ai miei genitori

Titolo	Sistemi di Telecomunicazione
Autore	Alessandro Falaschi
Rilascio	Edizione 2.0a, 10 luglio 2023
ISBN	Not yet
Copertina	Elaborazione via Nightcafe Creator da un soggetto di Marco Sebastiani - https://illustratoremarco.blogspot.com/
Licenza	 Creative Commons *Attribuzione - Non Commerciale - Condividi allo stesso modo* https://creativecommons.org/licenses/by-nc/4.0/deed.it
Sito Web	https://teoriadeisegnali.it/libro/ e /blog
Facebook	https://www.facebook.com/segnalisistemi - metti il *mi piace!*
LinkedIn	https://www.linkedin.com/company/teoriadeisegnali-it/
Strumenti editoriali	. Lyx - http://www.lyx.org/ . LaTeX- https://www.latex-project.org/ . Inkscape - http://www.inkscape.org/ . Gimp - https://www.gimp.org/ . Gnuplot - http://www.gnuplot.info/ . Octave - http://www.gnu.org/software/octave/ . Genius - https://www.jirka.org/genius.html . Linux - https://www.linux.it/
Donazione	Presso https://teoriadeisegnali.it/wiki/Libro/Donazioni puoi manifestare il tuo apprezzamento, stimolare la continuazione del lavoro, e garantirti la ricezione gratuita delle edizioni future. Il ricavato è in parte devoluto a progetti OpenSource
Liberatoria	L'inclusione accidentale di materiale protetto da copyright è da considerare momentanea, fino alla riproduzione di copie originali dello stesso. Ove possibile sono forniti riferimenti all'origine del materiale. L'autore si impegna alla rimozione dei contenuti che possano ledere gli altrui diritti
Dello stesso autore	**Lo strato applicativo di Internet** http://teoriadeisegnali.it/wiki/Didattica/LoStratoApplicativodiInternet **Signal Processing and Information Theory** https://blog.teoriadeisegnali.it/2022/06/an-abridged-translated-and-twisted-edition/

Prefazione

L'IDEA di autoprodurre questo testo trae forse origine da una sorta di desiderio di rivalsa rispetto alle fotocopie di appunti scritti a mano su cui da studente preparavo gli esami; con il passare degli anni dalla sua prima edizione del 2001, mi rendo sempre più conto di aver intrapreso un percorso *interminabile*. Un po' alla volta ho continuato infatti a migliorare la qualità del testo e ad integrare nuovi argomenti, incoraggiato dal gran numero di lettori raggiunti grazie alla sua accessibilità on-line, ed a distanza di più di venti anni dall'inizio dell'avventura annuncio il rilascio dell'edizione *2.0*, che nel formato a stampa si suddivide in due volumi, di cui questo *è il secondo*.

Gli aspetti affrontati abbracciano un ampio spettro di tematiche relative alle *telecomunicazioni*, il collante nascosto che definisce gli algoritmi implementati da programmi, a loro volta eseguiti mediante circuiti elettronici: tutti protagonisti assoluti della nostra vita immersa nella società *dell'informazione*. Lo sviluppo della trattazione, che nelle prime edizioni era orientato ad un approccio *bottom-up* collegando strettamente gli aspetti teorici con le rispettive applicazioni pratiche, si è via via strutturato ed riorganizzato, raggruppando tra loro argomenti affini secondo la sequenza logica sperimentata attraverso i cicli didattici che ne hanno accompagnato la stesura.

La forma espositiva è tuttora articolata *su due livelli*, con numerose note ed appendici dove vengono svolti i passaggi e sviluppate le osservazioni, mentre il testo principale tenta di mantenere il *filo logico* del ragionamento complessivo. Si fa uso sistematico di rimandi e collegamenti che letteralmente *attraversano* l'intero testo, consentendo di *ricucire assieme* argomenti correlati ed interdipendenti, in modo particolarmente interattivo nel caso del formato PDF *navigabile*. Ci si avvale inoltre di *numerosissime illustrazioni*, per mostrare 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 dei parametri di sistema. Sono infine presenti svariati rimandi a contenuti on-line per gli argomenti accennati solo in parte, principalmente verso *Wikipedia*, da cui il lettore interessato può iniziare un percorso di 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 prodotto un risultato con diversi aspetti di originalità. 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 italofona, 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[1]. Una terza considerazione riguarda l'elevata *qualità tipografica* per un testo autoprodotto, ottenuta con l'utilizzo esclusivo di strumenti *opensource*. Il quarto punto di forza è la scelta di non affidarsi ad un editore tradizionale, ma affiancare al formato elettronico ad accesso pubblico[2] quello cartaceo in modalità *stampa on-demand*. Infine, l'aspetto forse più nascosto ma a mio avviso realmente qualificante è l'attività di *revisione dinamica* a cui è continuamente sottoposto durante i periodi didattici, che lo rende materia in continua evoluzione e ad ogni revisione sempre più completo.

