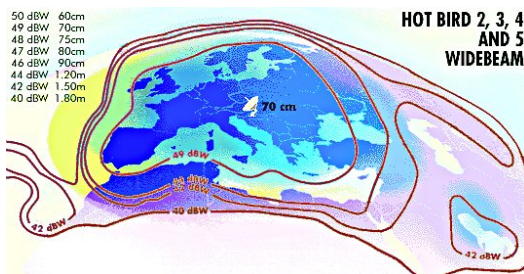
locale⁷⁰, ed infine amplificato una seconda volta⁷¹.

Quindi, i singoli canali FDM che compongono il segnale sono separati tra loro mediante il banco di filtri passa-banda indicati come IMUX (*input multiplexer*), e amplificati individualmente mediante dei TWTA⁷² che, se spinti alla massima potenza, presentano una caratteristica ingresso-uscita non lineare (vedi § 14.7), mostrata alla figura a lato. Nel caso di trasmissioni modulate angolarmente, la distorsione in ampiezza è ben tollerata, e quindi si può mantenere limitato il *back-off* necessario; d'altra parte, le componenti frequenziali spurie prodotte dalla non linearità devono essere rimosse per non provocare disturbo alle altre comunicazioni, e questo è il compito del banco di filtri passa banda OMUX (*output multiplexer*) posti di seguito ai TWTA.



Footprint e Downlink L'antenna trasmittente del satellite sagoma il proprio diagramma di radiazione in modo da concentrare la potenza trasmessa in una ben determinata area della terra, dando luogo alla cosiddetta *footprint* (impronta) raffigurata a lato, in cui le curve isomere individuano sia il livello di potenza ricevuto, che il diametro (e quindi il guadagno) necessario per l'antenna ricevente.

La tecnica che permette di distribuire la potenza emessa secondo una geometria diversa da una simmetria radiale prende il nome di *beamforming*, e si basa sull'utilizzo di più antenne trasmittenti, in modo da realiz-



⁷⁰Come descritto al § 11.2.5, l'oscillatore locale deve avere una frequenza f_e tale che $f_d = f_u - f_e$, in modo che il segnale di downlink sia centrato ad una frequenza pari alla differenza tra quella di uplink e quella di eterodina.

⁷¹La suddivisione della amplificazione in due stadi a frequenza diversa previene fenomeni di reazione positiva.

⁷²*Travelling Wave Tube Amplifier*, ovvero tubi amplificatori ad onda progressiva: http://it.wikipedia.org/wiki/Travelling_wave_tube.

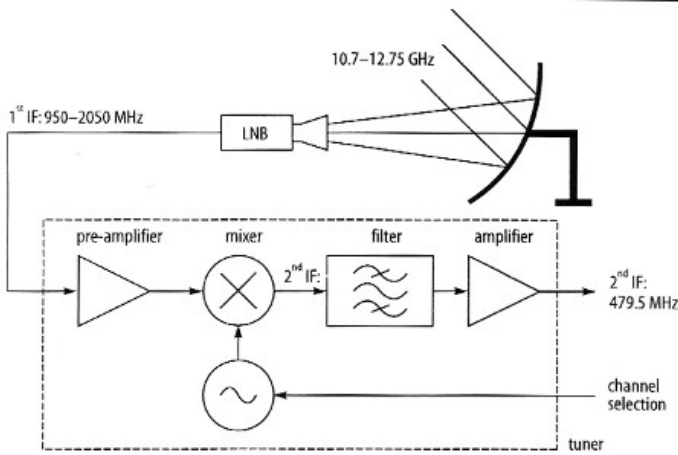


Figura 15.9: Ricevitore satellitare DVB

zare un *phased array*⁷³. Ad ogni antenna dell'array perviene lo stesso segnale modulato, ma con una fase tale da creare uno schema di interferenza con le altre antenne dell'array, in modo che alla distanza di ricezione, si determini la distribuzione spaziale desiderata.

Dal lato del ricevitore terrestre arriva dunque un segnale di potenza W_R^e , che ha subito l'attenuazione del down-link; questo è quindi riportato ad un livello di potenza appropriato, sia grazie al guadagno di antenna, che per mezzo di uno stadio di amplificazione.

Temperatura di antenna Come illustrato al § 15.3, una antenna ricevente è schematizzabile come un generatore controllato, ed al Cap. 16 si mostra come la sua impedenza interna sia la fonte del rumore additivo gaussiano in ingresso al ricevitore, caratterizzato da una densità di potenza disponibile $W_{dn}(f) = \frac{1}{2}kT_g$, in cui T_g ora viene detta *temperatura di antenna* T_a , e assume un valore inferiore ai 290 °K, e precisamente compreso tra i 15 ed i 60 °K. La fonte diretta di rumore, in questo caso, è il *rumore galattico*, la cui temperatura si abbatte a 10 °K sopra i 2,5 GHz, mentre i *lobi laterali* del diagramma di radiazione captano il rumore legato alla temperatura terrestre⁷⁴.

Ricevitore a terra La figura 15.9 mostra l'architettura del ricevitore satellitare per la trasmissione televisiva DVB. La parabola, puntata nella direzione del satellite desiderato, riceve il segnale in una di due bande 10.7-11.7 GHz, oppure 11.7-12.75 GHz, ed un dispositivo LNB (*low noise block*) provvede ad un primo stadio di amplificazione a basso rumore, e ad una prima conversione di frequenza che centra il segnale tra 0.95 e 2.05 GHz, in modo da ridurre le perdite introdotte dal cavo coassiale⁷⁵ che

⁷³http://en.wikipedia.org/wiki/Phased_array)

⁷⁴Per contro, nel caso in cui dietro al satellite verso cui è puntata l'antenna vi sia una stella luminosa, la T_a è più elevata.

⁷⁵Come descritto nel paragrafo che discute dell'*effetto pelle* (pag. 381), l'attenuazione in dB del cavo aumenta con l'aumentare della radice della frequenza.

<i>Intervallo</i>	λ	<i>Sigla</i>	<i>Denominazione</i>	<i>Uso</i>
30 - 300 Hz	$10^4 - 10^3$ Km	ELF	<i>Extremely Low</i>	
.3 - 3 KHz	$10^3 - 10^2$ Km	VF	<i>Voice Frequency</i>	
3 - 30 KHz	100 - 10 Km	VLF	<i>Very Low</i>	Radionavigazione a largo raggio. Attività nucleare.
30 - 300 KHz	10 - 1 Km	LF	<i>Low Frequency</i>	Radiolocalizzazione marittima ed aeronautica
.3 - 3 MHz	.1 - 1 Km	MF	<i>Medium Frequency</i>	Comunicazioni aeree e marittime. Radionavigazione. Broadcast AM
3 - 30 MHz	10 - 100 metri	HF	<i>High Frequency</i>	Collegamenti a lunga distanza fissi e mobili. Radioamatori.
30 - 300 MHz	1 - 10 metri	VHF	<i>Very High</i>	Broadcast FM e TV. Collegamenti in visibilità. Radiomobili civili e militari.
.3 - 3 GHz	.1 - 1 metro	UHF	<i>Ultra High</i>	Ponti radio e radiomobili ter- restri. Broadcast TV. Satelliti meteo e TV.
3 - 30 GHz	10 - 100 mm	SHF	<i>Super High</i>	Ponti radio terrestri. Satelliti. Radar.
30 - 300 GHz	1 - 10 mm	EHF	<i>Extremely High</i>	Radar

Tabella 15.1: Allocazione delle frequenze radio

collega l'antenna al ricevitore casalingo. Quindi, si ritrova un schema simile a quello del trasponder, ovvero amplificatore-mixer-amplificatore, in cui questo secondo stadio eterodina centra il canale desiderato alla frequenza intermedia di 479.5 MHz.

Polarizzazione Chi ha provato a sintonizzare un ricevitore TV satellitare, si sarà accorto che tra le varie opzioni possibili, si può indicare anche il *tipo di polarizzazione*, orizzontale o verticale. Questo termine si riferisce all'orientamento (rispetto all'orizzonte) del piano su cui varia il vettore di campo elettrico relativo alla trasmissione radio. Mentre per le trasmissioni terrestri, a causa delle molteplici possibili riflessioni, questo è imprevedibile al ricevitore, nelle comunicazioni satellitari il tipo di polarizzazione adottata dal trasmettitore (il satellite) si mantiene fino a terra. Dato che un segnale polarizzato in un senso, risulta attenuato di decine di dB se ricevuto da una antenna predisposta per la polarizzazione nell'altro senso, nella stessa banda di frequenze possono essere effettuate due trasmissioni contemporanee.

15.5.2 Allocazione delle frequenze radio

L'assegnazione generale dello spettro radio ai diversi utilizzi è riportata in tabella 15.1, che non pretende di essere completa né tanto meno esatta, così come per le tabelle che seguono.

Canali televisivi

VHF: Numerati da 1 a 6 a partire da 55.25 MHz, spaziati di 6 MHz, fino a 83.25 MHz; numerati da 7 a 13 a partire da 175.25 MHz, fino a 211.25 MHz, ancora spaziati di 6 Mhz. Nell'intervallo 88-108 Mhz è presente il broadcast FM.

UHF: Numerati da 14 a 69 a partire dalla portante video di 471.25 MHz, fino a 801.25 MHz, spaziati di 6 MHz.

Per le stesse frequenze, sono state attivate la trasmissioni televisive in *digitale terrestre.*, ad eccezione dei canali da 61 a 69, che saranno assegnati agli operatori di telefonia mobile di 4^a generazione.

Bande di frequenza Radar Oltre alle bande HF, VHF ed UHF, le trasmissioni radar che operano in SHF ed EHF distinguono tra i seguenti intervalli di frequenze:

GHz	1-2	2-4	4-8	8-12	12-18	18-27	27-40	40-75	75-110	110-300
Banda	L	S	C	X	K _u	K	K _a	V	W	millimetriche

Banda ISM ISM sta per *Industrial, Scientific and Medical*, per i cui usi sono state riservate le seguenti frequenze per le quali non occorre il rilascio di licenza. Gli intervalli più usati sono:

<i>Intervallo</i>	<i>utilizzo</i>
26.957–27.283 MHz	Banda cittadina dei radioamatori CB, ma anche dei camionisti
2.4–2.5 GHz	Forni a microonde, Bluetooth, WiFi 802.11b e g
5.725–5.875 GHz	WiFi 802.11a

Telefonia mobile

<i>Intervallo Uplink (MHz)</i>	<i>Intervallo Downlink(MHz)</i>	<i>utilizzo</i>
890,0 - 915,0	935,0 - 960,0	GSM 900
880,0 - 890,0	925,0 - 935,0	GSM 900 esteso
1710,0 - 1785,0	1805,0 - 1880,0	GSM 1800
1920 - 1980	2110 - 2170	UMTS

Capitolo 16

Rumore termico

Descriviamo ora la natura e le fonti del processo di rumore sempre presente negli apparati di telecomunicazione, e di come questo sia tenuto in considerazione nel progetto degli stessi.

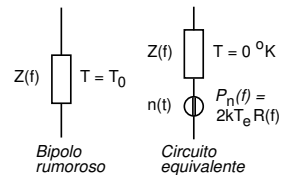
16.1 Rumore nei bipoli passivi

Ai capi di un resistore R a temperatura T è presente una *tensione a vuoto* $n(t)$, realizzazione di un processo gaussiano a media nulla, che è l'effetto del moto caotico degli elettroni all'interno della resistenza¹. Lo spettro di densità di potenza della tensione a vuoto ha espressione²

$$\mathcal{P}_n(f) = 2R \frac{\hbar f}{e^{\frac{\hbar f}{kT}} - 1} \simeq 2kTR \quad [\text{Volt}^2]$$

in cui $k = 1.38 \cdot 10^{-23}$ Joule/°K è la *costante di Boltzman* ed $\hbar = 6.62 \cdot 10^{-34}$ Joule·sec è la *costante di Plank*: questi valori³ fanno sì che l'approssimazione $\mathcal{P}_n(f) \simeq 2kTR$ sia valida ad ogni frequenza di interesse.

In un bipolo passivo di impedenza $Z(f) = R(f) + jX(f)$, solamente la parte reale (componente resistiva) concorre a generare il processo di rumore termico, che pertanto possiede una densità di potenza *di segnale* $\mathcal{P}_n(f) \simeq 2KT_e R(f)$. Nel caso in cui il bipolo contenga *più* resistori a temperature diverse, si può definire una temperatura equivalente $T_e(f)$; un bipolo passivo equivale pertanto allo stesso bipolo non rumoroso (a temperatura zero), con in serie un generatore di rumore con densità di potenza $\mathcal{P}_n(f) \simeq 2kT_e(f) R(f)$. Questo generatore equivalente, è



¹Possiamo pensare che gli elettroni, qualora si trovino in maggior misura in una metà della resistenza, producano una differenza di potenziale negativa in quella direzione. Allo zero assoluto (- 273 °C) il moto caotico degli elettroni cessa, e si annulla così la tensione di rumore. Di qui l'aggettivo *termico* per descrivere il fenomeno.

²Si tratta di una forma della legge di Plank, vedi http://en.wikipedia.org/wiki/Thermal_noise#Noise_at_very_high_frequencies

³Espandendo $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots$ si ottiene che per $x \ll 1$ risulta $e^x \simeq 1 + x$, e quindi $e^{\frac{\hbar f}{kT}} \simeq 1 + \frac{\hbar f}{kT}$.

quindi descritto da una *potenza disponibile di rumore*

$$\mathcal{W}_{dn}(f) = \frac{\mathcal{P}_n(f)}{4R(f)} = \frac{1}{2}kT_e(f) \left[\frac{\text{Watt}}{\text{Hz}} \right]$$

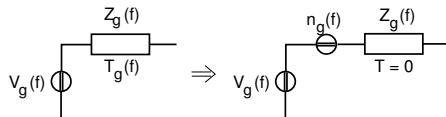
Nel caso in cui $T_e(f) = T_0 = 290$ °K (temperatura ambiente), il termine $kT_0 = N_0 = 2\mathcal{W}_{dn}(f)$ assume i valori riportati di seguito per diverse unità di misura, da adottare in alternativa, allo scopo di rendere la grandezza omogenea con le altre che compaiono nelle formule di progetto:

$$kT_0 = -204 \text{ [dBW/Hz]} = -174 \text{ [dBm/Hz]} = -114 \text{ [dBm/MHz]}$$

Ad esempio, all'uscita di un filtro passa-banda ideale non rumoroso⁴ di estensione 1 MHz, si ha una potenza disponibile di rumore pari a $10^{-11.4}$ mW .

16.2 Rapporto segnale rumore dei generatori

Un generatore di tensione $V_g(f)$, che possiede una propria impedenza interna $Z_g(f)$ a temperatura equivalente $T_g(f)$, produce anch'esso un processo di rumore in virtù della componente reale $R_g(f) = \Re\{Z_g(f)\}$ di $Z_g(f)$, e $Z_g(f)$ può quindi schematizzarsi con il circuito equivalente mostrato in figura.



Pertanto, oltre alla potenza disponibile di segnale $\mathcal{W}_{dg}(f) = \frac{\mathcal{P}_g(f)}{4R_g(f)}$, troviamo anche una potenza disponibile di rumore $\mathcal{W}_{dn}(f) = \frac{1}{2}kT_g(f)$, e dunque un rapporto segnale rumore disponibile

$$SNR_g(f) = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)}$$

che come osserviamo dipende da f , sia a causa di $\mathcal{W}_{dg}(f)$ che di $T_g(f)$. Infine, notiamo che lo stesso valore di SNR_g è esprimibile anche come rapporto tra le potenze *di segnale* anziché *disponibili*: infatti

$$SNR_g(f) = \frac{\mathcal{P}_g(f)}{4R_g(f)} \cdot \frac{1}{\frac{1}{2}kT_g(f)} = \frac{\mathcal{P}_g(f)}{2kT_g(f)R_g(f)} = \frac{\mathcal{P}_g(f)}{\mathcal{P}_n(f)}$$

16.3 Rumore nelle reti due porte

Se colleghiamo un generatore rumoroso a temperatura T_g all'ingresso di una rete due porte a temperatura T_Q , è lecito aspettarsi all'uscita della rete un processo di rumore dipendente sia dal generatore che dalla rete, e la cui potenza disponibile $\mathcal{W}_{dnu}(f)$ può essere espressa in funzione di una temperatura equivalente di uscita $T_{e_u}(f)$, tale che

$$\mathcal{W}_{dnu}(f) = \frac{1}{2}kT_{e_u}(f)$$

⁴Si intende dire che il filtro non introduce altro rumore oltre quello di natura termica.

D'altra parte a $T_{e_u}(f)$ concorrono sia la temperatura del generatore $T_g(f)$, che la rete con una propria $T_{Q_u}(f)$ "equivalente di uscita"; scriviamo dunque

$$W_{dn_u}(f) = \frac{1}{2}k \cdot [T_g(f) G_d(f) + T_{Q_u}(f)]$$

in cui la potenza disponibile in ingresso alla rete (che ha guadagno disponibile $G_d(f)$) è riportata in uscita, moltiplicata per $G_d(f)$. Se effettuiamo l'operazione inversa per il contributo di rumore dovuto a T_{Q_u} , otteniamo $W_{dn_u}(f) = \frac{1}{2}k G_d(f) \cdot [T_g(f) + T_{Q_i}(f)]$ (in cui $T_{Q_i}(f) = \frac{T_{Q_u}(f)}{G_d(f)}$), ovvero

$$W_{dn_u}(f) = \frac{1}{2}k G_d(f) T_{e_i}(f)$$

dove

$$T_{e_i}(f) = T_g(f) + T_{Q_i}(f) = T_g(f) + \frac{T_{Q_u}(f)}{G_d(f)} \tag{16.1}$$

è detta anche *temperatura di sistema* $T_s = T_{e_i}$, poiché riporta in ingresso alla rete tutti i contributi al rumore di uscita, dovuti sia al generatore che alla rete. Siamo però rimasti con un problema irrisolto: che dire a riguardo di T_{Q_i} e T_{Q_u} ?

16.3.1 Reti passive

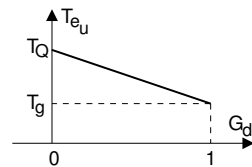
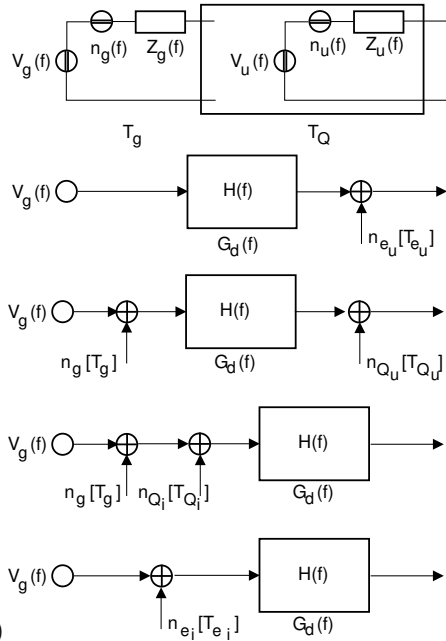
Supponiamo ora tutti i componenti della rete due porte alla stessa temperatura T_Q . In questo caso si può mostrare che risulta

$$\begin{cases} T_{Q_u}(f) &= [1 - G_d(f)] T_Q \\ T_{Q_i}(f) &= \frac{T_{Q_u}(f)}{G_d(f)} = [A_d(f) - 1] T_Q \end{cases}$$

in modo da poter scrivere:

$$\begin{cases} T_{e_u}(f) &= G_d(f) T_g(f) + T_{Q_u}(f) = G_d(f) T_g(f) + [1 - G_d(f)] T_Q \\ T_{e_i}(f) &= \frac{T_{e_u}(f)}{G_d(f)} = T_g(f) + [A_d(f) - 1] T_Q \end{cases}$$

Questo risultato evidenzia come per una rete passiva (con $0 \leq G_d \leq 1$), la temperatura di rumore equivalente in uscita sia una media pesata delle temperature del generatore e della rete. Nei casi limite in cui $G_d = 0$ oppure 1, la $T_{e_u}(f)$ è pari rispettivamente a T_Q e $T_g(f)$; infatti i due casi corrispondono ad una "assenza" della rete oppure ad una rete che non attenua.



16.3.1.1 Rapporto SNR in uscita

Se si valuta il rapporto segnale rumore in uscita alla rete, otteniamo

$$SNR_u(f) = \frac{\mathcal{W}_{dg}(f) G_d(f)}{\frac{1}{2}kT_{e_i}(f) G_d(f)} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}k \cdot [T_g(f) + [A_d(f) - 1]T_Q]}$$

Ricordando che il generatore in ingresso presenta un $SNR_i(f) = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)}$, possiamo valutare il peggioramento prodotto dalla presenza della rete:

$$\begin{aligned} \frac{SNR_i(f)}{SNR_u(f)} &= \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)} \cdot \frac{\frac{1}{2}k \cdot [T_g(f) + [A_d(f) - 1]T_Q]}{\mathcal{W}_{dg}(f)} = \\ &= 1 + \frac{T_Q}{T_g(f)} \cdot [A_d(f) - 1] \end{aligned} \quad (16.2)$$

16.3.1.2 Fattore di rumore per reti passive

Il coefficiente (16.2) $F(f) = 1 + \frac{T_Q}{T_g(f)} \cdot [A_d(f) - 1] \geq 1$ è chiamato *fattore di rumore*⁵ della rete passiva, e rappresenta il peggioramento dell' SNR dovuto alla sua presenza, potendo infatti scrivere

$$SNR_u(f) = \frac{SNR_i(f)}{F(f)} \leq SNR_i(f)$$

Notiamo subito che se $T_g(f) = T_Q$, allora $F = A_d$: pertanto una rete passiva che si trova alla stessa temperatura del generatore, esibisce un fattore di rumore pari all'attenuazione. Infatti, mentre la potenza disponibile di rumore è la stessa (essendo generatore e rete alla stessa temperatura), il segnale si attenua di un fattore A_d .

16.3.2 Reti attive

In questo caso il rumore introdotto dalla rete ha origine *non solo* dai resistori, e dunque *non è più vero* che $T_{Q_u}(f) = [1 - G_d(f)]T_Q$. Inoltre, il guadagno disponibile può assumere valori $G_d > 1$. In questo caso, si può esprimere l' SNR in uscita dalla rete come

$$SNR_u(f) = \frac{\mathcal{W}_{dg}(f) G_d(f)}{\frac{1}{2}k [G_d(f)T_g(f) + T_{Q_u}(f)]} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}k \cdot [T_g(f) + T_{Q_i}(f)]}$$

ed il peggioramento individuato in (16.2) come

$$\frac{SNR_i(f)}{SNR_u(f)} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)} \cdot \frac{\frac{1}{2}k \cdot [T_g(f) + T_{Q_i}(f)]}{\mathcal{W}_{dg}(f)} \quad (16.3)$$

$$= 1 + \frac{T_{Q_i}(f)}{T_g(f)} = F(f, T_g) \quad (16.4)$$

Quest'ultima espressione dipende ancora da T_g . Allo scopo di ottenere una grandezza che dipenda solamente dalla rete due porte, si definisce quindi il

⁵A volte si incontra anche il termine *figura di rumore*, derivato dall'inglese NOISE FIGURE (che in realtà si traduce *cifra di rumore*), e che si riferisce alla misura di F in decibel.

16.3.2.1 Fattore di rumore per reti attive

Viene posto pari a

$$F(f) = 1 + \frac{T_{Q_i}(f)}{T_0}$$

e rappresenta il peggioramento di SNR causato dalla rete quando il generatore è a temperatura ambiente $T_0 = 290 \text{ }^\circ\text{K} = 17 \text{ }^\circ\text{C}$. In realtà non ci è dato di conoscere $T_{Q_i}(f)$, mentre invece $F(f)$ può essere misurato a partire dal rapporto dei rapporti SNR , ed è proprio ciò che fa il costruttore della rete due porte. Questo ci permette dunque il calcolo di $T_{Q_i}(f) = T_0 [F(f) - 1]$ che, sostituito nella (16.1), permette finalmente di valutare la temperatura di sistema come

$$T_{e_i}(f) = T_g(f) + T_0 [F(f) - 1]$$

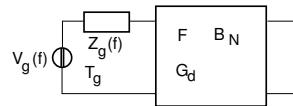
ed il peggioramento dell' SNR come

$$\frac{SNR_i(f)}{SNR_u(f)} = 1 + \frac{T_0}{T_g(f)} [F(f) - 1]$$

Riassunto

- il fattore di rumore è definito come il peggioramento di SNR dovuto alla presenza della rete tra generatore e carico, quando il generatore è a temperatura $T_0 = 290 \text{ }^\circ\text{K} = 17 \text{ }^\circ\text{C}$;
- dal fattore di rumore si deriva la temperatura di sistema $T_{e_i}(f) = T_g(f) + T_0 [F(f) - 1]$;
- Se $T_g = T_0$ allora $T_{e_i}(f) = F(f) T_0$, e dunque la temperatura di sistema T_{e_i} è $F(f)$ volte quella del generatore;
- Se la rete non è rumorosa si ottiene $F = 1$ (pari a 0 dB);
- Se la rete è passiva allora $F(f) = [A_d(f) - 1] \frac{T_Q}{T_0} + 1$, e se $T_Q = T_0$ allora $F = A_d$.

Esempio Sia data una rete due porte con assegnati guadagno disponibile G_d , banda di rumore B_N e fattore di rumore F . Valutare il rapporto segnale rumore disponibile in uscita nei due casi in cui il generatore si trovi ad una generica temperatura T_g oppure a T_0 .



Soluzione Sappiamo che la densità di potenza disponibile di rumore in uscita vale

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2} k T_{e_i} G_d = \frac{1}{2} k \cdot [T_g + T_{Q_i}] \cdot G_d$$

in generale $F = 1 + \frac{T_{Q_i}}{T_0}$ e quindi $T_{Q_i} = T_0 (F - 1)$, dunque

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2} k \cdot [T_g + T_0 (F - 1)] \cdot G_d$$

Pertanto, la potenza disponibile di rumore si ottiene integrando la densità sulla banda di rumore

$$\mathcal{W}_{dn_u} = k \cdot [T_g + T_0 (F - 1)] \cdot G_d B_N$$

che, nel caso in cui $T_g = T_0$, si riduce a $\mathcal{W}_{dn_u} = kT_0 F G_d B_N$. Per la potenza di segnale, si ha invece $\mathcal{W}_{ds_u} = \mathcal{W}_{dg} G_d$, e pertanto se $T_g = T_0$, risulta

$$SNR_u = \frac{SNR_i}{F} = \frac{\mathcal{W}_{dg}}{kT_0 F B_N} = \frac{\mathcal{W}_{dg}}{kT_{e_i} B_N}$$

ottenendo quindi lo stesso SNR in ingresso, ma con un rumore F volte più potente. Nel caso in cui T_g sia generico, considerando un fattore di rumore costante nella banda di rumore B_N , otteniamo:

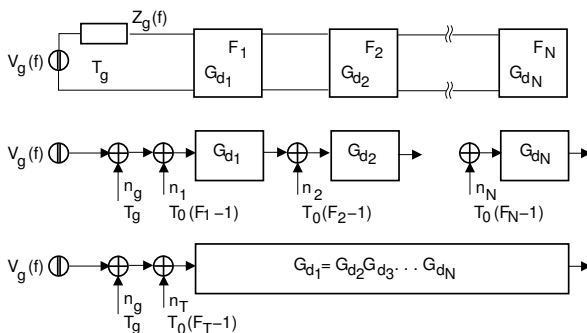
$$\begin{aligned} SNR_u &= \frac{SNR_i}{F(T_g)} = \frac{\mathcal{W}_{dg}}{kT_g B_N} \cdot \frac{1}{1 + \frac{T_{Q_i}}{T_g}} = \frac{\mathcal{W}_{dg}}{kT_g B_N} \cdot \frac{1}{1 + \frac{T_0(F-1)}{T_g}} = \\ &= \frac{\mathcal{W}_{dg}}{k[T_g + T_0(F-1)] B_N} = \frac{\mathcal{W}_{dg}}{kT_{e_i} B_N} \end{aligned}$$

16.3.3 Fattore di rumore per reti in cascata

Sappiamo che il guadagno disponibile dell'unica rete due porte equivalente alle N reti poste in cascata, è pari al prodotto dei singoli guadagni, ovvero $G_d = \prod_{n=1}^N G_{d_n}$. Come determinare invece il fattore di rumore equivalente complessivo ?

Con riferimento alla figura mostrata a lato, il singolo contributo di rumore dovuto a ciascuna rete può essere riportato all'ingresso della rete stessa, individuando così una temperatura

$$T_{Q_i}^{(n)} = T_0 (F^{(n)} - 1)$$



I singoli contributi possono

quindi essere riportati a monte delle reti che li precedono, dividendo la potenza (ovvero la temperatura) per il guadagno disponibile delle reti *scavalcate*. Dato che i contributi di rumore sono indipendenti, le loro potenze si sommano, e dunque è lecito sommare le singole temperature $T_{Q_i}^{(n)}$ riportate all'ingresso, in modo da ottenere un unico contributo complessivo di valore

$$T_{Q_i}^{(T)} = T_{Q_i}^{(1)} + T_{Q_i}^{(2)} \frac{1}{G_{d_1}} + T_{Q_i}^{(3)} \frac{1}{G_{d_1} G_{d_2}} + \dots + T_{Q_i}^{(N)} \frac{1}{\prod_{n=1}^{N-1} G_{d_n}}$$

in cui, sostituendo le espressioni per i $T_{Q_i}^{(n)}$ si ottiene

$$T_{Q_i}^{(T)} = T_0 \cdot \left[F_1 - 1 + \frac{F_2 - 1}{G_{d_1}} + \frac{F_3 - 1}{G_{d_1} G_{d_2}} + \dots + \frac{F_N - 1}{\prod_{n=1}^{N-1} G_{d_n}} \right]$$

Applicando la definizione $F^{(T)} = 1 + \frac{T_{Q_i}^{(T)}}{T_0}$, si ottiene

$$F^{(T)} = F_1 + \frac{F_2 - 1}{G_{d_1}} + \frac{F_3 - 1}{G_{d_1} G_{d_2}} + \dots + \frac{F_N - 1}{\prod_{n=1}^{N-1} G_{d_n}}$$

che costituisce proprio l'espressione cercata:

$$F^{(T)} = F_1 + \sum_{i=2}^N \frac{F_i - 1}{\prod_{j=1}^{i-1} G_{d_j}}$$

Il risultato si presta alle seguenti considerazioni:

- la prima rete due porte deve avere F più piccolo possibile, in quanto quest'ultimo non può essere ridotto in alcun modo e contribuisce per intero ad $F^{(T)}$;
- la prima rete due porte deve avere G_d più elevato possibile, in quanto quest'ultimo divide tutti i contributi di rumore delle reti seguenti.

Pertanto l'elemento che determina in modo preponderante il rumore prodotto da una cascata di reti due porte è la prima rete della serie, ed il suo progetto deve essere eseguito con cura particolare, anche tenendo conto del fatto che le due esigenze sopra riportate sono spesso in contrasto tra loro. E' inoltre appena il caso di ricordare che l'espressione ottenuta non è in dB, mentre spesso F è fornito appunto in dB; pertanto per il calcolo di $F^{(T)}$ occorre prima esprimere tutti gli F_i in unità lineari.

Esercizio Un trasmettitore con potenza di 50 mW e portante 30 MHz, modula AM BLD PS un segnale con banda $\pm W = \pm 10$ KHz. Qualora si desideri mantenere un SNR in ricezione di almeno 25 dB, determinare la distanza che è possibile coprire adottando antenne isotrope, ed un ricevitore caratterizzato da un fattore di rumore $F = 10$ dB.

Svolgimento Assumendo che si verifichino le condizioni di massimo trasferimento di potenza, il valore desiderato $SNR = \frac{W_R}{W_N}$ può essere ottenuto se $W_R = W_N \cdot SNR = 2W \cdot W_{dN_i} \cdot SNR = W \cdot kT_0 F \cdot SNR$, e quindi

$$\begin{aligned} W_{R_{min}} \text{ (dBm)} &= 10 \log_{10} 10^4 \text{ (Hz)} - 174 \text{ (dBm/Hz)} + F_{dB} + SNR_{dB} = \\ &= 40 - 174 + 10 + 25 = -99 \text{ dBm.} \end{aligned}$$

Il guadagno di sistema risulta allora pari a

$$G_s \text{ (dB)} = W_T \text{ (dBm)} - W_{R_{min}} \text{ (dBm)} = 10 \log_{10} 50 + 99 = 17 + 99 = 116 \text{ dB}$$

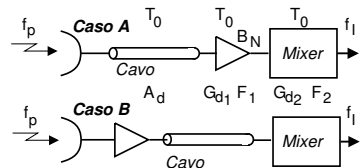
Non prevedendo nessun margine, l'attenuazione dovuta alla distanza può essere pari al guadagno di sistema, e pertanto scriviamo

$$\begin{aligned} A_d = 116 &= 32.4 + 20 \log_{10} f \text{ (MHz)} + 20 \log_{10} d \text{ (Km)} = \\ &= 32.4 + 29.5 + 20 \log_{10} d \text{ (Km)} \end{aligned}$$

e quindi $2.7 = \log_{10} d \text{ (Km)}$, da cui $d = 10^{2.7} = 501$ Km. Svolgendo nuovamente i calcoli nel caso in cui il fattore di rumore del ricevitore sia pari a 20 dB e 100 dB, si ottiene che la nuova massima distanza risulta rispettivamente di 158 Km e di 15 metri.

Esercizio Una trasmissione video modulata AM-BLU con portante $f_p = 2$ GHz viene ricevuta secondo uno dei due schemi in figura, indicati come caso A e B.

E' presente una discesa in cavo coassiale con $\phi = 1.2/4.4$ mm lunga 50 metri, un filtro-amplificatore con guadagno disponibile $G_{d1} = 20$ dB, fattore di rumore $F_1 = .4$ dB e banda di rumore $B_N = 7$ MHz, ed un mixer che converte il segnale a frequenza intermedia f_I , e che esibisce $G_{d2} = 0$ dB e $F_2 = 10$ dB. Tutti i componenti a valle dell'antenna si trovano alla stessa temperatura $T_0 = 290$ °K. Calcolare:



- 1) La minima potenza disponibile W_{dR} che occorre ricevere per ottenere $SNR_0 = 50$ dB nei due casi. Ripetere il calcolo supponendo l'antenna ricevente a temperatura $T_a = 10$ °K anziché T_0 .
- 2) La minima potenza che è necessario trasmettere per superare un collegamento terrestre lungo 50 Km, con antenne di guadagno $G_T = G_R = 30$ dB. Ripetere il calcolo per un down link satellitare in orbita geostazionaria, con $G_T = G_R = 40$ dB.
- 3) Il valore efficace della tensione ai capi del generatore equivalente di uscita dell'amplificatore di potenza del trasmettitore, per il caso migliore (tra **A** e **B**) del collegamento terrestre, nel caso di massimo trasferimento di potenza con $Z_u = Z_a = 50$ Ω, oppure con $Z_u = 50$ Ω e $Z_a = 50 - j 50$ Ω.

Svolgimento Determiniamo innanzitutto l'attenuazione del cavo coassiale, che risulta $A_d(f) = A_0 \sqrt{f}$ (MHz) dB/Km. Per il diametro indicato risulta $A_0 = 5.3$ dB/Km, ed alla frequenza di 2 GHz si ottiene $A_d(f)_{dB} = 5.3 \sqrt{2 \cdot 10^3} = 237$ dB/Km; e quindi in 50 metri si hanno $11.85 \simeq 12$ dB. Dato che il cavo è a temperatura T_0 , risulta anche $F_{cavo} = A_d = 12$ dB. Riassumendo:

	$A_d = F_{cavo}$	F_1	G_{d1}	F_2	G_{d2}
dB	11.85	.4	20	10	0
lineare	15.3	1.1	100	10	1

1)

A) Il fattore di rumore complessivo risulta

$$F^A = F_{cavo} + A_d(F_1 - 1) + \frac{A_d}{G_{d1}}(F_2 - 1) = 15.3 + 15.3 \cdot (.1) + \frac{15.3}{100}(9) = 18.2$$

ovvero pari a 12.6 dB. Dato che per la trasmissione televisiva AM-BLU si ha $SNR = SNR_0$, scriviamo

$$W_{dR} = SNR \cdot W_{dN} = SNR_0 \cdot F^A \cdot B_N \cdot kT_0 \text{ e quindi}$$

$$\begin{aligned} W_{dR}(\text{dBm}) &= SNR_0(\text{dB}) + F^A(\text{dB}) + B_N(\text{dBMHz}) + KT_0(\text{dBm/MHz}) = \\ &= 50 + 12.6 + 8.45 - 114 = -43 \text{ dBm} \end{aligned}$$

B) Il fattore di rumore complessivo risulta ora

$$F^B = F_1 + \frac{(F_{cavo}-1)}{G_{d1}} + \frac{A_d}{G_{d1}}(F_2 - 1) = 1.1 + \frac{14.3}{100} + \frac{15.3}{100}(9) = 2.26$$

ovvero pari a 3.5 dB. La differenza con il caso **A** è di 9.1 dB, e la potenza disponibile che occorre ricevere diminuisce pertanto della stessa quantità, e quindi ora risulta $W_{dR} = -52.1$ dBm.

Nel caso in cui $T_a = 10$ °K $\neq T_0$, non si ottiene più $T_{ei} = FT_0$, ma occorre introdurre la T_{Qi} della rete riportata al suo ingresso, e considerare la rete non rumorosa in modo da scrivere $T_{ei} = T_g + T_{Qi} = T_a + T_0(F - 1)$. Ripetiamo i calcoli per i due casi **A** e **B**:

$$\mathbf{A)} \quad W_{dRW} = SNR \cdot W_{dN} = SNR \cdot B_N \cdot k \cdot (T_a + T_{Qi}) =$$

$$= SNR \cdot B_N \cdot k \cdot (T_a + T_0(F^A - 1)), \text{ che espresso in dB fornisce}$$

$$\begin{aligned} W_{dRdBW} &= SNR_{dB} + 10 \log_{10} 7 \cdot 10^6 + 10 \log_{10} (1.38 \cdot 10^{-23} (10 + 290 \cdot 17.2)) \\ &= 50 + 68.5 - 191.61 = -73.11 \text{ dBW} = -43.11 \text{ dBm} \end{aligned}$$

$$\mathbf{B)} \quad W_{dR}(\text{dBW}) = 50 + 68.5 + 10 \log_{10} (1.38 \cdot 10^{-23} (10 + 290 \cdot 1.26)) = -84.3 \text{ dBW} = -54.3 \text{ dBm}$$

Notiamo che se la T_a è ridotta, le prestazioni per la configurazione **A** migliorano di soli 0.11 dB, mentre nel caso **B** il miglioramento è di circa 2.2 dB. Questo risultato trova spiegazione con il fatto che in **A** predomina comunque il T_{Qi} prodotto dal cavo.

2) In un collegamento radio terrestre si assume $T_a = 290$ °K. Inoltre, per il caso in esame si trova una attenuazione disponibile pari a

$$A_d = 32.4 + 20 \log_{10} f(\text{MHz}) + 20 \log_{10} d(\text{Km}) - G_T - G_R = 32.4 + 66 + 34 - 60 = 72.4 \text{ dB}$$

A) $W_{dT} = W_{dR} + A_d = -43.11 + 72.4 = 29.29 \text{ dBm} = 850 \text{ mW}$

B) $W_{dT} = W_{dR} + A_d = -54.3 + 72.4 = 18.1 \text{ dBm} = 66 \text{ mWatt}$

Per il downlink si ha $d = 36.000 \text{ Km}$, mentre $T_a = 10$ °K. Pertanto:

$$A_d = 32.4 + 20 \log_{10} f(\text{MHz}) + 20 \log_{10} d(\text{Km}) - G_T - G_R = 32.4 + 66 + 91.12 - 80 = 109.5 \text{ dB}$$

e quindi, utilizzando il valore W_{dR} ottenuto per il caso B, otteniamo

$$W_{dT} = W_{dR} + A_d = -54.3 + 109.5 = 55.2 \text{ dBm} = 25.2 \text{ dBW} \rightarrow 331 \text{ Watt}$$

3) Nel caso di adattamento, la potenza ceduta all'antenna T_x è proprio quella disponibile del generatore, e quindi si ha $W_{dT} = \frac{\sigma_g^2}{4R}$, da cui

$$\sigma_g = \sqrt{W_{dT} 4R} = \sqrt{66 \cdot 10^{-3} \cdot 4 \cdot 50} = 3.63 \text{ Volt.}$$

In caso di disadattamento, desiderando che la potenza ceduta all'antenna trasmittente rimanga la stessa, e supponendo le impedenze indipendenti dalla frequenza, scriviamo (in accordo alla relazione (14.1))

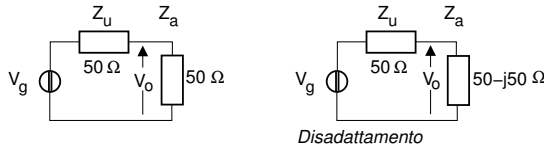
$$W_T = P_{v_o} \frac{R_a}{|Z_a|^2} = P_{v_o} \frac{50}{50^2 + 50^2} = P_{v_o} \cdot 10^{-2}$$

e quindi $P_{v_o} \simeq 6.6 \text{ (Volt}^2\text{)}$. Applicando ora la regola del partitore, si ottiene

$$P_{v_o} = P_{v_g} \left| \frac{Z_a}{Z_a + Z_u} \right|^2 = P_{v_g} \left| \frac{50 - j50}{50 + 50 - j50} \right|^2 = P_{v_g} \frac{50^2 + 50^2}{100^2 + 50^2} = P_{v_g} \cdot 0.4.$$

Dunque, $P_{v_g} = \frac{P_{v_o}}{0.4} = \frac{6.6}{0.4} = 16.5 \text{ Volt}^2$, ovvero $V_{g_{eff}} = \sqrt{16.5} \simeq 4 \text{ Volt}$.

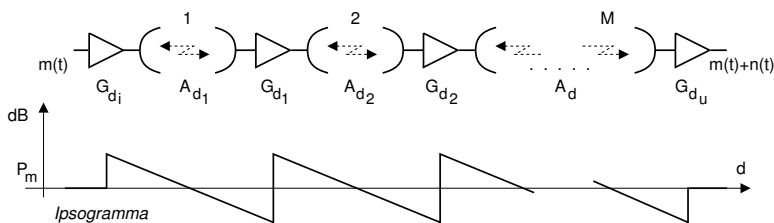
Evidentemente, il disadattamento produce un innalzamento del valore efficace, se si vuol mantenere la stessa potenza di uscita.



16.4 Rumore nei ripetitori

Analizziamo il problema con riferimento ad un collegamento radio, anche se la trattazione può essere estesa ad altre tecniche trasmissive. Esaminiamo il caso in cui il collegamento da effettuare sia molto lungo, tanto da impedirne la realizzazione mediante un'unica tratta, od a causa dell'eccessiva attenuazione disponibile, oppure per la mancanza di condizioni di visibilità. In tal caso, occorre suddividere il collegamento in più tratte (in numero di M) dimensionate in modo tale che ognuna ripristini il livello di segnale ad un valore pari a quello in ingresso alla tratta stessa (tranne ovviamente l'ultima).

Tra ogni coppia di tratte si trova un ripetitore, che amplifica il segnale in misura pari al proprio guadagno disponibile $G_{d_i} = \frac{1}{A_{d_i}}$, posto cioè pari all'inverso dell'attenuazione disponibile della tratta precedente. Il rumore termico accumulato alla fine



del collegamento può calcolarsi con i metodi tradizionali, ma considerando che il livello di segnale è lo stesso per tutti i ripetitori, si ritrova il risultato ottenuto al § 14.8.1, come ora illustreremo. Faremo quindi notare l'influenza della distorsioni di non linearità nel progetto.

16.4.1 Rumore termico accumulato

Osservando solamente il segnale in uscita dall'ultimo ripetitore, si può definire un SNR complessivo come $SNR_T = \frac{P_m}{P_n}$. D'altra parte, il rumore complessivo è dovuto ai contributi di rumore introdotti dai singoli ripetitori: essendo questi ultimi indipendenti tra loro, la potenza di rumore accumulata è la somma delle singole potenze di rumore:

$$P_n = \sigma_n^2 = E \{ n^2(t) \} = E \left\{ \left(\sum_i n_i(t) \right)^2 \right\} = E \left\{ \sum_i n_i^2(t) \right\} = \sum_i \sigma_{n_i}^2 = \sum_{i=1}^M P_{n_i}$$

Osserviamo ora che per ogni singolo ripetitore può essere definito un proprio $SNR_i = \frac{P_{m_i}}{P_{n_i}}$, e quindi $P_{n_i} = \frac{P_{m_i}}{SNR_i}$; pertanto, l' SNR complessivo risulta: $SNR_T = \frac{P_m}{\sum_i \frac{P_{m_i}}{SNR_i}}$.

Notiamo inoltre che se esprimiamo i singoli contributi P_{n_i} nei termini di uno stesso livello di segnale, nell'ultima espressione è possibile considerare $P_{m_i} = P_m$ per $\forall i$, e dunque per l' SNR complessivo si ottiene

$$SNR_T = \frac{P_m}{P_m \sum_i \frac{1}{SNR_i}} = \frac{1}{\sum_i \frac{1}{SNR_i}}$$

Questo risultato può essere espresso con la frase

l' SNR prodotto da più cause indipendenti è il parallelo degli SNR dovuti alle diverse cause di rumore

per via della analogia formale con l'espressione della resistenza equivalente di un parallelo di resistenze; l'analogia evidenzia, tra l'altro, che se una tratta è considerevolmente peggiore delle altre, SNR_T dipenderà essenzialmente da questa.