Formato cartaceo

Come per le precedenti edizioni, si è scelto un servizio di tipo *print-on-demand*, in qualche modo *snobbando* gli editori tradizionali e by-passando qualsiasi *vaglio editoriale*. Anche se i principi della cultura libera reclamano priorità per la massima diffusione dell'opera, conseguita con la disponibilità del formato PDF liberamente scaricabile presso `teoriadeisegnali.it`, il formato cartaceo sicuramente consente un approfondimento ed una memorizzazione dei contenuti assai più efficiente di quello elettronico, e dunque è giunto il momento di... *andare in stampa!*

In realtà l'operazione non è del tutto indolore, e dato che realizzare un unico volume con più di 900 pagine non è ne pratico ne fattibile, l'edizione stampata di *Trasmissione dei Segnali e Sistemi di Telecomunicazione* si suddivide in due volumi:

- Teoria dei Segnali che si compone di nove capitoli più uno introduttivo, e verte sugli argomenti *propedeutici* allo sviluppo successivo, mentre
- Sistemi di Telecomunicazione di ben 23 capitoli (ma l'ultimo è solo un *segnaposto*) che tratta (quasi) tutto il resto.

Quello che stai leggendo è il *secondo* volume, e tratta di tutti gli aspetti inerenti la *trasmissione dell'informazione*. Si inizia dai *segnali modulati*, illustrando la loro rappresentazione *complessa* e le tecniche di *modulazione di ampiezza* ed *angolare* con i rispettivi inversi, passando agli effetti causati da fenomeni di *distorsione* lineare (e non) ed alle *prestazioni* delle trasmissioni modulate in presenza di *rumore* additivo.

[1] In realtà il mondo universitario di afflizioni ne ha diverse, come ad esempio il fatto che un lavoro come questo ha valore zero per quanto riguarda la carriera accademica. Si, perché il mestiere del docente, a quanto pare, non è insegnare bene, ma scrivere *tanti articoli*, da far vendere alle riviste scientifiche, ovviamente a carico delle biblioteche universitarie.

[2] Presso `https://teoriadeisegnali.it/libro/` è disponibile il *download* del testo in PDF, il formato HTML che viene indicizzato dai motori di ricerca, gli esercizi di esame svolti, e... molto altro!

Viene quindi affrontata la trasmissione di *dati numerici*, sia in banda base che in forma modulata, illustrando in tal caso le tecniche a *portante singola*, quelle OFDM, e quelle a *spettro espanso*. Si passa poi a definire la *capacità di canale* e quindi ad illustrare le tecniche di *codifica* a blocco, *convoluzionale* ed *iterativa*, mettendole a confronto. Il *modello circuitale* della trasmissione su canale analogico permette quindi di caratterizzare la *rumorosità degli apparati* e dei ripetitori, mentre una rassegna di tecniche di *equalizzazione* illustra gli approcci adottati a contrasto degli *effetti filtranti*. Sono quindi approfonditi gli aspetti legati ai *mezzi fisici* di trasmissione quali *cavo*, *fibra ottica* e *canale radio*, e per quest'ultimo è approfondita l'analisi del caso delle *trasmissioni mobili*. Chiude questa parte l'esteso capitolo dedicato alle *trasmissioni multi antenna* o MIMO.

L'ultima parte del testo raccoglie argomenti per così dire *avanzati* (dalle precedenti edizioni), ma tuttora degni di interesse, come la *teoria del traffico* e le reti a *commutazione di pacchetto* (inclusa Internet) e di *circuito*.

All'interno di *questo volume* potrai trovare citazioni e riferimenti relativi a un numero di capitolo o di pagina *che compare nell'altro*, permettendo a chi ha accesso ad entrambi di *riassemblarli* in una unica *fonte coerente*.

Chi acquista una copia stampata ha il diritto di *scaricare* anche *il formato PDF navigabile*, riservato ai sostenitori del progetto, e senz'altro acquistare il libro significa sostenerlo. Chi lo desidera può inviare una foto del volume in suo possesso ad `alef@teoriadeisegnali.it`, e riceverà al suo indirizzo le istruzioni per il download. Ad ogni modo, l'intervallo di quasi un anno tra l'annuncio dell'edizione 2.0 in formato PDF e quello del suo formato cartaceo è anche servito a correggere alcuni piccoli refusi che ancora *aleggiavano*, oltre che ad includere la *data processing inequality* al capitolo nove.

Istruzioni per l'uso

Anche se un libro è esso stesso una spiegazione, e non dovrebbe averne di ulteriori, a volte qualche studente mi chiede: *ma quali sono le cose principali da sapere?* Purtroppo è una domanda senza risposta, o meglio, a cui potrebbe rispondere lo studente stesso dopo aver studiato, e infatti è quel genere di cose che vengono tramandate mediante passa parola. Io per primo non ho troppa simpatia per gli sviluppi analitici fini a se stessi, e ritengo più importante che siano afferrati i concetti e le modalità di procedere piuttosto che memorizzare i singoli passaggi senza avere al contempo una idea precisa del loro ruolo. D'altra parte *i conti* sono ciò che ci fa procedere nella conoscenza e che prova la sua esattezza: per questo *ce li ho messi* praticamente tutti, proprio li dove è giusto che siano, *dentro un libro*, in modo che all'occorrenza si sappia dove andarli a cercare. Forse in una qualche edizione futura riuscirò a corredare ogni capitolo con una sezione di riassunto finale, leggendo la quale si possa avere una visione sintetica degli argomenti trattati e dei risultati più rilevanti ottenuti. Ma come consiglio generale allo studio, posso solo suggerire di provare a scrivere per proprio conto un riassunto del genere, perché è proprio attraverso la scrittura e la ripetizione che si rinforza la

memoria; viceversa, leggendo e sottolineando è più probabile che si concili il sonno. Questa è una materia tanto affascinante quanto noiosa, non posso farci nulla.

Cultura libera

Nell'era di Internet un libro a carattere tecnico-scientifico non è un oggetto statico, bensì qualcosa che evolve per mantenere il passo con i progressi di ciò che descrive. Questo testo è inoltre espressione di un *progetto di cultura libera*, cultura che deve poter fluire liberamente dall'accademia al bagaglio di conoscenze di chiunque ne sia interessato, priva di vincoli di costo, intermediari, distribuzione, rating, e con la sola forza della libera circolazione delle idee; la sua disponibilità pubblica è regolata dalle norme di licenza Creative Commons *Attribuzione - Non commerciale - Condividi allo stesso modo*

`https://creativecommons.org/licenses/by-nc-sa/4.0/deed.it`

E' possibile contribuire allo sviluppo del progetto promuovendo la sua diffusione, acquistando una copia a stampa[3], o incoraggiandone lo sviluppo ulteriore attraverso una *donazione*[4], a cui corrisponde l'accesso *vitalizio* al formato PDF *navigabile* di tutte le edizioni presenti e future. Le donazioni ricevute sono poi in gran parte devolute ai progetti *open source* che hanno reso possibile il lavoro editoriale.

Prefazione all'edizione 2.0

Anche se ormai il testo si è fatto *maggiorenne*[5] i suoi contenuti non potevano che continuare ad espandersi, sia in termini di approfondenti che di nuovi argomenti. Il formato è un po' cambiato, da 16x23 siamo passati a 17x24 cm, ed il corpo dei font è passato da 9 ad 11, prima erano veramente *troppo piccoli*! Oltre a queste (ed altre) operazioni di *maquillage*, l'insieme dei contenuti è stato suddiviso in *quattro* parti[6]:

- teoria dei segnali (I) *(interamente contenuta nel primo volume)*
- modulazione (II) *(come le seguenti, inclusa nel secondo volume)*
- trasmissione dei segnali (III)
- sistemi di telecomunicazione (IV)

con un associato rimescolamento degli argomenti tra i capitoli rispetto alla precedente edizione. Se le primissime *uscite pubbliche* di questo testo (in cui gli argomenti, sia pur disparati, erano resi contigui) incontrassero questa loro progenie cresciuta, immagino che non la riconoscerebbero!

[3] Nella copia a stampa, dopo l'ultimo capitolo viene fornito il link al download del PDF completamente navigabile.

[4] Le donazioni sono raccolte a partire da `https://teoriadeisegnali.it/donazione.html`, mediante *Paypal* o carta.

[5] La prima edizione pubblica risale al 2001.

[6] Precedentemente erano tre, ora la modulazione ha acquisito lo status di parte a sé, dato che è un argomento che non necessariamente occorre spiegare (ad esempio) a studenti di (bio)informatica.

Principali novità

Questa edizione ha subito nel corso del suo sviluppo rallentamenti vari, ha attraversato il periodo del Covid per ritrovarsi con una guerra in Europa. Con l'auspicio che il buon senso possa prevalere sulla follia, vediamo cosa ci porta di nuovo l'edizione 2.0.