Il risultato a cui siamo giunti ha validità più generale del caso illustrato, e può essere invocato ogni volta che un sistema di comunicazione è affetto da più cause di disturbo additivo indipendenti tra loro, per ognuna delle quali si sia separatamente in grado di giungere ad una espressione di SNR .

Proseguiamo l'analisi ipotizzando ora che tutte le tratte siano uguali tra loro, ovvero con uguali A_d e G_d , uguali temperature di rumore, ed uguali SNR_i . In tal caso si ottiene

$$SNR_T = \frac{1}{\frac{M}{SNR_i}} = \frac{SNR_i}{M}$$

con $SNR_i = \alpha SNR_0 = \alpha \frac{\mathcal{P}_R}{\mathcal{P}_n}$, dove $\mathcal{P}_n = kT_{e_i}W$ è la potenza di rumore nella banda di messaggio W , \mathcal{P}_R è la potenza ricevuta da un ripetitore (uguale per tutti se le tratte sono uguali), e α è un fattore che dipende dal tipo di modulazione. Sembra dunque che per migliorare l' SNR complessivo sia sufficiente elevare il livello di trasmissione di tutti gli stadi, in modo da elevare la potenza ricevuta. In realtà la potenza trasmessa non può aumentare a piacere, in quanto intervengono fenomeni di non-linearità.

16.4.2 Compromesso tra rumore termico e di intermodulazione

Al § 14.7 si è osservato come per un segnale modulato la presenza di un elemento a comportamento non lineare (tipicamente l'amplificatore di potenza del trasmettitore) produce *interferenza in banda*, la cui potenza dipende con legge cubica dalla potenza del segnale trasmesso. Indicando quindi con $SNR_I = \frac{\mathcal{P}_m}{\mathcal{P}_I}$ il rapporto SNR complessivo del collegamento dovuto a cause di non linearità, *indipendenti dal rumore termico*, osserviamo che questo *diminuisce* all'aumentare della potenza trasmessa da ogni ripetitore. Pertanto, l' SNR complessivo deve tener conto sia del rumore termico che di quello di intermodulazione, e può essere espresso come "il parallelo" di entrambi, ossia

$$SNR = \frac{1}{\frac{1}{SNR_T} + \frac{1}{SNR_I}}$$

Dato che all'aumentare della potenza di trasmissione SNR_T diminuisce mentre SNR_I aumenta, l' SNR complessivo presenta *un massimo* per un certo valore di potenza trasmessa, ovvero esiste un dimensionamento *ottimo* in grado di fornire il miglior SNR complessivo.

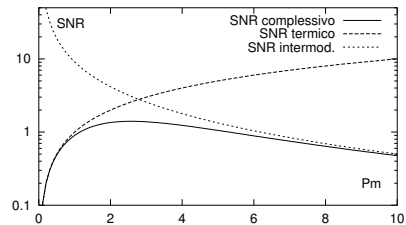
Esempio La figura a lato mostra l'andamento di

$$SNR = \frac{1}{\frac{1}{SNR_T} + \frac{1}{SNR_I}}$$

dovuto ai due termini

$$SNR_T = \mathcal{P}_m \text{ e } SNR_I = \frac{\mathcal{P}_m}{.1 \cdot \mathcal{P}_m^2 + .01 \cdot \mathcal{P}_m^3}$$

Come si vede, SNR presenta un massimo per $\mathcal{P}_m \simeq 2.5$.



Capitolo 17

Teoria dell'informazione e codifica

In questo capitolo sono approfonditi gli aspetti teorici relativi alla codifica di *canale* e di *sorgente* accennati al § 1.2, ovvero come *proteggere* dagli errori di trasmissione l'informazione trasmessa, e come *ridurre* la mole di dati da trasmettere senza pregiudicare la qualità del messaggio. Sono inoltre investigati *i limiti* entro i quali è possibile operare nel senso ora definito, e cioè viene stabilito quale sia *il massimo* teorico della velocità di trasmissione dell'informazione su un determinato canale, indicato come la sua *capacità*, e quale sia *la minima* velocità di trasmissione con cui è possibile rappresentare una determinata sorgente, indicata come la sua *entropia*. Infine, vengono illustrate alcune possibili tecniche con cui approssicare i suddetti limiti nel caso di sorgenti e canali numerici, mentre gli aspetti relativi alle sorgenti analogiche sono rimandati al capitolo seguente.

17.1 Codifica di sorgente

Una fonte informativa (o sorgente) può essere per sua natura di tipo discreto, come nel caso di un documento scritto, o di tipo continuo, come nel caso di un segnale analogico, ad esempio audio e video. In base a considerazioni di tipo statistico, la sorgente può essere caratterizzata da una grandezza, l'*entropia*, che indica il tasso di informazione (in bit/secondo) *intrinseco* per i messaggi prodotti dalla sorgente; d'altra parte, la descrizione in modo nativo dei messaggi prodotti può risultare in una *velocità di trasmissione* ben superiore!

Lo scopo della *codifica di sorgente* è quello di individuare rappresentazioni alternative per le informazioni prodotte dalla sorgente, in modo da ridurre la quantità di bit/secondo necessari alla trasmissione, a valori quanto più possibile prossimi a quelli dell'entropia, sfruttando le caratteristiche della sorgente, del processo di codifica, e del destinatario dei messaggi, ovvero

- la particolare distribuzione statistica dei simboli o dei valori emessi dalla sorgente, che può permettere l'uso di un minor numero di bit per rappresentare i simboli *più frequenti* di altri;
- la dipendenza statistica presente tra simboli successivi, ovvero la presenza di un fenomeno di memoria intrinseco della sorgente, che rende possibile entro certi limiti *la predizione* dei valori futuri;

- l'introduzione di un *ritardo di codifica*, che permette di analizzare un intero intervallo temporale del messaggio;
- nel caso di segnali multimediali, l'esistenza di *fenomeni percettivi* legati alla fisiologia dell'apparato sensoriale, che possono guidare il codificatore nella scelta delle componenti di segnale da sopprimere, in quanto percettivamente non rilevanti.

Nel caso di sorgenti nativamente discrete, come ad esempio documenti in formato elettronico, lo scopo della codifica di sorgente è quello di permettere la ricostruzione *integrale* di quanto trasmesso, che dunque viene detta *senza perdita di informazione*. Nel caso invece di sorgenti continue, si ottiene una sequenza numerica a seguito di un processo di campionamento e quantizzazione, che determina l'insorgenza di una prima causa di *distorsione* nel messaggio ricostruito, mentre la velocità binaria effettiva può essere ulteriormente ridotta grazie allo sfruttamento dei fenomeni percettivi: in tal caso il risultato della codifica viene detto *con perdita di informazione*.

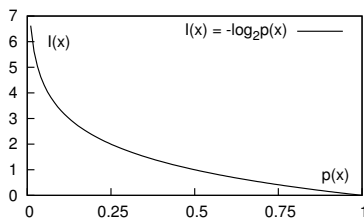
17.1.1 Codifica di sorgente discreta

Sorgente senza memoria Prendiamo in considerazione una sorgente discreta e stazionaria, che emette una sequenza $x(n)$ composta di simboli x_k appartenenti ad un alfabeto di cardinalità L (ossia con $k = \{1, 2, \dots, L\}$), ognuno contraddistinto dalla probabilità di emissione $p_k = Pr(x_k)$. Il termine *senza memoria* si riferisce al fatto che, se indichiamo con x_h, x_k una coppia di simboli emessi uno dopo l'altro (ossia $x(n) = x_h, x(n+1) = x_k$), la probabilità del simbolo emesso per secondo non dipende dall'identità di quello(i) emesso precedentemente, ossia $p(x_k/x_h) = p(x_k) = p_k$.

Misura dell'informazione Definiamo informazione (espressa in *bit*) associata all'osservazione del simbolo x_k il valore¹

$$I_k = I(x_k) = \log_2 \frac{1}{p_k} = -\log_2 p_k \text{ bit}$$

che rappresenta il grado di incertezza a riguardo del verificarsi dell'evento x_k (prima che questo si verifichi), ovvero di quanto possiamo ritenerci sorpresi nel venire a conoscenza dell'evento x_k , di cui riteniamo di conoscere la probabilità p_k . Osserviamo infatti che per come è fatta la funzione logaritmo, a bassi valori di probabilità è associata una informazione elevata. La scelta di usare il logaritmo in base 2 produce i seguenti risultati:



Prob. p_k	Informazione $-\log_2 p_k$	Commento
1	0	L'evento certo non fornisce informazione
0	∞	L'evento impossibile dà informazione infinita
$\frac{1}{2}$	1	In caso di scelta binaria (es. testa o croce) occorre una cifra binaria (<i>bit = binary digit</i>) per indicare il risultato

¹Per calcolare il logaritmo in base 2, si ricordi che $\log_2 \alpha = \frac{\log_{10} \alpha}{\log_{10} 2} \simeq 3.32 \log_{10} \alpha$.

17.1.1.1 Entropia

Come in termodinamica al concetto di entropia si associa il grado di *disordine* in un sistema, così per una sorgente informativa l'entropia misura il livello di *casualità* dei simboli emessi. Definiamo quindi *entropia* (indicata con H) di una sorgente discreta S , il valore atteso della quantità di informazione apportata dalla conoscenza dei simboli (scelti tra L possibilità) da essa generati

$$H_s = E \{I_k\} = \sum_{k=1}^L p_k I_k = \sum_{k=1}^L p_k \log_2 \frac{1}{p_k} \text{ bit/simbolo} \quad (17.1)$$

che pesando in probabilità il valore di informazione associato ai diversi simboli, rappresenta il *tasso medio* di informazione per simbolo delle sequenze osservabili. Come mostrato nell'esercizio che segue, risulta che:

- se i simboli sono *equiprobabili* ($p_k = \frac{1}{L}$ con $\forall k$), la sorgente è *massimamente informativa*, e la sua entropia è la massima possibile per un alfabeto ad L simboli, e pari a $H_{s_{Max}} = \frac{1}{L} \sum_{k=1}^L \log_2 L = \log_2 L$ bit/simbolo;
- se i simboli non sono equiprobabili, allora $H_s < \log_2 L$;
- se la sorgente emette sempre e solo lo stesso simbolo, allora $H_s = 0$.

Da queste osservazioni discende l'espressione riassuntiva

$$0 \leq H_s \leq \log_2 L \quad (17.2)$$

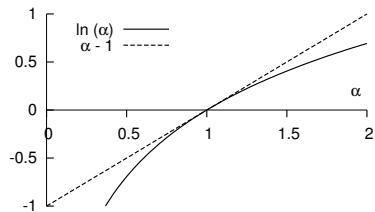
Esercizio Per dimostrare formalmente questo risultato, osserviamo innanzitutto che $H_s \geq 0$ in quanto la (17.1) comprende tutti termini positivi o nulli, essendo $\log_2 \alpha \geq 0$ per $\alpha = 1/p_k \geq 1$. Quindi, mostriamo che $H_s - \log_2 L \leq 0$. Innanzitutto riscriviamo il primo membro della disuguaglianza come

$$H_s - \log_2 L = \sum_k p_k \log_2 \frac{1}{p_k} - \log_2 L \cdot \sum_k p_k = \sum_k p_k \log_2 \frac{1}{L \cdot p_k}$$

dato che $\sum_k p_k = 1$. Quindi, esprimiamolo in termini dei logaritmi *naturali*, tenendo conto che $\log_2 \alpha = \frac{\ln \alpha}{\ln 2}$, e quindi

$$\sum_k p_k \log_2 \frac{1}{L \cdot p_k} = \frac{1}{\ln 2} \sum_k p_k \ln \frac{1}{L \cdot p_k}$$

A questo punto utilizziamo la relazione $\ln \alpha \leq \alpha - 1$ mostrata in figura, con l'uguaglianza valida solo se $\alpha = 1$, ottenendo così



$$\begin{aligned} H_s - \log_2 L &= \frac{1}{\ln 2} \sum_k p_k \ln \frac{1}{L \cdot p_k} \leq \frac{1}{\ln 2} \sum_k p_k \left(\frac{1}{L \cdot p_k} - 1 \right) = \\ &= \frac{1}{\ln 2} \left(\sum_k \frac{1}{L} - \sum_k p_k \right) = \frac{1}{\ln 2} (1 - 1) = 0 \end{aligned}$$

con il segno di uguale solo se $\frac{1}{L \cdot p_k} = 1$ ovvero $p_k = \frac{1}{L}$.

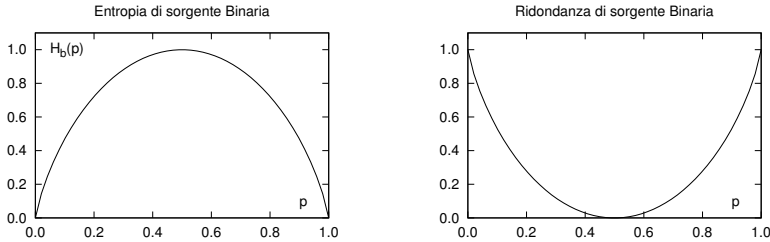


Figura 17.1: Entropia di sorgente binaria, e ridondanza associata

Entropia di sorgente binaria Nel caso particolare delle sorgenti binarie, che producono uno tra due simboli $\{x_0, x_1\}$ con probabilità rispettivamente $p_0 = p$, $p_1 = q = 1 - p$, la formula dell'entropia fornisce l'espressione

$$H_b(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \text{ bit/simbolo} \quad (17.3)$$

il cui andamento è mostrato nella figura 17.1, in funzione di p .

I due simboli $\{x_0, x_1\}$ possono essere rappresentati dalle 2 cifre binarie $\{0, 1\}$, che in questo caso chiamiamo *binit*, per non confonderli con la misura dell'informazione (il bit). Osserviamo quindi che se $p \neq .5$ risulta $H_b(p) < 1$, ossia la sorgente emette informazione con un tasso inferiore a un bit/simbolo, mentre a prima vista non potremmo usare meno di un binit per rappresentare ogni simbolo binario, introducendo una ridondanza pari a $1 - H_b(p)$ (graficata)².

Esempio Consideriamo il caso di una sorgente con $p_0 = 0.8$ e $p_1 = 0.2$. L'applicazione della (17.3) fornisce un valore $H_b(0.8) = .8 \log_2 \frac{1}{.8} + .2 \log_2 \frac{1}{.2} = 0.72$ bit/simbolo, minore del valore di 1 bit/simbolo che si sarebbe ottenuto nel caso di equiprobabilità.

Entropia di sorgente L-aria L'applicazione della (17.2) al caso di una sorgente che emette simboli *non* equiprobabili ed appartenenti ad un alfabeto di cardinalità L , determina per la stessa un valore di entropia $H_L < \log_2 L$ bit/simbolo: se i simboli sono codificati utilizzando $\lceil \log_2 L \rceil$ binit/simbolo³, otteniamo una ridondanza pari a $\lceil \log_2 L \rceil - H_L$.

Esempio Consideriamo il caso di una sorgente quaternaria con $p_0 = 0.5$, $p_1 = 0.25$, $p_2 = 0.125$, $p_3 = 0.125$. L'applicazione della (17.1) fornisce $H_4 = 1.75$ bit/simbolo, inferiore ai 2 bit/simbolo che si sarebbero ottenuti nel caso di simboli equiprobabili.

17.1.1.2 Intensità informativa e codifica binaria

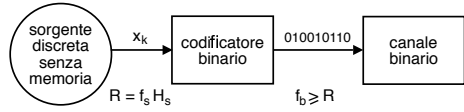
Continuiamo a riferirci ad una sorgente discreta e senza memoria, caratterizzata da una entropia H_s bit/simbolo, ed i cui simboli x_k sono emessi ad una frequenza f_s simboli/secondo: il *flusso informativo* prodotto è dunque caratterizzato da una *intensità* o *velocità informativa* pari a

$$R = f_s \cdot H_s \text{ bit/secondo}$$

²Si presti attenzione sulla differenza: la ridondanza della codifica *di sorgente* indica i binit/simbolo che eccedono il valore dell'entropia, mentre la ridondanza della codifica *di canale* indica il rapporto tra binit di protezione e quelli di effettivamente emessi dalla sorgente, come definito a pag. 98.

³La notazione $\lceil \alpha \rceil$ indica l'intero superiore ad α : ad esempio, se $\alpha = 3.7538$, si ha $\lceil \alpha \rceil = 4$.

Volendo trasmettere tale informazione attraverso un *canale binario* (vedi § 17.2.1), occorre che l'elemento indicato nella figura a lato come *codificatore binario* faccia corrispondere ad ogni simbolo x_k una univoca sequenza di N_k bit (nel seguito indicata come *codeword* appartenete ad un *codebook*)⁴, scelti in uno dei modi descritti nel seguito, producendo una *velocità di trasmissione* binaria di f_b *bin/bit/sec*. Dal punto di vista del canale, il messaggio è prodotto da una nuova sorgente, i cui simboli binari hanno probabilità p e $1-p$, e dunque caratterizzata da una entropia $H_b(p)$ *bit/bin/bit* ≤ 1 . Dato che l'intensità informativa in ingresso ed in uscita dal codificatore deve essere la stessa⁵, la velocità binaria f_b della nuova sorgente rispecchia il vincolo



$$f_b \geq f_b \cdot H_b(p) = f_s \cdot H_s = R$$

Il rapporto $\bar{N} = \frac{f_b}{f_s} \geq H_s$ rappresenta il numero *medio* di cifre binarie emesse per ciascun simbolo della sorgente, e può essere valutato a partire dalle probabilità p_k dei simboli x_k , come $\bar{N} = \sum_k p_k N_k$. Il primo teorema di Shannon⁶, o della *codifica di sorgente*, afferma che *esiste* un modo di scegliere gli N_k binit associati ai simboli x_k tale che⁷

$$H_s \leq \bar{N} \leq H_s + \epsilon \quad (17.4)$$

con ϵ piccolo a piacere, e che si annulla in corrispondenza della codifica *ottima*, per la quale risulta $\bar{N} = H_s$. Ma non dice come fare.

Rendimento del codice Rappresenta l'efficienza del processo di codifica binaria, ed è definito come

$$\eta = \frac{H_s}{\bar{N}} = \frac{R}{f_b} = H_b(p) \leq 1$$

ovvero quanti bit di sorgente sono trasportati da ogni binit di codifica⁸. D'altra parte, sappiamo già che se i binit emessi a velocità f_b assumono i valori 0 o 1 in modo equiprobabile, allora $H_b(\frac{1}{2}) = 1$, ovvero $f_b = R$ e $\bar{N} = H_s$, e dunque il problema di individuare un codice ottimo sembrerebbe quello di trovare un insieme di *parole di codice* (dette CODEWORD) tali da rendere le cifre binarie equiprobabili, con il vincolo di mantenere il codice *decifrabile*, ovvero tale da rispettare la *regola del prefisso*. Ma andiamo con ordine.

17.1.1.3 Codifica con lunghezza di parola variabile

Mostriamo mediante un esempio come scegliendo codeword *più lunghe* per rappresentare i simboli *meno probabili*, e *più corte* per i simboli *più frequenti*, si può subito ottenere un miglior rendimento del codice. Consideriamo infatti una sorgente con

⁴Mettere in corrispondenza i diversi simboli di sorgente con una loro codifica binaria è detta *codifica per blocchi*, discussa al § 17.1.1.4, dove si mostra anche la possibilità di emettere le *codeword* in corrispondenza di più di un simbolo di sorgente.

⁵Essendo la corrispondenza tra x_k ed N_k univoca, non vi è perdita di informazione.

⁶http://it.wikipedia.org/wiki/Primo_teorema_di_Shannon

⁷In effetti la (17.4) sussiste qualora il codificatore non operi indipendentemente su ogni simbolo di sorgente, ma più in generale possa emettere i binit in corrispondenza di sequenze di x_k *via via più lunghe*.

⁸Ad esempio, un valore $\eta = 0.33$ indica che ogni binit trasporta solo $1/3$ di bit.

alfabeto di cardinalità $L = 4$, ai cui simboli competono le probabilità riportate alla seconda colonna della tabella. In questo caso l'entropia vale

$$\begin{aligned} H_s &= \sum_k p_k \log_2 \frac{1}{p_k} \\ &= \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + \frac{2}{8} \log_2 8 = \frac{1}{2} + \frac{1}{2} + \frac{2}{8} \cdot 3 = 1.75 \text{ bit/simbolo} \end{aligned}$$

Simbolo	Prob.	Codeword	N_k
x_1	.5	0	1
x_2	.25	10	2
x_3	.125	110	3
x_4	.125	111	3

Se il codificatore di sorgente adotta le corrispondenze mostrate nella terza colonna della tabella, a cui competono le lunghezze in binit riportate quarta colonna, il numero medio di binit/simbolo prodotto dalla codifica binaria risulta pari a

$$\bar{N} = E\{L\} = \sum_k N_k p_k = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{2}{8} = 1.75 \text{ binit/simbolo}$$

Il risultato ottenuto $H_s = \bar{N}$, ovvero il migliore possibile, non è per nulla scontato, e mostriamo appresso come dipenda sia dalla particolare scelta fatta per le codeword, sia dal particolare tipo delle p_k dell'esempio, tutte potenze negative di due (essendo $0.5 = 2^{-1}$, $0.25 = 2^{-2}$, $0.125 = 2^{-3}$).

Regola del prefisso Perché un insieme di codewords possa costituire un *codice di sorgente*, esse devono poter essere riconosciute come *distinte* presso il ricevitore, e questo è possibile a patto che nessuna di esse sia *uguale all'inizio* di una codeword più lunga. Si può mostrare che ciò è possibile purché le lunghezze N_k delle codeword soddisfino la *diseguaglianza di Kraft*, espressa come

$$K = \sum_{k=1}^L 2^{-N_k} \leq 1 \quad (17.5)$$

Esempio Nella tabella sono riportati quattro possibili codici (A, B, C, D) per la sorgente quaternaria già discussa, assieme al corrispettivo valore di \bar{N} e K .

Simb.	p_k	A	B	C	D
x_1	.5	00	0	0	0
x_2	.25	01	1	01	10
x_3	.125	10	10	011	110
x_4	.125	11	11	0111	111
\bar{N}		2.0	1.25	1.875	1.75
K		1.0	1.5	0.9375	1.0

Il codice A corrisponde ad un codificatore particolarmente banale con $N_k = \bar{N}$ per tutti i k , dunque la (17.5) diviene $K = L2^{-\bar{N}} \leq 1$ ed è soddisfatta a patto che $\bar{N} \geq \log_2 L$: nel nostro caso essendo $L = 4$ ed $\bar{N} = 2$ si ottiene $\bar{N} = \log_2 L = 2$ e quindi $K = 1$, dunque il codice è decifrabile (anche perché a lunghezza fissa), ma non particolarmente efficiente, in quanto $\frac{H_s}{\bar{N}} \leq \frac{H_s}{\log_2 L}$ e nel nostro caso, anche se vale l'eguaglianza, si ottiene $\frac{H_s}{\bar{N}} = 0,875 < 1$; in questo caso, o in generale quando $H_s < \log_2 L$, si può realizzare una efficienza migliore ricorrendo ad un codice a lunghezza variabile.

Le codeword del codice B producono un valore $K = 1.25 > 1$, e dunque rappresentano un codice ambiguo⁹: difatti violano sia la regola del prefisso, che il limite (17.5). Il codice C invece è non ambiguo¹⁰, essendo $K < 1$, ma presenta una efficienza $\frac{H_s}{N} = 0,9\bar{3} < 1$ e dunque subottimale. Infine, il codice D è quello analizzato al precedente paragrafo, ed effettivamente risulta una scelta ottima, dato che $\frac{H_s}{N} = 1$ e le sue codeword soddisfano la (17.5).

Codice ottimo Si può dimostrare che un codice ottimo, ossia per il quale $\frac{H_s}{N}=1$, dove soddisfare la (17.5) con il segno di eguale, e perché questo accada, è necessario che le probabilità di simbolo abbiano valori $p_k = 2^{-N_k}$. In tal caso infatti, risultando $N_k = \log_2 \frac{1}{p_k}$, l'espressione che calcola $\bar{N} = \sum_k p_k N_k$ coincide con quella che fornisce $H_s = \sum_k p_k \log_2 \frac{1}{p_k}$, ovvero ogni simbolo è codificato con una parola di codice lunga tanti binit quanti sono i bit di informazione che trasporta. Per individuare un codice che *si avvicini* a questa proprietà si può realizzare la tecnica di *Huffman* presentata appresso, mentre per *modificare* le p_k si ricorre alla *codifica per blocchi* di simboli.

Codice di Huffman Si tratta di un algoritmo che permette di realizzare un codice a lunghezza variabile che soddisfa la regola del prefisso, adotta codeword più lunghe ai simboli meno probabili, ed uniforme, per quanto possibile, la probabilità delle cifre binarie. Si basa¹¹ sulla costruzione di un albero binario, i cui rami sono etichettati con 1 e 0, e può essere descritto come segue:

- crea una lista contenente i simboli della sorgente, ordinati per valore di probabilità decrescente, ed associa ad ognuno di essi un nodo-foglia dell'albero;
- finché c'è più di un nodo nella lista:
 - rimuovi dalla lista i due nodi con la probabilità più bassa;
 - crea un nuovo nodo interno all'albero con questi due nodi come figli, e con probabilità pari alla somma delle loro probabilità;
 - aggiungi il nuovo nodo alla lista;
- il nodo rimanente è la radice, e l'albero è completo;
- assegna cifre binarie diverse ad ogni coppia di rami a partire dalla radice, concatenando le quali ottieni le codeword per i simboli sulle foglie

Si può dimostrare che il codice di Huffman generato in questo modo è il migliore possibile nel caso in cui la statistica dei simboli di sorgente sia nota a priori, nel senso che produce una codifica con il minor numero possibile di binit/simbolo medi. La codifica di Huffman è ampiamente utilizzata nel contesto di altri metodi di compressione (metodo DEFLATE di PKZIP) e di codec multimediali (JPEG e MP3), in virtù della sua semplicità, velocità, ed assenza di brevetti.

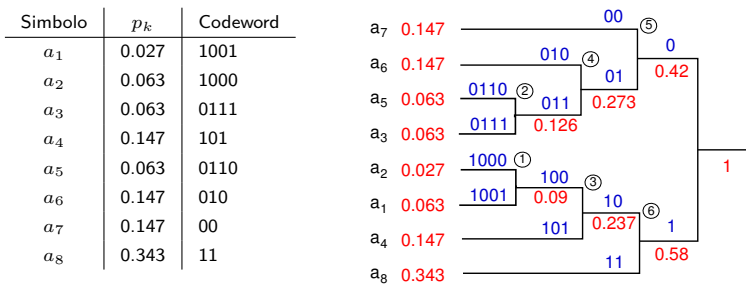
Ovviamente ci deve essere un accordo a priori tra sorgente e destinatario a riguardo delle corrispondenze tra parole di codice e simboli (o blocchi di simboli) della sorgente. Nel caso in cui ciò non sia vero, oppure nel caso in cui la statistica dei simboli della sorgente sia stimata a partire dal materiale da codificare, occorre inviare all'inizio della comunicazione anche la tabella di corrispondenza, eventualmente in forma a sua volta codificata.

⁹Ad esempio, la sequenza 10110010 potrebbe essere interpretata come $x_3x_4x_1x_1x_3$ oppure $x_2x_1x_4x_1x_1x_2x_1$ o $x_3x_2x_2x_1x_1x_3$

¹⁰Nonostante il codice C non soddisfi la regola del prefisso, non è ambiguo in quanto lo zero indica comunque l'inizio di una *nuova* codeword.

¹¹Vedi http://en.wikipedia.org/wiki/Huffman_coding

Esempio una sorgente con $L = 8$ simboli è caratterizzata dalle probabilità di simbolo riportate in figura, a partire dalle quali si realizza un codice di Huffman mediante la costruzione grafica riportata, in cui le probabilità sono scritte sotto i rami ed accanto ai simboli, mentre i binit sopra. Dopo aver ordinato i simboli in base alle probabilità, si individuano i due nodi a prob.



più bassa come a_1 e a_2 , che assommano prob. 0.09; dunque la coppia ora meno probabile è a_3 con a_5 , che cumulano prob. 0.126. Quindi, le due prob. minori divengono quelle di a_4 e della coppia a_1a_2 , per un totale di 0.237, e poi è il turno di raggruppare a_6 con a_3a_5 , per un totale di 0.273. Quest'ultimo nodo viene allora combinato con a_7 con prob. complessiva 0.42, quindi è il turno di a_8 con $a_1a_2a_4$ che totalizzano 0.58, ed infine questo nodo è collegato con il (5). A questo punto, partendo dalla radice a destra, si assegna un binit pari a 0 o 1 ad ogni coppia di rami, ripetendo l'assegnazione per le coppie discendenti, fino al risultato mostrato.

Dynamic Huffman coding Questa variante permette di costruire e modificare l'albero di codifica¹² man mano che i simboli sono trasmessi. In questo modo, se un carattere è già presente nel codebook, viene trasmessa la codeword corrispondente, mentre se non lo è, viene trasmesso il codice del carattere, ed aggiornato il codebook. Lo stesso processo si svolge anche dal lato ricevente, permettendo una codifica in tempo reale, e l'adattamento a condizioni di variabilità nei dati. Ovviamente, il metodo inizia ad essere efficiente solo dopo aver accumulato sufficienti informazioni statistiche.

17.1.1.4 Codifica per blocchi

Riprendiamo la discussione iniziata a pag. 437 relativa al codice binario ottimo per una sorgente L -aria senza memoria con simboli a_k a probabilità p_k , notando che se la lunghezza N_k delle codeword viene scelta in modo tale che

$$\log_2 \frac{1}{p_k} \leq N_k \leq \log_2 \frac{1}{p_k} + 1 \tag{17.6}$$

si può mostrare che la condizione (17.5) è soddisfatta. Inoltre, moltiplicando i membri di (17.6) per p_k e sommando su k si ottiene

$$H_s \leq \bar{N} \leq H_s + 1$$

da cui si osserva che è possibile ottenere $\frac{H_s}{\bar{N}} \simeq 1$ solo se $H_s \gg 1$ oppure se $N_k \simeq \log_2 \frac{1}{p_k}$. Altrimenti, possiamo ricorrere al *trucco* di trasformare la nostra sorgente ad L simboli in una equivalente con L^n simboli, ottenuti concatenando n simboli della sorgente originaria. Dato che questi ultimi sono indipendenti, la nuova sorgente esibisce una entropia $H_s^{blocco} = nH_s$, e la regola di codifica (17.6) produce ora il risultato $nH_s \leq$

¹²Presso Wikipedia si trova una descrizione dell'algoritmo di VITTER http://en.wikipedia.org/wiki/Adaptive_Huffman_coding

$n\bar{N} \leq nH_s + 1$ in cui $n\bar{N}$ è il numero medio di binit per blocco. Dividendo per n , otteniamo infine

$$H_s \leq \bar{N} \leq H_s + \frac{1}{n} \quad (17.7)$$

che rappresenta una diversa forma del teorema (17.4) con $\epsilon = \frac{1}{n}$, e che permette di ottenere $\bar{N} \rightarrow H_s$ se $n \rightarrow \infty$, avvicinandosi alle condizioni di codifica ottima per qualsiasi distribuzione delle p_k .

Per applicare questo metodo ad un caso pratico, consideriamo una sorgente binaria senza memoria con le p_k mostrate in tabella, e raggruppiamone i simboli a coppie, a cui competono probabilità di emissione ottenute moltiplicando le probabilità originarie¹³, e quindi codifichiamo i nuovi simboli quaternari con il codice a lunghezza variabile già esaminato.

Simbolo	Prob.	Codeword
x_1	.8	1
x_2	.2	0
x_1x_1	.64	0
x_1x_2	.16	10
x_2x_1	.16	110
x_2x_2	.04	111

Mentre il valore dell'entropia della sorgente binaria originale è ancora quello calcolato nel primo esempio a pag. 434 e pari a $H_b = 0.72$ bit a simbolo, la lunghezza media del nuovo codice risulta ora

$$\bar{N} = 1 \cdot 0.64 + 2 \cdot 0.16 + 3 \cdot 0.16 + 3 \cdot 0.04 = 1.58 \text{ binit}$$

ogni 2 simboli, ossia pari ad una media di 0.79 binit/simbolo binario, effettivamente più vicina al valore di $H_b = 0.72$.

Compromesso velocità-ritardo Come indicato dalla (17.7), realizzando blocchi via via più lunghi è possibile ridurre la velocità media di codifica \bar{N} (in binit/simbolo) rendendola sempre più vicina all'entropia, ovvero

$$\min [\bar{N}] = H_s + \epsilon$$

in cui $\epsilon \rightarrow 0$ se la lunghezza del blocco tende ad infinito. D'altra parte, all'aumentare della dimensione del blocco aumenta di egual misura il *ritardo* che intercorre tra l'emissione di un simbolo e la sua codifica, e di questo va tenuto conto, nel caso sussistano dei vincoli temporali particolarmente stringenti sulla consegna del messaggio.

Riassumendo Qualora una sorgente discreta ad L simboli esibisca un valore di entropia inferiore a $\log_2 L$, la velocità binaria in uscita dal codificatore di sorgente può essere ridotta adottando una codifica a blocchi, e utilizzando per i nuovi simboli un opportuno codice di Huffman.

Esercizio Si ripeta il calcolo del numero medio di binit/simbolo, adottando lo stesso codice a lunghezza variabile usato finora, per codificare i simboli emessi dalla sorgente binaria Markoviana di primo ordine analizzata all'esempio seguente, e mostrare come in questo caso si riesca ad ottenere una velocità media pari a 0.72 bit/simbolo. Sperimentare quindi la costruzione di un codice di Huffman basato sul raggruppamento di tre simboli di sorgente, per verificare se ci si riesce ad avvicinare di più al valore limite indicato dalla entropia, come vedremo pari a 0.58 bit/simbolo.

¹³operazione resa possibile in conseguenza della indipendenza statistica tra i simboli binari

17.1.1.5 Sorgente con memoria

Rimuoviamo ora l'ipotesi di indipendenza statistica tra i simboli emessi. In questo caso indichiamo con $\mathbf{x} = \{x(1), x(2), \dots, x(N)\}$ una sequenza di N di simboli, la cui probabilità congiunta si calcola come

$$p(\mathbf{x}) = p(x_1) p(x_2/x_1) p(x_3/x_1, x_2) \dots p(x_N/x_1, x_2, \dots, x_{N-1}) \neq \prod_{k=1}^N p(x_k)$$

dato che appunto la dipendenza statistica comporta l'uso delle probabilità condizionali. In questo caso, l'espressione dell'entropia si modifica in

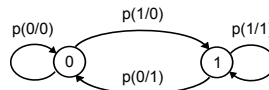
$$H_N = E_{\mathbf{x}} \{I(\mathbf{x})\} = -\frac{1}{N} \sum_{\text{tutti gli } \mathbf{x} \text{ possibili}} \sum \dots \sum p(\mathbf{x}) \log_2 p(\mathbf{x}) \text{ bit/simbolo}$$

in modo da eseguire la media statistica su tutte le possibili sequenze \mathbf{x} di lunghezza N . La grandezza H_N prende il nome di *entropia a blocco*, e si dimostra che al crescere di N il suo valore è *non crescente*, ossia $H_{N+1} \leq H_N \leq H_{N-1}$, mentre per $N \rightarrow \infty$, H_N tende ad un valore $H_{\infty} \leq H_s$, in cui l'uguaglianza è valida solo per sorgenti senza memoria.

Sorgente Markoviana Se oltre ad un certo valore $N_{Max} = M$ la sequenza H_N non decresce più, allora la sorgente è detta a *memoria finita* o *di Markov* di ordine M , caratterizzata dal fatto che le probabilità condizionate dipendono solo dagli ultimi M simboli emessi.

Esempio Analizziamo il caso di una sorgente binaria di Markov del primo ordine, per la quale sono definite le probabilità mostrate a lato, ed a cui corrisponde il diagramma di transizione mostrato. In questo caso, l'ultimo simbolo emesso determina lo stato della sorgente, condizionando così i valori delle probabilità di emissione di un

$$\begin{aligned} p(0/0) &= 0.9 & p(1/0) &= 0.1 \\ p(0/1) &= 0.4 & p(1/1) &= 0.6 \end{aligned}$$



nuovo simbolo: con i valori dell'esempio, si osserva come la sorgente *preferisca* continuare ad emettere l'ultimo simbolo prodotto, piuttosto che l'altro.

Sotto un certo punto di vista, è come se ora vi fossero L^M diverse sorgenti (nel caso dell'esempio, $2^1 = 2$), ognuna con delle probabilità di emissione differenti in base allo stato S_i (o memoria) in cui si trova. Perciò, in questo caso l'entropia di sorgente può essere calcolata applicando la (17.3) ad ognuno dei possibili stati, ottenendo dei valori di *entropia condizionata* $H(x/S_i)$, ed il valore dell'entropia di sorgente può essere ottenuto come *valore atteso* dell'entropia condizionata rispetto alle probabilità di trovarsi in ognuno degli stati del modello Markoviano.

Tornando all'esempio, i valori di entropia condizionata risultano pari a

$$H(x/0) = -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.47$$

$$H(x/1) = -0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.97$$

bit/simbolo, mentre il valore della probabilità di trovarsi in uno dei due stati si ottiene risolvendo il sistema

$$\begin{cases} p(0) &= p(0/0) p(0) + p(0/1) p(1) \\ 1 &= p(0) + p(1) \end{cases}$$

in cui la prima equazione asserisce che la probabilità di trovarsi in S_0 è pari alla somma di quella di esserci già, per quella di emettere ancora zero, più la probabilità di aver emesso uno, ed ora emettere zero. Sostituendo i valori, si ottiene $p(0) = 0.8$ e $p(1) = 0.2$, ossia gli stessi valori dell'esempio binario senza memoria. Ma mentre in quel caso il valore dell'entropia risultava pari a 0.72 bit/simbolo, ora si ottiene

$$H = p(0) H(x/0) + p(1) H(x/1) = 0.58 \text{ bit/simbolo}$$

mostrando come la presenza di memoria aumenti la predicibilità delle sequenze emesse dalla sorgente.

Esercizio Si ripeta il calcolo dell'entropia per un modello di Markov del primo ordine, caratterizzato dalle probabilità $p(0) = p(1) = 0.5$ e $p(1/0) = p(0/1) = 0.01$, mostrando che in questo caso si ottiene una entropia di 0.08 bit/simbolo.

17.1.1.6 Codifica per sorgenti con memoria

La discussione appena svolta ha evidenziato come nel caso di sorgenti con memoria i valori di entropia si riducano ulteriormente, e conseguentemente anche la velocità di codifica, a patto di accettare un maggior ritardo legato all'uso di codici a blocchi. A volte però la dimensione dei blocchi da prendere in considerazione può risultare eccessiva, producendo spropositate tabelle di codeword. Inoltre si può ritenere di *non* conoscere la statistica della sorgente, e di non desiderare effettuarne una stima, ed evitare di trasmettere la tabella di codeword. In questi casi, può essere opportuno adottare tecniche diverse dalle precedenti, come le due riportate appresso.

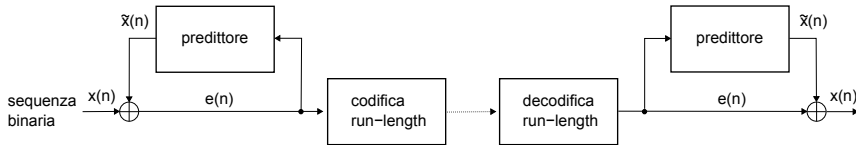
Codifica run-length Prendendo come esempio tipico il caso della trasmissione fax, in cui si ha a che fare con un segnale di immagine in bianco e nero, scansionato per righe, che è assimilabile ad una sorgente binaria che emetta uno zero per il bianco, ed un uno per il nero: per la natura delle immagini scansionate, tipicamente ci saranno lunghe sequenze di uni o di zeri, e dunque si può assumere valido un modello di sorgente Markoviano di primo ordine, con elevate probabilità condizionate di restare nello stesso stato.

Le lunghe sequenze di bit tutti uguali vengono dette *run* (corse), e la codifica *run-length* consiste effettivamente nel trasmettere una parola di codice che indica il numero (*length*) di questi bit uguali. In questo caso quindi, la codeword è di lunghezza fissa (ad esempio $k + 1$ binit, il primo dei quali indica se il run è tutto di uni o di zeri), e rappresenta un numero variabile (da 0 a $2^k - 1$) di binit di sorgente. Se ad esempio $k = 6$ binit, questi $6+1 = 7$ binit possono codificare fino a 64 bit di immagine: un bel risparmio!¹⁴

Codifica predittiva Questa ulteriore tecnica si basa sul fatto che un elevato grado di dipendenza statistica dei messaggi comporta la possibilità di *predire* in qualche modo i simboli a venire, in base all'identità di quelli già emessi. La differenza tra la sequenza predetta $\tilde{x}(n)$ e quella effettiva $x(n)$ è una nuova sequenza indicata come *errore di predizione* $e(n) = \tilde{x}(n) - x(n)$ che, se il predittore *ci azzecca* per la maggior parte del tempo, è quasi tutta nulla.

¹⁴In realtà, nel caso specifico del fax le cose non stanno esattamente in questi termini: infatti, anziché usare una parola di lunghezza fissa di k binit, l'ITU-T ha definito un apposito codebook <http://www.itu.int/rec/T-REC-T.4-199904-S/en> che rappresenta un codice di Huffman a lunghezza variabile, in modo da codificare le run length più frequenti con un numero ridotto di bit.

Nella figura seguente mostriamo l'applicazione della tecnica al caso di sequenze binarie, per le quali l'operazione di differenza è realizzata tramite una somma modulo due $e(n) = \tilde{x}(n) \oplus x(n)$, in modo che nel caso di predizione corretta, l'errore sia nullo; possiamo verificare l'effettiva invertibilità del processo di predizione binaria confrontando lo schema mostrato con quello proposto per la codifica differenziale al § 13.4.



Il predittore conserva uno *stato interno* che rappresenta gli ultimi bit di ingresso, in base ai quali determina¹⁵ la stima $\tilde{x}(n)$; l'errore $e(n)$ che è frequentemente nullo, e che viene sottoposto a codifica run-length, è re-inserito anche nel predittore, in modo che questo possa ri-determinare l'ultimo simbolo di ingresso $x(n) = e(n) \oplus \tilde{x}(n)$, ed aggiornare il proprio stato interno. La medesima formula di ricostruzione viene applicata anche in uscita del predittore di ricezione, che condivide con quello di trasmissione la conoscenza dello stato iniziale del segnale trasmesso, in modo da evolvere allo stesso modo.

Notiamo come la codifica run-length non preveda l'esistenza di un accordo a priori tra trasmettitore e ricevitore, a parte il comune stato di partenza ad inizio messaggio, mentre la codifica predittiva necessita solo di un accordo in merito alla struttura del predittore. Per contro, in presenza di errori di trasmissione i due predittori restano disallineati, finché non si inizia a co-decodificare un nuovo messaggio. Ma lo stesso problema è comune anche al caso di codifica a lunghezza di parola variabile, ed a quello di Huffman dinamico.

17.1.1.7 Compressione basata su dizionario

Nella comune accezione del termine, un dizionario è costituito da un *array* di stringhe, popolato con le parole esistenti per un determinato linguaggio. Anziché operare carattere per carattere, un codificatore di sorgente testuale può ricercare la posizione nel dizionario delle parole del messaggio, e quindi trasmettere l'indice della parola: per un dizionario di 25.000 termini bastano 15 bit di indirizzamento, ottenendo un rapporto di compressione variabile, in funzione della lunghezza della parola codificata.

Metodo di Lempel-Ziv-Welsh Per evitare di dover condividere la conoscenza dell'intero dizionario tra sorgente e destinatario, che tra l'altro potrebbe essere assolutamente sovradimensionato rispetto alle caratteristiche dei messaggi da trattare, il metodo LZW prevede che il codificatore generi il dizionario in modo graduale, man mano che analizza il testo, e che il decodificatore sia in grado di replicare questa modalità di generazione. Inoltre, il dizionario non è vincolato a contenere le reali *parole* del

¹⁵Il lettore più curioso si chiederà a questo punto, come è fatto il predittore. Molto semplicemente, *scommette* sul prossimo simbolo più probabile, in base alla conoscenza di quelli osservati per ultimi, ed ai parametri del modello markoviano: se questo simbolo possiede una probabilità a priori > 0.5 , allora la *maggioranza* delle volte la predizione sarà corretta, ed il metodo inizia a consentire una riduzione di velocità. Nel caso di sorgenti continue, troveremo invece alcune particolarità aggiuntive.

messaggio, ma semplicemente ospita le sequenze di caratteri effettivamente osservati, di lunghezza due, tre, quattro...

Operando su di un alfabeto ad L simboli, rappresentabili con $n = \lceil \log_2 L \rceil$ bit, il dizionario iniziale conterrà i simboli di sorgente alle prime L posizioni, e posti liberi

nelle restanti $2^n - L - 1$ posizioni¹⁶. Ogni carattere letto in ingresso viene accodato in una *stringa*, ed il risultato confrontato con le stringhe già presenti nel dizionario. Nel caso non si verifichi nessuna corrispondenza, viene aggiunta una nuova voce di dizionario, e quindi viene trasmesso il codice associato alla sua parte iniziale, escludendo cioè il simbolo concatenato per ultimo, e che ha prodotto

```
w = NIL;
while (read a char c) do
  if (wc exists in dictionary) then
    w = wc;
  else
    add wc to the dictionary;
    output the code for w;
    w = c;
  endif
done
output the code for w;
```

l'occorrenza della nuova voce. Nel caso invece in cui la stringa sia già presente (e questo in particolare è vero per la stringa di lunghezza uno corrispondente al primo simbolo analizzato) non si emette nulla, ma si continuano a concatenare simboli fino ad incontrare una stringa mai vista. Presso *Wikipedia*¹⁷ è presente un esempio di risultato della codifica.