- sono investigati gli aspetti implementativi del *campionamento*, a cui segue il nuovo capitolo dedicato esclusivamente al *filtraggio*, che ora affronta anche le architetture *polifase* e *integratore-pettine in cascata*, chiudendo così il cerchio con la tecnica di sovracampionare e decimare;
- aggiunte figure esplicative della relazione tra DFT, DTFT e trasformata *zeta*;
- nel capitolo di *probabilità* si è approfondito lo studio della d.d.p. *gaussiana*, menzionata la funzione $Q()$, e dimostrato il *teorema centrale del limite*; è stata inoltre sviluppata l'esposizione della *funzione caratteristica*, e la relazione tra momenti e serie di potenze della d.d.p;
- il capitolo sulla *correlazione* si arricchisce della sezione dedicata alla *regressione lineare*, sia semplice che *multipla*, e della relazione che intercorre tra questa ed il *metodo dei minimi quadrati*, introducendo la matrice *pseudo inversa* e svolgendo esemplificazioni figurate dei *concetti geometrici* associati;
- aggiunto un esempio figurato sulla *distorsione non lineare* presente in una sinusoide che va in *saturazione*;
- la *teoria dell'informazione* è ora suddivisa in due capitoli, il primo orientato alla codifica *di sorgente* ed il secondo a quella *di canale*:
  - approfondito lo studio della sorgente informativa *gaussiana*, con il calcolo della sua *entropia differenziale* (anche per il caso *multidimensionale*), e la dimostrazione che questa è massima applicando il metodo dei *moltiplicatori di Lagrange*, a cui è stata dedicata una apposita appendice corredata di grafico esplicativo;
  - aggiunta sezione sull'informazione per coppie di v.a.: *entropia congiunta* e *condizionale*, *informazione mutua media*, entropia *relativa* (o divergenza di *Kullback Leibler*) ed entropia di *Rényi*;
  - aggiunta sezione che, dopo aver evidenziato la formula di *calcolo a catena* per la probabilità di sequenze markoviane, ne estende le conseguenze al calcolo dell'entropia e dell'informazione mutua media, sino ad arrivare alla dimostrazione del *data processing inequality*;
  - estesa la trattazione della *teoria velocità-distorsione*, e dimostrato il *limite inferiore di Shannon* $R(D) \geq h(X) - \frac{1}{2}\log_2(2\pi \mathrm{e} D)$ con $h(X)$ pari all'entropia differenziale della sorgente e $\frac{1}{2}\log_2(2\pi \mathrm{e} D)$ quella di una v.a. gaussiana con varianza D;
  - valutata l'entropia di un processo *gaussiano con memoria* e l'associata misura di *piattezza spettrale*, a cui segue la definizione di *funzione distorsione-velocità* mediante il procedimento di *water filling*;

  - la trattazione della *codifica di canale* si allinea agli sviluppi intercorsi nelle ultime decadi, affrontando argomenti come il codice di Reed-Solomon *accorciato*, la codifica *concatenata* e *l'interleaving* da associarvi; è stato aggiunto lo *pseudo codice* della decodifica di *Viterbi*, illustrata la tecnica del *tail-biting* ed approfondito il principio di *decisione soffice* con la verosimiglianza associata, graficate le *prestazioni* di Viterbi per diverse *lunghezze di vincolo*. Dopo un accenno alle alternative di codifica convoluzionale, si descrive il codice *perforato* e la *concatenazione* Solomon-Viterbi, affrontando quindi la decodifica di Viterbi con *uscite soffici*;
  - un significativo avanzamento riguarda poi la trattazione dei codici TURBO e di quelli a bassa densità di controllo parità (LDPC): per i primi dopo aver illustrato la *codifica ricorsiva parallela* e la decodifica SISO, sono definiti il rapporto di verosimiglianza logaritmica LLR e l'*informazione estrinseca*, valutata la LLR di ingresso e di uscita al SISO, e sviluppato l'algoritmo di *decodifica turbo*;
  - l'elegante codifica LDPC viene affrontata dopo aver attinto a diverse fonti, tutte citate, realizzando una sintesi senza per questo sprofondare nei conti. Dopo aver illustrato le particolarità della *matrice di controllo* $\boldsymbol{H}$ e l'associato *grafo di Tanner*, si affronta le decodifica *iterativa* basata sullo scambio di messaggi tra nodi secondo un principio di *propagazione della credenza*, che può essere implementato mediante un algoritmo *somma-prodotto*. Come per le altre tecniche vengono forniti grafici prestazionali per diverse condizioni operative, e gli attuali contesti di adozione.

- alla trattazione della *ricezione ottima* di una trasmissione numerica tramite canale con distorsione lineare e rumore bianco, si è aggiunta la dimostrazione di come una *equalizzazione ripartita* tra i due estremi del collegamento sia quella in grado di garantire le migliori prestazioni, e valutata la perdita conseguente alla necessità di localizzare tutta l'equalizzazione al ricevitore;
- alla discussione sulla *ripartizione della potenza complessiva* tra le diverse portanti dell'OFDM si è aggiunto lo studio analitico della massimizzazione della *capacità aggregata* che porta all'espressione della soluzione ottima di tipo *water-filling*;
- anche il capitolo sui *mezzi trasmissivi* si è scisso in due, il primo dedicato a *cavo* e *fibra*, ed il secondo ai collegamenti *radio*;
- infine la novità *più poderosa* di questa edizione sono le 60 pagine del capitolo sui sistemi *multiantenna* o MIMO! La possibilità (offerta dalla tecnologia) di dotarsi di *più di una antenna* ha aperto le porte ad una *evoluzione* che giunge fino ai nostri giorni, e che è ripercorsa nei suoi diversi aspetti. Il *canale vettoriale* MIMO viene definito come una matrice complessa $\boldsymbol{H}$ i cui elementi rappresentano il *guadagno aleatorio* del canale passa basso equivalente associato ad ogni coppia di antenne ai due lati del collegamento. Si mostra quindi come nel caso SIMO di più antenne al solo lato ricevente, quest'ultimo sia in grado di combinare in modo

coerente le copie di messaggio ricevuto, e poi come nel caso MISO di più antenne al solo lato trasmittente si possano definire *codici spazio-tempo* STC che consentono di *diluire* l'informazione trasmessa oltre che nel tempo, anche sulle antenne, permettendo di ottenere un *guadagno di diversità*. Segue poi un impegnativo approfondimento sulla *capacità* del canale MIMO vero e proprio, da confrontare con quella per i casi SISO, SIMO e MISO. Il risultato che si ottiene è la possibilità di operare su di un set di *canali virtuali indipendenti*, ottenibile eseguendo un opportuno *signal processing* dal lato del trasmettitore, purché quest'ultimo sia a conoscenza dei valori di $\boldsymbol{H}$; in tal caso si conseguono le prestazioni *ottime* dopo aver ripartito la potenza sui diversi canali virtuali con la tecnica del *riempimento d'acqua*.

- Ma siamo solo a metà capitolo! Qui il discorso *si allarga* introducendo la tecnica della *multiplazione spaziale*, ossia dell'inviare diversi messaggi in simultanea dalle diverse antenne di trasmissione. Il ricevitore multiantenna può allora applicare diverse strategie di decodifica, di *massima verosimiglianza* o ML, *sphere decoding*, *zero forcing*, MMSE, VBLAST.
- Si passa quindi a trattare la trasmissione multiutente o MU - MIMO, in cui le precedenti tecniche vengono per così dire *ribaltate* al trasmettitore (purché conosca $\boldsymbol{H}$) che può così effettuare il *precoding* dei messaggi da trasmettere. Nel caso di una trasmissione FDD sussiste quindi il problema di comunicare $\boldsymbol{H}$ tra Rx a Tx, oppure di realizzare un *beamforming opportunistico*.
- Nella trasmissione MIMO - OFDM alla *molteplicità* delle antenne si aggiunge quella delle sottoportanti, risolvendo allo stesso tempo in modo semplice il problema della *equalizzazione* necessaria alle velocità più elevate. Il canale risultante acquisisce oltre alla *diversità* spaziale anche quella *di frequenza*, permettendo di definire *codici spazio-tempo-frequenza* in grado di trarre vantaggio da entrambe. La trattazione dei sistemi *multiutente* MU-MIMO-OFDM si focalizza al caso TDD che *non* comporta la trasmissione della matrice $\boldsymbol{H}$ tra utenti mobili e stazione radio base (BS) e permette la definizione dei *blocchi di coerenza* entro i quali le parti si alternano a trasmettere, mentre la BASE STATION stima il canale di *downlink* mediante opportune *sequenze pilota ortogonali*. Il caso viene quindi calato nel contesto delle moderne *reti cellulari*, per le quali si forniscono gli opportuni rimandi di approfondimento.
- Il capitolo si conclude con l'applicazione della tecnica MISO al caso della diffusione *broadcast televisiva digitale* DVB-T mediante una *single frequency network* o SFN, in cui dopo una introduzione all'architettura ed agli aspetti trasmissivi, si approfondisce quello della sincronizzazione da parte del ricevitore TV dei simboli OFDM trasmessi da molteplici ripetitori, mediante l'utilizzo del segnale GPS e l'inserimento nel flusso MPEG di uno speciale pacchetto *mega frame initialization packet* o MIP.