La parte iniziale del testo, ovviamente, ha una alta probabilità di contenere tutte coppie di caratteri mai viste prima, e quindi in questa fase vengono semplicemente emessi i codici associati ai simboli osservati. Con il progredire della scansione, aumenta la probabilità di incontrare stringhe già osservate e sempre più lunghe. Ogni volta che viene esaurito lo spazio residuo per i nuovi simboli, viene aggiunto un bit alla lunghezza della codeword, ovvero viene raddoppiata la dimensione del vocabolario. Man mano che viene analizzato nuovo materiale, aumenta la lunghezza delle stringhe memorizzate nel dizionario, che riflette l'effettiva composizione statistica del documento in esame, ivi compresa la presenza di memoria; allo stesso tempo, la dimensione del dizionario (e la lunghezza delle codeword) resta sempre la minima indispensabile per descrivere il lessico effettivamente in uso. Alla fine del processo, il dizionario ottenuto viene aggiunto *in testa* al file compresso, seguito dalle codeword risultanti dall'algoritmo.

L'algoritmo LZW è usato nel programma di compressione Unix **compress**, per la realizzazione di immagini GIF e TIFF, ed incorporato in altri software, come ad esempio *Adobe Acrobat*.

Algoritmo Deflate L'ultimo metodo di compressione senza perdite che esaminiamo è quello che è stato introdotto dal programma PKZIP, e quindi formalizzato nella RFC 1951¹⁸, e tuttora ampiamente utilizzato per le sue ottime prestazioni e l'assenza di brevetti. Usa una variante dell'algoritmo LZW, al cui risultato applica poi una codifica di Huffman. *Deflate* opera su blocchi di dati con dimensione massima 64 Kbyte, ognuno dei quali può essere replicato intatto (come nel caso in cui i bit siano già sufficientemente imprevedibili), oppure essere compresso con un codice di Huffman statico, oppure ancora dinamico.

Per quanto riguarda la variante di LZW, essa consiste nel *non costruire* esplicitamente il dizionario, ma nell'usare invece *dei puntatori all'indietro* per specificare che

¹⁶Ad esempio con $L=96$ simboli si ha $n=7$ bit, ed un dizionario iniziale con 128 posizioni, di cui 96 occupate e 32 libere.

¹⁷<http://en.wikipedia.org/wiki/Lempel-Ziv-Welch>

¹⁸<http://tools.ietf.org/html/rfc1951>

una determinata sotto-stringa di ingresso, è in realtà la ripetizione di un'altra già osservata in precedenza. In questo caso, anziché emettere il codice (di Huffman) associato al byte corrente, si emette (il codice di Huffman del) la lunghezza della stringa da copiare, e la distanza (nel passato) della stessa. Quindi in pratica, anziché usare una codeword di lunghezza fissa per indicizzare gli elementi del dizionario come per LZW, viene usato un puntatore di lunghezza variabile, privilegiando le copie della sottostringa corrente più prossime nel tempo, oppure quelle con un maggior numero di caratteri uguali.

17.1.2 Codifica con perdite di sorgente continua

Come verrà mostrato ai § 17.2.3 e 17.2.4, un segnale $r(t)$ di potenza S ricevuto mediante un canale alla cui uscita sia presente un filtro con banda W ed rumore $n(t)$ gaussiano bianco di potenza N non può convogliare un tasso di informazione R più elevato della *capacità di canale* pari a $C = W \log_2 \left(1 + \frac{S}{N}\right)$ bit per secondo. Nel caso discreto e senza perdite trattato finora, C pone semplicemente un limite massimo al tasso di informazione che può essere trasmesso. Viceversa, nel caso di una sorgente continua $x(t)$, abbiamo due strade possibili: intraprendere un processo di campionamento e quantizzazione per produrre un segnale numerico con velocità $R \leq C$, incorrendo così in una distorsione di quantizzazione D tanto minore quanto maggiore è la R consentita, oppure effettuare una trasmissione analogica in cui

- il rumore al ricevitore può essere visto come una distorsione $d(t) = n(t) = x(t) - r(t)$ di potenza $D = N$
- l'entità della distorsione D può essere ridotta aumentando $\frac{S}{N}$ oppure W , e quindi C , permettendo così l'aumento del tasso di informazione R trasferito.

Nel caso continuo pertanto, sia che la sorgente sia quantizzata o che sia trasmessa come segnale analogico, sussiste un legame diretto tra la distorsione ed il tasso informativo, di cui discutiamo ora.

17.1.2.1 Curva velocità-distorsione

La valutazione della dipendenza tra D ed R si avvale di una serie di sviluppi teorici¹⁹ che prendono come caso-tipo quello di una sorgente $x(t)$ gaussiana, stazionaria, ergodica e bianca, con potenza σ_x^2 . In tal caso si ottiene che la *minima distorsione* conseguibile in corrispondenza di una velocità di trasmissione (*rate*) pari ad R , assume un valore pari a

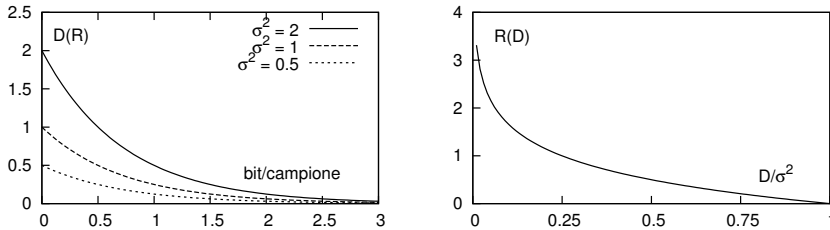
$$D(R)_G = 2^{-2R} \sigma_x^2 \quad (17.8)$$

Questa espressione risulta poi essere *il più grande valore minimo* possibile per una potenza σ_x^2 assegnata, dato che per sorgenti non gaussiane, oppure gaussiane ma non bianche, si possono ottenere valori inferiori. D'altra parte, invertendo la (17.8) si ottiene la curva $R(D)_G$, che descrive la minima velocità R necessaria per trasmettere i campioni di una sorgente gaussiana con distorsione D e potenza σ_x^2 assegnate

$$R(D)_G = \begin{cases} -\frac{1}{2} \log_2 \frac{D}{\sigma_x^2} & \text{se } 0 \leq D \leq \sigma_x^2 \\ 0 & \text{se } D \geq \sigma_x^2 \end{cases}$$

¹⁹http://en.wikipedia.org/wiki/Rate-distortion_theory

la cui seconda riga può essere interpretata osservando che, se la distorsione è superiore alla potenza di segnale, non occorre trasmettere proprio nulla, dato che tanto il ricevitore può ri-generare un segnale di errore, a partire da un rumore gaussiano di potenza D prodotto in loco.

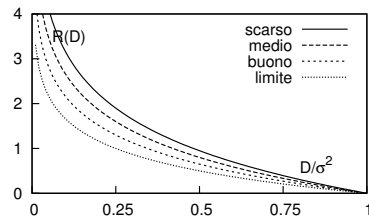


Valori limite Il valore $D(R)_G$ costituisce un *limite superiore* per ciò che riguarda la distorsione ottenibile ad una certa velocità R , utile per rapportare le prestazioni del codificatore della nostra sorgente con quelle migliori ottenibili per la sorgente più *difficile*, ossia la gaussiana. Allo stesso tempo, è definito un *limite inferiore* $D(R)_L$ (ossia, la minima distorsione sotto cui non si può scendere per un dato R) per sorgenti *non gaussiane e senza memoria*, in modo da poter scrivere

$$D(R)_L = 2^{-2R}Q \leq D(R) \leq D(R)_G = 2^{-2R}\sigma_x^2 \tag{17.9}$$

in cui Q è la *potenza entropica* (17.12) e risulta $Q < \sigma_x^2$ per sorgenti non gaussiane.

In definitiva, per una determinata sorgente per la quale sono disponibili diversi codificatori, potremmo ottenere una famiglia di curve del tipo di quelle mostrate in figura.



17.1.2.2 Entropia di sorgente continua

Nel caso discreto abbiamo apprezzato come l'entropia fornisca un potente strumento per valutare il tasso informativo intrinseco di una sorgente, supportando così la ricerca di metodi di riduzione della ridondanza al fine di avvicinare la velocità di trasmissione alla entropia effettiva. Ci chiediamo allora se anche nel caso continuo possa essere definita una entropia, e come questa possa aiutarci nello stabilire dei limiti prestazionali. Estendendo formalmente al caso continuo la definizione trovata per le sorgenti discrete, si ottiene l'espressione

$$h(X) = E \{-\log_2 p_x(x)\} = - \int p_x(x) \log_2 p_x(x) dx \tag{17.10}$$

che è indicata con la h minuscola per distinguerla dal caso discreto, e che viene detta *entropia differenziale* o *relativa* perché il suo valore può risultare positivo, negativo o nullo, in funzione della dinamica della variabile aleatoria X .

Esempio Se calcoliamo il valore di entropia differenziale per un processo i cui valori sono descritti da una variabile aleatoria a distribuzione uniforme $p_x(x) = \frac{1}{A} \text{rect}_A(x)$, otteniamo il risultato $h(X) = -\frac{1}{A} \int_{-\frac{A}{2}}^{\frac{A}{2}} \log_2\left(\frac{1}{A}\right) dx = \log_2 A$ il cui valore effettivo, appunto, dipende dal valore di A .

Sebbene inadatta ad esprimere il contenuto informativo *assoluto*²⁰ di una sorgente continua, l'entropia differenziale può comunque essere utile per confrontare tra loro due sorgenti con uguale varianza σ_x^2 ; in particolare, il massimo valore di $h(X)$ per σ_x^2 assegnata è ottenuto in corrispondenza di un processo gaussiano²¹, e risulta pari a

$$h(X)_G = \frac{1}{2} \log_2(2\pi e \sigma_x^2) > h(X) \quad (17.11)$$

ed è per questo motivo che, a parità di velocità R e di potenza σ_x^2 , il processo gaussiano incorre nella *massima distorsione minima* (17.8). Associata alla definizione di entropia differenziale, sussiste quella di *potenza entropica* Q , scritta come

$$Q = \frac{1}{2\pi e} 2^{2h(X)} \quad (17.12)$$

che per sorgenti gaussiane fornisce $Q_G = \sigma_x^2$, mentre per altri tipi di statistiche, si ottiene un valore minore. Applicando questa definizione alla (17.9) osserviamo come il limite inferiore di distorsione $D(R)_L$ si riduce al diminuire di $h(X)$.

17.1.2.3 Sorgente continua con memoria

Come per il caso di sorgenti discrete, anche per quelle continue l'esistenza di una dipendenza statistica tra i valori prodotti riduce la quantità di informazione emessa, al punto che a parità di distorsione, questa può essere codificata a velocità ridotta; oppure, a parità di velocità, può essere conseguita una distorsione inferiore. Anche stavolta, la sorgente più *difficile* (ossia a cui compete la *massima distorsione minima*) è quella gaussiana, per la quale risulta che la minima distorsione per σ_x^2 assegnata può esprimersi come

$$D(R)_G = 2^{-2R} \gamma_x^2 \sigma_x^2 \quad (17.13)$$

²⁰In effetti esiste una misura di entropia *assoluta* per sorgenti continue, che però ha la sgradevole caratteristica di risultare sempre infinita. Infatti, approssimando la (17.10) come limite a cui tende una sommatoria, e suddividendo l'escursione dei valori di x in intervalli uguali Δx , possiamo scrivere

$$\begin{aligned} h_{abs}(x) &= \lim_{\Delta x \rightarrow 0} \sum_i p(x_i) \Delta x \log_2 \frac{1}{p(x_i) \Delta x} \\ &= \lim_{\Delta x \rightarrow 0} \sum_i \left[p(x_i) \Delta x \log_2 \frac{1}{p(x_i)} + p(x_i) \Delta x \log_2 \frac{1}{\Delta x} \right] = h(x) + h_0 \end{aligned}$$

in cui $h(x)$ è proprio la (17.10) mentre $h_0 = -\lim_{\Delta x \rightarrow 0} \log_2 \Delta x \int_{-\infty}^{\infty} p(x) dx = -\lim_{\Delta x \rightarrow 0} \log_2 \Delta x = \infty$. D'altra parte, la differenza tra le entropie assolute di due sorgenti z e x risulta pari a $h_{abs}(z) - h_{abs}(x) = h(z) - h(x) + h_0(z) - h_0(x)$, in cui la seconda differenza tende a $-\log_2 \frac{\Delta z}{\Delta x}$ che, se z ed x hanno la medesima dinamica, risulta pari a zero.

²¹Questo risultato si ottiene massimizzando la (17.10) rispetto a $p(x)$ mediante il metodo dei moltiplicatori di LAGRANGE http://it.wikipedia.org/wiki/Metodo_dei_moltiplicatori_di_Lagrange, in modo da tener conto dei vincoli espressi dalle condizioni $\int p_x(x) dx = 1$ e $\int x^2 p_x(x) dx = \sigma_x^2$. Notiamo esplicitamente la differenza rispetto al caso continuo, in cui la d.d.p. che rende massima l'entropia, è invece quella uniforme.

in cui $0 \leq \gamma_x^2 \leq 1$ rappresenta una misura di *piattezza spettrale*²², che vale uno nel caso senza memoria, ovvero di processo *bianco*, e si riduce nel caso di un segnale i cui valori sono correlati tra loro, ed a cui corrisponde una densità spettrale *colorata*.

Nel caso *non gaussiano*, infine, la (17.13) si riscrive sostituendo al posto di σ_x^2 la potenza entropica Q espressa dalla (17.12), ottenendo valori $D(R)$ ancora inferiori.

L'applicazione dei principi relativi alla codifica di sorgente al caso specifico dei messaggi multimediali (audio e video) viene trattata al capitolo 18.

17.2 Codifica di canale

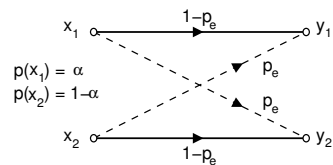
Come già fatto osservare al § 5.4.3, lo scopo della codifica di canale è quello di ridurre il tasso di errore di una trasmissione numerica ricorrendo alla aggiunta di ridondanza, ossia trasmettendo più binit di quanti necessari dopo l'attuazione della codifica di sorgente, e dunque (nel caso di una trasmissione *real-time*) occupare più banda dello stretto necessario.

E' quindi naturale chiedersi: fino a che punto si può arrivare, ovvero di quanto si può ridurre la P_e , e quanta ridondanza è necessario aggiungere? La risposta fornita in questa sezione è che, finché il tasso informativo R si mantiene inferiore al valore di una grandezza C denominata *capacità di canale*, definita ai § 17.2.3 e 17.2.4, il canale può trasportare l'informazione (teoricamente) *senza errori!* Mentre se al contrario $R > C$, non è possibile trovare nessun procedimento in grado di ridurre gli errori - che anzi, divengono praticamente *certi*. Infine per quanto riguarda la ridondanza che occorre aggiungere, pur senza spiegare come fare, la teoria assicura che questa può essere resa *trascurabile!* Ma prima di approfondire l'enunciazione di questi risultati, svolgiamo alcune riflessioni su come

- il processo di decisione svolto dal lato ricevente di un canale numerico può basarsi oltre che sulla conoscenza della statistica di *come avvengano* gli errori, anche su quella che descrive *come sono emessi* i simboli della sorgente;
- il verificarsi di errori costituisca una *perdita di informazione*.

17.2.1 Canale simmetrico binario e decisore Bayesiano

In figura è mostrato uno schema che rappresenta un canale numerico al cui ingresso si può presentare uno tra due simboli x_1 e x_2 , con probabilità rispettiva α e $1 - \alpha$, mentre in uscita si osserva il simbolo y_1 oppure y_2 .



²²Si può mostrare che γ_x^2 può essere interpretato come il rapporto tra la media aritmetica e la media geometrica della densità spettrale di potenza $\mathcal{P}_x(f)$ del processo $x(t)$: indicando con $S_k = \mathcal{P}_x(f_k)$, $k = 1, 2, \dots, N$, i campioni equispaziati della densità spettrale valutati a frequenze positive f_k tra zero e la massima frequenza del processo, si ha

$$\gamma_x^2 = \lim_{N \rightarrow \infty} \frac{\left(\prod_{k=1}^N S_k \right)^{1/N}}{\frac{1}{N} \sum_{k=1}^N S_k}$$

Nel caso di un processo bianco, per il quale i valori S_k sono tutti uguali, le due medie coincidono, e $\gamma_x^2 = 1$. Altrimenti, γ_x^2 risulta tanto più piccolo quanto più i valori S_k si discostano dal loro valore medio.

Il canale è detto *simmetrico* perché tali sono le probabilità condizionate *in avanti*: la probabilità di errore

$$p_e = p(y_2/x_1) = p(y_1/x_2)$$

e la probabilità (complementare) di non-errore

$$p_{ne} = 1 - p_e = p(y_1/x_1) = p(y_2/x_2)$$

Rapporto di verosimiglianza Qualora si osservi in uscita uno dei due valori (ad es. y_1), si possono confrontare le probabilità condizionate *in avanti* per le due possibili *ipotesi* che in ingresso sia presente x_1 od x_2 , valutando il *rapporto di verosimiglianza* (vedi § 7.5.3)

$$RV_{ML} = \frac{p(y_1/x_1)}{p(y_1/x_2)} = \frac{p_{ne}}{p_e} \quad (17.14)$$

e quindi si decide per l'ipotesi *più verosimile* in funzione del valore maggiore o minore di uno per RV_{ML} : ad esempio se $p_{ne} > p_e$ (ovvero $p_e < \frac{1}{2}$) si ha $RV_{ML} > 1$, e quindi si decide per x_1 , mentre nel caso fosse stato ricevuto y_2 , si sarebbe deciso per x_2 .

Se disponiamo della conoscenza delle probabilità *a priori* $p(x_1)$ e $p(x_2)$, ed i due simboli x_1 ed x_2 non sono equiprobabili²³, possiamo costruire il *rapporto di verosimiglianza* utilizzando le probabilità *a posteriori* $p(x_1/y_1)$ e $p(x_2/y_1)$, calcolabili applicando il teorema di Bayes (vedi § 7.1.4). Facendo di nuovo il caso di aver ricevuto il simbolo y_1 , il *rapporto di verosimiglianza* si scrive ora come

$$RV_{MAP} = \frac{p(x_1/y_1)}{p(x_2/y_1)} = \frac{p(y_1/x_1)p(x_1)}{p(y_1/x_2)p(x_2)} \cdot \frac{p(y_1)}{p(y_1)} \quad (17.15)$$

$$= \frac{p(y_1/x_1)p(x_1)}{p(y_1/x_2)p(x_2)} \quad (17.16)$$

Anche qui, RV_{MAP} può assumere valore $>$, $<$ od $=$ ad 1, a seconda di quale delle due probabilità *a posteriori* sia più grande, portando rispettivamente la decisione a favore di x_1 , x_2 , o l'indifferenza.

Verifica di ipotesi ML e MAP La metodologia ora descritta prende il nome di *verifica di ipotesi statistica* (vedi § 7.5.1), che si basa appunto sul confronto di quanto la grandezza osservata sia *verosimile*, compatibilmente con le *ipotesi* possibili. Se RV utilizza solamente le probabilità *in avanti* (17.14), la decisione si dice *di massima verosimiglianza* (vedi § 7.5.3) e viene indicata come *ML* o MAXIMUM LIKELIHOOD, mentre se si impiegano le probabilità *a posteriori* (17.15), si sta effettuando una *decisione di massima probabilità a posteriori*, indicata come MAP.

Riflessioni Il meccanismo con cui, nella decisione MAP, le probabilità *in avanti* si combinano con quelle *a priori*, può essere analizzato mediante alcune osservazioni: innanzi tutto, x_1 potrebbe essere *così raro* che, in presenza di una moderata probabilità di errore, il ricevitore potrebbe preferire di decidere sempre x_2 , attribuendo l'eventuale ricezione di y_1 dovuta più verosimilmente ad un errore del canale, piuttosto che all'effettiva trasmissione di x_1 . In assenza di canale poi, l'unico rapporto di verosimiglianza

²³In caso contrario (ovvero x_1 ed x_2 sono equiprobabili) la 17.15 è equivalente alla 17.14. Nei casi in cui *non si conosca* la statistica di sorgente, non si può quindi fare altro che attuare una *decisione di massima verosimiglianza*.

possibile sarebbe stato quello tra le probabilità a priori $p(x_1)$ e $p(x_2)$; la ricezione di un simbolo y_i dunque ha portato un miglior livello di informazione, alterando il RV , in misura tanto maggiore quanto più minore è la probabilità di errore.

Esempio Verifichiamo i ragionamenti appena svolti riscrivendo per esteso una probabilità a posteriori:

$$\begin{aligned} p(x_1/y_1) &= \frac{p(x_1, y_1)}{p(y_1)} = \frac{p(y_1/x_1)p(x_1)}{p(y_1/x_1)p(x_1) + p(y_1/x_2)p(x_2)} = \\ &= \frac{p_{ne} \cdot p(x_1)}{p_{ne} \cdot p(x_1) + p_e \cdot p(x_2)} \end{aligned}$$

Se $p_e = p_{ne} = \frac{1}{2}$, il canale è *inservibile* e non trasferisce informazione: infatti si ottiene $p(x_1/y_1) = p(x_1)$ in quanto $p(x_1) + p(x_2) = 1$. D'altra parte, se $p_e < p_{ne}$, risulta

$$p(x_1/y_1) = \frac{p(x_1)}{p(x_1) + \frac{p_e}{p_{ne}}p(x_2)} > p(x_1)$$

aumentando quindi la probabilità di x_1 rispetto alla sua probabilità a priori; se poi la probabilità di errore tende a zero ($p_e \rightarrow 0$), osserviamo che $p(x_1/y_1) \rightarrow 1$.

17.2.2 Informazione mutua media per sorgenti discrete

Abbiamo discusso di come l'entropia permetta di valutare la capacità informativa di una sorgente; estendiamo ora il concetto, per mostrare come *l'informazione condivisa* tra ingresso ed uscita di un canale consenta di determinare anche la quantità di informazione che viene persa a causa degli errori che si sono verificati.

Consideriamo una sorgente discreta che emette simboli x appartenenti ad un alfabeto finito di cardinalità L , ossia $x \in \{x_i\}$ con $i = 1, 2, \dots, L$, a cui è associata la distribuzione $p(x_i)$, ed indichiamo con $y \in \{y_j\}$ (con $j = 1, 2, \dots, L$) il simbolo ricevuto, in generale diverso da x , a causa di errori introdotti dal canale. Conoscendo le densità di probabilità $p(x_i)$, $p(y_j)$, e le probabilità congiunte $p(x_i, y_j)$, possiamo definire la quantità di informazione *in comune* tra x_i e y_j , denominata *informazione mutua*, come²⁴

$$I(x_i, y_j) = \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)} = \log_2 \frac{p(x_i/y_j)}{p(x_i)} = \log_2 \frac{p(y_j/x_i)}{p(y_j)} \quad \text{bit} \quad (17.17)$$

Notiamo che

- se ingresso ed uscita del canale sono indipendenti, si ha $p(x_i, y_j) = p(x_i)p(y_j)$, e quindi l'informazione mutua è *nulla*;
- se $p(y_j/x_i) > p(y_j)$, e quindi la conoscenza della emissione di x_i rende la ricezione di y_j più probabile di quanto non lo fosse a priori, allora l'informazione mutua è positiva;
- la definizione di informazione mutua è simmetrica, ovvero $I(x_i, y_j) = I(y_j, x_i)$.

Per giungere ad una grandezza $I(X, Y)$ che tenga conto del comportamento del canale per qualsiasi simbolo di sorgente e ricevuto, occorre pesare i valori di $I(x_i, y_j)$ con

²⁴Per ottenere (17.17) si ricordi che $p(x_i, y_j) = p(x_i/y_j)p(y_j) = p(y_j/x_i)p(x_i)$

le relative probabilità congiunte, ossia calcolarne il valore atteso rispetto a tutte le possibili coppie (x_i, y_j) , e quindi

$$I(X, Y) = E_{X, Y} \{I(x_i, y_j)\} = \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i/y_j)}{p(x_i)} \quad (17.18)$$

$$= \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(y_j/x_i)}{p(y_j)} \quad (17.19)$$

ottenendo così la quantità denominata *informazione mutua media*, misurata in bit/simbolo, e che rappresenta (in media) quanta informazione ogni simbolo ricevuto trasporta a riguardo di quelli trasmessi. In virtù della simmetria di questa definizione, ci accorgiamo²⁵ che il suo valor medio può essere espresso nelle due forme alternative

$$I(X, Y) = H(X) - H(X/Y) \quad (17.20)$$

$$= H(Y) - H(Y/X) \quad (17.21)$$

in cui l'entropia *condizionale*

$$H(X/Y) = \sum_{i,j} p(x_i, y_j) \log_2 \frac{1}{p(x_i/y_j)} \quad (17.22)$$

prende il nome di *equivocazione*, e rappresenta la quantità media di informazione *persa*, rispetto all'entropia di sorgente $H(X)$, a causa della rumorosità del canale. Nel caso in cui il canale non introduca errori, e quindi $p(x_i/y_j)$ sia pari a 1 se $i = j$ e zero altrimenti, è facile vedere che $H(X/Y)$ è pari a zero, e $I(X, Y) = H(X)$, ossia tutta l'informazione della sorgente si trasferisce a destinazione. D'altra parte

$$H(Y/X) = \sum_{i,j} p(x_i, y_j) \log_2 \frac{1}{p(y_j/x_i)} \quad (17.23)$$

prende il nome di *noise entropy* dato che considera il processo di rumore come se fosse un segnale informativo: infatti, sebbene si potrebbe dire che l'informazione media ricevuta è misurata dalla entropia $H(Y)$ della sequenza di osservazione, una parte di essa $H(Y/X)$ è *falsa*, perché in realtà è introdotta dagli errori.

Esercizio: calcolare $I(X, Y)$ per il canale binario simmetrico. Mantenendo la notazione introdotta al § 17.2.1, usiamo la (17.21) per calcolare l'informazione mutua in funzione di $p(x_1) = \alpha$ e p_e , e dunque iniziamo con il valutare $H(Y)$ e $H(Y/X)$. Dal punto di vista dell'uscita del canale, i simboli $y_{1,2}$ costituiscono l'alfabeto di una sorgente binaria senza memoria, la cui entropia si esprime in termini di $p(y_1)$ mediante la (17.3), ovvero $H(Y) = H_b(p(y_1))$, in cui

$$\begin{aligned} p(y_1) &= p(y_1/x_1)p(x_1) + p(y_1/x_2)p(x_2) = \\ &= (1 - p_e)\alpha + p_e(1 - \alpha) = p_e + \alpha - 2\alpha p_e \end{aligned}$$

²⁵Infatti

$$\begin{aligned} \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i/y_j)}{p(x_i)} &= \sum_i \sum_j p(x_i, y_j) \left[\log_2 \frac{1}{p(x_i)} - \log_2 \frac{1}{p(x_i/y_j)} \right] = \\ &= \sum_i \sum_j p(x_i, y_j) \log_2 \frac{1}{p(x_i)} - \sum_i \sum_j p(x_i, y_j) \log_2 \frac{1}{p(x_i/y_j)} \end{aligned}$$

e, saturando la sommatoria doppia del primo termine rispetto ad j , si ottiene la (17.20). Per la (17.21), il passaggio è del tutto simile.

e dunque $H(Y) = H_b(p_e + \alpha - 2\alpha p_e)$. Per quanto riguarda la *noise entropy* $H(Y/X)$, sostituendo $p(x_i, y_j) = p(y_j/x_i)p(x_i)$ nella (17.23) otteniamo

$$H(Y/X) = \sum_i p(x_i) \left[\sum_j p(y_j/x_i) \log_2 \frac{1}{p(y_j/x_i)} \right] = H_b(p_e)$$

dato che il termine tra parentesi quadre rappresenta appunto l'entropia di una sorgente binaria con simboli a probabilità p_e e $1 - p_e$. Possiamo quindi ora scrivere l'espressione cercata

$$I(X, Y) = H_b(p_e + \alpha - 2\alpha p_e) - H_b(p_e)$$

che dipende sia dalla probabilità di errore p_e , sia dalla statistica dei simboli della sorgente: osserviamo che se $p_e \ll 1$, il canale (quasi) non commette errori e risulta $I(X, Y) \simeq H_b(\alpha) = H(X)$, mentre se $p_e \rightarrow \frac{1}{2}$ allora $I(X, Y) \rightarrow 0$.

17.2.3 Capacità di canale discreto

I risultati ora mostrati, pur permettendo di valutare la perdita di informazione causata dai disturbi, dipendono sia dalle probabilità in avanti $p(y_j/x_i)$ che effettivamente descrivono il comportamento del canale, sia da quelle a priori $p(x_i)$, che invece attengono unicamente alle caratteristiche della sorgente. Al contrario, vorremmo trovare una grandezza che esprima esclusivamente l'attitudine (o *capacità*) del canale a trasportare informazione, indipendentemente dalle caratteristiche della sorgente. Questo risultato può essere ottenuto provando a variare la statistica della sorgente in tutti i modi possibili, fino a trovare il valore

$$C_s = \max_{p(x)} I(X, Y) \quad \text{bit/simbolo} \quad (17.24)$$

che definisce la *capacità di canale* come il massimo valore dell'informazione mutua media, ottenuto in corrispondenza della migliore sorgente possibile²⁶. Il pedice s sta per *simbolo*, e serve a distinguere il valore ora definito da quello che esprime il massimo *tasso* di trasferimento dell'informazione espresso in bit/secondo, ottenibile una volta nota la frequenza f_s con cui sono trasmessi i simboli, fornendo per la capacità di canale il nuovo valore²⁷

$$C = f_s \cdot C_s \quad \text{bit/secondo} \quad (17.25)$$

L'importanza di questa quantità risiede nel *teorema fondamentale per canali rumorosi*²⁸ già anticipato più volte, che asserisce

Per ogni canale discreto senza memoria di capacità C

- esiste una tecnica di codifica che consente la trasmissione di informazione a velocità R e con probabilità di errore per simbolo p_e piccola a piacere, purché risulti $R < C$;
- se è accettabile una probabilità di errore p_e , si può raggiungere (con la miglior codifica possibile) una velocità $R(p_e) = \frac{C}{1 - H_b(p_e)} > C$ in cui $H_b(p_e)$ è l'entropia di una sorgente binaria (17.3);
- per ogni valore di p_e , non è possibile trasmettere informazione a velocità maggiore di $R(p_e)$

²⁶In definitiva, questo modo di ottenere una grandezza rappresentativa del solo canale ricorda un po' la via per la quale si è definita ad es. la potenza *disponibile* di un generatore (vedi § 14.2.2.1), al variare di tutti i possibili valori di impedenza di carico.

²⁷Osserviamo l'invarianza di (17.25) rispetto alla modifica del numero di livelli: se M bit sono raggruppati per generare simboli ad $L = 2^M$ livelli, come noto f_s si riduce di M volte, mentre C_s aumenta della stessa quantità, dato che ogni simbolo trasporta ora M bit anziché uno.

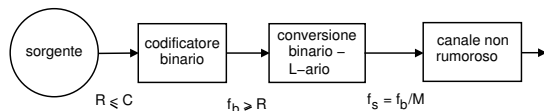
²⁸http://it.wikipedia.org/wiki/Secondo_teorema_di_Shannon

Il teorema non suggerisce come individuare la tecnica di codifica, né fa distinzioni tra codifica di sorgente e di canale, ma indica le prestazioni limite ottenibili mediante la migliore tecnica possibile, in grado di ridurre a piacere la p_e purché $R < C$, mettendoci al tempo stesso in guardia a non tentare operazioni impossibili. Da questo punto di vista, le prestazioni conseguibili adottando le tecniche di codifica note possono essere valutate confrontandole con quelle *ideali* predette dal teorema. Inoltre, dato che la capacità di canale è definita come massimo valore di $I(X, Y)$ per la migliore $p(x)$, qualora la statistica dei messaggi prodotti dal codificatore di sorgente differisca da quella ottima per il canale, l'informazione mutua media risulterà ridotta, così come la massima velocità R .

Illustriamo l'applicazione di questi risultati con un paio di esempi:

Canale L -ario non rumoroso Consideriamo il caso mostrato alla figura seguente, che rappresenta un canale che trasporta *senza errori* simboli con $L = 2^M$ livelli: in tal caso l'equivocazione $H(Y/X)$ è nulla, e la (17.20) permette di scrivere $I(X, Y) = H(X)$, che è massima se $P(x_i) = 1/L$ per tutti gli i , risultando così $C_s = \log_2 L = M$ bit/simbolo, e $C = f_s \cdot C_s = f_s \cdot M$ bit/secondo.

I simboli ad L livelli sono ottenuti a partire da M bit prodotti da una codifica binaria a velocità f_b , risultando $f_b \geq R = H_x$ in funzione della ottimalità



o meno del codificatore; pertanto, risulta $R \leq f_b = f_s M = C$ con l'uguaglianza valida nel caso in cui il codificatore riesca a rimuovere tutta la ridondanza dei messaggi della sorgente, conseguendo in tal caso il massimo trasferimento di informazione.

Al contrario, volendo realizzare una velocità $R > C$, il codificatore di sorgente dovrebbe produrre codeword con lunghezze tali da violare la disuguaglianza di Kraft (17.5)²⁹, e quindi la regola del prefisso non sarebbe rispettata, causando in definitiva errori di decodifica anche in assenza di rumore!

Capacità del canale simmetrico binario Esaminiamo l'effetto della presenza di rumore per questo caso già studiato, e per il quale abbiamo valutato che l'espressione dell'informazione mutua media risulta

$$I(X, Y) = H_b(p_e + \alpha - 2\alpha p_e) - H_b(p_e)$$

in cui $H_b(p_e)$ dipende solo dalla probabilità di errore, mentre il termine $H_b(p_e + \alpha - 2\alpha p_e)$ dipende anche dalla statistica di sorgente, e risulta massimizzato e pari ad 1 se $p_e + \alpha - 2\alpha p_e = \frac{1}{2}$, come risulta per qualunque p_e se $\alpha = \frac{1}{2}$, ossia per simboli equiprobabili. Pertanto, la capacità in questo caso è

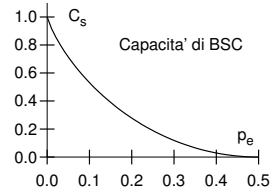
$$C_s = 1 - H_b(p_e)$$

ed il suo andamento è rappresentato alla figura che segue³⁰, evidenziando che $C_s \simeq 1$ bit/simbolo se $p_e \simeq 0$, ma che poi decade rapidamente a zero se $p_e \rightarrow 0.5$.

²⁹Infatti, potrebbe risultare $R > C$ solo se $f_b < R$, ovvero il codificatore dovrebbe produrre *meno* bit/secondo di quanti bit/secondo produca la sorgente

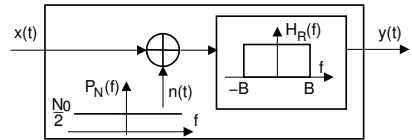
³⁰Sono mostrati solo i valori per $0 \leq p_e \leq 0.5$ dato che successivamente l'andamento di C_s si riflette in modo speculare.

Quest'ultimo esempio in particolare ci conferma l'esigenza, in presenza di un canale rumoroso, di attuare tecniche di codifica di canale in grado di ridurre la probabilità di errore, e di preferire tra queste le tecniche che vi riescono mantenendo al minimo la quantità dei bit aggiuntivi. Infatti spesso il canale impone una velocità di trasmissione, parte della quale è impiegata per trasmettere i soli binit di protezione e non l'informazione della sorgente, riducendo di fatto il tasso R effettivamente trasmesso.



17.2.4 Capacità di canale continuo

Come anticipato fin da pag. 3, un canale numerico è in realtà una astrazione che ingloba internamente un codificatore di linea o *modem* che, a partire da una sequenza numerica, produce un segnale trasmissibile su di un canale analogico, che a sua volta può essere caratterizzato da un valore di capacità, espresso nei termini dei parametri che descrivono la trasmissione analogica soggiacente. Una situazione tipica è quella rappresentata in figura, in cui al segnale ricevuto è sommato un rumore $n(t)$ gaussiano, bianco e a media nulla, mentre il filtro di ricezione $H_R(f)$ impone una limitazione di banda $2B$, in modo che la potenza di rumore in ingresso al decisore vale $P_n = \sigma_n^2 = N_0 B$. Una tale descrizione viene indicata come *canale AWGN (additive white gaussian noise) limitato in banda*.



Indicando ora con $p(x)$, $p(y)$, $p(x/y)$, $p(y/x)$ le densità di probabilità marginali e condizionali che descrivono *un campione* dei processi di ingresso $x(t)$ ed uscita $y(t)$, entrambi limitati in banda $\pm B$, l'applicazione formale della (17.18) al caso continuo porta a scrivere l'espressione dell'informazione mutua media come

$$I(X, Y) = \int \int_{-\infty}^{\infty} p_{XY}(x, y) \log_2 \frac{p_Y(y/x)}{p_Y(y)} dx dy \quad \text{bit/campione} \quad (17.26)$$

che è una misura *assoluta*³¹ del trasferimento di informazione per campione di uscita. Il massimo valore di (17.26) al variare di $p_X(x)$ consente anche questa volta di definire la capacità di canale per campione $C_s = \max_{p(x)} I(X, Y)$; in virtù della limitazione di banda, i campioni prelevati ad una frequenza di campionamento $f_c = 2B$ risultano indipendenti tra loro (vedi § 9.2.3), cosicché la capacità di canale risulta definita come

$$C = 2B \cdot \max_{p(x)} \{I(X, Y)\} \quad \text{bit/secondo} \quad (17.27)$$

Riscrivendo la (17.26) nella forma

$$I(X, Y) = h(Y) - h(Y/X) \quad (17.28)$$

si ottiene una espressione analoga alla (17.21) ma i cui termini sono ora da intendersi come entropia differenziale, definita in (17.10). Osserviamo ora che il termine di *noise entropy* $h(Y/X) = \int \int_{-\infty}^{\infty} p_{XY}(x, y) \log_2 \frac{1}{p_Y(y/x)} dx dy$ dipende esclusivamente

³¹Per il fatto di avere una ddp di y sia a numeratore che a denominatore del logaritmo, la (17.26) non soffre dei problemi discussi alla nota 20

dal rumore additivo, in quanto $y(t) = x(t) + n(t)$ e quindi $p_Y(y/x) = p_N(x+n)$: infatti $p_Y(y/x)$ altro non è che la gaussiana del rumore, a cui si somma un valor medio fornito dal campione di x ; pertanto $h(Y/X)$ si riduce all'entropia differenziale di un processo gaussiano, che non dipende dal valor medio, ma solo dall'andamento di $p_N(n)$; pertanto

$$h(Y/X) = \int_{-\infty}^{\infty} p_N(n) \log_2 \frac{1}{p_N(n)} dn = \frac{1}{2} \log_2 (2\pi e \sigma_n^2)$$

come risulta per l'entropia differenziale di sorgenti gaussiane (17.11). Quindi ora il termine della (17.28) che deve essere massimizzato rispetto a $p(x)$ è solo il primo, ossia $h(Y)$, che come sappiamo, è massimo se $y(t)$ è gaussiano. Dato che il processo ricevuto $y(t)$ è composto da due termini $x(t) + n(t)$ di cui il secondo è già gaussiano, si ottiene $y(t)$ gaussiano a condizione che anche $x(t)$ sia gaussiano. Indicando con σ_x^2 la potenza di quest'ultimo, ed in virtù della indipendenza statistica tra $x(t)$ e $n(t)$, risulta $\sigma_y^2 = \sigma_x^2 + \sigma_n^2$, e quindi

$$h(Y) = \frac{1}{2} \log_2 [2\pi e (\sigma_x^2 + \sigma_n^2)]$$

cosciché la (17.27) si riscrive come

$$\begin{aligned} C &= 2B \cdot \left\{ \frac{1}{2} \log_2 [2\pi e (\sigma_x^2 + \sigma_n^2)] - \frac{1}{2} \log_2 (2\pi e \sigma_n^2) \right\} = \\ &= B \cdot \log_2 \frac{\sigma_x^2 + \sigma_n^2}{\sigma_n^2} = B \cdot \log_2 \left(1 + \frac{P_x}{P_n} \right) \quad \text{bit/secondo} \end{aligned}$$

che è proprio il risultato tanto spesso citato, che prende il nome di *legge di Shannon-Hartley*³² e che esprime la capacità di canale per un canale additivo gaussiano. Tenendo conto che $P_n = \sigma_n^2 = N_0 B$ e che P_x è la potenza del segnale ricevuto P_s , riscriviamo l'espressione della capacità nella sua forma più nota:

$$C = B \cdot \log_2 \left(1 + \frac{P_s}{N_0 B} \right) \quad \text{bit/secondo} \quad (17.29)$$

che, associata al teorema fondamentale della codifica espresso al § 17.2.3, stabilisce il massimo tasso informativo trasmissibile senza errori su di un canale AWGN limitato in banda come $R < B \cdot \log_2 (1 + P_s/N_0 B)$. Discutiamo ora delle conseguenze di questo risultato.

Sistema di comunicazione ideale Una volta noto il massimo tasso di informazione $R < C$ che il canale può trasportare senza errori, come fare per evitare, appunto, questi ultimi? Il metodo suggerito da Shannon, anziché introdurre ridondanza come avviene per le tecniche di codifica di canale classiche, effettua invece la trasmissione semplicemente ripartendo l'informazione in blocchi codificati mediante simboli di durata elevata. In pratica, si tratta di realizzare una sorta di *trasmissione multilivello* (vedi § 5.1.2.4) come mostrato alla figura 17.2 dove l'informazione generata ad una velocità R bit/secondo viene trasmessa mediante simboli emessi con periodo T_s secondi, ognuno dei quali convoglia quindi una quantità di informazione pari a $M = RT_s$ bit, e dunque occorrono $L = 2^M$ diversi simboli.

³²http://en.wikipedia.org/wiki/Shannon-Hartley_theorem

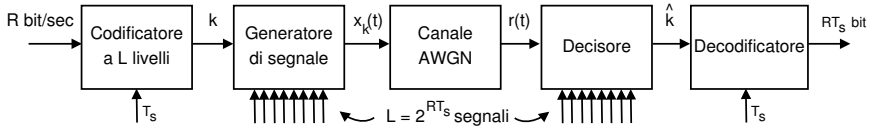


Figura 17.2: Schema ideale di codifica di canale ad errore asintoticamente nullo

Nella dimostrazione di Shannon ogni simbolo, anziché essere rappresentato da un livello costante, è costituito da un segnale $x_k(t)$, $k = 1, 2, \dots, L$ di durata T_s , ottenuto prelevando una finestra temporale T_s da una realizzazione di processo gaussiano bianco limitato in banda. Il ricevitore possiede una copia di tali forme d'onda, e per ogni periodo di simbolo calcola l'errore quadratico $\varepsilon_k = \frac{1}{T_s} \int_0^{T_s} (r(t) - x_k(t))^2 dt$ tra il segnale ricevuto $r(t)$ ed ognuna delle forme d'onda associate ai simboli, decidendo per la trasmissione del simbolo \hat{k} la cui forma d'onda $x_{\hat{k}}(t)$ fornisce l'errore ε_k minimo. Mantenendo R fisso e pari al tasso informativo della sorgente, all'aumentare di T_s anche $M = RT_s$ aumenta di pari passo, mentre il numero di simboli $L = 2^M$ aumenta esponenzialmente. Claude Shannon ha dimostrato³³ che, per $T_s \rightarrow \infty$, lo schema indicato riesce effettivamente a conseguire una $P_e \rightarrow 0$, tranne per il piccolo particolare che... occorre attendere un tempo che tende a infinito!

In realtà, uno schema di trasmissione numerica che approssima piuttosto bene questo ideale appena discusso esiste veramente, ed è quello esposto al § 13.5.1 e indicato come FSK ortogonale. Infatti, il grafico delle sue prestazioni a pag. 324 mostra come, aumentando L , lo stesso valore di $\frac{E_b}{N_0}$ permette di conseguire valori di P_e via via più piccoli.

Minima energia per bit Lo stesso grafico di pag. 324 consente anche di verificare come allo stesso tempo il valore di $\frac{E_b}{N_0}$ necessario a conseguire una ben determinata P_e diviene sempre più piccolo all'aumentare di L , anche se non può ridursi a meno di un valore limite, ossia deve comunque risultare³⁴

$$\frac{E_b}{N_0} \geq \ln 2 = 0,693 \quad \text{ovvero} \quad \left. \frac{E_b}{N_0} \right|_{dB} \geq -1.6 \text{ dB} \quad (17.30)$$

D'altra parte nell'FSK l'aumento di L comporta l'aumento, oltre che di T_s , anche della banda occupata per la trasmissione, e questo ci dà lo spunto per le osservazioni che seguono.

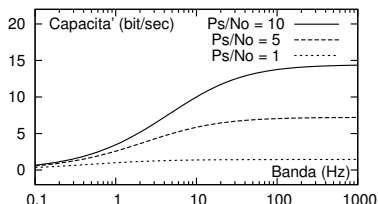
³³Senza pretendere di svolgere l'esatta dimostrazione, tentiamo di dare credibilità a questo risultato. Osserviamo quindi che se $r(t) = x_k(t) + n(t)$, il valore atteso dell'errore ε_k si riduce a $\frac{1}{T_s} \int_0^{T_s} [n(t)]^2 dt \rightarrow \sigma_n^2$, dato che essendo $n(t)$ stazionario ergodico, le medie di insieme coincidono con le medie temporali. Viceversa, se il segnale trasmesso è $x_h(t)$ con $h \neq k$, allora il relativo errore quadratico vale $\varepsilon_k^{(h)} = \frac{1}{T_s} \int_0^{T_s} (x_h(t) + n(t) - x_k(t))^2 dt$, ed il suo valore atteso $E \left\{ \varepsilon_k^{(h)} \right\} \rightarrow \sigma_n^2 + 2\sigma_x^2$ essendo le forme d'onda dei simboli ortogonali tra loro e rispetto al rumore. I valori limite mostrati sono in realtà grandezze aleatorie, ma la loro varianza diviene sempre più piccola all'aumentare di T_s , e quindi in effetti con $T_s \rightarrow \infty$ risulta sempre $\varepsilon_k < \varepsilon_k^{(h)}$, azzerando la probabilità di errore.

³⁴La (17.30) si ottiene considerando che se la capacità di canale per $B \rightarrow \infty$ fornita dalla (17.31) vale $C_\infty = \frac{P_s}{N_0 \ln 2}$, e se deve risultare $R \leq C$, allora $\ln 2 = \frac{P_s}{N_0 C_\infty} \leq \frac{P_s}{N_0 R} = \frac{E_b}{N_0}$.

Compromesso banda-potenza e capacità massima Il valore limite (17.30) trae origine da una conseguenza della (17.29) già fatta notare a pag. 90, ovvero la possibilità di risparmiare potenza aumentando l'occupazione di banda (o viceversa), dato che a ciò corrisponde un aumento di C , che però *non può oltrepassare* un valore massimo. Infatti, se nella (17.29) si aumenta B , aumenta anche la potenza di rumore, e l'effetto finale è che per un canale con *banda infinita* non si ottiene una capacità infinita, bensì il valore³⁵

$$\begin{aligned} C_\infty &= \lim_{B \rightarrow \infty} B \cdot \log_2 \left(1 + \frac{P_s}{N_0 B} \right) = \\ &= \frac{P_s}{N_0 \ln 2} \simeq 1.44 \frac{P_s}{N_0} \end{aligned} \quad (17.31)$$

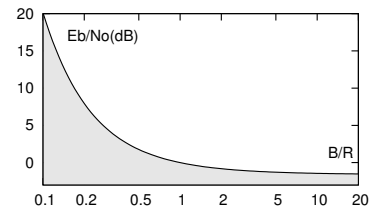
che individua anche il limite *assoluto* al massimo tasso informativo R trasmissibile. In figura è mostrato l'andamento effettivo della (17.29) in funzione di B , per alcuni valori di $\frac{P_s}{N_0}$ di esempio.



Limite inferiore per $\frac{E_b}{N_0}$ Una volta assegnato il tasso informativo $R < C$ della sorgente e la banda B del canale, partendo dalla (17.29) si può ottenere³⁶ una relazione che esprime l' $\frac{E_b}{N_0}$ necessario a conseguire una trasmissione senza errori (nel caso ideale):

$$\frac{E_b}{N_0} > \frac{B}{R} \left(2^{\frac{R}{B}} - 1 \right) \quad (17.32)$$

e che, espressa in dB, è graficata nella figura a lato, in cui l'area grigia indica i valori di $\frac{E_b}{N_0}$ vietati, ossia per i quali è impossibile ottenere una trasmissione senza errori.



Compromesso banda-potenza per un sistema ideale

Notiamo innanzitutto che, mentre per $\frac{B}{R} = 1$ il sistema ideale richiede un valore di $\frac{E_b}{N_0}$ pari ad almeno 0 dB, questo si riduce nel caso in cui la trasmissione occupi una banda maggiore del tasso informativo R , fino a raggiungere (già per valori $B > 10R$) il limite (17.30) di -1.6 dB. D'altra parte, qualora la trasmissione impegni una banda inferiore ad R , il valore di $\frac{E_b}{N_0}$ necessario aumenta in modo piuttosto brusco.

³⁵La (17.31) si ottiene riscrivendo la (17.29) nella forma

$$C = \frac{P_s}{N_0 \frac{P_s}{N_0 B}} \cdot \frac{\ln \left(1 + \frac{P_s}{N_0 B} \right)}{\ln 2} = \frac{P_s}{N_0 \ln 2} \cdot \frac{\ln(1 + \lambda)}{\lambda}$$

in cui \ln è il logaritmo *naturale* in base e , e si è posto $\frac{P_s}{N_0 B} = \lambda$. Ricordando ora lo sviluppo di Maclaurin $f(x) = f(0) + \sum_{n=1}^{\infty} \left(\frac{\partial^n f(x)}{\partial x^n} \Big|_{x=0} \cdot \frac{x^n}{n!} \right)$ e che $\frac{d}{dx} \ln x = \frac{1}{x}$, il termine $\ln(1 + \lambda)$ può essere espanso in serie di potenze come $\ln(1 + \lambda) = \lambda - \frac{1}{2}\lambda^2 + \frac{1}{3}\lambda^3 + \dots$; notando infine che per $B \rightarrow \infty$ si ha $\lambda \rightarrow 0$, e che $\lim_{\lambda \rightarrow 0} \frac{\ln(1+\lambda)}{\lambda} = 1$, si giunge in definitiva al risultato (17.31).

³⁶Riscrivendo la (17.29) come $2^{\frac{C}{B}} - 1 = \frac{P_s}{N_0 B}$, moltiplicando ambo i membri per $\frac{B}{R}$, e semplificando il risultato, si ottiene $\frac{B}{R} \left(2^{\frac{C}{B}} - 1 \right) = \frac{P_s}{N_0 R}$. L'uguaglianza individua la circostanza limite in cui $R = C$, mentre se nell'esponente di 2 a primo membro sostituiamo C con R , e $R < C$, il primo membro diviene più piccolo, e pertanto $\frac{B}{R} \left(2^{\frac{R}{B}} - 1 \right) < \frac{P_s}{N_0 R}$. Infine, notiamo che $\frac{P_s}{N_0 R} = \frac{E_b}{N_0}$, da cui il risultato mostrato (17.32).

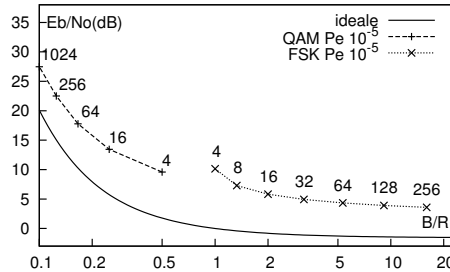


Figura 17.3: Prestazioni di QAM ed FSK confrontate con quelle ideali

Prestazioni di sistemi di comunicazione reali La verifica dei comportamenti appena evidenziati può essere svolta confrontando le prestazioni ideali (17.32) con quelle ottenibili adottando le tecniche di modulazione numerica già discusse, e per le quali si riesce a ridurre la banda occupata come nel caso della trasmissione multilivello³⁷, oppure la si aumenta, come nel caso dell'FSK.

La figura 17.3 riporta i valori di $\frac{E_b}{N_0}$ vs $\frac{B}{R}$ per le tecniche di modulazione numerica QAM (§ 13.3.1) e FSK ortogonale (pag. 322): a partire dai rispettivi andamenti della P_e in funzione di $\frac{E_b}{N_0}$ ed L , si sono ricavati i valori di $\frac{E_b}{N_0}$ necessari ad ottenere una P_e pari a 10^{-5} per diversi valori di L , e questi sono stati riportati nel grafico assieme alla banda occupata, valutata come segue.

Considerando di adottare per il QAM un impulso di Nyquist a banda minima, la banda occupata risulta pari a $B_{QAM} = \frac{f_b}{\log_2 L}$, e pertanto

$$\left. \frac{B}{R} \right|_{QAM} = \frac{1}{\log_2 L}$$

mentre come riportato a pag. 324, per l'FSK ortogonale si ha $B_{FSK} \simeq \frac{f_b}{2} \frac{L}{\log_2 L}$, e dunque

$$\left. \frac{B}{R} \right|_{FSK} = \frac{L}{2 \log_2 L}$$

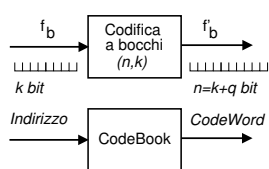
Possiamo osservare come per le due tecniche di trasmissione l'andamento dei valori di $\frac{E_b}{N_0}$ in funzione di $\frac{B}{R}$ ricalchi abbastanza fedelmente quello ideale, a parte una perdita di efficienza, che si riduce per L crescente.

17.3 Codici di canale

Dopo aver esposto i risultati che la teoria dell'informazione fornisce a riguardo delle migliori prestazioni ottenibili, proseguiamo il discorso iniziato al § 5.4.3.1 relativo a come aggiungere ridondanza ad un flusso binario a velocità f_b da trasmettere su di un canale numerico, in modo da realizzare una protezione FEC capace di ridurre la probabilità di errore sul bit P_e in ricezione. Notiamo subito che, mentre la soluzione teorica basata su segnalazione asintoticamente ortogonale di fig. 17.2 prevede un aumento

³⁷Vedi ad es. il caso di banda base al § 5.3.5 o quello del QAM al § 13.3.1.

esponenziale della banda occupata, le tecniche che verranno illustrate consentono di mantenere la relazione tra grado di protezione e banda occupata di tipo *proporzionale*.



Riprendiamo la notazione introdotta a pag. 98 per i codici a blocchi, in cui per ogni k bit in ingresso sono prodotte *codeword* con lunghezza $n = k + q > k$ mediante *l'aggiunta* di q bit di *protezione* in funzione dei primi k : indichiamo questo procedimento come *un codice* (n, k) , la cui efficienza è misurata dal *tasso di codifica* (CODE RATE)

$$R_c = \frac{k}{n} < 1$$

che rappresenta la frazione di bit informativi sul totale di quelli trasmessi. La nuova velocità di trasmissione vale pertanto

$$f'_b = \frac{f_b}{R_c}$$

La teoria esposta al § 17.2.3 afferma che la probabilità di errore può essere resa piccola a piacere purché $f_b = R < C$, ma se questo limite è applicato alla effettiva velocità di segnalazione f'_b , allora il massimo tasso di trasferimento dell'informazione si riduce a $f_b = R_c f'_b < R_c C$. Inoltre, aumentando la velocità di segnalazione (essendo $f'_b > f_b$) diminuisce di pari misura il rapporto $E_b/N_0 = \frac{P_s}{N_0 f'_b}$, e conseguentemente peggiora anche la probabilità di errore del decisore, in modo che la capacità correttiva del codice deve essere tale da compensare anche quest'altro fattore. In definitiva, vorremmo trovare codificatori per cui R_c sia il più possibile vicino ad uno.

Al tempo stesso, la capacità di correzione del codice è direttamente legata alla *minima* distanza di Hamming d_m definita a pag. 98 come il minimo numero di bit diversi tra due parole di codice, sussistendo le relazioni

- per accorgersi di l (o meno) errori per codeword occorre $d_m \geq l + 1$
- per correggere t (o meno) errori per codeword occorre $d_m \geq 2t + 1$

Un codice è tanto più *potente* quanti più errori è in grado di correggere, e dunque deve possedere d_m elevato. In un codice a blocchi (n, k) i k bit del messaggio originale assumono tutte le configurazioni possibili, e quindi contribuiscono alla distanza tra codeword per un solo bit; per ottenere $d_m > 1$ occorre pertanto sfruttare gli $n - k = q$ bit di protezione, portando a scrivere

$$d_m \leq n - k + 1 = q + 1$$

che evidenzia la relazione tra d e la quantità di bit aggiunti q . L'uguaglianza sussiste solo per una particolare classe di codici³⁸ tra cui il codice a ripetizione, discusso a pag. 99, che adottando una dimensione di blocco in ingresso $k = 1$ ha un tasso di codifica $R_c = k/n = 1/n$ molto inefficiente. Mostriamo allora delle soluzioni che consentono di ottenere un adeguato potere di detezione senza per questo aumentare di molto la velocità di trasmissione del flusso codificato.

³⁸Indicati come codici MDS, vedi http://en.wikipedia.org/wiki/Singleton_bound#MDS_codes.

17.3.1 Codice lineare a blocchi

Le proprietà di questa classe di codici di canale possono essere meglio analizzate interpretando l'insieme delle possibili codeword da un punto di vista algebrico, che ci porta ad adottare una notazione matriciale idonea a descrivere la classe di codici a blocchi, mentre per i codici *ciclici* interverrà una notazione polinomiale.

Distanza d_m per codici lineari Scegliamo di rappresentare una codeword arbitraria \mathbf{x} di un codice a blocchi (n, k) mediante un vettore ad elementi binari

$$\mathbf{x} = (x_1 \quad x_2 \quad \cdots \quad x_n)$$

che può assumere solo 2^k diversi valori tra i 2^n possibili, mentre le ricezioni di una delle rimanenti $2^n - 2^k$ combinazioni di bit segnala la presenza di almeno un errore. Le 2^k codeword costituiscono uno *spazio lineare* (pag. 25) se comprendono la codeword nulla, e se la somma di due vettori è anch'essa una parola di codice. La somma tra due codeword è definita in base alla matematica binaria modulo due, espressa mediante l'operatore di OR esclusivo \oplus come

$$\mathbf{x} + \mathbf{y} = (x_1 \oplus y_1 \quad x_2 \oplus y_2 \quad \cdots \quad x_n \oplus y_n)$$

Indicando ora come *peso* $w(\mathbf{z})$ di un vettore \mathbf{z} il numero di suoi elementi non zero, una conseguenza della linearità è la possibilità di valutare la minima distanza di Hamming d_m tra parole di codice come il *minimo peso* tra tutte le codeword non zero³⁹, ossia

$$d_m = \min_{\mathbf{x} \neq \mathbf{0}} [w(\mathbf{x})]$$

Rappresentazione matriciale per codici sistematici Finora abbiamo assunto che gli n bit delle codeword rappresentino per primi i k bit da proteggere, a cui seguono i $q = n - k$ bit di protezione, anche se questa ripartizione non è per nulla scontata: qualora effettivamente si verifici, il codice risultante viene detto *sistematico*. In tal caso le codeword possono essere scritte nella forma

$$\mathbf{x} = (m_1 \quad m_2 \quad \cdots \quad m_k \quad c_1 \quad c_2 \quad \cdots \quad c_q)$$

ovvero in forma partizionata $\mathbf{x} = (\mathbf{m} \mid \mathbf{c})$, che consente di rappresentare \mathbf{x} a partire dai bit da proteggere \mathbf{m} e dalla definizione di una matrice *generatrice* $k \times n$ con struttura generale $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$, in cui \mathbf{I}_k è una matrice identità $k \times k$ e \mathbf{P} è una sotto-matrice di elementi binari $k \times n$, permettendo di scrivere $\mathbf{x} = \mathbf{m} \cdot \mathbf{G}$, ovvero

$$\left[m_1 \cdots m_k \quad c_1 \cdots c_q \right] = \left[m_1 \cdots m_k \right] \cdot \begin{bmatrix} 1 & \cdots & 0 & p_{11} & \cdots & p_{1q} \\ \vdots & 1 & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & p_{k1} & \cdots & p_{kq} \end{bmatrix} \quad (17.33)$$

in modo che \mathbf{P} produca⁴⁰ i q bit di protezione come $\mathbf{c} = \mathbf{m} \cdot \mathbf{P}$. Dato che il valore della (generica) j -esima componente di \mathbf{c} si calcola come

$$c_j = m_1 p_{1j} \oplus m_2 p_{2j} \oplus \cdots \oplus m_k p_{kj}$$

³⁹ Infatti dalla definizione di somma otteniamo che la distanza tra due codeword \mathbf{x} e \mathbf{y} è pari al peso della codeword $\mathbf{z} = \mathbf{x} + \mathbf{y}$: infatti \mathbf{z} presenterà componenti $z_j = 1$ solo in corrispondenza di elementi $x_j \neq y_j$. Ma per la linearità anche \mathbf{z} appartiene al codebook, e dunque la ricerca su tutte le coppie si trasforma in una ricerca su tutte le codeword.

⁴⁰ Sono valide le normali regole di moltiplicazione tra matrici, tranne per l'accortezza di usare la somma modulo due anziché quella convenzionale.

osserviamo che ciascuna colonna di \mathbf{P} individua un sotto-insieme di componenti di \mathbf{m} su cui calcolare una *somma di parità*, ed è per questo che il sotto-blocco di matrice è rappresentato dalla lettera \mathbf{P} . Ma non è ancora stato definito nulla che ci possa aiutare a scegliere i coefficienti p_{ij} allo scopo di ottenere i valori d_m e R_c desiderati: il *codice di Hamming* ci fornisce una possibile soluzione.

17.3.1.1 Codice di Hamming

È un codice a blocchi (n, k) sistematico con $q \geq 3$ bit di controllo, e per il quale si definisce

$$n = 2^q - 1 \quad \text{e} \quad k = n - q$$

per cui il tasso di codifica vale

$$R_c = \frac{k}{n} = \frac{n - q}{n} = 1 - \frac{q}{2^q - 1}$$

q	n	k	R_c
3	7	4	0.57
4	15	11	0.73
5	31	26	0.84
6	63	57	0.9
7	127	120	0.94

che aumenta con il crescere di q , come mostrato in tabella. Le codeword si individuano ponendo *le righe* della sottomatrice \mathbf{P} pari a tutte le parole di q bit con *due o più* uni, in qualsiasi ordine. Ma la cosa *più simpatica*, è che per un codebook siffatto si ottiene una distanza di Hamming minima $d_m = 3$, indipendentemente dalla scelta di q .

Esempio: codice di Hamming $(7, 4)$. Consideriamo un codice di Hamming con $q = 3$, e quindi $n = 2^3 - 1 = 7$ e $k = 7 - 3 = 4$. Una possibile matrice generatrice è quindi pari a

$$\mathbf{G} = \left[\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{array} \right]$$

a cui corrispondono le seguenti $2^4 = 16$ codeword, confermando che $d_m = 3$.

\mathbf{m}	\mathbf{c}	$w(\mathbf{x})$	\mathbf{m}	\mathbf{c}	$w(\mathbf{x})$
0 0 0 0	0 0 0	0	1 0 0 0	1 0 1	3
0 0 0 1	0 1 1	3	1 0 0 1	1 1 0	4
0 0 1 0	1 1 0	3	1 0 1 0	0 1 1	4
0 0 1 1	1 0 1	4	1 0 1 1	0 0 0	3
0 1 0 0	1 1 1	4	1 1 0 0	0 1 0	3
0 1 0 1	1 0 0	3	1 1 0 1	0 0 1	4
0 1 1 0	0 0 1	3	1 1 1 0	1 0 0	4
0 1 1 1	0 1 0	4	1 1 1 1	1 1 1	7

Correzione basata sulla distanza Indichiamo ora con \mathbf{y} la parola di codice ricevuta; in presenza di errori, risulta $\mathbf{y} \neq \mathbf{x}$. Il metodo *diretto* per rivelare ed eventualmente correggere gli errori presenti è quello di confrontare gli n bit ricevuti con tutte le possibili 2^k codeword, e se nessuna di queste risulta uguale ad \mathbf{y} , scegliere la $\hat{\mathbf{x}}$ con la minima distanza di Hamming, ossia quella per la quale il peso $w(\mathbf{y} \oplus \hat{\mathbf{x}})$ è minimo, ovvero

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \{w(\mathbf{y} \oplus \mathbf{x})\}$$

Correzione basata sulla sindrome Un metodo che non richiede una ricerca esaustiva si basa invece sul calcolo della cosiddetta *sindrome*, ottenuta mediante moltiplicazione del vettore \mathbf{y} ricevuto per una matrice $n \times q$ di *controllo parità* \mathbf{H} , definita come $\mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I}_q \end{bmatrix}$ in cui \mathbf{P} è la stessa matrice di parità utilizzata nella matrice generatrice \mathbf{G} , e \mathbf{I}_q è una matrice identità di dimensioni $q \times q$. La matrice \mathbf{H} esibisce la simpatica proprietà che, se moltiplicata per una qualunque codeword valida, fornisce un vettore *nullo* di dimensione q , ossia

$$\mathbf{x} \cdot \mathbf{H} = (0 \quad 0 \quad \dots \quad 0)$$

Al contrario, se moltiplicata per un vettore \mathbf{y} non appartenente al codebook, fornisce un vettore detto *sindrome* $\mathbf{s} = \mathbf{y} \cdot \mathbf{H}$ non nullo, e quindi il suo calcolo permette la *rivelazione* (nei limiti consentiti da d_m) dell'occorrenza di errori.

Esempio considerando di nuovo il caso di $q = 3$, la corrispondente matrice di controllo parità $\mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I}_q \end{bmatrix}$ è mostrata a lato. È facile verificare che per tutte le possibili codeword \mathbf{x} (ad es. $\mathbf{x} = [0100111]$) si ottiene $\mathbf{x} \cdot \mathbf{H} = [000]$. Poniamo ora che si verifichi una sequenza di errore $\mathbf{e} = [0010010]$, dando luogo alla ricezione della parola $\mathbf{y} = \mathbf{x} \oplus \mathbf{e} = [0110101]$: per essa si ottiene una sindrome $\mathbf{s} = \mathbf{y} \cdot \mathbf{H} = [100]$.

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ \hline 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

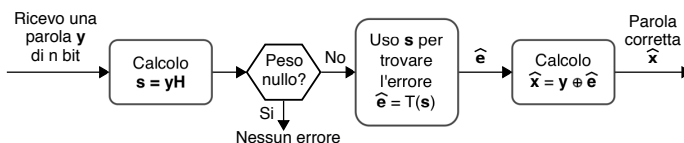
Per quanto riguarda la *correzione*, se scriviamo il vettore ricevuto come $\mathbf{y} = \mathbf{x} \oplus \mathbf{e}$, dove \mathbf{e} è un vettore di n bit le cui componenti sono diverse da zero in corrispondenza dei bit errati di \mathbf{y} , il calcolo della sindrome risulta

$$\mathbf{s} = \mathbf{y} \cdot \mathbf{H} = (\mathbf{x} \oplus \mathbf{e}) \cdot \mathbf{H} = \mathbf{x} \cdot \mathbf{H} \oplus \mathbf{e} \cdot \mathbf{H} = \mathbf{e} \cdot \mathbf{H}$$

visto che come detto sopra, la sindrome delle codeword è nulla. Dato però che la sindrome ha dimensione di q elementi, tutte le 2^n possibili sequenze di errore \mathbf{e} danno luogo a sole 2^q diverse sindromi, e quindi la conoscenza della sindrome non consente di risalire direttamente a \mathbf{e} . Ma riprendendo i risultati esposti a pag. 94, risulta che la probabilità $P(i, n)$ che si siano verificati i errori su n bit decresce al crescere di i , e pertanto il vettore $\hat{\mathbf{e}}$ che con maggior probabilità ha prodotto ognuna delle 2^q sindromi $\mathbf{s} \neq 0$, è quello (tra tutti quelli che producono la stessa \mathbf{s}) con il minor peso:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}: \mathbf{e} \cdot \mathbf{H} = \mathbf{s}}{\operatorname{argmin}} \{w(\mathbf{e})\}$$

Aniché effettuare questo calcolo ad ogni codeword ricevuta, si può *precalcolare*, per ogni possibile \mathbf{e} , la relativa sindrome, e compilare una tabella $T(\mathbf{s})$ in cui memorizzare per ogni possibile \mathbf{s} , il vettore $\hat{\mathbf{e}}$ di peso minore che la produce. La stessa tabella può quindi essere consultata usando come chiave di ricerca la sindrome $\mathbf{s} \neq 0$ calcolata per ogni \mathbf{y} ricevuto, ottenendo il vettore di errore $\hat{\mathbf{e}} = T(\mathbf{s})$ più probabile, e che se sommato ad \mathbf{y} , permette la correzione degli errori. La figura che segue riepiloga i passi necessari alla correzione mediante sindrome.



Osservazioni Notiamo che se si verificano più errori di quanti d_m permetta di correggere, è inutile (anzi dannoso) cercare di eseguire la correzione, perché il numero di errori complessivo può risultare ancora più elevato. Nel caso del codice di Hamming in cui $d_m = 3$ si può correggere un solo errore per codeword, ed esistono solo n vettori e di peso unitario, a cui corrispondono sindromi esattamente pari alle $n = 2^q - 1$ righe di \mathbf{H} . Ma se ad esempio \mathbf{e} contiene *due* errori, la sua moltiplicazione per \mathbf{H} produce comunque una delle 2^q possibili sindromi, ed il tentativo di correzione produce un vettore $\hat{\mathbf{x}}$ contenente tre errori.

Esempio Un gruppo di bit 1001 è protetto dal codice di Hamming di pag. 460 producendo $\mathbf{x} = 1001110$, mentre in ricezione si osserva $\mathbf{y} = 1101110$. Il calcolo della sindrome $\mathbf{s} = \mathbf{y} \cdot \mathbf{H}$ fornisce il risultato $\mathbf{s} = 111$, che corrisponde alla seconda riga di \mathbf{H} , ovvero al vettore di errore $\hat{\mathbf{e}} = 0100000$, cioè proprio quello che si è verificato. Se invece avvengono due errori, e si riceve ad es. $\mathbf{y} = 1111110$, si ottiene $\mathbf{s} = 101$, a cui corrisponde $\hat{\mathbf{e}} = 1000000$, e quindi il calcolo $\hat{\mathbf{x}} = \mathbf{y} \oplus \hat{\mathbf{e}}$ produce ora $\hat{\mathbf{x}} = 0111110$, che appunto contiene tre errori.

Per il motivo illustrato, nel caso in cui gli errori non siano indipendenti e dunque isolati nel tempo, ma avvengano in sequenza come nel caso del *fast fading* (§ 15.3.4.4), il solo codice di Hamming è del tutto inadeguato alla loro correzione, mentre una soluzione comunemente adottata è quella di ricorrere alla tecnica dell'*interleaving* (§ 5.4.3.3) in cui m codeword di lunghezza n uscenti dal codificatore sono scritte per righe, e quindi lette per colonne prima di trasmettere il flusso numerico risultante: ciò consente di correggere fino ad m errori consecutivi, in quanto in ricezione il processo inverso di *de-interleaving* ridistribuisce i bit errati su m diverse codeword. In alternativa, si possono utilizzare le soluzioni descritte nel seguito.

Soft decision decoding Finora la decodifica di canale è stata applicata *a valle* del processo di decisione⁴¹, in modo da individuare le codeword come quelle con la minima *distanza di Hamming* rispetto al risultato prodotto dal decisore. Se invece le due operazioni sono *fuse* in un unico elemento che opera su interi blocchi di n bit $\mathbf{y} = (y_1, y_2, \dots, y_n)$, e che tenta di individuare le codeword trasmesse come quelle con la minima *distanza euclidea* rispetto a quanto ricevuto⁴², si ottengono prestazioni migliori a spese di una maggior complessità di calcolo. Questa seconda tecnica valuta le distanze $d_E(\mathbf{y}, \mathbf{x}^j) = \sum_{i=1}^n (y_i - x_i^j)^2$ tra i valori y_i effettivamente ricevuti, e quelli che avrebbero dovuto essere nell'ipotesi che sia stata trasmessa una delle possibili codeword $\mathbf{x}^j = (x_1^j, x_2^j, \dots, x_n^j)$, decidendo quindi per la codeword $\hat{\mathbf{x}}^j$ tale che $d_E(\mathbf{y}, \hat{\mathbf{x}}^j)$ è minima⁴³. In tal modo si ottengono prestazioni equivalenti a quelle dell'approccio *hard decoding* operante con un E_b/N_0 di circa 2 dB più elevato.

17.3.1.2 Codici ciclici

Sono un sottoinsieme dei codici lineari a blocchi, con la condizione aggiuntiva che se $\mathbf{x} = (x_1, x_2, \dots, x_n)$ è una codeword, lo sono anche tutti i suoi *scorrimenti ciclici*⁴⁴, ovvero le codeword $\mathbf{x}^{(1)} = (x_2, x_3, \dots, x_1)$, $\mathbf{x}^{(2)} = (x_3, x_4, \dots, x_2)$, \dots , $\mathbf{x}^{(n)} =$

⁴¹Per questo detta *hard decision decoding* in quanto opera su decisioni già prese.

⁴²Tale possibilità è indicata come *soft decision decoding* (decodifica con decisioni *soffici*) in quanto richiede operazioni in virgola mobile; si veda ad es. la spiegazione fornita presso <http://www.gaussianwaves.com/2009/12/hard-and-soft-decision-decoding-2/>.

⁴³Si veda lo sviluppo a pag. 367 per mettere in relazione la distanza $d_E(\mathbf{y}, \mathbf{x}^j)$ con la verosimiglianza logaritmica $\ln(p_n(\mathbf{y}/\mathbf{x}^j))$ della v.a. n -dimensionale \mathbf{y} rispetto all'ipotesi che sia stata trasmessa la codeword \mathbf{x}^j , in presenza del rumore gaussiano n in sede di decisione.

⁴⁴Ad esempio, il codice [000, 110, 101, 011] è un codice ciclico, mentre [000, 010, 101, 111] no.

$(x_n, x_1, \dots, x_{n-1})$. In tal caso sussistono delle proprietà algebriche aggiuntive che si basano sulla teoria dei *campi di Galois*⁴⁵, i cui elementi sono polinomi $x(p) = x_1 p^{n-1} + x_2 p^{n-2} + \dots + x_{n-1} p + x_n$ nella variabile p , con coefficienti definiti a partire dagli elementi delle codeword \mathbf{x} ; per tale motivo, i codici ciclici sono detti anche codici *polinomiali*⁴⁶: senza volerci addentrare nei particolari della teoria⁴⁷, citiamo direttamente i principali risultati.

Un codice ciclico (n, k) è completamente definito a partire da un *polinomio generatore* $g(p) = p^{n-k} + g_2 p^{n-k-1} + \dots + g_{n-k} p + 1$ di grado $n-k$ e che deve risultare un divisore di $p^n + 1$, ovvero essere tale che $p^{n+1}/g(p) = q(p)$ con resto nullo⁴⁸. Una volta definito $g(p)$, è possibile ottenere la matrice generatrice \mathbf{G} del codice in forma *sistematica* $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$, calcolando le k righe di \mathbf{P} come i coefficienti del polinomio *resto* della divisione tra p^i e $g(p)$, con $i = n-1, n-2, \dots, n-k$.

Esempio Troviamo la matrice generatrice in forma sistematica per un codice ciclico $(7, 4)$. Osserviamo innanzitutto che $p^n + 1$ si fattorizza come $p^n + 1 = (p+1)(p^3 + p^2 + 1)(p^3 + p + 1)$, e scegliamo $g(p) = p^3 + p^2 + 1$ come polinomio generatore. Il calcolo di $p^i/p^3 + p^2 + 1$ per $i = 6, 5, 4, 3$ fornisce come resti i polinomi $p^2 + p, p + 1, p^2 + p + 1, p^2 + 1$, pertanto

$$\text{la matrice generatrice risulta } \mathbf{G} = \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{array} \right]. \text{ Osserviamo che, a}$$

parte una permutazione, le righe di \mathbf{P} sono le stesse di quelle che definiscono il codice di Hamming di pag. 460: infatti, anche quest'ultimo rientra nella classe dei codici ciclici.

D'altra parte nel caso di un codice ciclico le codeword \mathbf{x} associate ai bit \mathbf{m} da proteggere possono ottenersi non solo utilizzando \mathbf{G} come descritto dalla (17.33), ma anche in base alla seguente proprietà: indicando con $m(p) = m_1 p^{k-1} + m_2 p^{k-2} + \dots + m_{k-1} p + m_k$ il polinomio associato ai k bit da codificare, il polinomio associato alla corrispondente codeword \mathbf{x} si ottiene come *il prodotto* $x(p) = m(p) \cdot g(p)$, e quindi gli elementi x_i della codeword sono individuati come la *convoluzione discreta*⁴⁹ tra i coefficienti di $m(p)$ e $g(p)$: $x_i = \sum_{j=1}^k m_j g_{i-j+1}$ per $i = 1, 2, \dots, n$, operazione che può essere realizzata mediante un filtro FIR con coefficienti dati dai valori g_i .

Notiamo infine che nel testo si è già incontrato un codice polinomiale (e quindi ciclico) al § 5.4.2.3 a proposito del CRC: i q bit di protezione possono quindi essere anche ottenuti mediante l'utilizzo di un registro a scorrimento controeazionato⁵⁰, e lo stesso può essere usato anche per il calcolo della sindrome dal lato ricevente.

Codice BCH Prende il nome dalle iniziali dei rispettivi inventori, e rappresenta una sottoclasse dei codici ciclici, in grado di correggere fino a $t < n$ errori: per ottenere questo risultato occorre scegliere un numero di bit di protezione $q \leq mt$ in cui $m \geq 3$ è un intero, una lunghezza delle codeword pari a $n = 2^m - 1$, ed un polinomio generatore

⁴⁵http://it.wikipedia.org/wiki/Campo_finito

⁴⁶Si veda http://en.wikipedia.org/wiki/Polynomial_code

⁴⁷http://en.wikipedia.org/wiki/Cyclic_code

⁴⁸Si applicano le regole di divisione tra polinomi http://it.wikipedia.org/wiki/Divisione_dei_polinomi

⁴⁹Il valore dei coefficienti del polinomio prodotto è indicato anche come *prodotto di Cauchy* (vedi http://it.wikipedia.org/wiki/Prodotto_di_Cauchy), e che si ottenga come una convoluzione è facilmente verificabile: dati ad es. $a(p) = a_0 + a_1 p + a_2 p^2$ e $b(p) = b_0 + b_1 p$, si ottiene $c(p) = a_0 b_0 + (a_0 b_1 + a_1 b_0) p + a_2 p^2 = \sum_{i=0}^2 p^i \sum_{j=0}^2 a_j b_{i-j}$. La notazione lievemente diversa del testo, è dovuta al diverso modo di indicizzare i coefficienti.

⁵⁰Forse ho trovato dove si spiega come faccia un registro a scorrimento controeazionato a svolgere questo genere di compiti: penso che in una prossima edizione sarà interessante aggiungerlo.

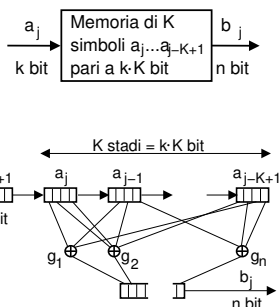
$g(p)$ di grado massimo mt le cui radici sono potenze dell'elemento primitivo α del campo di Galois $GF(2^m)$ ⁵¹. In tal caso si ottiene $d_m \geq 2t + 1$, e la capacità di correggere fino a t errori. Nel caso in cui $t = 1$, si ricade nel caso dei codici di Hamming. La fase di correzione degli errori può essere realizzata mediante il calcolo di una sindrome, per mezzo di una matrice di controllo \mathbf{H} realizzata a partire dalle potenze dell'elemento primitivo α , o sfruttando nuovamente proprietà algebriche, che eventualmente si traducono in soluzioni circuitali basate su registri a scorrimento controeazionati. Inoltre, è anche possibile applicare l'algoritmo iterativo di *Berlekamp-Massey*⁵².

Codice di Reed-Solomon Si tratta di un sottoinsieme dei codici BCH, caratterizzato⁵³ da parole di codice ad elementi *non binari* ma bensì L -ari; scegliendo $L = 2^k$ si possono mappare k bit della sequenza informativa \mathbf{m} in un singolo elemento del codice. Un codice di *Reed Solomon* (N, K) pone quindi in corrispondenza K simboli ad L livelli prelevati da \mathbf{m} , in N simboli ad L livelli della codeword, che viene poi trasmessa mediante un sistema di modulazione eventualmente anche anch'esso ad L livelli. I parametri del codice di Reed-Solomon sono espressi dalle relazioni $N = L - 1 = 2^k - 1$, $K = 1, 3, \dots, N - 2$ e $d_m = N - K + 1$, ed è in grado di correggere fino a $t = \frac{d_m - 1}{2}$ simboli L -ari per codeword; il codice è quindi particolarmente indicato nel caso in cui gli errori siano raggruppati e loro, dato che probabilmente incidono su di un solo simbolo (o due).



17.3.2 Codice convoluzionale

Individua un modo di realizzare la codifica di canale che, a differenza dei codici a blocchi, produce una sequenza binaria i cui valori dipendono da gruppi di bit di ingresso *temporalmente sovrapposti*, in analogia all'integrale di convoluzione che fornisce valori che dipendono da *intervalli* dell'ingresso, pesati dai valori della risposta impulsiva. Un generico codice convoluzionale è indicato con la notazione $CC(n, k, K)$, che lo descrive capace di generare gruppi di n bit di uscita (sequenza $\{b_j\}$) in base alla conoscenza di K simboli di ingresso (sequenza $\{a_j\}$), ognuno composto da k bit.



La memoria dei K ingressi è usualmente rappresentata mediante un registro a scorrimento che ospita gli ultimi $K \cdot k$ bit di ingresso, dove per ogni nuovo simbolo a_j che entra da sinistra, i precedenti scorrono a destra, ed il più "vecchio" viene dimenticato. Ognuno degli n bit di uscita $b_j(i)$, $i = 1, 2, \dots, n$ è ottenuto a partire dai $k \cdot K$ bit di memoria, eseguendo una somma modulo 2 tra alcuni di essi⁵⁴. Resta valido quindi il concetto di *coding rate* $R_c = \frac{k}{n}$ che rappresenta *quanti bit di informazione* sono presenti per ogni bit di uscita dal codificatore.

⁵¹http://en.wikipedia.org/wiki/BCH_code

⁵²http://en.wikipedia.org/wiki/Berlekamp-Massey_algorithm

⁵³http://en.wikipedia.org/wiki/Reed-Solomon_error_correction

⁵⁴Gli n modi di scegliere quali dei $k \cdot K$ bit sommare, per ottenere ognuno degli n bit di uscita, sono determinati mediante n vettori generatori g_i , $i = 1, 2, \dots, n$, di lunghezza $k \cdot K$, contenenti una serie di cifre binarie zero od uno, a seconda che l' i -esimo sommatore modulo due sia connesso (o meno) al corrispondente bit di memoria.

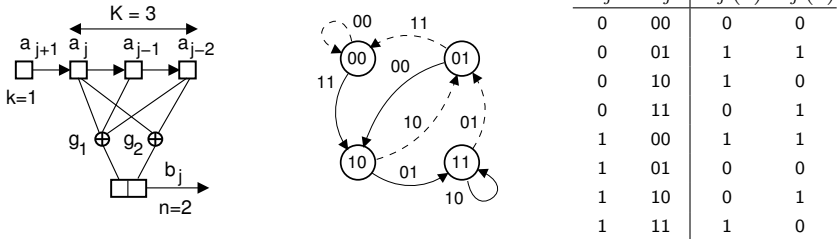


Figura 17.4: Architettura del codice convoluzionale $CC(2, 1, 3)$, diagramma di transizione e tabella della verità.

Diagramma di transizione Una volta definito il meccanismo di calcolo di b_j a partire da a_j , il codificatore può essere descritto mediante un *diagramma di transizione*, costituito da $2^{(K-1)k}$ stati S associati a tutte le possibili combinazioni di bit degli ultimi $K - 1$ simboli di ingresso; ad ogni stato competono quindi 2^k transizioni, una per ogni possibile nuovo a_j di ingresso, ed ognuna delle quali termina sul nuovo stato che si è determinato. Ad ogni transizione, è infine associato in modo univoco il gruppo di n bit $b_j(a_j, S_j)$ da produrre in uscita⁵⁵.

Esempio: CC(2,1,3) Per fissare le idee, definiamo un $CC(n, k, K) = CC(2, 1, 3)$ con $g_1 = [1 \ 1 \ 1]$ e $g_2 = [1 \ 0 \ 1]$, caratterizzato da un coding rate $\frac{1}{2}$, e mostrato nella parte sinistra della figura 17.4. Per questo, costruiamo il diagramma di transizione mostrato al centro, in cui le transizioni tra stati sono tratteggiate o piene, in corrispondenza rispettivamente dei valori a_j pari a zero o ad uno, e sono etichettate con la coppia di bit in uscita b_j , calcolabile mediante i vettori generatori $g_{1,2}$, come risulta dalla tabella della verità mostrata a destra. Come si può notare, ogni stato ha *solo due* transizioni, e quindi solo *due valori* di uscita (la metà dei 4 possibili con due bit); inoltre, questi valori differiscono in *entrambi i bit*⁵⁶.

Qualora si ponga in ingresso la sequenza $\{a\} = \{\dots 010100\}$ ⁽⁵⁷⁾, è possibile osservare che la sequenza di stati risulta $\{S\} = \{\dots, 01, 10, 01, 10, 00\}$, mentre quella di uscita è $\{b\} = \{\dots, 10, 00, 10, 11\}$, semplicemente seguendo il diagramma di transizione, a partire dallo stato iniziale 00. D'altra parte, è possibile anche il procedimento opposto: conoscendo $\{b\}$, si può risalire ad $\{a\}$, percorrendo di nuovo le transizioni etichettate con i simboli b_j . In definitiva, osserviamo come ad ogni coppia $(\{a\}, \{b\})$ sia biunivocamente associata una sequenza di stati $\{S\}$.

Diagramma a traliccio Per meglio visualizzare le possibili sequenze di stati, costruiamo il *diagramma a traliccio* (TRELLIS) del codificatore, mostrato a sinistra di fig. 17.5 riportando sulle colonne i possibili stati attraversati ai diversi istanti j , col-

⁵⁵Lo stesso valore di b_i potrebbe essere prodotto da più di una delle $2^{k \cdot K}$ diverse memorie del codificatore.

⁵⁶La dipendenza di b_j da (a_j, S_j) è legata alla scelta dei generatori g_i . Nel caso in cui un valore $b_j(i)$ sia sempre uguale ad uno dei k bit di a_j , il codice è detto *sistematico*. La scelta dei g_i può essere effettuata mediante simulazioni numeriche, per individuare il gruppo che determina le migliori prestazioni.

⁵⁷Per comodità di rappresentazione, il bit più a destra nella sequenza $\{a\}$ è il primo in ordine di tempo ad entrare nel codificatore.

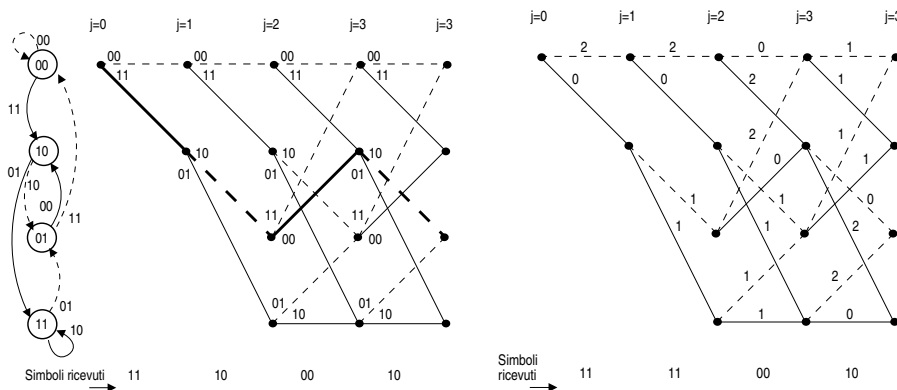
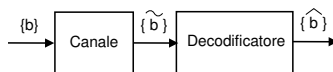


Figura 17.5: Diagramma a traliccio e costi d_H per la sequenza ricevuta

legando i *nod*i del traliccio con transizioni piene o tratteggiate nei casi in cui siano relative ad ingressi pari ad 1 o 0, e riportando sulle transizioni stesse i valori di uscita. Con riferimento alla sequenza *codificata* $\{b\}$ riportata in basso, l'effettiva successione di stati $\{S\}$ è rappresentata dalle linee piú spesse.

Decodifica di Viterbi Consideriamo ora il caso in cui la sequenza codificata $\{b\}$ venga trasmessa su di un canale, e che sia ricevuta *con errori*.



In generale, non sarà piú possibile rintracciare una sequenza di stati tale da produrre *esattamente* la sequenza ricevuta $\{\tilde{b}\}$, ed il problema diviene quello di individuare la sequenza di stati $\{S\}$ tale da produrre una $\{\hat{b}\}$ la piú vicina a $\{\tilde{b}\}$. Allo scopo di misurare questa differenza, utilizziamo la *distanza di Hamming* $d_H(\hat{b}, \tilde{b})$ pari al numero di bit diversi tra $\{\tilde{b}\}$ e le possibili $\{\hat{b}\}$ ⁵⁸, e etichettiamo gli archi del traliccio con le $d_H(\hat{b}_j, \tilde{b}_j)$ tra il simbolo da emettere \hat{b}_j e quello osservato in ricezione \tilde{b}_j (vedi lato destro di fig. 17.5). In tal modo, per ogni particolare sequenza di stati $\{S\}$ è possibile determinare un *costo* pari alla somma delle $d_H(\hat{b}_j, \tilde{b}_j)$ relative alle transizioni attraversate dal traliccio⁵⁹. Pertanto, la sequenza $\{\hat{b}\}$ piú vicina a $\{\tilde{b}\}$, ossia tale che

$$d_H(\hat{b}, \tilde{b}) = \min_{\{S\}} \left\{ \sum_j d_H(b_j, \tilde{b}_j) \right\}$$

può essere individuata come quella associata alla sequenza di stati $\{\hat{S}\}$ di *minimo costo*⁶⁰.

Dato che da ogni stato si dipartono 2^k archi, ad ogni istante il numero di percorsi alternativi aumenta di un fattore 2^k , crescendo molto velocemente all'aumentare di

⁵⁸Questo caso viene indicato con il termine *hard-decision decoding* in quanto il ricevitore ha già operato una decisione (quantizzazione) rispetto a \tilde{b} . Al contrario, se i valori ricevuti sono passati *come sono* al decodificatore di Viterbi, questo può correttamente valutare le probabilità $p(\tilde{b}/\hat{b})$ ed operare in modalità *soft decoding*, conseguendo prestazioni migliori.