Queste le novità in cui l'edizione 2.0 differisce rispetto alla 1.7 (si, ho saltato due decimali, ma ci voleva!) raccontate per esteso presso `https://teoriadeisegnali.it/libro/NEWS.txt`. Oltre, come sempre, alla miriade di altri aggiustamenti e precisazioni.

Posso dire di essere riuscito a mettere tutto dentro? Di certo gli ultimi sforzi pongono il testo su di un livello ancora più ambizioso! Anche se c'è ancora la situazione della quarta parte, che non ho toccato, e che potrebbe crescere includendo almeno i tratti essenziali di *cose di tutti giorni* come la *telefonia mobile*, l'IPv6, il *bluetooth*, il wireless ottico, il GPS... ma probabilmente queste possono ancora attendere: prima vorrei riuscire ad affrontare due argomenti più di base, il primo già acquisito dalla tecnologia, ed il secondo emergente. Si tratta rispettivamente della *software defined radio* o SDR, ossia le particolarità che assumono i processi di mo-demodulazione quando realizzati su dati campionati, e dei *segnali sui grafi*, che sfruttano le relazioni non topologiche[7] definite da una matrice di adiacenza, per sviluppare su questo tipo di segnali una analisi spettrale e definire approcci a filtraggio, sottocampionamento, inferenza e predizione. A chi può chiedersi *Si, ma cosa sono questi segnali sui grafi?* rispondo *di tutto*, dalle reti di sensori ai social network alle reti reputazionali, fino alle reti biologiche, di interazione proteica e malattia-farmaco, per arrivare alla medicina di precisione. E ti pare poco?

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

Alessandro Falaschi, Settembre 2022

[7] Mentre per i campioni di segnale nel tempo la topologia associata è definita dalla relazione di vicinanza unidirezionale *uno contro l'altro*, e per le immagini la topologia corrisponde a quella di una mappa spaziale bidimensionale, per i dati sui grafi la topologia è definita a partire dalla *matrice di adiacenza*, che si fa beffe di mappe e sequenze.

Elenco dei simboli

AAL	ATM Adaptation Layer
ACK	acknowledgment
ADM	Add and Drop Multiplexer
ADPCM	Adaptive Differential Pulse Coded Mudulation
ADSL	Asymmetric Digital Subscriber Line
AM	Amplitude Modulation
AMI	Alternate Mark Inversion
ARP	Address Resolution Protocol
ARQ	Automatic Repeat reQuest
AS	Autonomous System
ASCII	American Standard Code for Information Interchange
ATM	Asynchronous Transfer Mode
AWGN	Additive White Gaussian Noise
BCH code	codice Bose Chaudhuri Hocquenghem
BGP	Border Gateway Protocol
BLR	Banda Laterale Ridotta
BLU	Banda Laterale Unica
BPSK	Bi-Phase Shift Keying
BRAS	Broadband Remote Access Server
BRI	Basic Rate Interface
BS	Base Station
BSC	Binary Symmetric Channel
CAS	Channel Associated Signaling
CB	Coherence Block
CC	codice convoluzionale
CCS	Common Channel Signaling
CDM	Code Division Multiplex
CDMA	Code Division Multiple Access
CDN	Circuito Diretto Numerico
CELP	Code Excited Linear Prediction
CIC	Cascaded Integrator-Comb (filtro)
CIDR	Classless Interdomain Routing