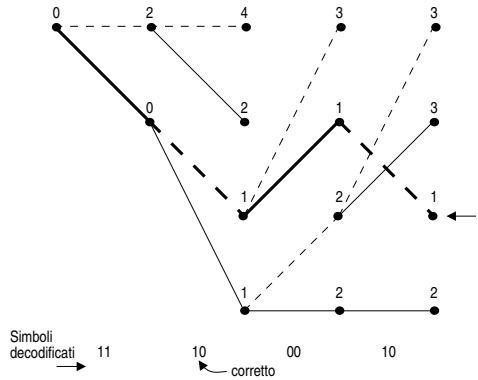
⁵⁹Ad esempio, con riferimento alla fig. 17.5, la $\{S\} = \{00, 10, 11, 01, 10\}$ ha un *costo* pari a 3.

⁶⁰Qualora la distanza tra \tilde{b}_j ed un possibile \hat{b}_j sia espressa come probabilità condizionata $p(\tilde{b}_j/\hat{b}_j)$, il processo di decodifica è detto *di massima verosimiglianza*.

j. L'enumerazione *completa* dei percorsi può essere però evitata, notando che quando due percorsi con costi diversi si incontrano in uno stesso nodo, quello di costo *maggiore* sicuramente *non* è la parte iniziale del percorso di minimo costo, e quindi può essere eliminato.

Questa filosofia si applica al caso in questione con riferimento alla figura che segue, che mostra come il calcolo dei *costi parziali* avvenga *per colonne* da sinistra a destra, scrivendo sopra ad ogni nodo il costo del *miglior* percorso che lo raggiunge. Ad ogni colonna sono scartati i percorsi che si incontrano con uno migliore, cosicché il numero di percorsi *sopravvissuti* è sempre pari al numero di stati 2^k .

All'estremità destra della figura una freccia indica la minima $d_H(\hat{b}, \tilde{b})$, associata al percorso a tratto spesso, e che permette di individuare la $\{\hat{b}\}$, che come si vede è *quella esatta*.



Tralasciamo ora di approfondire la teoria che consente l'analisi dettagliata dell'algoritmo, e ci limitiamo alle seguenti

Riflessioni

- l'esempio fornito si mostra in grado di correggere un errore pur impiegando un coding rate pari ad $\frac{1}{2}$, migliore (ad es.) di quello ($\frac{1}{3}$) del codice a ripetizione;
- la d_H del miglior percorso corrisponde al numero di bit errati (nel caso in cui siano stati corretti) nella \tilde{b} ricevuta;
- si verifica errore (cioè $\{\hat{b}\} \neq \{b\}$) se $d_H(\hat{b}, \tilde{b})$ è *minore* di $d_H(b, \tilde{b})$;
- le capacità di correzione del codice migliorano aumentando la d_H tra le possibili sequenze $\{b\}$ ⁶¹;
- la d_H tra diverse $\{b\}$ aumenta con $\frac{K}{k}$, in quanto la matrice di transizione tra stati diviene più sparsa, ed i valori di $\{b\}$ sono più interdipendenti;
- se il miglioramento di cui sopra è ottenuto aumentando K , ciò equivale ad estendere nel tempo la memoria del codificatore, ma senza per questo alterare il tasso di codifica $R_c = \frac{k}{n}$.

⁶¹La minima distanza tra le sequenze codificate è indicata come d_{min} , e può essere trovata come la d_H tra una $\{b^0\}$ tutta nulla ($\{b^0\} = \{\dots 00000000\}$) e quella con il minor numero di uni, che si diparte e ritorna (nel traliccio) dallo/allo stato 00.

Capitolo 18

Codifica di sorgente multimediale

Affrontiamo ora gli aspetti applicativi della teoria esposta al capitolo 17 e riguardante la codifica di sorgente continua, tipicamente applicata a *segnali audio-visivi*, e che dunque si particolarizza in base alle loro peculiari caratteristiche, ed alle caratteristiche degli organi sensoriali umani che ne sono i destinatari naturali. Per quanto riguarda l'*audio*, sono illustrate le tecniche orientate a riprodurre il segnale nel tempo come ad es. nel PCM, quindi quelle più orientate al segnale vocale e basate su di un modello del sistema fonatorio di produzione, ed infine le tecniche basate su di un modello del sistema uditivo come per l'MP3. Si passa quindi alla codifica di *immagini* fisse, alle quali possono applicarsi le tecniche nate per la codifica di sorgente discreta come ad es. GIF, oppure altre più ispirate alla fisiologia del sistema visivo come JPEG. La trattazione della codifica *video* si articola attraverso l'evoluzione storica dei vari standard che se ne sono occupati, arricchendo le tecniche di codifica di immagine con la rappresentazione del movimento e di come questo possa essere predetto, fino a generare il multiplex numerico con i programmi televisivi che pervadono la nostra esistenza.

18.1 Codifica audio

Al § 4.2.1.1 abbiamo svolto una valutazione approssimata della distorsione introdotta dal processo di quantizzazione di segnale audio, ricavando che l'utilizzo di M bit/campione si traduce in $SNR_q(M)|_{dB} \simeq 6 \cdot M$ dB. Quindi, al § 4.2.2 si è mostrato come adottando una caratteristica di quantizzazione logaritmica anziché lineare, ci si può adattare meglio alla effettiva densità di probabilità del segnale vocale, rendendo inoltre SNR_q relativamente poco sensibile alla sua effettiva dinamica, dando luogo alla cosiddetta codifica PCM con *legge A* o *legge μ* , standardizzata nel 1988 da ITU-T come G.711,¹. Mentre questa costituisce un formato universale di scambio permettendo la compatibilità tra dispositivi e tecnologie, nel seguito sono state sviluppate diverse tecniche alternative², capaci di offrire la stessa (o migliore) qualità di ascolto con velocità di trasmissione contenute, non solo per segnali vocali in banda telefonica, ma anche per segnali a banda larga, musicali, e multicanale, di cui tentiamo ora una sommaria rassegna.

¹<http://www.itu.int/rec/T-REC-G.711/e>

²Una raccolta di riferimenti a risorse relative a codec audio orientati alle applicazioni multimediali può essere trovata presso <http://labtel.ing.uniroma1.it/codecs>

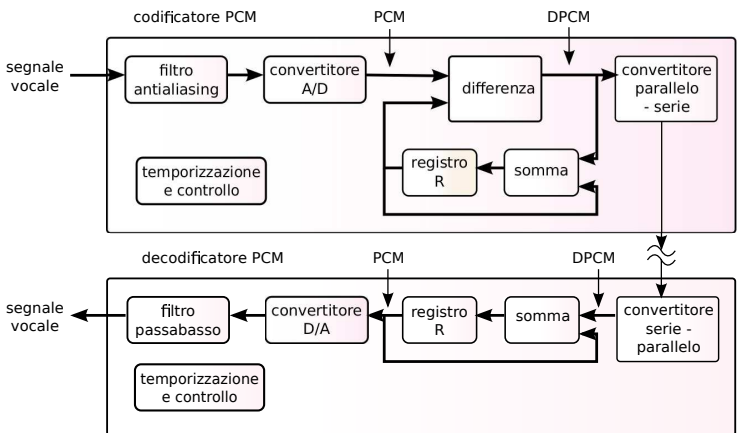


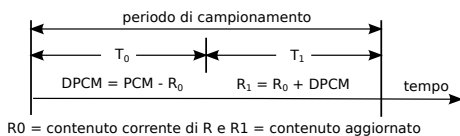
Figura 18.1: Codec audio Differential PCM o DPCM

18.1.1 Codifica di forma d'onda

Questa classe di codificatori opera esclusivamente nel dominio del tempo, agendo campione per campione, e ottiene una qualità comparabile o superiore a quella del PCM sfruttando le caratteristiche di memoria presenti nel segnale, e/o adattando alcuni parametri di funzionamento alle caratteristiche tempo varianti del segnale.

18.1.1.1 DPCM o PCM Differenziale

La prima variazione rispetto al PCM è stata quella di applicare il principio della codifica predittiva (pag. 441), semplicemente adottando il precedente campione di ingresso come *predizione* di quello successivo. Il corrispondente schema di elaborazione è mostrato in fig. 18.1, ed il suo funzionamento è suddiviso in due fasi: nella



prima il codificatore sottrae il campione precedente (all'inizio nullo) all'attuale, e nella seconda questa differenza è risommata al valore di differenza precedente (all'inizio nullo) in modo da ri-calcolare il valore attuale, e salvarlo nel registro di memoria. Il segnale differenza è caratterizzato da valori di ampiezza ridotti rispetto all'originale, e può essere codificato con 7 bit/campione, producendo ora una velocità per la qualità telefonica di 56 kbps. Il decodificatore si limita quindi a sommare alla differenza ricevuta il valore ricostruito del campione precedente, ed effettuare l'operazione di restituzione analogica. Osserviamo che il codificatore calcola il valore precedente mediante un circuito identico a quello presente al decodificatore, e per questo l'operazione è perfettamente invertibile.

18.1.1.2 ADPCM o DPCM Adattivo

Questo metodo differisce dal precedente per due aspetti: da un lato il processo di predizione tiene conto di più di un campione passato e non di uno solo come nel

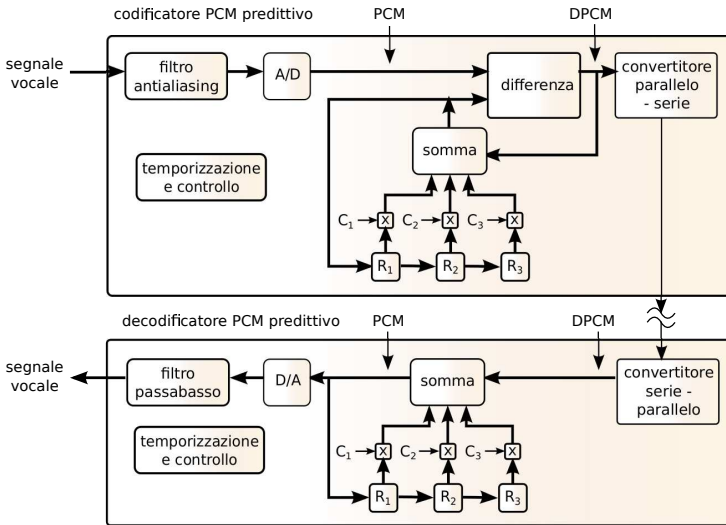


Figura 18.2: Codec DPCM con predittore a tre coefficienti costanti

DPCM, come descritto in fig. 18.2 in cui è mostrato un predittore del terzo ordine che in pratica consiste in un filtro trasversale i cui coefficienti sono fissati in base alle caratteristiche statistiche medie del segnale vocale. Il secondo aspetto è che ora il quantizzatore *modifica nel tempo* la propria dinamica di azione (da cui il termine *adattativo*) in base ad una stima della dinamica del segnale.

Nel lato sinistro della fig. 18.3 è mostrata una caratteristica di quantizzazione uniforme operante su di una dinamica di ingresso $\phi_x \hat{\sigma}_x$, con $\phi_x > 1$ scelto in modo da rendere trascurabile la probabilità che un valore di ingresso troppo elevato determini

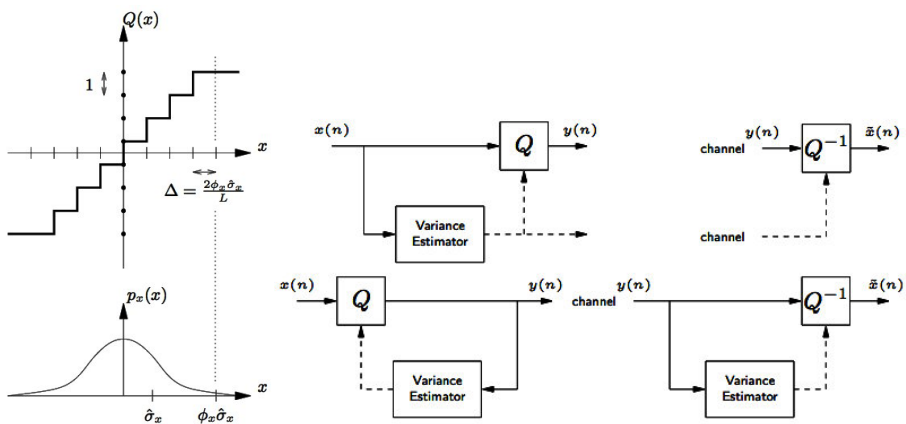


Figura 18.3: Dinamica di quantizzazione e sua stima diretta o *backward* - tratto da <http://cnx.org/content/m32074/latest/>

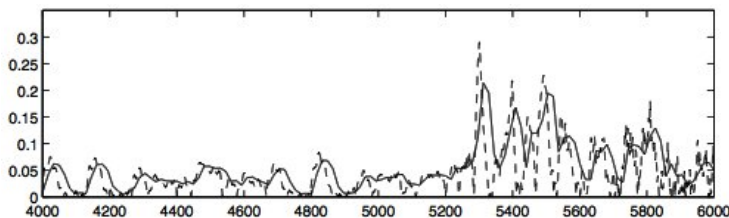


Figura 18.4: Stima recursiva backward della varianza, confrontata con i valori di $y^2(n)$, per $\alpha = 0.9$

la saturazione del quantizzatore. Utilizzando una stima a breve termine della varianza $\hat{\sigma}_x^2$ calcolata sugli ultimi campioni di segnale (a media nulla), ossia ad es. calcolando $\hat{\sigma}_x^2(n) = \frac{1}{N} \sum_{i=1}^N x^2(n-i)$, si possono rendere gli intervalli di decisione Δ piccoli nelle fasi di segnale piccolo, in modo da mantenere l'SNR costante anche per segnali con ampiezze molto variabili. Inoltre, è possibile *omettere* la trasmissione della stima di varianza se quest'ultima è calcolata in modalità *backward*, ossia a partire dai valori $y(n) = Q[x(n)]$, dato che la stessa operazione è eseguibile in modo indipendente anche dal lato del decodificatore. Infine, la stima della varianza è ulteriormente semplificata se realizzata mediante una formula recursiva, ossia

$$\hat{\sigma}_x^2(n) = \alpha \hat{\sigma}_x^2(n-1) + (1-\alpha) y^2(n)$$

il cui risultato è mostrato in fig. 18.4, dove la linea tratteggiata rappresenta il valore istantaneo di $y^2(n)$, mentre quella continua mostra i valori di $\hat{\sigma}_x^2(n)$ ottenuti in modo recursivo. Infine, la fig. 18.5 mostra i due estremi del codec ADPCM, che rimangono sincronizzati anche nel caso di saturazione del quantizzatore adattativo.

Il miglioramento della qualità ottenibile ha determinato la possibilità di ridurre il numero di bit (e di conseguenza di livelli) del quantizzatore a 5, 4, 3, 2 bit/campione, a cui corrispondono rispettivamente velocità di codifica di 40, 32, 24, 16 kbps. Questi sono i valori a cui si riferisce lo standard ITU-T G.721, successivamente confluito nel G.726.

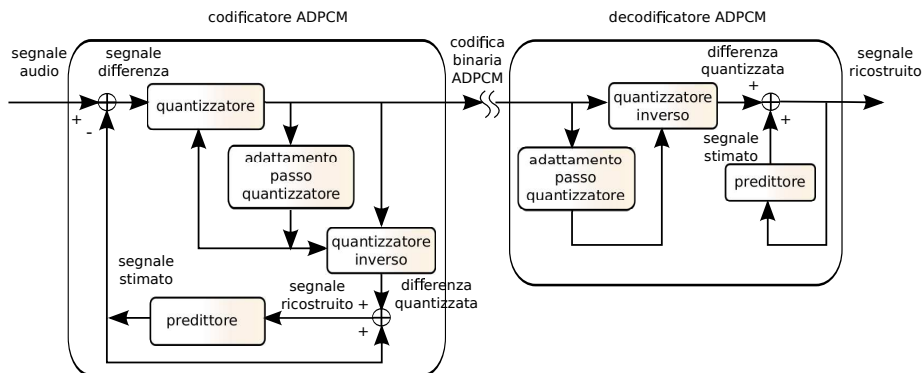


Figura 18.5: Architettura di un codec ADPCM

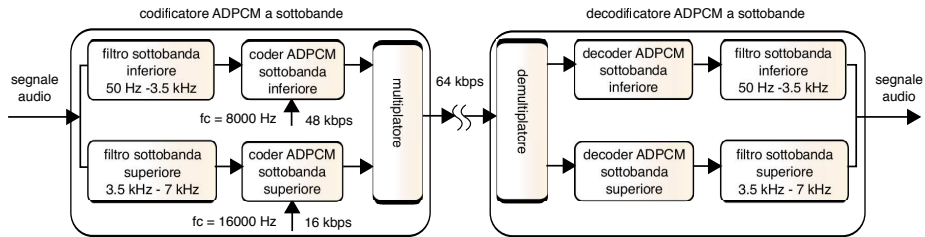


Figura 18.6: Architettura di un codec ADPCM a sottobande

18.1.1.3 Codifica per sottobande

Anche la raccomandazione G.722 è basata sulla codifica ADPCM, ma applicata ad un segnale con una banda audio più larga, riproducendo correttamente frequenze fino a 7 KHz. Ciò avviene dopo aver suddiviso le componenti frequenziali del segnale in due sottobande, come mostrato in fig. 18.6, mediante una coppia di filtri passa-basso e passa-alto con comune frequenza di taglio di 3.5 KHz. Il canale relativo alla semi banda superiore è quindi campionato a frequenza di 16 kHz, mentre l'altro è praticamente equivalente al segnale in banda telefonica preso in esame fino ad ora. Per entrambi i canali è applicata la codifica ADPCM, ma le velocità dei due sono impostate in modo differente, dando più importanza alla componente di bassa frequenza, percettivamente più rilevante: ad esempio, si può scegliere di assegnare 16 kbps alle alte frequenze e 48 alle basse, ottenendo un totale di 64 kbps per una qualità risultante migliore del G.711, in quanto ora si opera su di un segnale a larga banda, con risultati idonei ad applicazioni come la videoconferenza.

Lo stesso schema di codifica per sottobande più ADPCM è proposto anche dallo standard G.726, ma applicato ad un segnale a qualità telefonica, offrendo le velocità di 40, 32, 24 e 16 kbps.

18.1.2 Codifica basata su modello

I metodi fin qui discussi non tengono particolarmente conto della natura del segnale da codificare. Restringendo viceversa il campo al solo caso di segnale vocale, le conoscenze relative alla sua particolare modalità di produzione possono essere usate per ridurre le informazioni da trasmettere, costituite ora dai parametri che caratterizzano un suo modello di generazione. Essendo questo il dominio delle scienze linguistiche e fonetiche, svolgiamo una piccola digressione in tal senso.

18.1.2.1 Produzione del segnale vocale

L'apparato fonatorio viene idealizzato per mezzo del cosiddetto *modello a tubi* (vedi fig. 18.7), in cui sia il *tratto vocale* (compreso tra le corde vocali e le labbra) che il *tratto nasale* (dal velo alle narici) sono pensati come una concatenazione di tubi di diversa sezione. Nei suoni vocalici l'aria che attraversa le corde vocali ne determina la chiusura periodica, dando origine ad un *segnale di eccitazione* anch'esso periodico detto *onda glottale*, la cui frequenza è detta *pitch*³; la differenza di area delle diverse sezioni

³[http://en.wikipedia.org/wiki/Pitch_accent_\(intonation\)](http://en.wikipedia.org/wiki/Pitch_accent_(intonation))

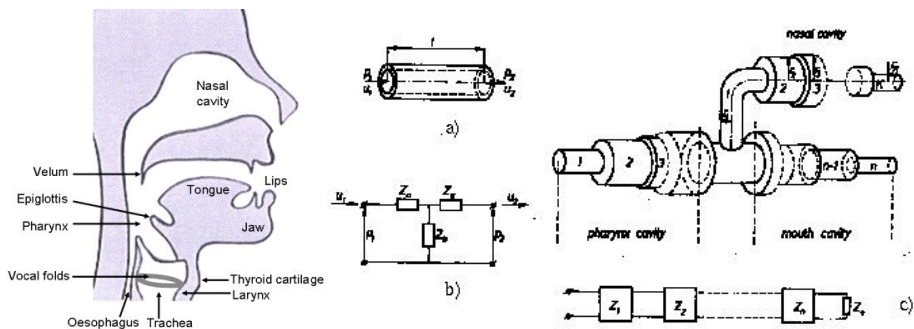


Figura 18.7: Rappresentazione schematica del tratto vocale e relativo modello a tubi

provoca un *disadattamento di impedenza acustica*⁴ e la conseguente formazione di onde riflesse (vedi fig. 18.8), che per lunghezze d'onda in relazione intera con la lunghezza del tratto vocale, determinano fenomeni di *onde stazionarie*, ovvero di *risonanze*⁵, che in fonetica sono indicate come *formanti*. La collocazione in frequenza delle formanti produce un *effetto filtrante* che modifica lo spettro dell'onda glottale, producendo così *il timbro* corrispondente ai diversi suoni della lingua⁶. Il tratto vocale termina quindi con l'apertura delle labbra, che nel modello a tubi corrisponde ad una *impedenza di*

⁴Si applica in pratica la stessa teoria valida per le linee elettriche, in cui al posto di tensione e corrente, ora si considerano rispettivamente pressione p e velocità u

⁵Si tratta di un fenomeno in qualche modo simile a quello che si verifica soffiando in una bottiglia, e producendo un suono che dipende dalla dimensione della stessa.

⁶Infatti i diversi suoni vocalici e/o consonantici sono prodotti mediante diverse posture articolatorie (la posizione di lingua, mascella e labbra), ovvero diversi profili d'area del tratto vocale, nonché l'attivazione o meno del tratto nasale. Pertanto ai diversi fonemi corrispondono differenti frequenze formanti, e dunque una diversa risposta in frequenza.

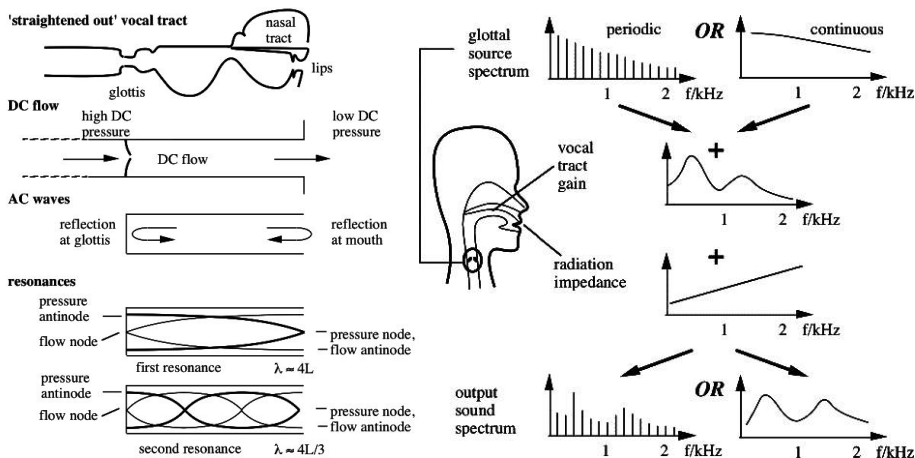


Figura 18.8: Natura delle risonanze del tratto vocale e loro effetto filtrante sull'onda glottale

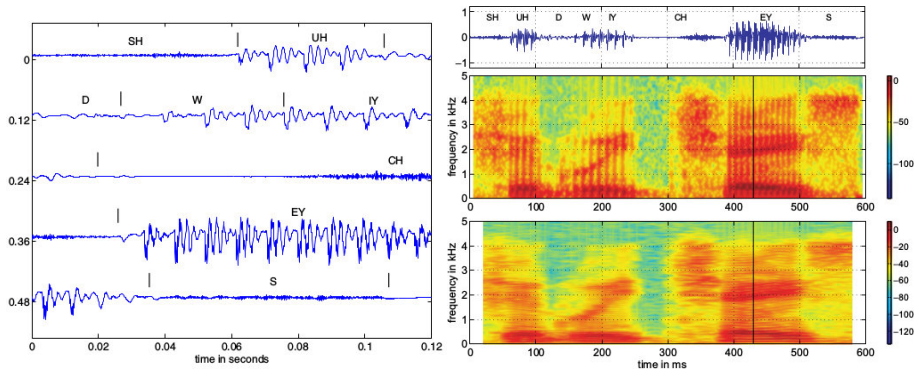


Figura 18.9: Forma d'onda e spettrogramma per la frase *should we chase*

radiazione che produce un effetto di derivata, e dunque *un'enfasi* delle alte frequenze per lo spettro complessivo del parlato. Infine, si assume il modello valido anche per i suoni *fricativi*, prodotti anziché mediante le corde vocali, mediante una occlusione che causa una turbolenza nel flusso d'aria.

Caratteristiche tempo-frequenza del segnale vocale La parte sinistra di fig. 18.9 mostra la forma d'onda relativa alla frase inglese “should we chase” (*dovremmo inseguire?*) assieme alla relativa trascrizione fonetica⁷, mettendone in luce il carattere quasi periodico in corrispondenza delle vocali e quello tipo rumore per le consonanti, nonché la diversa durata dei vari suoni, l'assenza di confini temporali precisi tra gli stessi, e la diminuzione del periodo di pitch a fine frase, corrispondente all'intonazione crescente tipica di una frase interrogativa. In particolare, notiamo come per i suoni vocalici i singoli periodi di pitch siano caratterizzati da una brusca discontinuità prodotta dall'onda glottale, seguita da oscillazioni smorzate legate alle risonanze del tratto vocale.

Il segnale viene quindi campionato a 10 KHz e suddiviso in *finestre di analisi*, per le quali vengono calcolate delle DFT, la cui densità di energia in dB è riprodotta *in verticale* mediante una scala cromatica come mostrato nei *sonogrammi*⁸ presenti al lato destro di fig. 18.9, che permettono di valutarne la variabilità temporale delle caratteristiche spettrali. Il differente aspetto dei due diagrammi è dovuto alla diversa lunghezza di finestra, pari rispettivamente a 10 e 40 msec per il grafico superiore ed inferiore⁹. In entrambe le rappresentazioni sono ben evidenti le *traiettorie delle formanti*, che evolvono in modo *continuo*, coerentemente con la velocità di articolazione del parlante.

⁷I simboli usati sono noti come *arphabet*, vedi <http://en.wikipedia.org/wiki/Arphabet>, e la pronuncia dovrebbe essere qualcosa del tipo *sciuduiceis*.

⁸Tratti da <https://books.google.it/books?id=Z60tr8Hj1WsC>

⁹Una finestra di 10 msec ha durata comparabile con il periodo di pitch, e ciò produce l'effetto a striature *verticali* del primo diagramma, meno pronunciato verso la fine, dove il pitch è più elevato. Una finestra di 40 msec si estende su più periodi di pitch, e determina una migliore risoluzione in frequenza, cosicché nel diagramma inferiore si possono notare delle striature *orizzontali* che corrispondono alle *armoniche* della frequenza di pitch.

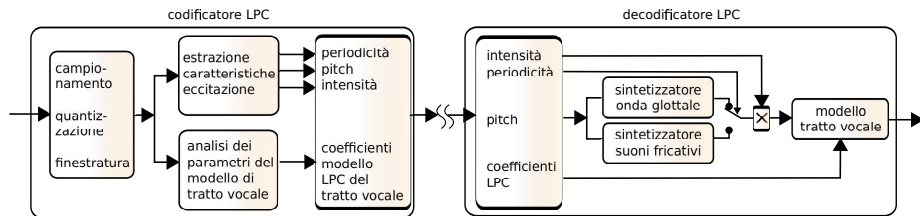


Figura 18.10: Schema di codificatore e decodificatore LPC

18.1.2.2 Codifica a predizione lineare - LPC

Il modello di produzione e le caratteristiche illustrate portano a formulare un processo di codifica basato sulla suddivisione del segnale vocale in intervalli (o finestre di analisi) di estensione tra i 10 ed i 30 msec, durante i quali il segnale può essere considerato praticamente stazionario¹⁰, e condurre una *analisi* (o stima) dei parametri del modello, che consistono in

- il tipo di eccitazione (periodica o caotica), la sua frequenza fondamentale (o *pitch*) se periodica, e la sua intensità;
- i parametri che caratterizzano l'effetto filtrante del tratto vocale

e quindi trasmettere questi valori, in modo che in ricezione sia possibile riprodurre un segnale simile all'originale mediante un decodificatore del tipo illustrato in fig. 18.10. In particolare, il modello *a tubi* (e quindi basato sulle risonanze) del tratto vocale si presta a considerare un filtro *di sintesi* di tipo numerico e *recursivo*, ossia che calcola il valore dei campioni di uscita y_n a partire da una combinazione lineare di p campioni passati $\hat{y}_n = \sum_{i=1}^p a_i y_{n-i}$, più un *errore di predizione* e_n che rappresenta il processo di eccitazione, ovvero

$$y_n = \hat{y}_n + e_n = \sum_{i=1}^p a_i y_{n-i} + e_n \quad (18.1)$$

Per ogni finestra di analisi, i coefficienti a_i (o coefficienti LPC del predittore di ordine p) si ottengono come quelli che rendono *minimo* il valore atteso dell'errore quadratico $E\{e_n^2\} = (y_n - \sum_{i=1}^p a_i y_{n-i})^2$ (ovvero, l'energia dell'errore), e sono individuati eguagliando a zero l'espressione delle sue derivate parziali rispetto ai coefficienti a_j : scriviamo dunque

$$\frac{\partial}{\partial a_j} E \left\{ \left(y_n - \sum_{i=1}^p a_i y_{n-i} \right)^2 \right\} = 2E \left\{ \left(y_n - \sum_{i=1}^p a_i y_{n-i} \right) y_{n-j} \right\} = 0$$

da cui si ottiene

$$E\{y_n y_{n-j}\} = \sum_{i=1}^p a_i E\{y_{n-i} y_{n-j}\} \quad (18.2)$$

¹⁰La durata di una sillaba può estendersi da 10-15 msec per le vocali *ridotte*, fino a più di 100 msec per quelle *accentate*.

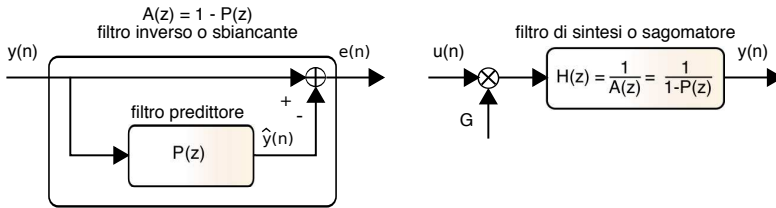


Figura 18.11: Filtro predittore, filtro inverso associato, e filtro di sintesi LPC

Il valore atteso $E\{y_n y_{n-j}\}$ è quindi *stimato*¹¹ a partire dai campioni temporali presenti nella finestra di analisi, ovvero

$$\begin{aligned} E\{y_{n-i} y_{n-j}\} &\doteq R_{yy}(|i-j|) \quad \text{e ponendo } k = |i-j| \\ &= R_{yy}(k) = \sum_{n=1}^{N-k} y_n y_{n+k} \end{aligned} \quad (18.3)$$

dove l'estremo superiore della sommatoria varia in modo da includere solo i campioni effettivamente presenti nella finestra di analisi¹². La (18.3) permette di riscrivere (18.2) come

$$R_{yy}(j) = \sum_{i=1}^p a_i R_{yy}(|i-j|)$$

che valutata per $j = 1, \dots, p$ individua un sistema di p equazioni¹³ in p incognite

$$\begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} = \begin{bmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \quad (18.4)$$

che può essere risolto nei termini dei coefficienti a_i mediante metodi particolarmente efficienti¹⁴; ed i coefficienti utilizzati dal decodificatore per applicare la 18.1.

Il filtro trasversale che esegue il calcolo $\hat{y}_n = \sum_{i=1}^p a_i y_{n-i}$ è indicato come *predittore*, ed associato ad un polinomio¹⁵ $P(z) = \sum_{i=1}^p a_i z^{-i}$; mentre il filtro FIR che valuta l'errore di predizione (o *residuo*) $e_n = y_n - \sum_{i=1}^p a_i y_{n-i}$ è indicato come *filtro inverso* o *sbiancante*, associato al polinomio $A(z) = 1 - P(z)$, e mostrato nel lato sinistro della fig. 18.11. Indicando ora con $G \cdot u_n$ una *codifica* del residuo e_n , il segnale di partenza può essere (quasi) ri-ottenuto come mostrato nella parte destra della fig. 18.11, ossia facendo passare u_n attraverso il filtro IIR $H(z) = \frac{1}{A(z)} = \frac{1}{1-P(z)}$.

¹¹Sottintendendo una ipotesi di stazionarietà ed ergodicità non vera, ma molto comoda per arrivare ad un risultato.

¹²La (18.3) è effettivamente una stima della *autocorrelazione* del segnale a durata limitata che ricade nella finestra di segnale, mentre l'inclusione nella sommatoria di un numero di termini pari al numero di campioni disponibili porta ad un diverso tipo di risultato, detto *metodo della covarianza*, ed un diverso modo di risolvere il sistema (18.4).

¹³dette di *Yule-Walker*.

¹⁴In base alle assunzioni adottate, $R_{yy}(j)$ risulta una funzione pari dell'indice j , e la corrispondente matrice dei coefficienti viene detta di *Toeplitz*, consentendone l'inversione mediante il metodo di *Levinson-Durbin* (vedi http://en.wikipedia.org/wiki/Levinson_recursion), che presenta una complessità $O(n^2)$ anziché $O(n^3)$, come sarebbe necessario per invertire la matrice dei coefficienti.

¹⁵Una breve analisi della relazione tra DFT e trasformata *zeta* è svolta al § 4.4.1.

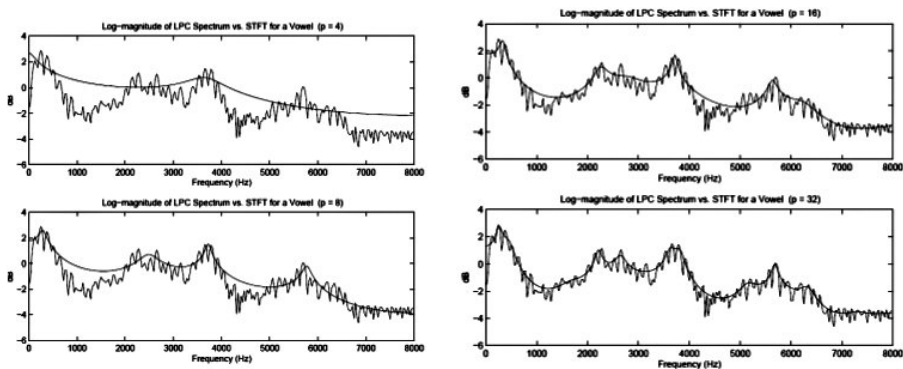
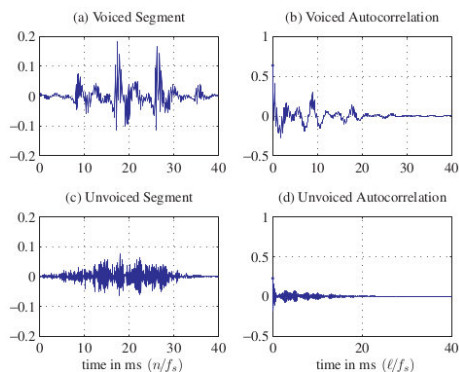


Figura 18.12: Approssimazione spettrale LPC per diversi ordini di predizione

Dato che, in base a considerazioni che non svolgiamo, e_n risulta possedere una densità spettrale *bianca*, $|H(z)|^2$ (calcolato per $z = e^{i\omega}$) rappresenta una vera e propria *stima spettrale* del segnale di partenza, come mostrato in fig. 18.12 per diversi valori di p , verificando che per suoni vocalici si ottengono risultati accettabili già per valori di p tra 8 e 14, mentre per le fricative l'ordine può essere ancora inferiore.

Stima del periodo di pitch Resta ora da illustrare il modo di decidere se la finestra di analisi contenga un suono sordo o sonoro, e nel secondo caso, il suo periodo. Osserviamo che *in media* la frequenza di pitch risulta pari a circa 120 e 210 Hz nel caso rispettivamente di voci maschili e femminili, con una estensione che varia approssimativamente da metà al doppio del pitch medio¹⁶. La stima del periodo di



pitch può essere realizzata a partire dalla funzione di autocorrelazione a breve termine 18.3, mostrata nella colonna di destra della figura a lato, a fianco delle finestre di segnale su cui è stata calcolata¹⁷, per un suono vocalico (sopra) e fricativo (sotto). Come evidente, nel caso del suono vocalico l'autocorrelazione presenta un primo picco a 9 msec ed un secondo a 18 msec, corrispondenti al periodo di pitch ed al suo doppio; viceversa nel caso del suono simile al rumore, non sono visibili picchi, come da aspettarsi nel caso di una segnale incorrelato. Pertanto, l'autocorrelazione può essere usata per

¹⁶Il pitch varia durante la pronuncia di una frase in accordo alla sua semantica, alla lingua, ed all'enfasi emotiva impressa dal parlatore. Da un punto di vista musicale, la dinamica dei valori (da metà al doppio) si estende quindi su di un intervallo di due ottave. L'intera gamma dei registri dell'opera si differenzia per 22 semitoni, dal Mi2 del basso al Do4 del soprano, ovvero un rapporto di frequenze pari a 3,6.

¹⁷In realtà prima del calcolo della autocorrelazione, il segmento di segnale è stato moltiplicato per una finestra di Hamming, che provoca lo smussamento visibile ai bordi.



Figura 18.13: Schema di codifica vocale ABS - *Analysis by Synthesis*

indicare la presenza o meno di un suono vocalico, e nel caso affermativo, stimare il suo pitch.

Nella pratica per i suoni sordi si ottengono buoni risultati di sintesi usando come eccitazione un vero e proprio rumore bianco; d'altra parte, per i suoni sonori l'uso di forme d'onda impulsive con periodo pari al pitch stimato, sebbene capaci di produrre un bit rate ridicolo fino a 2.4 kbps, non fornisce risultati particolarmente utilizzabili, producendo un voce piuttosto robotica. Per questo motivo, si sono sviluppate le tecniche seguenti.

18.1.2.3 Predizione lineare ad eccitazione residuale - RELP

Per ovviare alla sovra-semplificazione dello schema di sintesi riportato in fig. 18.10, dopo aver svolto l'analisi spettrale LPC, il residuo di predizione relativo alla finestra di analisi viene effettivamente calcolato, e applicata una tecnica di codifica di forma d'onda¹⁸: questo modo di operare è indicato come codifica RELP (*Residual Excited LP*).

Analysis by synthesis - ABS Anziché *calcolare* il residuo di predizione, codificarlo, e trasmetterlo in tale forma, la tecnica di *analisi via sintesi* adotta una tecnica *anello chiuso*, cercando di trovare quale segnale di eccitazione¹⁹ fornire al filtro di sintesi in modo che il risultato sia quanto più possibile simile al segnale originale (vedi fig. 18.13); quindi, i parametri del filtro di sintesi e della eccitazione sono trasmessi al decoder. La funzione di minimizzazione opera dunque una vera e propria *ricerca tra i possibili segnali* di eccitazione.

Filtraggio percettivo Sempre in fig. 18.13 si mostra come il processo di minimizzazione prende in considerazione un segnale di errore ottenuto filtrando l'errore effettivo mediante un filtro di *pesatura percettiva*, il cui andamento frequenziale è sostanzialmente *reciproco* rispetto a quello stimato del segnale²⁰ (vedi fig. 18.14), in modo da

¹⁸In questo modo si evita anche di dover operare una esplicita decisione *sonoro/sordo*, visto che in realtà le due fonti di eccitazione possono essere presenti contemporaneamente, come per i cosiddetti suoni *affricati*.

¹⁹Generato per tentativi, oppure da scegliere in un dizionario di sequenze di eccitazione già codificate.

²⁰Il filtro di pesatura percettiva si ottiene a partire dagli stessi coefficienti di predizione a_i che descrivono l'andamento spettrale della finestra di segnale, definendo la sua trasformata zeta come $W(z) = \frac{A(z/\alpha_1)}{A(z/\alpha_2)} = \frac{H(z/\alpha_2)}{H(z/\alpha_1)}$ in cui, se $\alpha_{1,2}$ sono numeri reali, i poli di $W(z)$ si trovano alle stesse frequenze di quelli di $H(z)$ ma con raggio α_2 volte maggiore, così come gli zeri di $W(z)$ hanno modulo α_1 volte maggiore. Scegliendo $0 < \alpha_{1,2} < 1$ e $\alpha_1 > \alpha_2$ per la $W(z)$ si ottiene l'effetto desiderato, e mostrato in fig. 18.14

attenuare la rilevanza dell'errore di predizione nelle regioni dove c'è più segnale²¹ ed esaltarla invece nelle regioni con meno segnale, sfruttando così il fenomeno percettivo noto come *mascheramento uditivo* (vedi pag. 483). Anche se per questa via l'energia totale del rumore è maggiore, l'effetto soggettivo è migliore.

Multi pulse linear prediction - MPLP Lo schema operativo suggerito dalla tecnica ABS è stato inizialmente realizzato cercando di *costruire* la sequenza di eccitazione ottima (ossia in grado di minimizzare l'errore pesato percettivamente) come una sequenza di pochi impulsi sparsi, decidendone uno alla volta: pertanto viene inizialmente trovata l'ampiezza e la posizione *ottime* per un primo impulso, poi per un secondo (con il primo fisso), e così via, fino al numero di impulsi desiderati, tipicamente 4-5 ogni 5 msec, ottenuti suddividendo una finestra di 20 msec in quattro sotto-trame, ognuna con 40 campioni, se $f_c = 8000$ Hz.

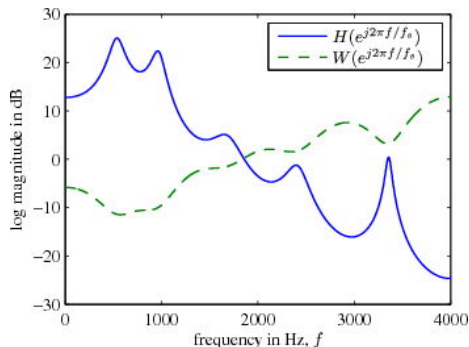


Figura 18.14: Spettro LPC vocalico e relativo filtro di pesatura percettiva dell'errore di predizione

Regular pulse excitation with long-term prediction - RPE-LTP o GSM 6.10

Il metodo MPLP presentava una complessità proibitiva, ma ha dato luogo alla versione semplificata RPE-LTP usata inizialmente nella telefonia GSM per fornire una velocità di 13 kbps. In questo caso dopo aver determinato la posizione del primo impulso nella sottofinestra, ne sono piazzati altri 9 ad intervalli regolari (un campione sì e tre no), e quindi l'ottimizzazione riguarda solo i valori delle ampiezze.

Rispetto allo schema di fig. 18.13 viene aggiunto un *predittore a lungo termine* o LTP, utilizzato per rimuovere dal segnale di eccitazione l'eventuale periodicità caratteristica dei suoni vocalici, e stimato a partire da sotto-finestre consecutive (vedi fig. 18.15). Il filtro LTP in essenza consiste in un semplice ritardo pari al periodo di pitch (e dunque $\gg p$), ed il predittore LTP relativo (vedi lo schema di decodifica) *ripropone* in uscita una copia ritardata ed attenuata dell'uscita stessa. Il codificatore GSM pertanto determina ritardo e attenuazione dell'LTP in base all'analisi del residuo di predizione LPC²², e lo usa per reintrodurre la componente periodica nella sequenza RPE di cui si sta valutando l'idoneità. Una volta che al residuo LPC viene sottratta la componente predicibile per tramite del LTP, ciò che rimane risulta effettivamente assimilabile ad un rumore, ed è indicato anche come *processo di innovazione*.

18.1.2.4 Quantizzazione vettoriale

Dato che la codifica del segnale di eccitazione è la parte che impegna più bit da trasmettere, l'evoluzione successiva della codifica vocale adotta la *quantizzazione vet-*

²¹La procedura di minimizzazione determina una eccitazione tale da rendere bianco il residuo al suo ingresso; dato però che questo ha subito il filtraggio da parte di $W(z)$, significa che le frequenze da questo deprese sono in realtà enfatizzate per il segnale di errore reale.

²²In effetti, mentre i coefficienti spettrali (denominati *parcor* in questo caso) sono determinati a partire dall'analisi dell'intera finestra di 20 msec, l'eccitazione RPE ed i parametri LTP sono ottenuti a partire da *sottofinestre* di 40 campioni, pari a 5 msec.

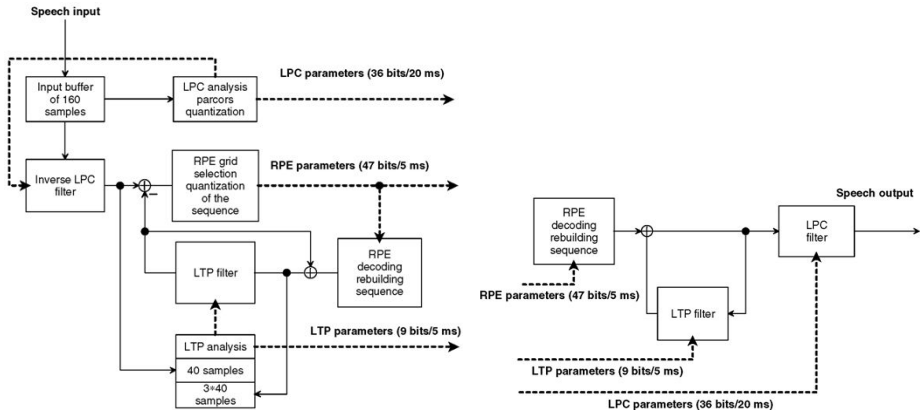


Figura 18.15: Codifica e decodifica GSM 6.10 full rate o RPE-LTP

toriale per evitare la trasmissione di tutti i parametri dell'eccitazione, che sono invece visti come *un vettore* e quindi rappresentati per mezzo di *una unica codeword* che identifica un elemento scelto in un *codebook* tale da rappresentare l'eccitazione con un errore minimo.

La costruzione del codebook è realizzata *partizionando* la distribuzione dei vettori in più regioni di decisione come quelle mostrate nell'esempio di fig. 18.16, in modo che un nuovo vettore possa essere classificato²³ come contenuto in una di esse, e venire quindi rappresentato dal *centroide* (i punti rossi) associato alla regione. I centroidi ed i confini di decisione sono determinati mediante un procedimento iterativo tale da minimizzare l'errore quadratico medio di rappresentazione²⁴.

I coefficienti dei vettori associati ai centroidi sono quindi memorizzati in un dizionario (o *codebook*) noto sia al codificatore che al decodificatore, in modo che ogni vettore può essere rappresentato, anziché da tutti i suoi coefficienti, dal solo indice della *codeword* del centroide più vicino: al solito, utilizzando M bit per rappresentare l'indice, il codebook sarà formato da 2^M diverse codeword. Oltre al codebook utilizzato per rappresentare i possibili vettori di innovazione, la codifica del segnale vocale si può avvantaggiare anche di un secondo codebook, usato per approssimare il vettore dei possibili coefficienti spettrali.

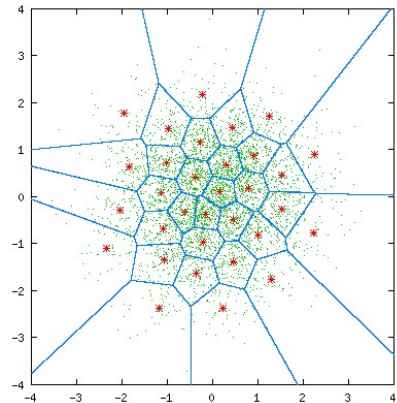


Figura 18.16: Regioni di decisione e centroidi per un quantizzatore vettoriale bidimensionale

²³Per questa classificazione, così come per poter definire l'insieme dei centroidi, occorre che sia definita una funzione di *distanza* tra vettori.

²⁴Vedi <http://www.data-compression.com/vq.html>, ma anche la nota 15 a pag. 58.

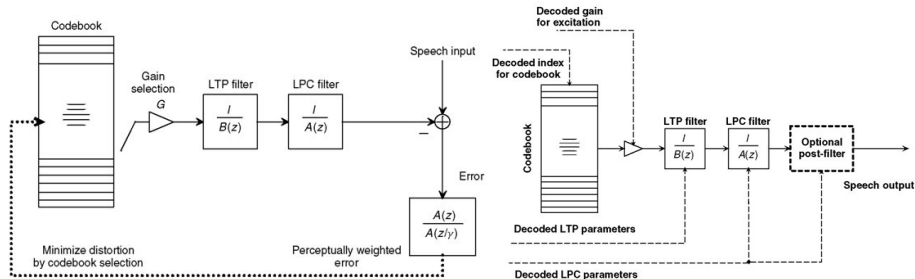


Figura 18.17: Codificatore e decodificatore CELP

18.1.2.5 Predizione lineare ad eccitazione codificata - CELP

La fig. 18.17 mostra lo schema realizzativo di un codificatore CELP, in cui sono evidenziati il filtro di predizione a lungo termine ed il filtro LPC, stimati in modalità *ad anello aperto*, ed il *filtro percettivo* che fa in modo che la densità spettrale dell'errore di predizione sia concentrata nelle regioni dove è presente segnale.

Per ogni codeword di eccitazione selezionata dal codebook, ed il guadagno associato, viene calcolata l'energia dell'errore ottenuto, ed il risultato confrontato con quello ottenibile mediante le altre codeword, finché non si trova la codeword che minimizza l'errore. Ovviamente questo modo di procedere è estremamente oneroso, ma si sono trovati metodi di ricerca più efficienti adottando tecniche di costruzione del codebook come combinazione di sequenze elementari, dando luogo alla famiglia dei codificatori *algebrici* o ACELP.

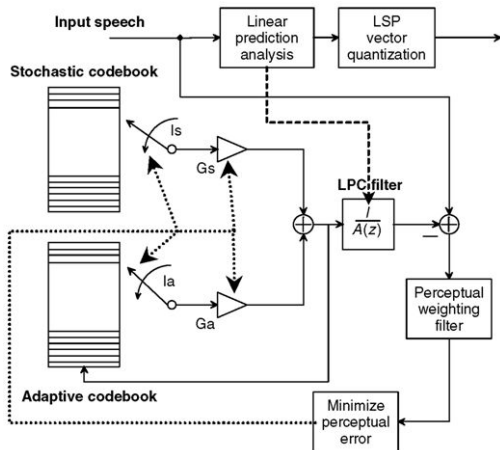


Figura 18.18: Codificatore CELP con codebook adattivo per la predizione a lungo termine

Infine, viene mostrato come anche i coefficienti spettrali LPC sono trasmessi mediante una codeword derivata da un processo di quantizzazione vettoriale. Possiamo elencare i seguenti standard che adottano una tecnica di questo tipo:

- Federal Standard 1016 (4800-16000 bit/s) CELP

D'altra parte, anche l'identificazione del LTP può essere ricondotta ad una ricerca ad anello chiuso, svolta ora nell'ambito di un *codebook adattivo*, costruito a partire dalla precedente sequenza di eccitazione ottima, replicata in forma traslata di un campione alla volta, come illustrato in fig. 18.18, che mostra appunto l'uso della eccitazione per la trama precedente per popolare il codebook adattivo: da questo viene quindi individuata la codeword I_a ed il guadagno G_a ottimi, e quindi individuata la codeword di innovazione I_s e G_s ottimi, riferiti ad un codebook detto *stocastico* perché costituito da sequenze pseudo casuali.

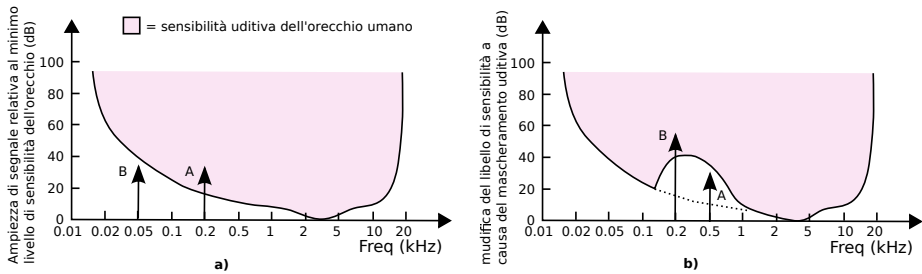


Figura 18.19: a) sensibilità uditiva alle diverse frequenze; b) mascheramento uditivo

- ITU-T 8-kbit/s G.729 CS-ACELP (*conjugate-structure algebraic CELP*);
- dual-rate multimedia ITU-T G.723.1 a 5.3 kbit/s con ACELP e 6.3 kbit/s con MP-MLQ (*multi-pulse maximum likelihood quantization*);
- ITU-T low-delay CELP 16-kbit/s G.728 - usa finestre di analisi molto brevi e una predizione lineare all'indietro per conseguire un ritardo di 2 msec;
- ETSI enhanced full-rate EFR-GSM e half-rate HR-GSM, con velocità di 12.2 e 5.6 kbps, così come i codec AMR (*adaptive multirate*) e WB-AMR, con velocità da 7.95 a 4.75 kbps;
- Speex²⁵ - un insieme di codecs open source esenti da brevetti e liberamente utilizzabili, con velocità (a banda stretta) da 5,95 a 24,6 kbps, e da 5.75 a 42,4 kbps per segnali con banda di 16 kHz

18.1.3 Codifica psicoacustica

Mentre la codifica di forma d'onda (§ 18.1.1) non fa assunzioni a riguardo della natura del segnale, i metodi esposti al § 18.1.2 sono tutti fortemente orientati a rappresentare segnali vocali. Viceversa, il gruppo di lavoro MPEG di ISO si è dedicato ad individuare metodi di codifica idonei alla trasmissione di segnali multimediali di natura qualsiasi, come ad esempio brani musicali. Inoltre, i vincoli relativi al basso ritardo necessario ad assicurare un buon grado di interattività vengono meno, e si possono dunque intraprendere elaborazioni più complesse, e che richiedono un tempo maggiore. Infine, vengono trascurati rigidi vincoli sulla velocità risultante, accettando invece che questa *vari* nel tempo in funzione del tipo di segnale da rappresentare.

Come vedremo tra breve, per queste tecniche si fa di nuovo uso di una codifica per sottobande, introdotta nella discussione dell'ADPCM, tenendo però anche conto di caratteristiche molto importanti della percezione sonora, il cui sfruttamento è già stato illustrato nella discussione del filtro di pesatura percettiva, ma che ora hanno un impatto ancora maggiore sulla realizzazione del codificatore. I codificatori che fanno uso di queste caratteristiche sono l'MPEG *layer 3* o MP3, il *Dolby AC*, e l'*advanced audio coding* o AAC.

Sensibilità uditiva e mascheramento in frequenza La fig. 18.19a mostra la curva di sensibilità del sistema uditivo, ovvero il livello di intensità minimo perché possa essere percepito un suono: come si vede, questo è molto variabile con la frequenza, per cui anche se il suono B (sinusoide o tono puro) ha la stessa intensità di A non può essere udito, mentre invece A sì. Ma ad una analisi più approfondita, si scopre che la

²⁵vedi <http://en.wikipedia.org/wiki/Speex>

presenza di un suono in una determinata regione di frequenza ha l'effetto di modificare la curva di sensibilità per le frequenze vicine, di fatto *mascherando* suoni a frequenze vicine che altrimenti avrebbero superato la soglia di sensibilità, come mostrato in fig. 18.19b: la presenza del suono B rende A non più udibile.

In realtà, l'estensione in frequenza per cui si verifica l'effetto di mascheramento dipende sia dalla frequenza del tono mascherante (come mostrato dalle curve in fig. 18.20 ottenute con toni a 1, 4 ed 8 kHz) che dalla sua intensità. In particolare, la banda delle frequenze mascherate viene detta *banda critica* ed ha una estensione differente alle diverse frequenze: si trova che sotto i 500 Hz la banda critica ha una estensione di circa 100 Hz, mentre a frequenze superiori aumenta (circa) linearmente per multipli di 100 Hz. Ad esempio, un segnale ad 1 KHz (2×500) produce una banda critica di 200 Hz (2×100), mentre a 5 kHz (10×500) questa vale circa 1 kHz (10×100).

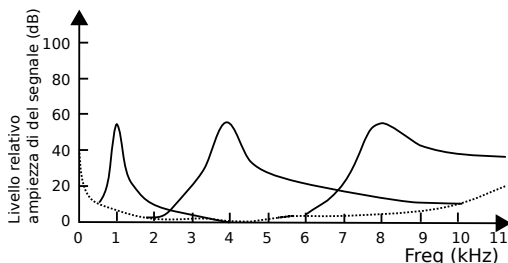


Figura 18.20: Variazione della banda critica in funzione della frequenza

Mascheramento temporale Il secondo effetto percettivo riguarda ancora una modifica alle curve di sensibilità, stavolta in modo *non selettivo* in frequenza, ma che coinvolge tutte le frequenze: si verifica infatti che dopo aver udito un suono forte, per il tempo necessario all'estinzione del suono e che tipicamente dura qualche decina di millisecondi (vedi fig. 18.21), l'orecchio non è più in grado di percepire suoni con intensità minore a quello che si sta estinguendo.

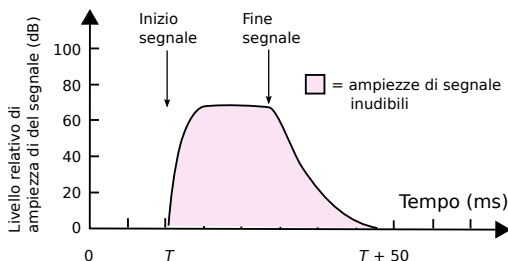


Figura 18.21: Mascheramento temporale

MPEG layer 3 Il gruppo di lavoro MPEG di ISO ha definito uno standard di codifica audio basato su tre livelli di complessità (e potere di compressione) crescente, ed il terzo (o MP3) è quello di gran lunga più popolare, anche grazie alla diffusione che ha avuto via Internet.

Lo schema di funzionamento di principio è mostrato in fig. 18.22: il segnale campionato in ingresso (a 32, 44.1 o 48 kHz) transita attraverso un PCM *encoder* che esegue un filtraggio²⁶ in 32 sottobande di eguale ampiezza, le cui uscite sono campionate a frequenza $1/32$ di quella di ingresso. Ogni 384 campioni di ingresso (pari a 12 msec se $f_c = 32$ kHz) sono quindi prodotti $384/32 = 12$ campioni per ogni sotto-banda, e per ognuna di esse è individuato il valore del campione più grande, che contribuisce sia ad

²⁶Eseguito mediante un banco di *filtri polifase*, vedi ad es. http://en.wikipedia.org/wiki/Polyphase_quadrature_filter o <http://cnx.org/content/m32148/latest/>. Le uscite dei filtri polifase, anche se campionate a frequenza inferiore a quella di Nyquist, sono esenti da aliasing, che viene cancellato dall'effetto delle altre sottobande.

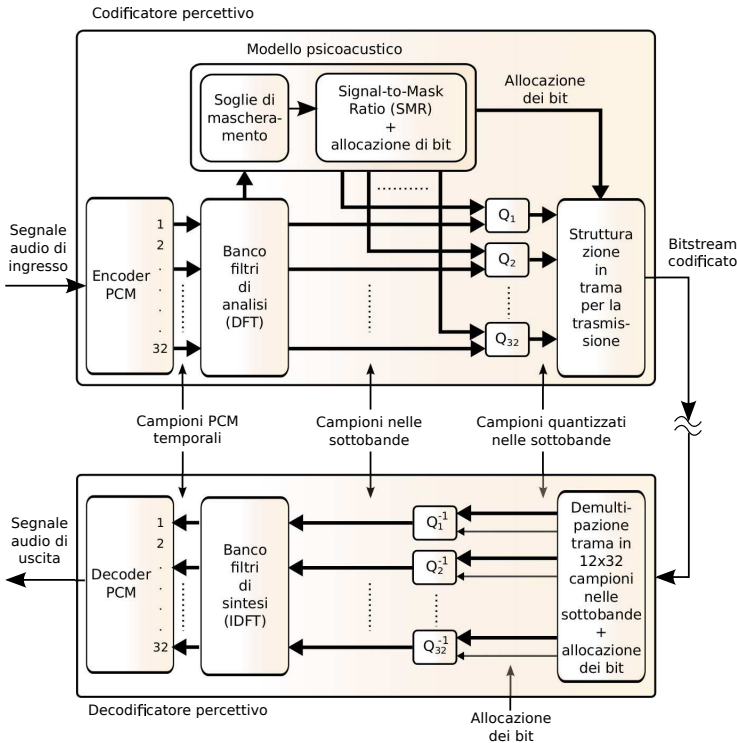


Figura 18.22: Codec percettivo MPEG

impostare la dinamica del quantizzatore per quella banda, sia come parametro per il modello psicoacustico.

Il modello psicoacustico riceve le informazioni prodotte da un banco di filtri di analisi realizzati mediante una MDCT,²⁷ che produce una stima spettrale con risoluzione maggiore di quella del primo banco di filtri, su cui basare le valutazioni di mascheramento uditivo, che a loro volta determinano per ogni sottobanda l'indicazione di una *signal to mask ratio (SMR)*, che a sua volta determina *quanti bit utilizzare* (e quindi quanti livelli) per la quantizzazione Q dei campioni relativi alle singole sottobande. Quelle contraddistinte da una maggiore sensibilità (ovvero nelle quali si percepiscono anche suoni deboli) saranno quantizzate con più accuratezza, e quindi con più bit e meno rumore; mentre le sottobande caratterizzate da una sensibilità inferiore possono essere quantizzate con meno bit, almeno finché l'*SNR* di quantizzazione si mantiene superiore all'*SMR*, dato che in tal caso il rumore è mascherato, e dunque non viene udito. Quindi, i 12 campioni delle 32 sottobande sono quantizzati tenendo conto sia della dinamica effettiva, che del numero di livelli in cui suddividere la dinamica. Infine, viene prodotta una struttura di trama che contiene, oltre ai campioni, anche le informazioni sulla effettiva allocazione dei bit.

Ad una futura edizione, un trattazione più approfondita.

²⁷Vedi http://en.wikipedia.org/wiki/Modified_discrete_cosine_transform

Riferimenti Si citano dei riferimenti essenziali sulla codifica audio, da cui sono anche tratte alcune illustrazioni

- Introduction to Digital Speech Processing, L. R. Rabiner and R. W. Schafer, <https://books.google.it/books?id=Z60tr8Hj1WsC>
- Beyond VoIP Protocols: Understanding Voice Technology And Networking, O. Hersent, J.P. Petit, D. Gurle <http://what-when-how.com/category/voip-protocols/>
- <http://www.data-compression.com/index.shtml>
- Voice Acoustics: an introduction - University of New South Wales, J. Wolfe, M. Garnier, J. Smith <http://www.phys.unsw.edu.au/jw/voice.html>
- Let's build an MP3-decoder! - Björn Edström <http://blog.bjrn.se/2008/10/lets-build-mp3-decoder.html>

18.2 Codifica di immagine

Un segnale di immagine può essere di tipo *vettoriale*²⁸, come nel caso di un disegno prodotto da un *plotter*, e rappresentato mediante un linguaggio descrittivo che codifica le operazioni grafiche necessarie alla sua realizzazione; al contrario, un segnale di immagine è detto di tipo *bitmap*, o *raster* (griglia, reticolo), quando è il risultato di un campionamento spaziale, come nel caso di una foto digitale, di un fax, o del risultato di un processo di scansione elettronico. Mentre le immagini vettoriali sono pienamente scalabili e ridimensionabili senza perdita di definizione, quelle bitmap sono ottimizzate per essere riprodotte nelle loro dimensioni originali, avendo già operato un processo di distorsione tale da sfruttare al più possibile le caratteristiche di predicibilità e di sensibilità percettiva.

18.2.1 Dimensioni

Per quanto riguarda le immagini bitmap, queste sono definite nei termini di una matrice di elementi di immagine o PIXEL (*picture elements*)²⁹, che sono l'equivalente bidimensionale dei campioni estratti da un segnale unidimensionale. Per ogni pixel è definito un valore associato alla intensità con la quale deve essere riprodotto: nel caso di immagini a colori, sono necessari tre valori di intensità, per cui una immagine è in realtà descritta da tre matrici, come approfondiamo di seguito.

Sebbene le dimensioni della matrice di pixel possano essere qualunque, nel corso del tempo si sono affermati una serie di valori di riferimento, associati ad altrettante serie di sigle, legate al tipo di dispositivo che deve poi riprodurre l'immagine, ma anche a quello da cui l'immagine viene acquisita; la tabella 18.1 riassume tali corrispondenze.

Ad esempio, la risoluzione VGA (640 x 480) trae origine dai parametri dello standard NTSC della televisione analogica, i cui quadri sono composti da una serie di 525 linee, di cui solo 480 visibili: volendo mantenere una risoluzione orizzontale pari a quella verticale, con un rapporto d'aspetto di 4:3, ogni linea deve essere campionata su $480/3 \times 4 = 640$ punti. Già prima dell'uso broadcast della TV digitale, la raccomandazione BT 601³⁰ stabilisce le regole per la conversione tra standard video differenti,

²⁸Esempi di formati per la grafica vettoriale sono SVG, EPS, PDF, e VRML.

²⁹Per alcuni anni, si è usato come sinonimo anche il termine PEL <http://www.foveon.com/files/ABriefHistoryofPixel2.pdf>.

³⁰Il sito di ITU-R <http://www.itu.int/ITU-R/index.asp?category=information&link=rec-601&lang=en> non consente l'accesso pubblico alla raccomandazione. Un approfondimento può essere svolto presso Wikipedia <http://it.wikipedia.org/wiki/BT.601>.

	banda	linee	fps	aspetto	colonne	righe	colore
PAL	6 MHz	625	25 int	4:3			
NTSC	5 MHz	525	30 int	4:3			
HDTV		1080		4:3	1440	1152	
				16:9	1920	1152	
SVGA				4:3	1024	768	
4:2:2		625/525	50/60 non int	4:3	720	576/480	360 × 576/480
4:2:0		625/525	25/30 int	4:3	720	576/480	360 × 288/240
VGA				4:3	640	480	
SIF		625/525	25/30 non int	4:3	360	288/240	180 × 144/120
CIF			30 non int	4:3	360	288	180 × 144
QCIF			15:7.5 non int	4:3	180	144	90 × 72

Tabella 18.1: Griglia dei parametri corrispondenti ai formati video

mediante l'uso di una comune frequenza di campionamento del segnale video a 13.5 MHz, individuando così nei $52 \mu\text{sec}$ ⁽³¹⁾ di una linea $52 \times 10^{-6} \times 13.5 \times 10^6 = 702$ campioni per linea, a cui si aggiungono 9 campioni neri in testa ed in coda per ottenere 720 campioni per linea; per un segnale a 525 linee si ottiene quindi la matrice 720×480 del formato 4:2:2, che approfondiremo tra breve.

Le matrici più grandi di 1024×768 sono spesso descritte in termini di *Megapixel* (es $1600 \times 1200 = 1,9$ Mpixel), spesso usati per confrontare la risoluzione (ma non necessariamente la qualità) dei mezzi di fotografia digitale; inoltre, i *grandi formati* traggono origine anche dalla tecnologia delle schede video per computer da un lato, e da quella della televisione ad alta definizione da un altro, come riassunto nella figura 18.23³².

Il formato SIF (*source intermediate format*) è ottenuto a partire dal 4:2:2, conservando la metà dei pixel sia in verticale che in orizzontale, e trascurando la metà dei quadri di immagine; il suo uso è orientato alla memorizzazione, e quindi usa una scansione non interallacciata. Il formato CIF (*common intermediate format*) è simile al SIF, tranne per aver perso il riferimento al numero di linee analogiche da cui deriva; il suo uso è orientato ai sistemi di videoconferenza, e da questo sono definiti formati a maggior risoluzione, come il 4CIF ed il 16CIF, equivalenti al 4:2:2 ed all'HDTV. Il formato QCIF (*quarter CIF*) è orientato alla videotelefonia, dimezzando ancora sia la risoluzione spaziale che quella temporale. Da questo è a sua volta derivato il formato SUB-QCIF (o S-QCIF) di 128×96 pixel, orientato a collegamenti lenti come quelli via modem.

³¹ Vedi fig. 11.3 a pag. 293.

³² La figura è tratta da Wikipedia, dove possono essere approfonditi gli altri aspetti legati a queste risoluzioni video http://it.wikipedia.org/wiki/Risoluzioni_standard.

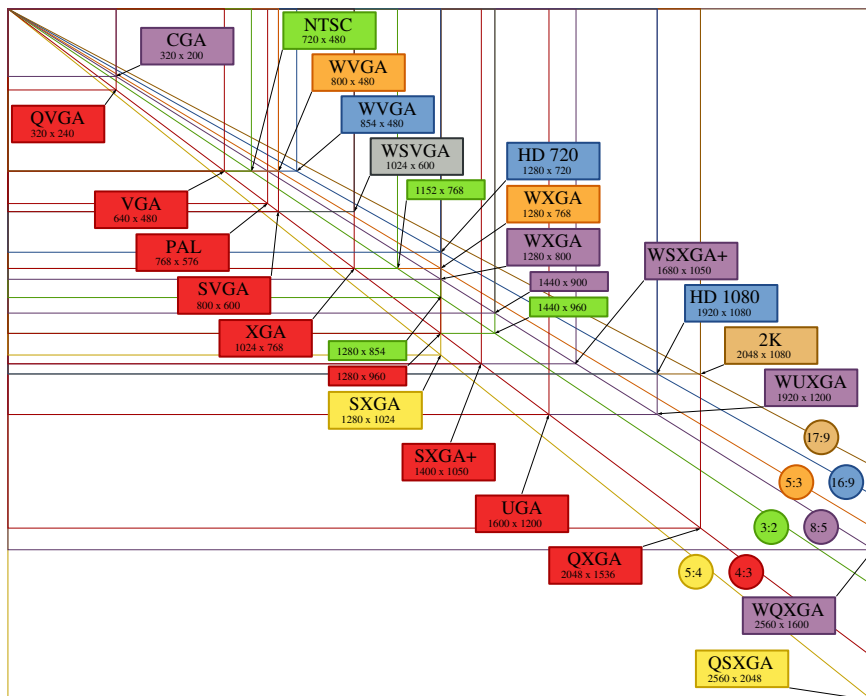


Figura 18.23: Risoluzioni standard o modalità video digitale

18.2.2 Spazio dei colori

I dispositivi di acquisizione e riproduzione di immagini a colori operano su tre diverse matrici di pixel, che rappresentano i tre colori di base della *sintesi additiva*, ossia *rosso*, *verde*, e *blu*, o RGB (dalle iniziali inglesi *Red*, *Green* e *Blue*). In figura 18.24 viene mostrato il principio di funzionamento di un *prisma dicroico*, che devia le tre componenti di colore verso tre diversi dispositivi di acquisizione. Variando quindi la proporzione con cui si sommano gli stimoli dei tre colori, si ottiene, oltre al bianco, anche qualunque altro colore. Sebbene dalle figure riportate sembra che il bianco risulti dal contributo in parti uguali delle tre componenti RGB, in realtà la scala di grigi della immagine *monocromatica* corrispondente si ottiene calcolando un segnale Y di *luminanza* secondo la formula

$$Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \quad (18.5)$$

che è quella usata per modulare il segnale video analogico³³. Come già discusso, in tale ambito la componente di colore viene trasmessa utilizzando due altri segnali, C_b o *crominanza blu* e C_r o *crominanza rossa*, secondo la formula

$$C_b = B - Y \quad \text{e} \quad C_r = R - Y \quad (18.6)$$

Disponendo dei segnali Y , C_b e C_r , si possono riottenere i valori RGB inserendo la (18.5) nelle (18.6), e risolvendo il sistema di tre equazioni in tre incognite risultante.

³³Vedi nota 42 a pag. 294.

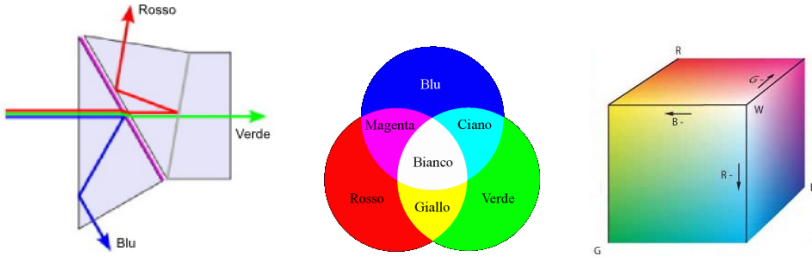


Figura 18.24: Prisma dicroico, sintesi cromatica additiva, cubo dei colori

Segnale video composito Al § 11.4.4 abbiamo descritto come nel segnale televisivo analogico la componente di colore sia trasmessa assieme alla luminanza, su di una diversa portante, con modulazione in fase e quadratura. In realtà, per diversi motivi, le componenti trasmesse non sono direttamente quelle individuate dalle (18.6), ma piuttosto componenti denominate U , V oppure I , Q , e così definite:

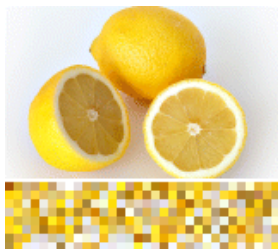
$$\begin{aligned} PAL : \quad U &= 0.493 \cdot C_b & NTSC : \quad I &= 0.74 \cdot C_r - 0.27 \cdot C_b \\ & & & Q &= 0.48 \cdot C_r + 0.41 \cdot C_b \\ & & & V &= 0.877 \cdot C_r \end{aligned}$$

Pertanto, in funzione delle diverse modalità di rappresentazione, un segnale video a colori può essere descritto indifferentemente da una delle seguenti quattro terne di segnali: RGB, $Y_C C_b$, YUV, YIQ.

Una descrizione alternativa dello spazio di colore è fornita dai parametri di *tinta*, *saturatione* e *chiarezza*, ovvero HUE, SATURATION e LIGHTNESS, o HSL: si tratta di attributi più legati alla descrizione percettiva che non alle tecnologie della riproduzione dell'immagine. Mentre la tinta descrive una famiglia di colori (es tutti i rossi), la saturazione ne indica il grado di purezza, ossia la presenza congiunta di altre tonalità; la chiarezza, infine, denota la luminosità del colore, rispetto ad un punto bianco. La terna HSL viene a volte usata per descrivere un colore nell'ambito di programmi di *computer graphic*, mediante i quali è fornito anche l'equivalente RGB.

Profondità di colore Dato che l'occhio umano non distingue più di 250 tinte diverse, e di 100 livelli di saturazione, si ritiene che utilizzare 8 bit per ogni componente dello spazio di colore RGB sia più che sufficiente. Con $8 \times 3 = 24$ bit per pixel (bpp) si possono infatti rappresentare $2^{24} - 1$ diversi colori, ovvero più di 16 milioni, molti dei quali indistinguibili ad occhio nudo. Modalità più spinte di quella a 24 bpp (detta *truecolor*) adottano 10, 12, 16 bit/componente, o rappresentazioni in virgola mobile, e sebbene non migliorino la qualità visiva, possono comunque essere usate in contesti professionali, per non perdere precisione nelle operazioni di editing ripetuto. Al contrario, profondità inferiori sono comunemente usate per risparmiare memoria, come nel caso di 15 bpp, che usa 5 bit per componente, o 16 bpp, che usa 6 bit per il verde, offrendo 65.536 colori diversi.

Palette Nel caso si decida di adottare profondità molto ridotte, come 8 bpp, si preferisce ricorrere ad una modalità detta a *colore indicizzato*: l'insieme dei colori presenti



nell'immagine viene *quantizzato*³⁴ in un insieme ridotto, i cui valori a 24 bpp sono memorizzati in una tavolozza (la *palette* detta anche *colour look-up table* o CLUT), che viene quindi utilizzata come un dizionario. La figura a lato mostra una immagine di esempio, assieme alla palette dei colori che usa. In questo modo, per ogni pixel dell'immagine è ora sufficiente specificare l'indice della palette dove è memorizzata la rappresentazione a 24 bpp del colore più prossimo.

Esempio Consideriamo una immagine in formato VGA rappresentata mediante una palette di 256 elementi da 24 bit: ognuno dei $640 \times 480 = 307.200$ pixel può quindi assumere uno tra 256 diversi colori, scelti tra $2^{24} = 16$ milioni. La dimensione di memoria occupata si ottiene considerando che per ogni pixel occorrono 8 bit per l'indice nella palette, e che la palette stessa ha dimensioni $256 \times 24 = 6144$ bit = 768 byte, e quindi in totale 307.968 byte.

Sottocampionamento del colore Nella tabella riportata a pag. 487 è presente la colonna *colore*, che mostra come la dimensione riservata alle matrici di pixel che codificano le informazioni di cromaticità sia ridotta di metà rispetto a quella della luminanza. Questo fatto trae origine da due buoni motivi: il primo è che l'acutezza visiva dell'occhio umano per ciò che riguarda le variazioni cromatiche è ridotta rispetto a quella relativa alle variazioni di luminosità; il secondo è che il segnale di cromaticità presente nel segnale video composto occupa una banda circa metà di quella del segnale di luminanza. Pertanto, le componenti di luminanza sono generalmente campionate con una risoluzione spaziale inferiore a quella del segnale di luminanza. Il tipo di sottocampionamento spaziale adottato per le componenti di cromaticità è generalmente caratterizzato da quattro numeri, in accordo allo schema di fig. 18.25:

- **4:4:4** - Non si effettua sottocampionamento, e le tre componenti hanno lo stesso numero di campioni. Applicato principalmente a segnali RGB trattati in studio di produzione.
- **4:2:2** - Questo schema si applica tipicamente alle rappresentazioni $YCbCr$, memorizzando per ogni 4 campioni di luminanza, 2 campioni della componente C_b e 2 della componente C_r , ed è utilizzato in ambito professionale e broadcast.
- **4:1:1** - In questo caso ogni quattro campioni di luminanza su una riga, ne viene preso uno per C_b ed uno per C_r . È lo schema usato nello standard DV NTSC.
- **4:2:0** - Ogni 4 campioni di luminanza, ne vengono salvati uno per C_b ed uno per C_r come per il caso 4:1:1, ma ora la cromaticità è campionata su righe alterne. In particolare, la versione utilizzata per l'MPEG-1 campiona assieme entrambi i segnali di cromaticità, una riga sì ed una no, mentre quella usata con il DV PAL li campiona a righe alternate, e prevede una riproduzione in modalità interlacciata.

18.2.3 Formato GIF

Il *Graphics Interchange Format* è un formato ad 8 bpp definito da *CompuServe* nel 1987³⁵ e da allora ha continuato ad essere molto popolare. Usa una *palette* con cui

³⁴Per una breve introduzione alla *quantizzazione cromatica*, può essere consultata Wikipedia http://en.wikipedia.org/wiki/Color_quantization

³⁵Il documento di specifica può essere trovato presso W3C: <http://www.w3.org/Graphics/GIF/spec-gif89a.txt>

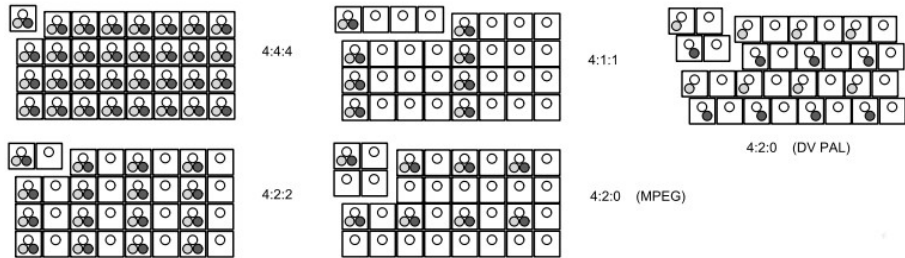
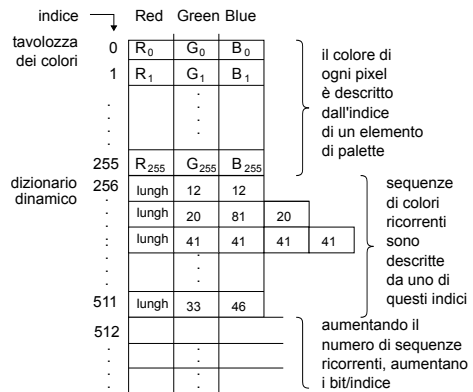


Figura 18.25: Sottocampionamento delle componenti di colore

rappresentare 256 colori scelti tra 16 milioni, e quindi comprime l'immagine mediante l'algoritmo LZW, individuando sequenze ricorrenti dei valori di colore. Un singolo file può contenere più immagini (ognuna con la sua palette) in modo da realizzare brevi animazioni. Il numero ridotto di colori rende il formato poco idoneo alla riproduzione di fotografie, ma più che adatto ad immagini più semplici, come ad es. un logo di pagina web. Per rappresentare i colori assenti dalla palette, il codificatore può ricorrere ad una operazione di *dithering*, alternando colori che, osservati da lontano, ricreano l'effetto della tonalità mancante.

Il metodo di compressione è illustrato con l'ausilio della figura che segue, e adotta come anticipato l'algoritmo LZW³⁶, il cui dizionario è inizialmente composto dalla palette, o meglio dai 256 valori ad 8 bit che indicizzano la terna RGB a 24 bit nella palette. Quando si incontra una sequenza di codici di colore già osservata, viene aggiunta una riga al dizionario, ed il valore dell'indice corrispondente viene usato per rappresentare tutta la sotto-sequenza; eventualmente, il numero di bit usati per indicare le righe del dizionario viene aumentato di uno. Per disegnare le sequenze di pixel rappresentate da indici inclusi nella sezione dinamica della tabella, occorre dunque individuare prima le rispettive terne RGB nella tavolozza.



PNG Dato che la compressione LZW era stata brevettata, venne sviluppata una codifica alternativa, denominata *Portable Network Graphics*. Al giorno d'oggi i brevetti relativi al formato GIF sono tutti scaduti, ed il formato PNG è stato standardizzato nella RFC 2083³⁷. Come per GIF, anche PNG è di tipo *lossless* (senza perdite), ossia individua una compressione invertibile, capace di replicare in modo identico l'immagine di partenza, ovviamente senza considerare il processo di quantizzazione che porta alla generazione della palette. Oltre alla modalità di colore indicizzato, PNG offre anche

³⁶Vedi § 17.1.1.7.

³⁷Reperibile presso il sito di IETF: <http://tools.ietf.org/html/rfc2083>

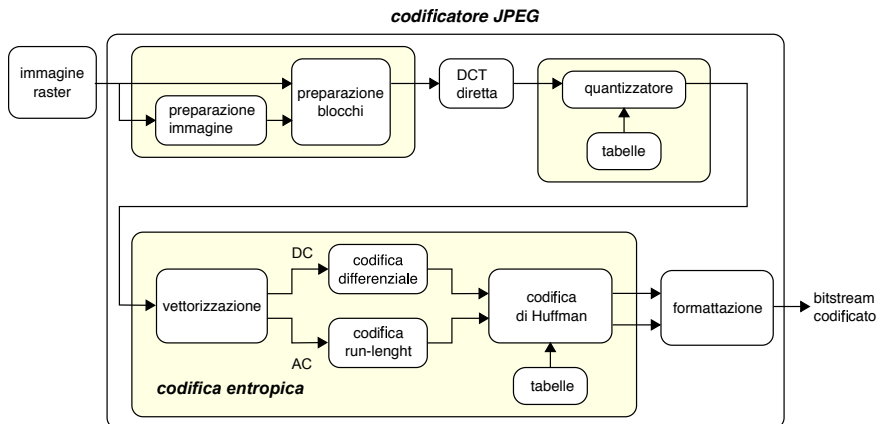


Figura 18.26: Stadi di elaborazione nella compressione jpeg

una modalità *truecolor* a 24 o 32 bpp, e per questo può correttamente rappresentare anche materiale fotografico, al punto da consigliare l'uso di PNG (anziché JPEG) nel caso si prevedano successive operazioni di editing dell'immagine.

Per quanto riguarda la compressione, PNG fa uso dell'algoritmo *deflate*, preceduto da un passaggio di compressione differenziale, in cui al valore che rappresenta il colore di un pixel viene sottratto il valore predetto a partire dai pixel adiacenti: in tal modo l'algoritmo *deflate* riesce a conseguire rapporti di compressione più elevati, riuscendo quasi sempre a battere le prestazioni di GIF.

18.2.4 Codifica JPEG

Il *Joint Photographic Experts Group* è un comitato congiunto ISO/ITU che ha definito lo standard internazionale per la compressione di immagini ISO 10918-1³⁸, particolarmente adatto alla codifica di immagini fotografiche. Descriviamo di seguito il funzionamento della modalità operativa detta *baseline*, o *lossy sequential mode*, che è quella che offre il migliore grado di compressione, e che prevede cinque stadi di elaborazione, mostrati alla fig. 18.26: preparazione dei blocchi, Discrete Cosine Transform (DCT), quantizzazione, codifica entropica, e formattazione.

Preparazione dell'immagine e dei blocchi L'immagine *raster* di partenza è formata da una o più matrici bidimensionali di valori (scala di grigi, oppure a colori indicizzati, o RGB, $Y_C C_b$, YUV, ...), eventualmente di dimensioni differenti (come nel caso $Y_C C_b$). Sebbene sia possibile elaborare direttamente una rappresentazione RGB, le migliori prestazioni si ottengono nello spazio $Y_C C_b$ con sotto-campionamento spaziale 4:2:2 o (meglio) 4:2:0, e dunque il primo passo è quello di convertire l'immagine in questa modalità di rappresentazione.

³⁸Scaricabile presso il W3C: <http://www.w3.org/Graphics/JPEG/itu-t81.pdf>

Ogni matrice viene quindi suddivisa in *blocchi* della dimensione di 8x8 pixel³⁹, ognuno dei quali è elaborato in sequenza in modo indipendente dagli altri.

DCT diretta Prima di procedere, la matrice Y (oppure le tre matrici R , G e B) che contiene valori ad 8 bit tutti positivi, viene normalizzata sottraendo ad ogni pixel il valore 128, in modo da ottenere valori tra -128 e 127. Quindi, per ogni blocco di 8x8 pixel, i cui valori indichiamo con $p(x, y)$, viene calcolata una nuova matrice di 8x8 valori $D(i, j)$ ottenuti come coefficienti di una *trasformata coseno discreta* (DCT) bidimensionale (vedi § 4.4.3):

$$D(i, j) = \frac{1}{4} c_i c_j \sum_{x=0}^7 \sum_{y=0}^7 p(x, y) \cos \frac{(2x+1)i\pi}{16} \cos \frac{(2y+1)j\pi}{16}$$

in cui c_i e c_j sono ognuno pari a $1/\sqrt{2}$ con indice i o j pari a zero, oppure $c_i = c_j = 1$ negli altri casi, mentre gli indici i e j variano tra zero e sette. Tralasciando di approfondire le relazioni esistenti tra DCT e DFT⁴⁰, consideriamo invece come i coefficienti $D(i, j)$ così ottenuti permettano la ricostruzione della matrice originaria nei termini di una somma pesata delle superfici rappresentate (per mezzo di una scala di grigi) nel diagramma riportato alla figura 18.27, mediante l'applicazione della DCT *inversa*

$$p(x, y) = \frac{1}{4} \sum_{i=0}^7 \sum_{j=0}^7 c_i c_j D(i, j) \cos \frac{(2x+1)i\pi}{16} \cos \frac{(2y+1)j\pi}{16}$$

Ma se fosse tutto qui, non avremmo realizzato la funzione di compressione! Questa è infatti realizzata dalle elaborazioni successive, a partire dalla rappresentazione in termini di blocchi DCT, di cui ora approfondiamo il significato. Osserviamo quindi che ognuna delle superfici elementari rappresentate in fig. 18.27 è legata ad una coppia i, j associata ad un coefficiente della DCT calcolata, in modo che tale coefficiente esprime il contenuto di frequenze spaziali descritto da quella particolare funzione della base. Per questo l'elemento $(i, j) = (0, 0)$ in alto a sinistra, ad andamento costante, è indicato come *coefficiente DC*, o componente continua, dato che essendo calcolato come somma di tutti i pixel, riflette un valore che è legato alla intensità media dell'intero blocco. I coefficienti legati alle funzioni della prima riga rappresentano contenuti di frequenza spaziale orizzontale, con un periodo via via minore spostandosi verso il margine destro, mentre quelli della prima colonna, frequenze verticali. I coefficienti localizzati all'interno della matrice esprimono contenuti di frequenze spaziali in entrambe le direzioni, con valori di frequenza tanto più elevati, quanto più ci si sposta verso l'angolo in basso a destra. Pertanto, i coefficienti descritti da indici diversi da $(0, 0)$ sono indicati come *coefficienti AC*.

L'esperienza pratica mostra come quasi sempre i coefficienti $D(i, j)$ presentino nella regione in alto a sinistra valori ben più elevati di quelli riscontrabili in basso a destra, come conseguenza della predominanza dei blocchi posti in corrispondenza ad aree dell'immagine quasi costanti, rispetto a quelli associati alla presenza di contorni netti e particolari dettagliati.

³⁹Notiamo incidentalmente come le dimensioni definite nella tabella di pag 487 siano multipli interi di 8. Se questo non è il caso, i blocchi ai bordi destro ed inferiore vengono riempiti con pixel scelti in modo da minimizzare le distorsioni risultanti.

⁴⁰Potremmo tentare comunque di estendere le considerazioni svolte al § 4.4.3 al caso bidimensionale...

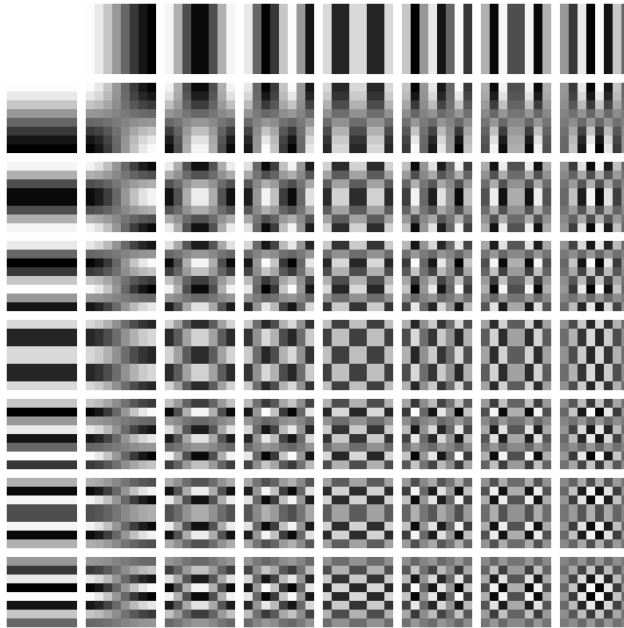


Figura 18.27: Rappresentazione grafica delle superfici DCT

Quantizzazione Questo passo della elaborazione JPEG mira a sfruttare il fenomeno percettivo della ridotta sensibilità dell'occhio umano alle frequenze spaziali più elevate, ovvero la capacità di *filtrare percettivamente* le componenti di errore corrispondenti ai dettagli più minuti. Per questo, il processo di quantizzazione è orientato a ridurre, ed eventualmente sopprimere, le componenti di immagine legate alle frequenze spaziali più elevate, introducendo di fatto *una soglia* sotto la quale si stabilisce di non trasmettere quelle informazioni che tanto non sarebbero percepibili. A questo scopo, ogni coefficiente $D(i, j)$ viene diviso per un coefficiente $Q(i, j)$ dipendente da (i, j) , ed il risultato viene arrotondato:

$$B(i, j) = \text{round} \left(\frac{D(i, j)}{Q(i, j)} \right)$$

Il risultato corrisponde ad un processo di quantizzazione, perché quando in ricezione il processo viene invertito (ri-moltiplicando il coefficiente per la stessa quantità), viene persa la precisione legata all'arrotondamento, e pari alla metà del coefficiente di divisione. La scelta dei $Q(i, j)$ è fatta in modo tale da utilizzare valori più elevati per gli indici (i, j) più elevati, in modo da ottenere due risultati: ridurre le componenti ad alta variabilità *spaziale* dell'immagine, e poter usare meno bit per codificare questi valori (più piccoli). Inoltre, molti dei coefficienti con (i, j) elevato, già piccoli di per se, quando divisi per un coefficiente di quantizzazione più elevato, non *sopravvivono* all'operazione di arrotondamento, in modo che tipicamente la parte in basso a destra della matrice $B(i, j)$ sarà tutta pari a zero, facilitando il compito della codifica run-length dello stadio successivo.