COFDM	Coded OFDM
CPK	Continous Phase Keying
CRC	Cyclic Redundancy Check
CSI	Channel State Information
CSMA/CD	Carrier Sense Multiple Access - Collision Detect
DAC	Digital to Analog Converter
DBPSK	Differential Bi Phase Shift Keying
DCT	Discrete Cosine Transform
DFE	Decision Feedback Equalizer
DFT	Discrete Fourier Transform
DL	DownLink
DLL	Delay Locked Loop
DMT	Discrete Multi Tone
DNS	Domain Name Service
DPC	Dirty Paper Coding
DPCM	Differential Pulse Coded Modulation
DPLL	Digital Phase Locked Loop
DSLAM	Digital Subscriber Line Access Multiplexer
DSSS	Direct Sequence Spread Spectrum
DTFT	Discete Time Fourier Transform
DWDM	Dense Wavelength Division Multiplex
f.d.t.	funzione di trasferimento
FDM	Frequency Division Multiplex
FEC	Forward Error Correction
FEXT	Far End Crosstalk
FFT	Fast Fourier Transform
FIFO	First In First Out
FIR	Finite Impulse Response
FM	Frequency Modulation
FSK	Frequency Shift Keying
FTTH	Fiber To The Home
GIF	Graphics Interchange Format
GMSK	Gaussian Minimum Shift Keying
GOB	Group Of (macro)Blocks
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
IDFT	Inverse Discrete Fourier Transform
IEEE	Institute of Electrical and Electronics Engineers
IGP	Interior Gateway Protocols
IIR	Infinite Impulse Response
IP	Internet Protocol
ISDN	Integrated Service Data Network
ISI	Inter Symbol Interference
JPEG	Joint Photographic Experts Group
LAN	Local Area Network

LDPC	Low Density Parity Check
LLC	Logical Link Control
LMS	Least Mean Square
LOS	Line Of Sight
LPC	Linear Predictive Coding
MAC	Media Access Control
MFN	Multiple Frequency Network
MIMO	Multiple Input Multiple Output
MIP	Megaframe Initialization Packet
MISO	Multiple Input Single Output
ML	Maximum Likelihood
MLSD	Maximum Likelihood Sequence Detection
MMSE	Minimum Mean Square Error
MPEG	Moving Pictures Expert Group
MRC	Maximal Ratio Combining
MSK	Minimum Shift Keying
MU-MIMO	Multiple User MIMO
NEXT	Near End Crosstalk
OFDM	Ortogonal Frequency Division Multiplex
OFDMA	Ortogonal Frequency Division Multiple Access
PCM	Pulse Coded Modulation
PDH	Plesiochronous Digital Hierarcy
PLL	Phase Locked Loop
PON	Passive Optical Network
POTS	Plain Old Telephony Services
PPS	Pulse Per Second
PSTN	Public Switched Telephone Network
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
QV	Quantizzazione Vettoriale
RELP	Residual Excited Linear Prediction
SAP	Service Access Point
SDH	Synchronous Digital Hierarcy
SDMA	Space Division Multiple Access
SFBC	Space Frequency Block Code
SFN	Single Frequency Network
SIMO	Single Input Multiple Output
SINR	Signal / Interferent and Noise Ratio
SISO	Single Input Single Output
SISO	Soft Input, Soft Output
SNAP	Subnetwork Access Protocol
SOVA	Soft Output Viterbi Algorithm
STBC	Space Time Block Code
STC	Space Time Code

STFBC	Space Time Frequency Block Code
STM	Synchronous Transport Module
STS	Synchronization Time Stamp
STTC	Space Time Trellis Code
SVD	Singular Value Decomposition
TCP	Transport Control Protocol
TDD	Time Division Duplex
UDP	User Datagram Protocol
UE	User Equipment
UL	UpLink
VBLAST	Vertical Bell Laboratories Layered Space-Time
VCO	Voltage Controlled Oscillator
VLAN	Virtual LAN
ZF	Zero Forcing

Parte II

Modulazione per segnali analogici

Prefazione alla seconda parte

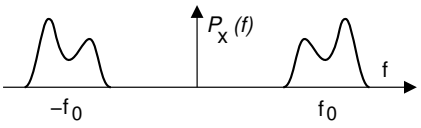
I QUATTRO capitoli dal 11 al 14 sono dedicati alla caratterizzazione, generazione e ricezione dei *segnali modulati* o *passa-banda*, gli unici idonei a trasportare informazione su di un canale radio. Come dire che se non ci fosse la modulazione, lo sviluppo dell'umanità sarebbe fermo a due secoli fa, e priva di radio, TV, WIFI, Bluetooth...

Dopo aver arricchito l'arsenale degli strumenti analitici con i concetti di *inviluppo complesso*, *filtro di Hilbert* e *componenti analogiche di bassa frequenza*, si giunge ad individuare l'espressione della *densità spettrale* di processi stazionari ergodici, ovvero del relativo inviluppo complesso e delle sue componenti di bassa frequenza. Il cap. 12 passa quindi a descrivere le varie tecniche di *modulazione di ampiezza*, assieme agli altrettanto vari modi di effettuarne la *demodulazione*; dopodiché sono affrontate le modulazioni *di fase* e *di frequenza*, mostrando come in questo caso l'occupazione di banda possa essere variata entro ampi margini, mentre invece l'inviluppo di ampiezza si mantiene costante. Mentre nel cap. 13 si analizzano le conseguenze del *passaggio* dei segnali modulati attraverso i sistemi fisici, e viene definito il concetto di *segnale a banda stretta*, al cap. 14 si investiga su come il processo di demodulazione possa alterare (o meno) la qualità di un segnale ricevuto attraverso un canale rumoroso e caratterizzato dal relativo rapporto segnale-rumore, permettendo di approfondire come il *compromesso banda-potenza* sia immediatamente applicabile al caso della modulazione di frequenza. Infine, al § 14.4 si approfondisce lo studio della detezione *incoerente* di sinusoide immersa nel rumore, portando con se la definizione delle v.a. di Rayleigh e di Rice, ed introducendo i principi su cui è basata la *teoria della decisione*.