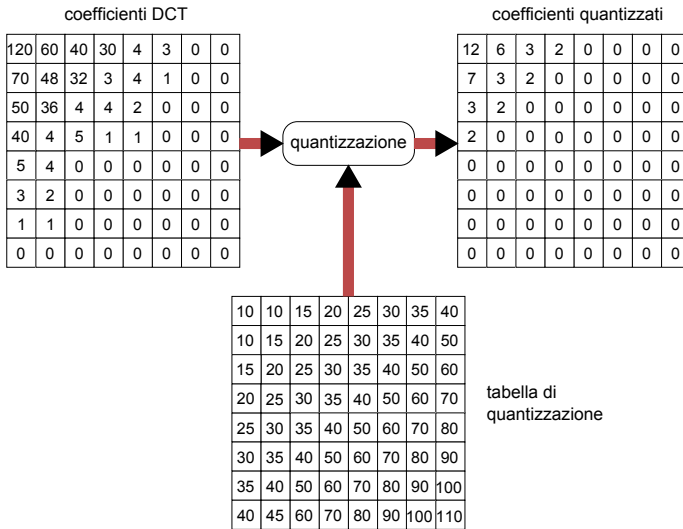


Figura 18.28: Processo di quantizzazione dei coefficienti DCT

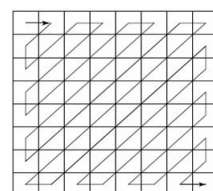
Esempio La figura 18.28 mostra un esempio di matrice di coefficienti DCT, assieme alla tabella di quantizzazione, ed al risultato dell'operazione. Notiamo come il valore dei coefficienti di quantizzazione aumenti allontanandosi dal coefficiente DC, e come nella matrice dei coefficienti quantizzati siano *sopravvissuti* solo i coefficienti relativi alle frequenze spaziali più basse.

Sebbene esistano delle tabelle di quantizzazione predefinite, i valori effettivi possono essere variati in base ad un compromesso tra qualità che si intende conseguire e fattore di compressione; tali valori vengono poi acclusi assieme al bitstream codificato durante la fase di formattazione, in modo che il processo di quantizzazione possa essere invertito in fase di riproduzione dell'immagine.

Codifica entropica Questo passo è un processo senza perdita, nel senso che non aggiunge altre distorsioni oltre a quelle introdotte dal passo di quantizzazione, ma è essenziale ai fini della compressione, e sfrutta le caratteristiche statistiche del risultato delle elaborazioni precedenti. Come posto in evidenza nello schema di pag. 492, la codifica entropica adotta due diverse procedure per i coefficienti DC e AC, che in entrambi i casi culminano con uno stadio di codifica a lunghezza variabile mediante codici di Huffman.

Vettorizzazione Le matrici 8x8 relative ai blocchi di elaborazione visti fin qui vengono ora trasformate in sequenze lineari da un processo di scansione a *zig zag* delle stesse, come mostrato dalla figura a lato.

La sequenza così ottenuta presenta il coefficiente DC in testa, a cui fanno seguito i rimanenti 63 coefficienti AC, ordinati in base al massimo valore di frequenza spaziale



che rappresentano. Se applichiamo la scansione zig-zag ai valori riportati nell'esempio precedente, otteniamo come risultato la sequenza

12 6 7 3 3 3 2 2 2 2 0 0 0 0 0

Codifica differenziale I blocchi adiacenti generalmente possiedono coefficienti DC molto simili tra loro, in virtù dell'omogeneità di ampie zone dell'immagine (pensiamo ad un porzione di cielo). Per questo motivo, anziché codificarli in modo indipendente, i singoli coefficienti DC di blocchi consecutivi vengono sottratti l'uno all'altro, e viene codificata solo la loro differenza. Ad esempio, se una sequenza di coefficienti DC risultasse pari a 12 13 11 11 10 ..., il risultato di questo processo di codifica differenziale darebbe luogo alla sequenza 12 1 - 2 0 - 1 ... (infatti, il valore *precedente* al primo coefficiente si assume pari a zero). Dato che differenze in valore assoluto piccole sono relativamente più frequenti di differenze grandi, si è scelto di adottare per queste una codifica a lunghezza di parola variabile, realizzata prima descrivendo ogni valore di differenza mediante la coppia (sss, valore), in cui sss rappresenta il numero di bit necessario per rappresentare il valore, e quindi concatenando una codeword di Huffman corrispondente ad sss, al codice binario che rappresenta il valore.

Esempio Per chiarire le idee, mostriamo le corrispondenze citate mediante due tabelle, che poi applichiamo all'esempio precedente.

differenza	N. di bit sss	valore codificato		sss	codeword di Huffman
0	0				
-1, 1	1	1=1	-1=0	0	010
-3, -2, 2, 3	2	2=10	-2=01	1	011
		3=11	-3=00	2	100
-7...-4, 4...7	3	4=100	-4=011	3	00
		5=101	-5=010	4	101
		6=110	-6=001	5	110
		7=111	-7=000	6	1110
-15...-8, 8...15	4	8=1000,	-8=0111	7	11110
⋮				⋮	
				11	11111110

Tornando dunque al nostro esempio della sequenza differenziale 12 1 - 2 0 - 1 ..., in termini di coppie (sss, valore) questa diviene (4, 12), (1, 1), (2, -2), (0, 0), (1, -1),... e quindi, sostituendo ad sss il relativo codice di Huffman preso dalla seconda colonna della seconda tabella, ed ai valori la loro rappresentazione indicata dalla terza colonna della prima tabella, otteniamo la sequenza di bit 101 1100, 011 1, 100 01, 010, 011 0,... in cui si sono mantenute le virgole per chiarezza. In definitiva, abbiamo usato un totale di 23 bit per rappresentare 5 differenze, che ne avrebbero richiesti 45 se codificate con 9 bit.

Codifica run-length Viene applicata alla sequenza di coefficienti AC che è il risultato dello *zig-zag scan*. In base all'effetto congiunto delle caratteristiche dei coefficienti della DCT, e del processo di quantizzazione, la sequenza degli AC in uscita dal vettoreizzatore presenta lunghe sequenze di zeri, consentendo di conseguire buoni rapporti di compressione mediante l'uso di una codifica *run-length*, realizzata scrivendo gli AC

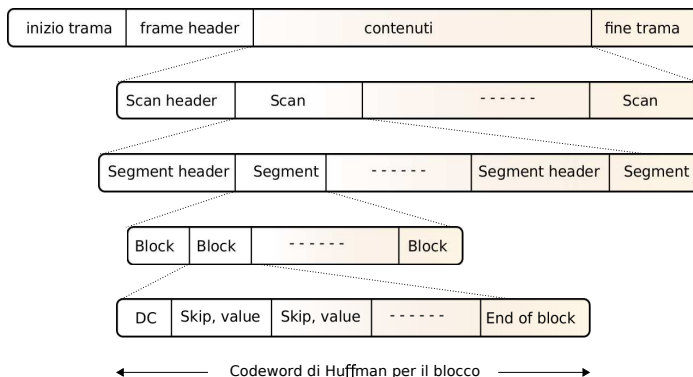


Figura 18.29: Formato del bitstream per la codifica JPEG

come una sequenza di coppie (*skip*, *ACN*), in cui *skip* rappresenta il numero di zeri nel run, e *ACN* è il coefficiente AC non nullo che viene dopo la sequenza di zeri. Quindi, il campo *ACN* viene espresso a sua volta nella forma *sss*, *valore*, come indicato dalla prima tabella riportata nell'ultimo esempio. Infine, la coppia *skip*, *sss* viene rappresentata con una codeword di Huffman individuata in un nuovo codebook appositamente definito.

Esempio Applicando la codifica run-length alla sequenza dei coefficienti AC individuati nell'esempio di vettorizzazione, ossia alla sequenza 6 7 3 3 3 2 2 2 2 0 0 0 ... 0 0, si ottiene una sequenza di coppie (*skip*, *ACN*), pari a (0,6), (0,7), (0,3), (0,3), (0,3), (0,3), (0,2), (0,2), (0,2), (0,2), (0,0) in cui l'ultima coppia (0,0) indica la fine del blocco, che in fase di decodifica viene quindi ricostruito riempiendolo di zeri. Anziché usare questa, proseguiamo adottando una diversa sequenza di coppie (*skip*, *ACN*), pari a (0,6), (0,7), (3,3), (0,-1), (0,0)⁴¹: sostituendo ai termini *ACN* di questa, la coppia *sss*, *valore*, e codificando quindi il termine *valore* come indicato nella prima tabella dell'esempio precedente, si ottiene (0, 3, 110), (0, 3, 111), (3, 2, 11), (0, 1, 0), (0,0). Il *bitstream* finale viene quindi realizzato sostituendo alle attuali coppie *skip*, *sss*, le rispettive codeword individuate alla colonna *Run/Size* della tabella a pagina 150 e segg. delle specifiche ITU-T T.81 <http://www.digicamssoft.com/itu/itu-t81-154.html>, ottenendo (100, 110), (100, 111), (111110111, 11), (00, 0), (1010), e producendo così un totale di 30 bit per rappresentare i 63 coefficienti AC.

Formattazione Lo standard JPEG definisce, oltre alla sequenza di operazioni indicata, anche il formato di trama con il quale devono essere memorizzato il bitstream finale. La struttura risultante è gerarchica, e mostrata alla figura 18.29. Al livello superiore troviamo una *frame header* che contiene le dimensioni compressive dell'immagine, il numero ed il tipo di componenti usate (CLUT, RGB, YCbCr, etc), ed il formato di campionamento (4:2:2, 4:2:0, etc.). Al secondo livello, troviamo uno o più *Scan*, ognuno preceduto da una intestazione in cui viene riportata l'identità del componente (R, G, B, o Y, C_b, C_r), il numero di bit usato per rappresentare ogni coefficiente di DCT, e la tabella di quantizzazione usata per quella componente. Ogni *Scan* è composta da uno o più *segmenti*, preceduti da un'ulteriore intestazione, che contiene il codebook di

⁴¹La nuova sequenza di coppie corrisponde ad una sequenza di coefficienti AC pari a 6 7 0 0 0 3 -1 0 0 0



Figura 18.30: Due fotogrammi consecutivi, e la differenza tra i rispettivi valori di luminanza

Huffman usato per rappresentare i valori dei blocchi del segmento, nel caso non siano stati usati quelli standard. Infine, nel segmento trovano posto le sequenze di blocchi dell'immagine, così come risultano dopo lo stadio di codifica entropica.

18.3 Codifica video

In accordo al metodo di realizzazione dei segnali video analogici, in cui i singoli quadri sono codificati indipendentemente gli uni dagli altri, la codifica video digitale può essere realizzata semplicemente applicando tecniche di codifica di immagine (come JPEG) ad ognuno dei quadri che costituiscono la sequenza video: questo tipo di approccio prende il nome di *moving JPEG* o MJPEG.

D'altra parte, i quadri relativi ad istanti temporali vicini sono spesso molto simili tra loro, anche se quanto siano simili, e per quanto tempo, dipende dal tipo di filmato. La presenza di *memoria* nella sorgente determina quindi la possibilità di ridurre il tasso informativo prodotto dalla codifica ricorrendo a tecniche predittive, tentando quindi di *stimare il movimento* presente in quadri contigui, e trasmettere solo l'informazione necessaria a *compensare* l'errore di predizione. Nella parte destra di fig. 18.30 è mostrata l'immagine differenza ΔY tra la componente di luminanza di due quadri consecutivi, consentendo di apprezzarne la relativa semplicità. Considerando poi che alcune regioni si sono mosse più di altre, il quadro da codificare è scomposto in sottoimmagini, per ognuna delle quali si ha un diverso spostamento, e viene calcolata una specifica differenza rispetto alla sotto-immagine precedente (e spostata); questa tecnica prende il nome di *compensazione del movimento*.

Tipo di quadro Come abbiamo fatto notare al § 17.1.1.6, le tecniche di codifica predittiva sono particolarmente sensibili agli errori di trasmissione, che possono causare una perdita di sincronismo tra i predittori di trasmissione e ricezione, e quindi l'impossibilità di ricostruire la restante parte di segnale. Per questo, nella codifica video sono presenti dei quadri *di riferimento* in corrispondenza biunivoca con un unico quadro di partenza, detti *intra-coded frames* o **I-frames**, che permettono al ricevitore di ri-partire da una condizione nota. Tra due quadri **I** sono poi presenti un certo numero di quadri **P** (*predicted*) come in fig. 18.31a, oltre che quadri **B** (*bidirectional*) come in fig. 18.31b, e che corrispondono rispettivamente alla codifica della compensazione del movimento calcolato a partire da un unico quadro precedente, o da una coppia di quadri passato e futuro.

I quadri **I** sono codificati mediante l'algoritmo JPEG, usando lo stesso coefficiente di quantizzazione per tutti i pixel delle DCT, conseguendo un rapporto di compressione relativamente basso, e sono inseriti a cadenza fissa con un periodo N tipicamente

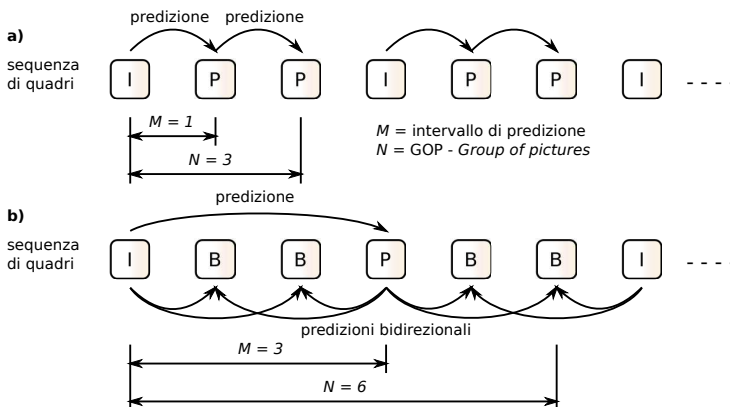


Figura 18.31: Esempi di sequenze di quadri con **a)** solo quadri di tipo I e P; **b)** quadri di tipo I, P e B

compreso tra 3 e 12: la sequenza di quadri compresi tra due quadri **I** è detta *group of pictures* o GOP. Come mostrato in figura 18.31, la codifica di quadri **P** può dipendere dal quadro **I** immediatamente precedente, o dalla ricostruzione di un precedente quadro **P**, ottenendo un fattore di compressione maggiore che per i quadri **I**; la distanza temporale tra **P** e l'originale **I** è detta *intervallo di predizione*, indicato con M .

Per realizzare la compensazione del movimento, ogni regione del nuovo quadro è confrontata con regioni *limitrofe* del quadro precedente, riducendo così la complessità di ricerca. Nel caso dei quadri **B** la ricerca delle regioni simili è invece svolta rispetto ai quadri **I** (o **P**) situati sia nel passato che al futuro, migliorando la precisione della stima di movimento, e conseguendo rapporti di compressione ancora maggiori, a patto di subire un aumento del ritardo di codifica, legato al dover attendere un quadro futuro.

Allo scopo di ridurre il ritardo di decodifica, la sequenza di quadri viene trasmessa con un ordine diverso da quello dei quadri originali, consentendo ai quadri **B** di essere riprodotti non appena ricevuti, e non dopo la ricezione del quadro *futuro* da cui dipendono. Pertanto, se la sequenza originale è ad esempio

IBBPBBPBBIBBP...

questa verrà trasmessa nell'ordine

IPBBPBBIBBPBB...

Stima di movimento e compensazione Specifichiamo innanzitutto cosa intendere con il termine *regione* prima usato per definire il dominio dell'operazione di confronto necessaria alla stima di movimento. Come mostrato in fig. 18.32a, considerando una suddivisione in componenti $Y C_b C_r$ ed un sottocampionamento 4:1:1, il quadro originale è suddiviso in N righe e M colonne di *macroblocchi* di 16x16 pixel, ed ogni macroblocco è rappresentato da sei blocchi 8x8 pixel, di cui quattro blocchi per la luminanza, più due blocchi per le componenti di cromaticità; ogni blocco 8x8 corrisponde quindi ad un equivalente numero di coefficienti DCT, ed è individuato all'interno del quadro, in base al suo indirizzo di riga e colonna.

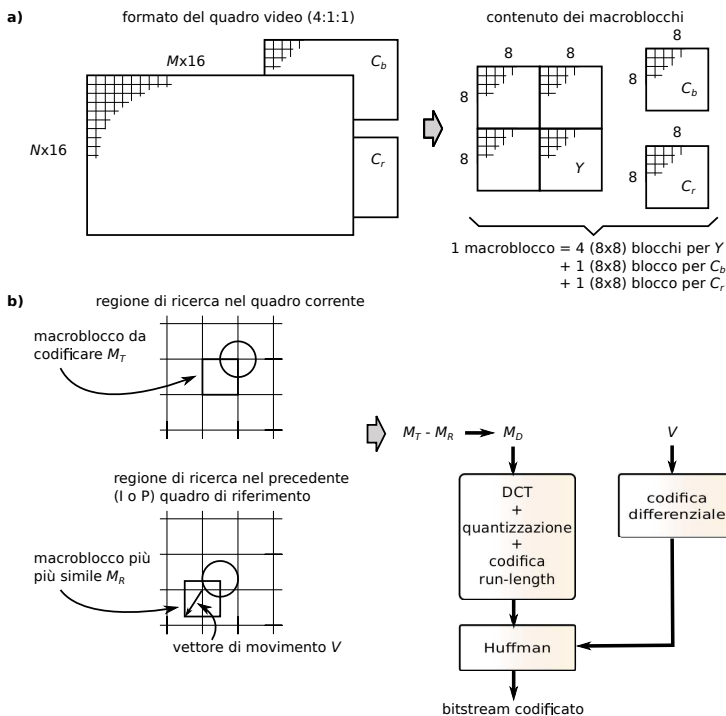


Figura 18.32: Codifica di un quadro P: a) struttura del macroblocco; b) procedura di codifica

Nella codifica dei quadri **P**, ogni macroblocco M_T del quadro corrente (*target*) è confrontato pixel per pixel con il corrispondente macroblocco M_R del quadro di riferimento, e nel caso sia riscontrata una sufficiente similitudine⁴² complessiva, viene trasmesso solo l'indirizzo del blocco. Altrimenti, il confronto viene ripetuto per tutti i possibili spostamenti del macroblocco target nell'ambito dei macroblocchi contigui⁴³, e qualora sia individuata una buona corrispondenza, il macroblocco viene codificato dal vettore di movimento V e dall'errore di predizione M_D . Con riferimento alla fig. 18.32b in cui l'immagine è simboleggiata da un cerchio, V rappresenta lo spostamento da applicare a M_T per portarlo a coincidere al meglio con il quadro precedente, ed è codificato come una coppia (x, y) corrispondente ad una *risoluzione di un pixel*. Al contrario M_D è composto dalle tre matrici ($Y C_b C_r$) dei valori differenza tra quelli di M_T spostato di V , ed M_R . I valori di V e di M_D relativi ai diversi macroblocchi di un quadro seguono poi due diversi percorsi di codifica, come specificato appresso.

Nel caso in cui la regione di ricerca sia estesa, i valori V possono risultare relativamente grandi; d'altra parte è probabile che macroblocchi vicini esibiscano vettori di spostamento molto simili tra loro. Per questi motivi, la sequenza dei V calcolati per

⁴²Il confronto è svolto considerando i soli valori di luminanza, e la similitudine valutata come media tra i valori assoluti delle differenze di luminanza.

⁴³l'effettiva estensione dell'area di ricerca non è oggetto di standardizzazione, mentre lo è la rappresentazione del risultato della ricerca.

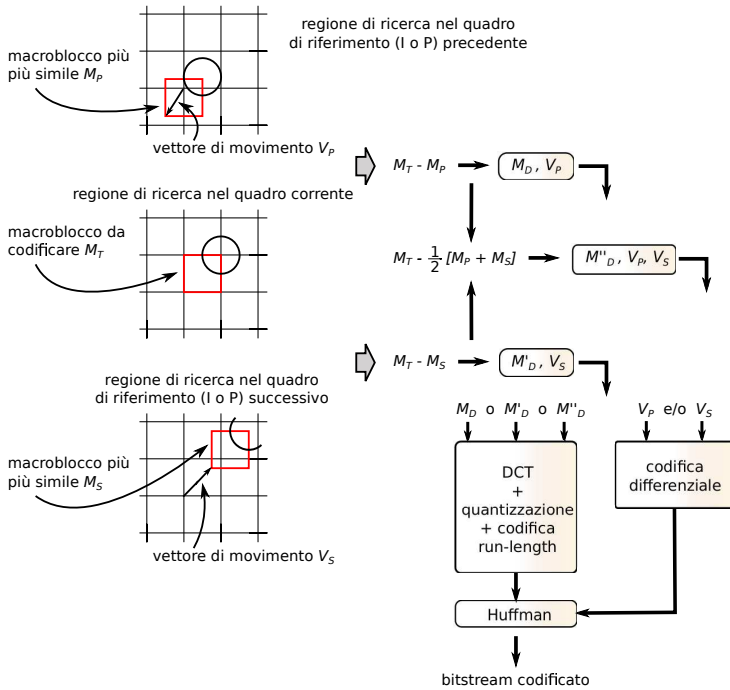


Figura 18.33: Procedura di codifica dei quadri B

macroblocchi contigui viene prima sottoposta ad un processo di codifica differenziale, e quindi i valori di differenza sono rappresentati da codeword a lunghezza variabile di Huffman. D'altra parte, le tre matrici differenza sono invece sottoposte alla stessa sequenza di operazioni dei quadri I (DCT, quantizzazione, codifica entropica), conseguendo però un fattore di compressione più elevato, essendo il macroblocco differenza con valori quasi tutti molto piccoli.

Nel caso in cui la stima di movimento fallisca⁴⁴ (o a causa di una estensione di ricerca insufficiente, oppure per un reale cambio di scena), il macroblocco è codificato in modo indipendente come avviene per i quadri I.

I macroblocchi dei quadri B (vedi fig. 18.33) sono invece confrontati sia con il precedente quadro M_P che con il successivo M_S , ottenendo due possibili insiemi di matrici differenza M_D e M'_D ed associati vettori V_P e V_S ; viene inoltre calcolato un ulteriore insieme M''_D come differenza tra M_R e la media dei macroblocchi (spostati) di riferimento, e determinato infine quale delle tre possibilità fornisca il minimo errore di predizione. In base a questa scelta, si individua quale macroblocco differenza codificare, assieme ai rispettivi vettori di movimento. Nel caso prevalga la predizione basata sulla media tra macroblocchi di riferimento, il vettore di movimento complessivo può determinare un potere di risoluzione a livello di *sub-pixel*.

⁴⁴Viene decretato il fallimento quando anche la migliore compensazione di movimento possibile non determina una riduzione della quantità di bit, rispetto ad una codifica JPEG.

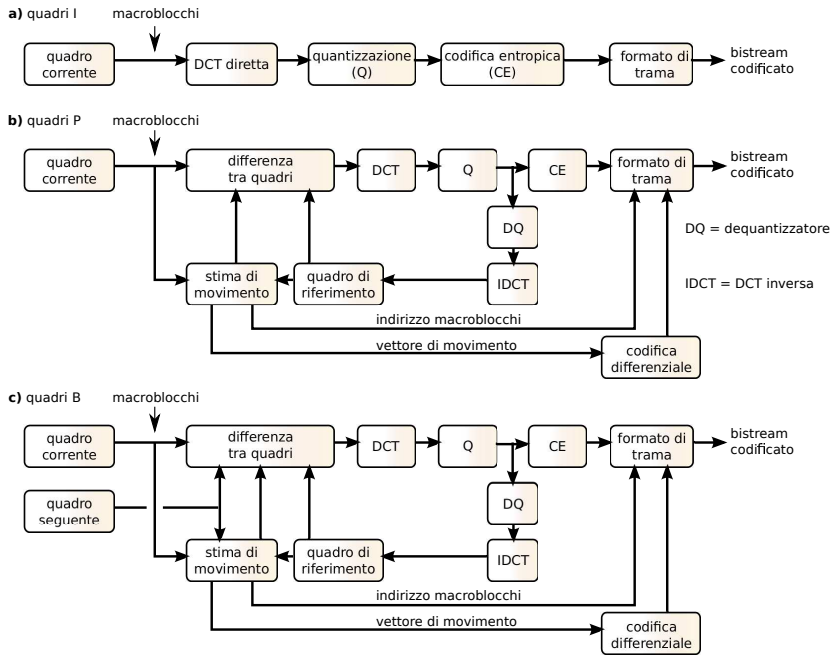


Figura 18.34: Stadi di elaborazione nella codifica di: a) quadri I; b) quadri P; c) quadri B

Questioni realizzative La fig. 18.34 riassume la sequenza di operazioni applicate alle tre tipologie di quadro **I**, **P** e **B**. Mentre nel primo caso queste seguono lo schema previsto dalla codifica JPEG, i quadri **P** meritano qualche commento: allo scopo di alimentare correttamente il componente di stima di movimento, il codificatore mantiene memoria del quadro di riferimento, all'inizio posto pari ad un quadro **I**, e quindi sostituito da una copia dell'ultimo quadro **P**, ottenuto risommando il quadro differenza al precedente quadro di riferimento. Lo stesso schema di calcolo è svolto nel caso di quadri **B**, tenendo ora conto anche del quadro successivo.

Rimarchiamo ora il fatto che, in funzione dell'esito del processo di stima di movimento, esistono tre diverse possibilità di rappresentazione per ogni macroblocco dei quadri **P** e **B**:

- se non vi è movimento, viene trasmessa solo la sua posizione;
- se vi è movimento e si trova un riferimento abbastanza simile, sono trasmessi il vettore di movimento e le matrici differenza;
- se non si è trovato un riferimento abbastanza simile, viene effettuata una codifica *inter* come per il caso dei quadri **I**.

Ciò determina l'esigenza di disporre di formato di trama di dimensione (e velocità) variabile, come realizzato nell'esempio mostrato in fig. 18.35, in cui ad ogni macroblocco è associato un tipo (**I**, **P** o **B**), il suo indirizzo nell'ambito del quadro, il coefficiente di quantizzazione relativo ai termini della DCT, ed il vettore di movimento (se presente). Quindi si dichiara l'identità dei blocchi presenti (che potrebbero essere assenti in caso



Figura 18.35: Esempio di formato trama per i macroblocchi di un bitstream video

di immagini statiche), e per questi viene infine prodotta la sequenza di informazioni previste dalla codifica JPEG.

18.3.1 Standard video

18.3.1.1 H.261

E' lo standard di codifica video definito da ITU-T a fine anni '80 per le applicazioni di videotelefonia su ISDN, ed anche se oggi tecnicamente superato, resta comunque un valido sistema di riferimento che consente la retro-compatibilità tra apparati⁴⁵. Il suo principale limite è la necessità di produrre una velocità ridotta e vincolata a multipli di 64 kbps.

La scelta del formato di immagine è limitata a quanto mostrato in tabella⁴⁶, mentre la scansione è non interallacciata e la velocità di rinfresco di 30 quadri/secondo per CIF oppure 15 o 7.5 per QCIF. Sono usati solo quadri di tipo **I** e **P**, con un GOP di 4 (ossia 3 **P** ogni **I**), e sono usate le procedure descritte alla sezione precedente per rappresentare ogni quadro nei termini di macroblocchi composti da 16x16 pixel (4 blocchi di 8x8) di luminanza e 2 blocchi 8x8 per ogni componente di colore C_b, C_r .

<i>Formato</i>	<i>Y</i>	<i>C_b, C_r</i>
CIF	352 x 288	176 x 144
QCIF	176 x 144	88 x 72

Ogni macroblocco segue la tipica formattazione mostrata in fig. 18.36a; tre file di 11 macroblocchi sono poi raggruppati in una nuova struttura sintattica detta GOB (*Group of (macro)Blocks*), che si articola in un contenuto (fig. 18.36b) ed una intestazione (fig. 18.36c), in cui troviamo un *codice di inizio* scelto in modo da non poter essere presente nella sequenza di codici di Huffman che seguono, e che permette la risincronizzazione nel caso di GOB *mancanti* (vedi appresso), in modo da poter tornare a riprodurre un quadro in corrispondenza del primo GOB disponibile. L'intero quadro è quindi realizzato con il formato di fig. 18.36d), in cui compare un *codice di inizio quadro*, un *riferimento temporale* necessario alla sincronizzazione con la traccia audio, e l'indicazione del tipo di quadro (**I** o **P**); a cui segue la sequenza dei GOB, in numero di 3 oppure 12 a seconda se il quadro rappresenti una immagine QCIF o CIF, in modo da permettere l'interoperabilità tra formati come mostrato in fig. 18.36e.

Controllo di velocità Dato che la codifica video produce una velocità di trasmissione variabile, questa può eccedere la capacità del canale a disposizione, ed un modo *drastico* per risolvere il problema è di scartare alcuni GOB. Il campo *Group number*

⁴⁵Vedi ad es. <http://www0.cs.ucl.ac.uk/teaching/GZ05/08-h261.pdf> (una presentazione di *Mark Handley*), o la trattazione su <http://en.wikipedia.org/wiki/H.261>.

⁴⁶Il *Common Intermediate Format* è stato pensato per facilitare la compatibilità con PAL e NTSC; il *Quarter-CIF* ha una superficie di $1/4$. Sono poi stati anche definiti il 4CIF e 16CIF, oltre che il SIF (352 x 240) che interopera con flussi MPEG.

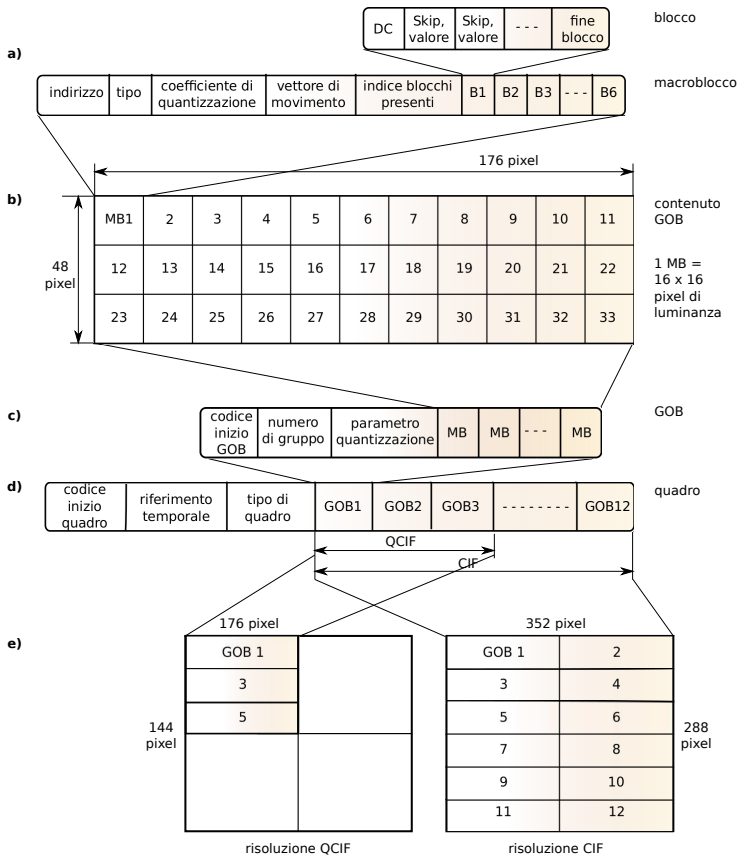


Figura 18.36: Formato codifica H.261: a) macroblocco; b) costruzione e c) formato di un GOB; d) trama di quadro; e) interoperabilità tra formati

dell'intestazione dei GOB permette quindi di collocare il nuovo GOB anche in mancanza dei suoi predecessori.

Un approccio più articolato è quello mostrato dalla figura 18.37a, che ripercorre le tappe già discusse e relative al calcolo del vettore di movimento ed alla codifica degli errori di predizione, ma pone in evidenza il campo di intervento di un componente di *controllo quantizzazione*, che variando l'entità dei coefficienti di quantizzazione della DCT, permette di ridurre e/o aumentare la velocità di codifica complessiva. In particolare, il controllo di quantizzazione opera in base allo stato di riempimento del *buffer FIFO*⁴⁷ mostrato in fig. 18.37b, alimentato dal risultato del processo di codifica e formattazione video, e da cui sono prelevati i dati da inviare a velocità costante. Nel caso in cui la velocità media di codifica ecceda quella disponibile, l'aumento della occupazione del buffer determina l'aumento del coefficiente di quantizzazione, e quindi una riduzione di qualità ma anche della velocità media di codifica; ovviamente, anche

⁴⁷ *First in First out*, è la disciplina di coda del primo arrivato primo servito, opposta a LIFO *Last In First Out*, realizzata come uno *stack*.

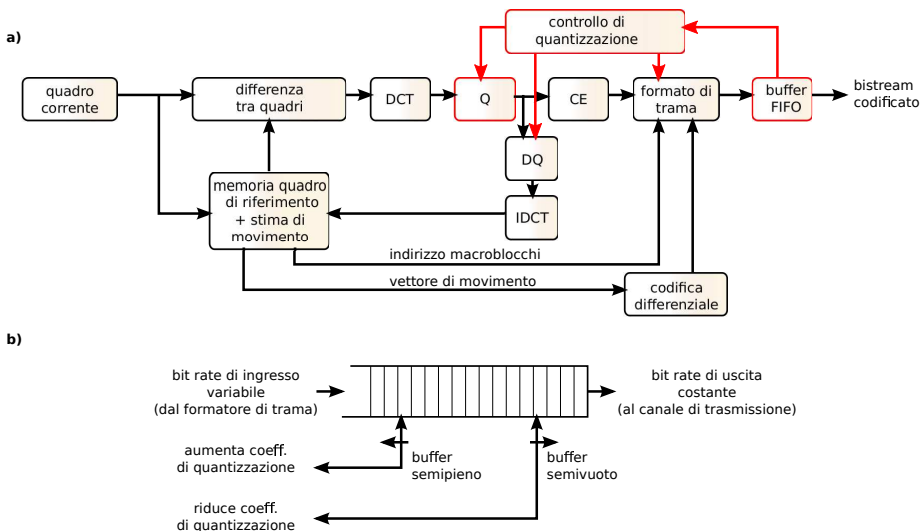


Figura 18.37: Principi della codifica H.261: a) schema del codificatore b) funzionamento del buffer FIFO

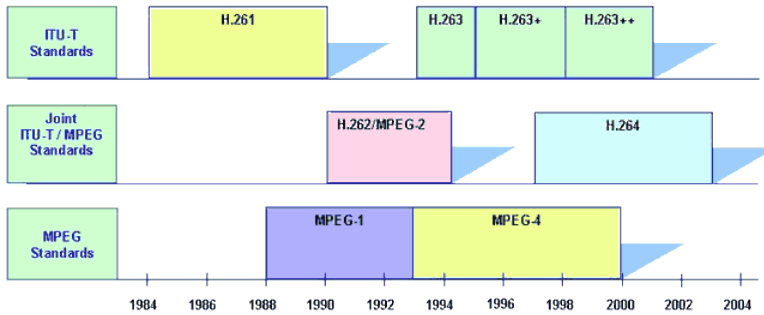
l'inverso è possibile, ossia un miglioramento di qualità mediante riduzione del coefficiente di quantizzazione, nel caso in cui la scena sia statica, e la codifica produca un basso bit rate che consente alla FIFO di svuotarsi.

Nel caso di un aumento improvviso di velocità, come anticipato si possono addirittura *scartare* alcuni GOB, mentre per i successivi si adottano coefficienti di quantizzazione ridotti, comunicati anche al lato ricevente per mezzo dell'apposito campo della intestazione GOB, come mostrato in fig. 18.36c.

18.3.1.2 H.263

Anche questo definito da ITU-T a partire dal 1995, nasce per risolvere i problemi di bassa qualità dell'H.261 a velocità molto ridotte, come quelle offerte dai collegamenti modem *dial-up* precedenti all'introduzione dell'ADSL, ovvero per migliorare la gestione delle possibili condizioni di errore sia sul canale *dial-up* che *wireless*. Le specifiche originarie si sono in seguito arricchite⁴⁸ di estensioni, favorendo l'adozione del codec da parte di altre applicazioni (inclusi i filmati di *youtube*), ed aggiungendo il supporto oltre che ai formati nativi CIF e QCIF, anche a S-QCIF, 4CIF, 16CIF, SIF e 4SIF. A partire dal 2003 si è formato un gruppo di lavoro congiunto tra ITU-T VCEG (*Video Coding Expert Group*) e ISO/IEC MPEG (*Moving Pictures Experts Group*), che segue la definizione del suo successore, l'H.264 detto anche AVC (*Advanced Video Coding*) o MPEG-4 *part 10*, determinando l'arresto dello sviluppo di H.263, che resta comunque (assieme all'H.261) supportato da un gran numero di applicazioni multimediali. Sebbene la

⁴⁸Nel 1998 viene rilasciato l'H.263v2, noto anche come H.263+ o H.263 1998, e nel 2000 è emesso l'H.263v3 noto anche come H.263++ o H.263 2000; inoltre l'MPEG-4 Part 2 è compatibile con l'H.263, in quanto un bitstream H.263 di base viene correttamente riprodotto da un decodificatore MPEG-4.



struttura generale del codificatore e del bitstream ricalchi quella vista per l'H.261, sono state introdotte alcune novità significative, che tentiamo di elencare appresso.

Tipi di quadro In H.263 sono usati, oltre ai quadri di tipo **I** e **P**, anche quelli bidirezionali **B**, consentendo di ottenere fattori di compressione maggiori, a parità di qualità percepita.

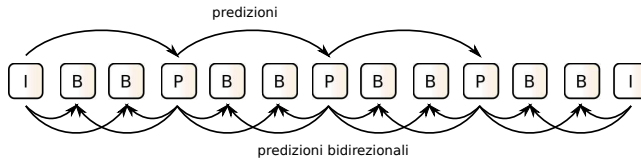
Slice I GOB sono ridefiniti come singole *strisce* di macroblocchi, quindi ad esempio per i formati CIF e QCIF un GOB è ora formato da 11 macroblocchi in fila, anziché 33 come avveniva per l'H.261.

Vettori di movimento estesi La stima di movimento dell'H.261 si arresta in corrispondenza dei bordi del quadro, per cui anche se un oggetto è solo parzialmente uscito di scena, il macroblocco corrispondente viene codificato in modalità *intra*. Al contrario H.263 permette di estendere la ricerca anche a vettori di spostamento che cadono al di fuori del quadro, alla ricerca di una corrispondenza parziale, consentendo al contempo maggiore efficienza e minor distorsione.

Predizione avanzata Anziché determinare il vettore di movimento in base al confronto di un intero macroblocco, i 4 blocchi 8x8 che lo costituiscono sono confrontati in modo indipendente con il quadro di riferimento, permettendo una migliore compensazione del movimento anche per l'immagine di oggetti che non solo traslano, ma si deformano. In definitiva, sono prodotti 4 diversi vettori di movimento per ogni macroblocco.

Resistenza agli errori La presenza di un errore nella ricezione⁴⁹ di un GOB, oltre ad impedire la corretta riproduzione dello stesso, ostacola la riproduzione anche dei quadri successivi che dipendono dai pixel presenti nel GOB, e peggio ancora l'errore finisce per

⁴⁹Qualcuno potrebbe aver notato che nella definizione degli standard fin qui discussi, non sono previsti controlli di tipo *checksum* nel bitstream prodotto. D'altra parte essendo le informazioni codificate di natura auto-sincronizzante, la presenza di errori determina presto presso il ricevitore una condizione di disallineamento, e la decodifica di valori non previsti, come ad esempio la ricezione di vettori di movimento o coefficienti DCT fuori dinamica, o codeword di Huffman non valide, od un numero eccessivo di coefficienti. Per tale via, il ricevitore diviene in grado di accorgersi dell'errore che si è verificato.



Esempio di sequenza di quadri MPEG-1

estendersi anche ad altri GOB, in virtù degli effetti dell'errore sulla ricostruzione dei macroblocchi *predetti* in presenza di movimento.

Per ridurre l'estensione temporale dell'effetto dell'errore, e non dover attendere fino alla ricezione del successivo quadro **I**, si può usare il canale di ritorno presente nei collegamenti punto-punto, consentendo al decodificatore di inviare dei NACK che notificano al mittente la coppia (quadro, GOB) per la quale si è rilevato un errore. Il codificatore è quindi in grado di valutare esso stesso le conseguenze sui quadri successivi, e può provvedere a fornire una codifica *intra* per tutti i blocchi che necessitano di essere rapidamente risincronizzati.

18.3.1.3 MPEG-1

Il *Moving Pictures Expert Group* di ISO emette una serie di standard ognuno orientato ad un particolare dominio applicativo di segnali multimediali, come

- MPEG-1 adotta un formato SIF di 352x288 pixel inteso per la *memorizzazione* audio-video a qualità VHS su CDROM, a velocità fino a 1.5 Mbps;
- MPEG-2 è orientato alla *memorizzazione e trasmissione* audio-video secondo quattro livelli di risoluzione, per ognuno dei quali diversi profili individuano tecniche alternative di codifica;
- MPEG-4 è stato inizialmente concepito per applicazioni simili a quelle dell'H.263, ma il suo uso si è successivamente esteso ad un'ampia gamma di applicazioni Internet.

Anche MPEG-1 adotta tecniche del tutto simili a quelle dell'H.261, con una scansione dell'immagine progressiva ed un sottocampionamento delle componenti di colore 4:1:1, una frequenza di quadro di 25 Hz, l'adozione di quadri di tipo **I**, **P** e **B**, la rappresentazione dei quadri in termini di macroblocchi composti da 16x16 pixel di luminanza, più due blocchi 8x8 per ciascuna componente di colore. Le principali differenze sono che

- possono essere inseriti riferimenti temporali *all'interno* di un quadro, permettendo al decodificatore di sincronizzarsi più rapidamente. L'intervallo tra due marche temporali è chiamato *slice* e comprende una sequenza orizzontale di macroblocchi, tipicamente che copre una intera riga⁵⁰, o meno, ma non di più;
- l'uso dei quadri di tipo B aumenta la distanza temporale tra i quadri di tipo P ed il loro riferimento, e quindi determina una maggiore distanza coperta dalle porzioni di immagine in movimento, cosicché l'ampiezza della finestra di ricerca adottata dal componente di detezione di movimento è stata estesa.

⁵⁰ad es., 22 macroblocchi in risoluzione CIF

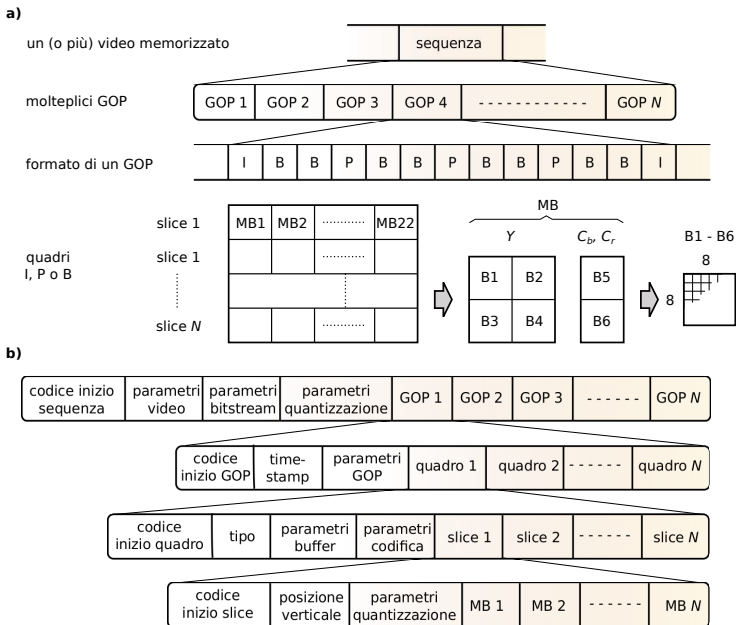


Figura 18.38: Struttura del bitstream MPEG-1: a) composizione; b) formato gerarchico

La figura 18.38a illustra la *struttura gerarchica* del bitstream risultante, secondo il quale l'intero filmato (*sequenza*) è costituito da una successione di GOP, ed ogni GOP da una sequenza di quadri, ognuno costituito da una successione di *slice* che comprendono ognuno 22 macroblocchi, ognuno con 6 blocchi. La sezione b di fig. 18.38 entra più nel dettaglio del formato del bitstream.

18.3.1.4 MPEG-2

Allo scopo di poter usare questo stesso standard per diversi contesti applicativi, sono stati definiti i quattro *livelli* qualitativi mostrati in tabella, e per ogni livello sono quindi definiti cinque profili (*simple, main, spatial resolution, quantization accuracy, high*) in modo da permettere lo sviluppo di nuove tecnologie. Il livello *Low* è compatibile con MPEG-1. Affrontiamo ora la descrizione di ciò che è offerto dal profilo *Main* al livello *Main* (MP@ML).

Livello MPEG-2	formato	bit rate (Mbps)	applicazione
Low	SIF	< 1.5	registrazione qualità VHS
Main	4:2:0	< 15	DVB - MP@ML
	4:2:2	< 20	
High 1440	4:2:0	< 60	HDTV 4/3
	4:2:2	< 80	
High	4:2:0	< 80	HDTV 16/9
	4:2:2	< 100	

MP@ML L'obiettivo è la diffusione televisiva DVB, con scansione interlacciata, risoluzione 720x576 a 25 quadri/secondo (PAL), sottocampionamento 4:2:0 per una velocità risultante tra i 4 ed i 15 Mbps. La principale differenza rispetto all'MPEG-

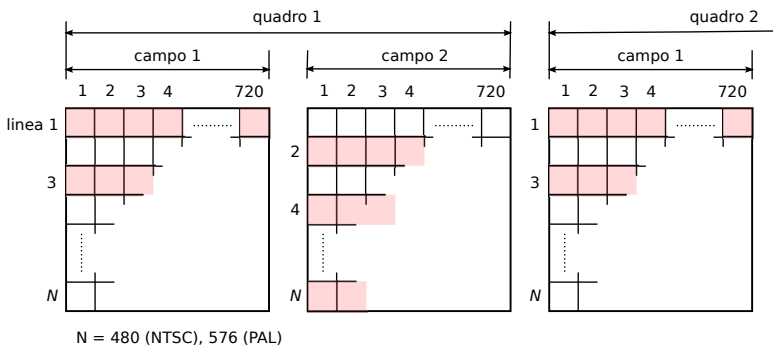


Figura 18.39: Effetto della scansione interlacciata in MPEG-2

1 è legata alla modalità di scansione *interlacciata*, in modo che come mostrato in fig. 18.39, ogni quadro è costituito da due sottoquadri (o *campi*) con le righe rispettivamente dispari e pari, ponendo la questione: come comporre i blocchi da 8x8 pixel su cui eseguire la DCT? Sono possibili due alternative:

- la *modalità campo* (fig. 18.40a) in cui i 16x16 pixel di un macroblocco sono ripartiti tenendo assieme prima le sole righe dispari del primo campo, e quindi le sole righe pari del secondo campo, oppure
- la *modalità quadro* (fig. 18.40b) in cui si usa la stessa suddivisione già vista per il caso non interlacciato, mescolando i due campi in ognuno dei blocchi.

La scelta migliore su quale tra le due modalità adottare dipende dal tipo di scena che si sta rappresentando. Se è presente molto movimento, è meglio adottare la modalità *campo*: essendo infatti i pixel di uno stesso blocco collezionati in un tempo pari a metà dell'intervallo di quadro (mentre nella seconda metà si collezionano i pixel della seconda serie di blocchi), si ottiene un *fotogramma meno mosso*; viceversa in presenza di una scena con poco movimento, può essere adottata la modalità *quadro*.

Per quanto riguarda la stima di movimento, sono ora previste tre possibilità: la *modalità campo* prevede che i campi dispari usino come riferimento i campi pari del quadro precedente, ed i campi pari quelli dispari dello stesso quadro: in tal modo l'intervallo temporale su cui è valutato il movimento è metà del periodo di quadro. Nella *modalità quadro* invece, i campi pari e dispari usano come riferimento i rispettivi campi pari e dispari del quadro precedente, ed è più idoneo nel caso di movimenti lenti. Il meglio di entrambi i modi si ottiene con la *modalità mista*, in cui sono attuati entrambi gli approcci, e viene scelto per la trasmissione quello in grado di dar luogo alla distorsione minore.

HDTV Sono definiti tre standard, ATV, DVB e MUSE, rispettivamente per il Nord America, l'Europa ed il Giappone, a cui si aggiunge la specifica HDTV di ITU-R relativa a studi di produzione e scambio internazionale, e che definisce un rapporto di aspetto 16/9 con 1920 colonne per 1152 righe (di cui solo 1080 visibili), con scansione interlacciata e sottocampionamento 4:2:2. ATV include le specifiche di ITU-R, oltre che un formato ridotto, sempre con aspetto 16/9 ma risoluzione 1280x720, e che adotta la codifica video MPEG-2 MP@ML e quella audio AC-3. Il DVB è basato su di un rapporto

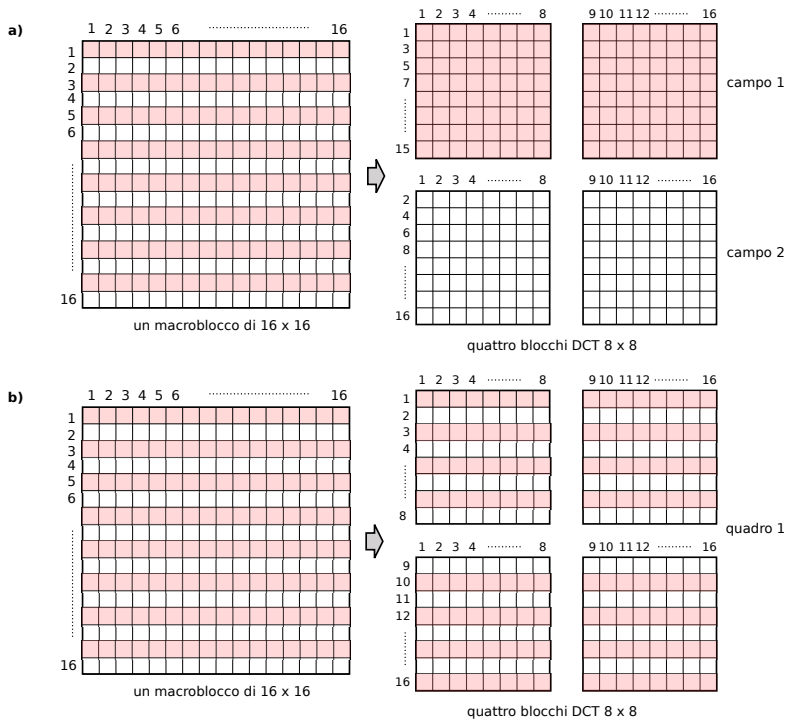


Figura 18.40: Composizione dei blocchi DCT per i quadri I di MPEG-2: a) modalità campo; b) modalità quadro

di aspetto 4/3 con 1440x1152 pixel (di cui 1080 visibili), pari cioè al doppio della risoluzione PAL di 720x576. La codifica video è MPEG-2 SSP@H1440 (*Spatially Scalable Profile at High 1440*), simile all'MP@ML, mentre la codifica audio è MPEG audio layer 2.

18.3.2 Contenitori

I flussi binari prodotti dai diversi codificatori (audio, video) prendono il nome di *Elementary Stream* (ES), e possono essere multiplati assieme per produrre un nuovo formato idoneo alla registrazione di un contenuto multimediale completo, eventualmente arricchito da un flusso di *dati privati*, come mostrato in fig. 18.41a per il caso di MPEG-1.

Prima di effettuare la moltiplicazione, gli ES sono suddivisi in *pacchetti* di dimensione variabile denominati *Packetized ES* (PES) in cui trova posto un *payload* contenente⁵¹ il risultato della codifica (ad es. un intero quadro), preceduto da una *intestazione* contenente un codice univoco di inizio PES ed un codice che individua il tipo di payload del pacchetto (audio, video o dati). L'intestazione PES può inoltre contenere un riferimento temporale necessario alla sincronizzazione audio-video, con risoluzione 33 bit, prodotto

⁵¹Per un approfondimento, vedi http://en.wikipedia.org/wiki/Packetized_elementary_stream

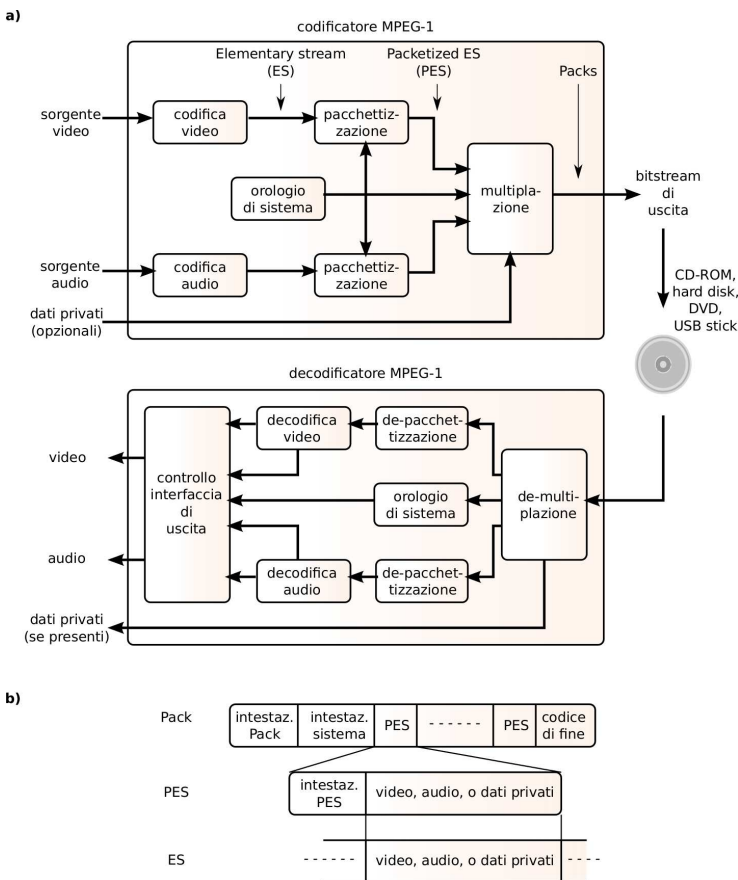
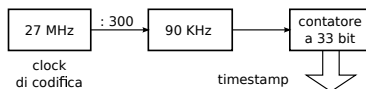


Figura 18.41: Generazione di un contenuto multimediale MPEG-1: a) co-decodificatore; b) formato del bistream di uscita

dal clock a 90 kHz descritto in fig. 18.41 come *orologio di sistema*, e ottenuto a partire da un oscillatore a 27 MHz come mostrato alla figura a lato.



Nel momento in cui un ES è pacchettizzato, viene inserito un *presentation timestamp* (PTS) che ne individua l'istante di riproduzione; per i flussi video è inserito anche⁵² un *decode timestamp* (DTS) perché come anticipato a pag. 499 l'ordine di trasmissione (e quindi di decodifica) può differire dall'ordinamento naturale.

I PES derivanti da diversi ES possono essere quindi inseriti in una struttura di trama detta *Pack*, che può infine essere memorizzata ai fini di una successiva riproduzione.

⁵²In realtà PTS e DTS non sono inseriti in tutti i pacchetti, ma una volta ogni tanto (con intervalli fino a 700 msec per i PS e 100 msec per i TS): il decoder rigenera infatti localmente il clock, ed i timestamp ricevuti servono a mantenerlo al passo con quello trasmesso.

18.3.2.1 Transport Stream

Nel caso di trasmissione dei contenuti mediante un sistema broadcast, tipicamente il singolo *Program Stream* PS (equivalente a quello prima indicato come *Pack*) viene ulteriormente multiplato assieme ad altri, in modo da realizzare un *Transport Stream* (TS), come mostrato in fig. 18.42. In particolare, la parte b) della figura mostra come i PES siano ora suddivisi in segmenti di lunghezza fissa e pari a 184 byte, intestati con 4 byte, producendo una struttura di trama con pacchetti di 188 byte⁵³; l'ultimo pacchetto del TS che origina da uno stesso PES è riempito con un fino a raggiungere i 188 byte. L'intestazione contiene, oltre ad un byte di inizio con pattern unico, un *Packet Identification code* (PID) di 13 bit, che identifica il PES a cui appartiene il pacchetto, permettendo al decoder di recuperare il programma a cui è interessato.

Alcuni PID sono riservati, come il PID 8191 che indica assenza di payload⁵⁴, ed il PID 0, che annuncia l'inserimento nel payload della *Program Association Table* (PAT), la cui ricezione permette al decoder di conoscere quali PID sono utilizzati per individuare i diversi PES (audio, video) di uno stesso programma⁵⁵. Questo avviene per mezzo delle *Program Map Table* (PMT) rappresentate in fig. 18.43: ogni riga della PAT individua infatti un nuovo PID, alla ricezione del quale, viene estratta dal payload la PMT che descrive i PID dei flussi che compongono il programma. Pertanto, quando uno spettatore seleziona un programma, il decodificatore cerca nella PAT il PID della PMT associata, e quindi inizia a prelevare i pacchetti intestati con ognuno dei PID trovati nella PMT, per ricostruire i relativi PES, sincronizzarli, ed iniziare la riproduzione.

⁵³in realtà le intestazioni dei pacchetti del TS possono essere estese e contenere più di 4 byte: in questo caso, la dimensione del payload si riduce, in modo che il totale sia ancora 188.

⁵⁴Un payload vuoto è in realtà comunque riempito di 184 bytes inutili, e viene inserito da parte del multiplatore che realizza il TS per mantenere una riserva di banda che consenta di assecondare le fluttuazioni di velocità dei tributari.

⁵⁵Altri PID riservati sono l'uno, che annuncia la presenza di una *Conditional Access Table* (CAT) contenente i parametri crittografici per visualizzare contenuti a pagamento, ed il PID 18, che annuncia la presenza della *Network Information Table* (NIT), che descrive altri TS disponibili. Per approfondimenti, vedi http://en.wikipedia.org/wiki/Program-specific_information.

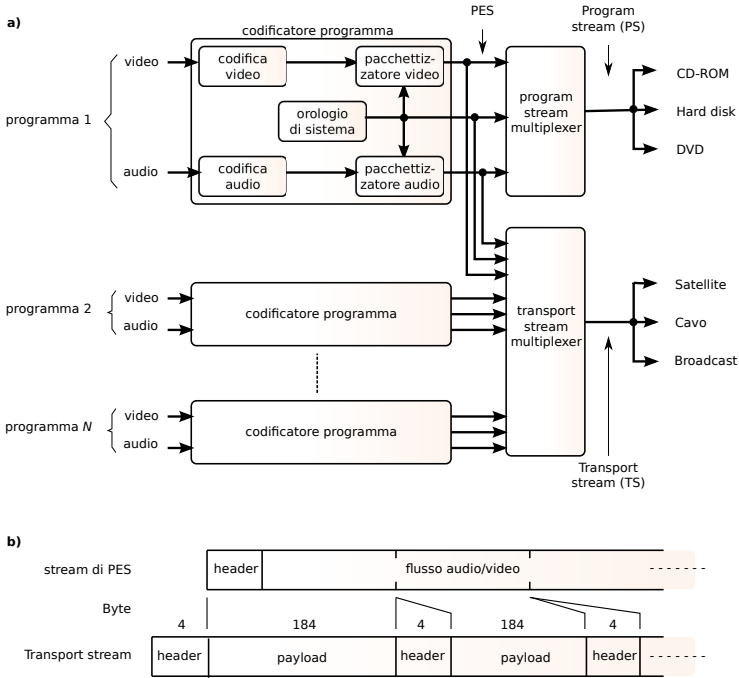


Figura 18.42: Multiplazione di programmi: a) generazione dei PS e TS; b) formato del Transport Stream

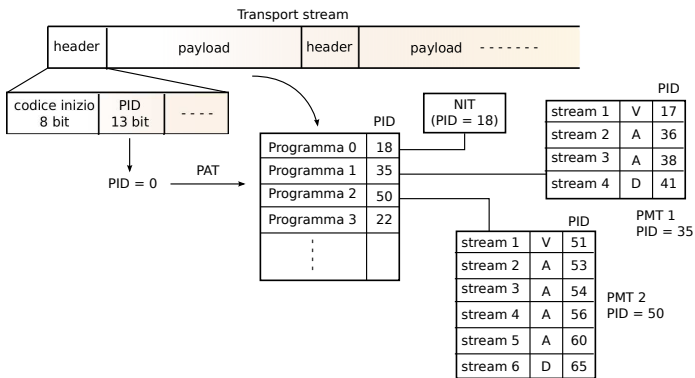


Figura 18.43: Estrazione della PAT e delle PMT dal Transport Stream

Bibliografia

- A.A. V.V.**, *Wikipedia, L'enciclopedia libera e collaborativa*, <http://it.wikipedia.org>
- S. Barbarossa, T. Bucciarelli**, *Teoria dei Segnali*
- S. Benedetto, E. Biglieri**, *Teoria della Probabilità e Variabili casuali*, Quaderni di Elettronica, 1980 Boringhieri
- C. A. Bentivoglio, A. Caldarelli**, *Tecniche e tecnologie multimediali*, 2007 EUM edizioni università di macerata
- J. Bellamy**, *Digital Telephony*, 1991 John Wiley and Sons, New York
- G. Cancellieri**, *Telecomunicazioni*, 2000 Pitagora editrice - Bologna
- A. B. Carlson**, *Communication Systems*, 3rd Edition, 1986 Mc Graw Hill
- F. Cuomo**, *Telematica*, 2001 <http://net.infocom.uniroma1.it/tlem/lucidi/lucidi.php3>
- R. Cusani, T. Inzerilli**, *Teoria dell'Informazione e Codici*, Ed. Ingegneria 2000, 2007 Roma
- M. Decina, A. Roveri**, *Code e Traffico nelle Reti di Comunicazione*, 1991 Editrice La Goliardica - Roma
- M. Decina, A. Roveri**, *Introduzione alle Reti Telefoniche Analogiche e Numeriche*, 1989 Editrice La Goliardica - Roma
- M. G. Di Benedetto, P. Mandarini**, *Comunicazioni Elettriche*, 2000 Editrice La Goliardica - Roma
- G. Fedele**, *Complementi ed applicazioni di Teoria dei Segnali*, Ed. Ingegneria 2000, 1996 Roma
- R. L. Freeman**, *Telecommunication System Engineering*, 2nd Edition, John Wiley & Sons
- F. Halsall**, *Multimedia Communications*, 2001 Pearson Education Limited
- C. W. Helstrom**, *Probability and Stochastic Processes for Engineers*, 2nd, 1991 Macmillan Publishing Company
- N. S. Jayant, P. Noll**, *Digital Coding of Waveforms*, 1984 Prentice-Hall, N.J.

- S. Lin, D.J. Costello Jr.**, *Error Control Coding: Fundamentals and Applications*, 1983 Prentice-Hall, Englewood Cliff, N.J.
- M. Listanti, A. Roveri**, *Comunicazioni Dati*, Appunti
- M. Luise, G. M. Vitetta**, *Teoria dei segnali, seconda edizione*, 2003 McGraw-Hill
- P. Mandarini**, *Teoria dei Segnali*, 1979 Editrice La Goliardica - Roma
- A. V. Oppenheim, R. W. Shafer**, *Digital Signal processing*, 1975 Prentice Hall, NJ
- A. Papoulis**, *Probability, Random variables, and Stochastic Processes*, 1991 McGraw-Hill Int.Eds.
- B. Peroni**, *Comunicazioni Elettriche*, Ed. Scientifiche Siderea, 1973 Roma
- G. M. Poscetti**, *Elementi di teoria dell'informazione*, Ed. Ingegneria 2000, 1996 Roma
- J.G. Proakis, M. Salehi**, *Communication systems engineering 2nd Ed.*, 2002 Prentice-Hall
- U. Reimers**, *Digital Video Broadcasting*, Springer-Verlag Berlin Heidelberg 2001
- A. Roveri**, *Reti di telecomunicazione*, Appunti
- M. Schwartz**, *Information Transmission, Modulation, and Noise*, 4th Edition, 1990, Mc Graw Hill
- C. Shannon, W. Weaver**, *La teoria matematica delle comunicazioni*, 1949 Univ. of Illinois, 1971 Gruppo Editoriale Fabbri
- W. Stallings**, *Trasmissione Dati e Reti di Computer*, Jackson Libri 2000, titolo originale Data & Computer Communications 6th Edition, 2000 Prentice Hall
- R. Steele**, *Mobile Radio Communications* 1992 Pentech Press London, 1994 IEEE press NJ
- F. G. Stremler**, *Communication Systems*, 1990 Addison-Wesley
- A. S. Tanenbaum**, *Reti di Computer*, 1989 Gruppo Editoriale Jackson
- H. Taub, D. L. Schilling**, *Principles of Communication Systems*, 1986 Mc Graw Hill
- J. Watkinson**, *The MPEG handbook*, 2001 Focal Press

Indice analitico

- accesso multiplo, 250, 342
- ACK, 100
- ADM (Add and Drop Multiplexer), 127
- ADPCM, 470
- ADSL, 144
- aliasing, 53
 - temporale, 65
- AMI, 79
- analisi via sintesi, 479
- antenna, 387
- architettura protocollare, 190
- ARP, 204
- ARQ, 94
- ASCII, 116
- assorbimento
 - atmosferico, 391
 - terrestre, 389
- attenuazione
 - chilometrica (cavo), 381
 - chilometrica (fibra), 406
 - di spazio libero, 388
 - disponibile, 378, 389
 - supplementare, 395
- autocorrelazione, 220, 238, 478
 - all'uscita di un filtro, 247
 - proprietà, 221
- AWGN, 86, 453
- back-off, 369
- banda
 - di coerenza, 400
 - di guardia, 251
 - di rumore, 297
 - larghezza di, 252
 - laterale doppia, 272
 - laterale ridotta, 275
 - laterale unica, 252, 274
- baud, 74
- Bayes (teorema di), 149
- Bessel
 - funzione di, 286
 - funzione modificata di, 164
- bit, 432
 - di start e di stop, 108
 - stuffing, 112, 127
- Boltzman (costante di), 419
- BPSK, 307
- BRAS, 145
- BRI, 142
- broadcast
 - dominio di, 206
 - Ethernet, 203
 - FM, 295
 - trasmissione, 250
- buffer, 190
 - fifo, 504
- cammini multipli, 344, 391, 392, 397, 401
- campionamento, 4, 8
 - del colore, 490
 - filtro adattato, 230
 - in frequenza, 65
 - PCM, 123
 - precisione, 81
 - restituzione D/A, 54
 - spaziale, 486
 - teorema del, 51
- canale
 - AWGN, 453
 - numerico, 3
 - trasmissione su, 73
 - perfetto, 349
 - simmetrico binario, 447
 - telefonico, 251
 - variante nel tempo, 401
 - virtuale, 188
- canali televisivi, 416
- capacità di canale
 - continuo, 453
 - discreto, 451
 - OFDM, 335
 - simmetrico binario, 452
- cardinale
 - interpolazione, 51
 - seno, 52
- Carson (regola di), 288
- CAS, 124
- cavo
 - coassiale, 385
 - linea aerea, 384

- ritorto, 384
- CCS, 125
- CDMA, 342
- CDN, 121
- CELP, 482
- centralinista, 122
- centralino, 171
- cerchio unitario, 64
- circuito
 - commutazione di, 188
 - elettrico, 350
 - virtuale, 188
- classless interdomain routing, 202
- Clos (condizione di), 138
- coda di attesa, 180
 - lunghezza media, 182
 - tempo medio, 184
- codebook, 98, 458
 - adattivo, 482
 - di Hamming, 460
 - di Huffman, 498
 - di sorgente, 435
 - quantizzazione vettoriale, 481
- codeword, 98, 435, 458
- codice
 - a ripetizione, 99
 - BCH, 463
 - bipolare, 79
 - ciclico, 462
 - convoluzionale, 464
 - di Gray, 84, 91
 - di linea, 78
 - di Reed-Solomon, 464
 - differenziale, 80
 - lineare a blocchi, 98, 458, 459
 - Manchester, 79
 - polinomiale, 96, 463
 - sistematico, 459
- codice di canale, 457
- codifica
 - a predizione lineare (LPC), 476
 - a velocità variabile, 503
 - audio, 469
 - binaria, 434
 - di canale, 98, 447
 - di forma d'onda, 470
 - di immagine, 486
 - di linea, 74
 - di sorgente, 56, 431
 - continua, 444
 - multimediale, 469
 - differenziale, 319
 - DPCM, 470
 - JPEG, 496
 - OFDM, 335
 - entropica (JPEG), 495
 - PCM, 58
 - per blocchi, 438
 - per sottobande, 473
 - predittiva, 441
 - psicoacustica, 483
 - run-length, 441
 - JPEG, 496
 - tasso di, 99
 - video, 498
- collegamento
 - bilancio di, 377
 - in cavo, 379
 - in diversità, 394
 - in fibra ottica, 404
 - radio, 387
 - radiomobile, 394
 - satellitare, 412
- collisione, 205
- colore
 - profondità di, 489
- commutazione
 - a divisione di spazio, 137
 - a divisione di tempo, 138
 - di circuito, 122, 188
 - di pacchetto, 121, 180, 188, 189
 - principio di, 187
 - telefonica, 137
- componenti analogiche di bassa frequenza, 255
- compressione basata su dizionario, 442
- compromesso
 - banda - potenza, 90, 311, 324, 456
 - velocità - distorsione, 91, 94, 304
 - curva, 444
 - velocità - ritardo, 439
- congestione, 189, 190
- controllo
 - di congestione, 199
 - di errore, 199
 - di flusso, 104, 189, 199
- convoluzione, 37
 - circolare, 68
 - con l'impulso traslato, 39
 - discreta, 68, 115
 - via DFT, 69
- correlatore, 234
 - banco di, 323
- correlazione, 218
 - coefficiente di, 242
 - di processo ergodico, 220
- costante di propagazione, 379
- covarianza, 219
- CRC, 96
- crominanza, 488
- DAC, 55
- datagramma, 189
- DBPSK, 319
- DCT, 67, 493
- decibel, 357
- decisione
 - Bayesiana, 447

- di massima prob. a posteriori, 448
- di massima verosimiglianza, 448
- decodifica
 - di massima verosimiglianza, 366, 466
 - di Viterbi, 466
- demodulazione
 - a correlazione, 322
 - a discriminatore, 284, 301
 - BLU e BLR, 279
 - coerente o omodina, 276
 - di ampiezza, 275
 - di frequenza, 283
 - effetto soglia, 305
 - di involuppo, 278
 - di rumore, 298
 - eterodina, 279
 - incoerente, 168
 - incoerente in fase e quadratura, 277
 - sincrona o omodina, 262
 - supereterodina, 281
- densità di energia
 - di un rect, 44
 - spettro di, 31
- densità di probabilità, 151
- densità spettrale, 222
 - del rumore demodulato FM, 302
 - di FM alto indice, 288
 - di processo dati, 244
 - di segnale AM, 272
 - di segnali passa-banda, 258
 - per modulazione angolare, 284
 - basso indice, 289
- detezione di sinusoide, 168
- DFT, 61, 329
- diafonia, 382
- diagramma a traliccio, 465
- diagramma ad occhio, 76
 - apertura orizzontale, 84
 - in presenza di rumore, 91
- diagramma di transizione, 465
- diffrazione, 390
- diffusione troposferica, 391
- discesa del gradiente, 365
- dispersione
 - cromatica, 407
 - modale, 406
- distanza del codice, 98, 458, 459
- distorsione, 249
 - di non linearità, 369
 - lineare, 359
 - assenza di, 242, 353
- distribuzione (funzione di), 151
- disturbi additivi (combinazione dei), 350
- DMT, 145
- DNS - domain name server, 195
- Doppler (dispersione), 402
- downlink, 414
- DPLL, 110
- DSLAM, 145
- DSSS, 339
- DTFT, 60
- DWDM, 411
- DXC, 134
- Eb/No, 88, 310
- effetto pelle, 381
- efficienza
 - codifica di canale, 99
 - dell'OFDM, 329
 - di AM-PI-PPS, 273
 - di giunzione, 178
 - selective repeat, 103
 - send and wait, 101
 - spettrale, 309
- elastic store, 128
- elementary stream, 510
- energia
 - di segnale limitato in banda, 54
 - mutua, 31, 220
- enfasi e de-enfasi, 306
- entropia, 433
 - a blocco, 440
 - condizionale, 450
 - condizionata, 440
 - di sorgente binaria, 434
 - di sorgente continua, 445
 - differenziale, 445
- equalizzatore
 - DFE, 366
 - LMS, 365
 - MLSD, 366
 - MMSE, 364
 - zero forcing, 115
- equalizzazione, 362
 - delle componenti analogiche, 261
 - OFDM, 335
- equivocazione, 450
- erfc{ }, 88, 89
 - definizione, 155
- Erlang, 172
 - formula B, 178
- errore
 - controllo di, 93
 - correzione di, 98
 - detezione, 95
 - di fase o frequenza, 276
 - di predizione, 476, 500
 - in trasmissioni numeriche di banda
 - base, 86
 - su parole, 94
- Ethernet, 203
 - gigabit, 207
- Eulero, formule di, 16
- evento, 147
 - completamente casuale, 173
 - statisticamente indipendente, 150
- fading

- di Rayleigh, 398
- di Rice, 399
- piatto, 397
- selettivo in frequenza, 399
- su larga scala, 396
- su piccola scala, 397
- fasori, 16
 - coeff. Fourier come, 19
- fattore di rumore, 423
 - per reti in cascata, 424
- FEC, 93
- fenomeno aleatorio, 147
- FEXT, 382
- FFT, 66, 328
- filtraggio, 40
 - di segnali e processi, 226
 - numerico, 67
 - perceptivo, 479
- filtro
 - a banda minima, 83
 - a coseno rialzato, 83
 - a radice di coseno rialzato, 113
 - adattato, 112, 229
 - analogico, 241
 - anti-aliasing, 53
 - di canale, 251
 - di Hilbert, 255
 - risposta impulsiva, 266
 - di restituzione, 52, 54, 55
 - digitale o numerico, 238
 - FIR, 238
 - IIR, 240
 - passa banda ideale, 262
 - passa basso, 241
 - sbiancante, 477
 - trasversale, 238
- finestra scorrevole, 104, 198
- finestratura, 42
 - nella stima spettrale, 48, 226
- flag byte, 112, 123
- footprint, 414
- formanti, 474
- Fourier
 - serie di, 17
 - troncamento, 22
 - trasformata di, 29
 - proprietà, 31
 - trasformata discreta di, 61
- frequenza
 - deviazione di, 303
 - di interarrivo, 174
 - di riga, 292
 - di servizio, 176
 - di simbolo, 74, 77
 - di taglio, 242
 - Doppler, 402
 - immagine, 280
 - istantanea, 282
- frequenze radio, 416
- Fresnel
 - ellissoidi di, 390
- FTTH, 120
- funzione caratteristica, 168
- funzione di trasferimento, 241
- Gibbs, fenomeno di, 23
- GIF, 490
- GOB, 503
- GOP, 499, 503
- GPRS, 120
- grado di servizio, 379
- GSM, 120, 250
 - codec, 480
- guadagno
 - di processo, 340
 - di sistema, 378
 - disponibile, 355
- H.261, 503
- H.263, 505
- Hamming
 - codice di, 460
 - distanza di, 98
- HDTV, 509
- Heaviside (condizione di), 380
- Hilbert
 - trasformata di, 256
- Huffman
 - codifica di, 437
- ibrido telefonico, 140
- IDFT, 63
- immagazzinamento e rilancio, 189
- impedenza caratteristica, 379
- impulso matematico, 35
- incapsulamento, 192
- incorrelazione, 219
- indice di modulazione
 - angolare, 285, 289, 304
 - di ampiezza, 273
- indice di rifrazione, 405
- indipendenza statistica, 150, 219, 243
- informazione mutua media, 449
- instradamento, 187, 189
- integrate and dump, 232
- intelligent network, 143
- intercorrelazione, 220
- interferenza intersimbolica, 76, 82
- interleaving, 99, 462
- intermodulazione
 - componenti analogiche di b. f., 261
 - fattore di, 370
- Internet, 193
 - indirizzi, 194
- intervallo di predizione, 499
- intracoded frame, 498
- involuppo complesso, 253
 - traiettorie, 312

- IP, 200
 - sottoreti, 195
- ipotesi, verifica di, 165
- ISDN, 142
- ISI, *vedi* interferenza intersimbolica⁸²
- istogramma, 152

- giacobiano, 162
- JPEG, 492

- Kraft, disuguaglianza di, 436

- L-ASK, 309
- L-FSK, 322
- L-PSK, 311, 314
- LAN switch, 206
- Laser, 407
- LED, 407
- Lempel-Ziv-Welsh, compressione di, 442
- Little, risultato di, 181
- livello di significatività, 165
- Lloyd-Max, algoritmo di, 58
- luminanza, 488
- lunghezza d'onda, 252
 - ottica, 404

- marginale, 378
- mascheramento uditivo, 483
- massima verosimiglianza
 - decisione di, 87, 165, 169, 230
 - per sequenze, 367
 - stima di, 166
- massimo trasferimento di potenza, 352, 373
- matrice di controllo parità, 461
- matrice generatrice, 459, 463
- media
 - campionaria, 167
 - di insieme, 156
 - frequenza, 280
 - quadratica, 157, 158
 - temporale, 156, 157
- mezzo trasmissivo, 377
- minima energia per bit, 455
- mixer, 291
- MJPEG, 498
- modem, 252
- modo di propagazione, 405
- modulazione, 34, 41, 254
 - angolare, 281
 - BPSK, 307
 - coerente, 345
 - di ampiezza, 271
 - prestazioni, 299
 - FM a basso indice, 295
 - incoerente, 346
 - L-FSK, 322
 - numerica, 307
 - OFDM, 324
 - QAM, 316
 - QPSK, 311
- momento, 152
 - centrato, 153
 - misto, 218
- MP3, 484
- MPEG-1, 507
- MPEG-2, 508
- MPLP, 480
- multicast, 203
- multipath, 393
- multiplazione, 121
 - a divisione di frequenza, 249, 251
 - a divisione di tempo, 121, 187
 - add and drop, 127, 129
 - asincrona, 126
 - schema di, 187

- NACK, 100
- NEXT, 382
- NRZ, 78
- Nyquist
 - condizioni di, 82
 - filtro di, 83
 - frequenza di, 51
 - impulso di, 82

- OAM, 130
- OFDM, 324
 - coded, 336
 - densità di potenza, 327
 - equalizzazione, 335
 - modem, 328
 - prestazioni, 330
- onda
 - diretta, 389
 - PAM, 75, 146
 - densità spettrale, 243
- ortogonalità
 - degli esponenziali complessi, 24
 - tra simboli sinusoidali, 345
 - tra sinc, 54, 327
- overlap and add, 70

- p-value, 165
- pacchetto
 - dati, 185
 - fuori sequenza, 198
- palette, 489
- parità
 - bit di, 95, 108
 - matrice di controllo, 460
- Parseval, teorema di
 - per segnali di energia, 31
 - per segnali di potenza, 23
- path loss, 396
- PCM
 - G.711, 59, 469
 - struttura di trama, 123
- PDH, 126

- PDU - protocol data unit, 192
 periodogramma, 225
 PES, 510
 pitch, 473
 pixel, 486
 Plank, costante di, 419
 PLL, 276
 ricevitore a, 283, 322
 PNG, 491
 Poisson, somma di, 46
 polinomio generatore, 96, 463
 ponte radio, 251
 portante
 di colore, 293
 intera, 272
 parzialmente soppressa, 273
 pilota, 337
 ricostruzione per quadratura, 291
 soppressa, 272
 potenza
 AM BLU, 290
 assorbita da un bipolo, 351, 372
 attiva, 373
 ceduta ad un carico, 374
 del rumore, 86
 di segnale, 350
 di un coseno, 36
 di un processo, 157, 158
 di un segnale AM, 275
 disponibile, 352
 di rumore, 420
 entropica, 446
 istantanea, 273
 ricevuta, 89
 POTS, 140
 predittore, 476
 a lungo termine, 480
 bidirezionale, 498
 PRI, 142
 probabilità
 a posteriori, 149
 assiomi, 148
 condizionata, 87, 148
 come verosimiglianza, 165
 congiunta, 149
 densità di, 151
 di blocco, 172
 di detezione, 168
 di errore
 filtro adattato, 233
 L-ASK, 310
 OFDM, 330
 QPSK, 314
 sul bit, 92, 113
 sul simbolo, 87, 89
 di falso allarme, 169
 di perdita, 169
 di rifiuto, 176
 di ritardo, 182, 183
 teoremi di base, 148
 teoria delle, 147
 processo
 ad aleatorietà parametrica, 159
 aleatorio, 156
 armonico, 160, 223
 di ingresso, 175
 di innovazione, 480
 di nascita e morte, 176
 ergodico, 158
 gaussiano (demodulazione), 298
 gaussiano bianco limitato in banda,
 223, 266
 in banda traslata, 263
 prodotto e somma, 236
 stazionario, 157
 in senso lato, 157
 prodotto banda-lunghezza, 408
 propagazione
 condizioni di, 389
 luminosa, 405
 prove ripetute, 172
 pseudo-noise (sequenza), 338
 PSTN, 140
 punto-multipunto, 250
 punto-punto, 251

 QAM, 316
 QPSK, 311, 313
 quantizzazione, 4, 56
 adattiva, 471
 della DCT, 494
 lineare, 56
 logaritmica, 58
 SNR di, 57
 vettoriale, 480

 Rayleigh
 fading di, 398
 variabile aleatoria, 163
 regola del prefisso, 436
 RELP, 479
 rete
 ATM, 208
 di accesso, 120
 di trasporto, 134
 in fibra ottica, 133
 anello, 136
 protezione, 135
 ottica, 411
 ottica sincrona, 129
 plesiocrona, 122
 telefonica, 119
 topologia, 132
 reti
 due porte, 353
 ingegneria delle, 171
 RGB, 488
 ricevitore ottimo, 112

- ridondanza, 3
 - ciclica, 97
 - codifica di canale, 98, 447, 457
 - delle fibre ottiche, 136
 - di diversità, 394
 - di sorgente binaria, 434
 - OFDM, 329, 336
- riflessione ionosferica, 391
- ringback, 141
- riscontro, 199
- risoluzione spaziale, 294
- risonanza, 474
- risposta impulsiva, 37
 - costruzione grafica, 38
- risposta in frequenza, 235
- round trip time, 104
- routing, 201
- RPE-LTP, 480
- rumore
 - additivo, 74
 - bianco limitato in banda, 86
 - dopo demodulazione FM, 301
 - nei ripetitori, 427
 - nelle reti due porte, 420
 - termico, 419
- RZ, 78
- sample and hold, 55
- scattering, diagramma di, 218
- Schwartz, disuguaglianza di, 27
- SDH, 128
 - trama, 129
- SDU - service data unit, 192
- segnalazione
 - a canale comune, 125
 - antipodale, 233
 - associata al canale, 124
 - ortogonale, 233
- segnale
 - a banda stretta, 361
 - analitico, 257
 - dati, 73, 74
 - generazione, 78
 - limitato in banda, 80
 - processo di, 160, 224
 - di cromaticità, 294
 - di energia, 13
 - di potenza, 12
 - impulsivo, 13
 - modulato, 253
 - modulato (filtraggio), 260
 - passa-banda, 258
 - periodico, 12
 - come treno di periodi, 45
 - rettangolare, 20
 - utile, 349
 - video composito, 489
 - vocale, 473, 475
- selective repeat, 102
- send and wait, 100
- sensibilità, 377
- serie trigonometrica, 20
- sfarfallamento, 292
- shadowing, 396
- Shannon
 - primo teorema di, 435
 - secondo teorema di, 451
- Shannon-Hartley
 - legge di, 454
 - limite di, 324
- signaling system n. 7, 142
- simmetria coniugata
 - attorno ad fo, 262
 - o Hermitiana, 18
- simmetria Hermitiana o coniugata, 31, 222
- sinc
 - definizione, 21
 - ortogonalità, 54
- sincronizzazione, 344
 - di bit, 112
 - di bit e di parola, 109
 - di carattere, 111
 - di centrale, 125
 - di rete, 128
 - di simbolo, 110
 - di trama, 109
- sindrome, 461
- sistema
 - a coda
 - finita, 183
 - infinita, 181
 - autonomo, 202
 - di servizio
 - orientato al ritardo, 180
 - orientato alla perdita, 175
 - lineare e permanente, 234
- Snell, legge di, 405
- SNR, 1, 88, 297
 - di riferimento, 300, 310
 - dopo demodulazione AM, 299
 - dopo demodulazione FM, 303
- soft decision decoding, 462
- soglie di decisione, 87
- somma di controllo, 95
- SONET, 129
- sonogramma, 475
- sorgente
 - con memoria, 440
 - continua con memoria, 446
 - Markoviana, 440
 - senza memoria, 432
- sovracampionamento, 54
- spazio campione, 147
- spazio normato, 25
- spettro di potenza
 - per segnali dati, 75
 - per segnali periodici, 24
- spread spectrum, 337

- frequency hopping, 343
- sequenza diretta, 339
- statistica, 164
 - del secondo ordine, 217
- stima
 - della varianza, 472
 - di autocorrelazione, 238, 477
 - di movimento, 499
 - di parametro, 165
 - di periodo, 478
 - spettrale, 225
 - LPC, 478
- STM-1, 130
- stratificazione ISO-OSI, 191
- subnetting, 201
- SYN, 111
- tasso di codifica, 458
- TCP, 197
- temperatura
 - di antenna, 415
 - di sistema, 421
- tempo
 - di coerenza, 403
 - di guardia, 326
 - di ritardo di gruppo, 359, 375
- teorema
 - centrale del limite, 154
 - del campionamento, 51
- teoria dell'informazione, 431
- three way handshake, 198
- traffico
 - a valanga, 179
 - dolce, 172
 - intensità media, 176
 - offerto, 172
 - smaltito, 179
- transponder, 413
- transport stream, 512
- trasformata
 - di Fourier, 29
 - discreta coseno, 493
 - zeta, 64
- trasformata di
 - costante, 35, 47
 - derivata, 42
 - integrale, 43
 - segnale periodico, 35, 45
 - sequenze, 60
 - un gradino, 49
 - un rettangolo, 30
 - un sinc, 32
 - un treno di impulsi, 45
 - un triangolo, 43
- trasmissione
 - a circuito, 119
 - asincrona, 108
 - dati (in banda base), 73
 - dati (reti per), 185
 - FM broadcast, 295
 - multilivello, 77
 - numerica (dimensionamento), 92
 - sincrona, 110
 - televisiva, 292
- treno
 - di impulsi, 44, 52
 - di impulsi rettangolari, 55
- TSI, 139
- UDP, 199
- UNICODE, 116
- uplink, 412
- valore
 - atteso, 152, 156
 - efficace, 158
 - medio, 157
- variabile aleatoria, 150
 - di Bernoulli, 172
 - di Poisson, 173
 - di Rayleigh, 163, 168
 - di Rice, 164, 168
 - esponenziale negativa, 174
 - gaussiana, 87, 154
 - multidimensionale, 243
 - indipendente e identicamente distribuita, 154
 - trasformazioni di, 160
 - uniforme, 153
- varianza, 153, 158
 - campionaria, 167
- VCO, 276, 283
- verifica di ipotesi statistica, 448
- verosimiglianza
 - funzione di, 165
 - rapporto di, 448
- VGA, 486
- virtual container, 131
- Viterbi, algoritmo (o decodifica) di, 466
- VLAN, 206
- VPN, 121
- WDM, 410
- Wiener (teorema di), 222
- Wiener-Hopf (equazioni di), 364
- Yule-Walker (equazioni di), 477

La prima edizione a stampa di questo testo risale al 2001, come materiale di supporto didattico per un corso di Telecomunicazioni erogato nell'ambito del corso di Laurea in Ingegneria Elettrica presso l'Università di Roma La Sapienza, con l'intento di affrontare in modo coerente temi generalmente sviluppati in corsi separati.

Di anno in anno la didattica svolta in aula, per corsi diversi, ha fornito l'opportunità per sviluppare nuovo materiale e migliorare l'esistente, arricchito di approfondimenti ed innumerevoli illustrazioni di schemi e curve. La varietà degli argomenti affrontati ha posto fin da subito una sfida di completezza ed esattezza, affrontata cercando di mantenere un buon livello di rigore analitico, di alternare la teoria alle applicazioni, di bilanciare i concetti con gli esempi concreti. Ne risulta un testo *a più livelli*, il cui corpo principale dona la giusta rilevanza a concetti, principi, argomenti e tecniche, mentre gli aspetti più rigorosi e speculativi sono approfonditi nell'ambito di numerose note. Allo scopo di evitare inutili ripetizioni, si fa molto uso di rimandi e collegamenti che letteralmente *attraversano* l'intero testo, e che *tengono assieme* argomenti correlati ed interdipendenti ma trattati in capitoli diversi, mentre per gli argomenti solo accennati sono aggiunti riferimenti diretti a Wikipedia, invitando il lettore all'approfondimento.

Si giunge così all'attuale edizione, per la quale si è accumulato materiale su *segnali, sistemi, probabilità, statistica, campionamento, quantizzazione, elaborazione numerica, modulazione analogica e numerica, moltiplicazione, commutazione, traffico, trasmissione dati, reti, trasferimenti energetici, equalizzazione, bilancio di collegamento, collegamenti radio, via satellite, in mobilità, in cavo ed su fibra ottica, rumore, teoria dell'informazione, compressione dati, codifica di immagine, codifica audio e video, capacità di canale e codici correttori*.

Il testo è rimasto fin dall'inizio disponibile per il download presso <http://infocom.uniroma1.it/alef/wiki/libro>, totalizzando circa 2000 scaricamenti/anno, permettendo il feedback diretto da parte dei lettori, ed incoraggiando così il continuo perfezionamento ed ampliamento dell'opera.

Alessandro Falaschi è nato nel 1959 a Roma, e si laurea con lode nel 1983 in Ingegneria Elettronica presso l'Università di Roma La Sapienza, dove in seguito consegue il Dottorato di Ricerca. Diviene quindi ricercatore presso l'Università di Perugia, e dopo essere tornato al Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni (DIET) di Roma, attualmente insegna presso il polo didattico di Latina.

Si è occupato di elaborazione del segnale vocale come riconoscimento automatico e sintesi automatica; quindi, si è interessato di aspetti relativi alle trasmissioni numeriche, alla rete Internet, ed alle applicazioni telematiche di comunicazione multimediale, come VoIP e WebRTC, video conferenze di gruppo, e live streaming Internet; inoltre, ha seguito gli sviluppi delle reti peer to peer, e delle reti veicolari o VANET.