Capitolo 11

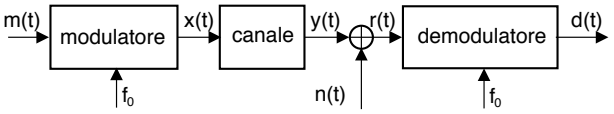
Segnali modulati

Indicati anche come segnali *passa banda* o *in banda traslata*, dato che occupano una banda circoscritta ad una regione di frequenze contigua ad un valore f_0, detta *frequenza portante.*



I segnali passa banda sono quasi sempre il risultato di una operazione di *modulazione* (§ 3.5.2) che trasforma un segnale $m(t)$ *modulante* (o di *banda base*) in un secondo segnale $x(t)$ *modulato,* allo scopo di renderlo idoneo alla trasmissione mediante il canale a disposizione, in base alle considerazioni discusse nel seguito.

Il processo inverso viene indicato come *demodulazione*, che se eseguita direttamente su $x(t)$ permette di riottenere $m(t)$; al contrario, in ingresso al lato ricevente è presente il segnale $r(t) = y(t) + n(t)$, ovvero la somma tra ciò che esce dal canale (che può introdurre una *distorsione*) ed un disturbo additivo o rumore (*noise*), vedi cap. 8; pertanto il segnale *demodulato* $d(t)$ può essere espresso come $d(t) = m(t) + \varepsilon(t)$ in cui il termine $\varepsilon(t) = \text{Dem}\{n(t)\} + \text{Dem}\{y(t) - x(t)\}$ tiene conto sia del risultato della demodulazione del rumore in ingresso, sia dagli effetti prodotti dalla demodulazione sulle alterazioni introdotte dal canale sul segnale modulato.



Il processo di modulazione è quasi sempre associato ad una trasmissione radio, ma può rendersi necessario e/o utile anche per trasmissioni via cavo. In generale, individuiamo almeno tre situazioni in cui è necessario l'impiego di segnali modulati:

- il canale non permette la trasmissione di frequenze contigue all'origine, presenti invece nel segnale;
- il canale presenta un comportamento ideale (modulo costante e fase lineare, vedi cap 8) solo in determinati intervalli di frequenza;
- il canale presenta disturbi additivi (ovvero altre trasmissioni) solo in determinate regioni di frequenza.

11.1 Contesti applicativi e prime definizioni

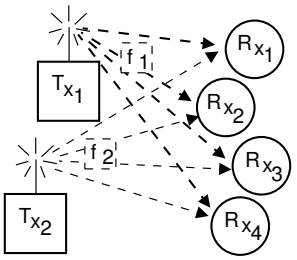
Prima di addentrarci nei dettagli analitici del § 11.2, descriviamo le principali modalità di trasmissione dei segnali modulati, assieme alle relative motivazioni.

11.1.1 Multiplazione a divisione di frequenza

Consiste in una tecnica di trasmissione in cui più comunicazioni avvengono in contemporanea, condividendo lo stesso mezzo fisico, ma impegnando ognuna una diversa banda di frequenze, per il semplice motivo che se utilizzassero tutte la stessa banda, costituirebbero termini di *interferenza* reciproca[1]. 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 tipici.

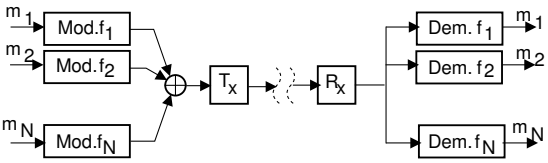
11.1.1.1 Collegamenti punto-multipunto

Si tratta della modalità 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 cosiddetti *canali* della TV), mentre altre emittenti utilizzano contemporaneamente lo stesso mezzo trasmissivo, occupando canali centrati ad altre frequenze. Nel caso di trasmissione TV analogica (§ 25.1) e di radio FM (§ 25.2) vi è una corrispondenza 1:1 tra frequenza portante ed emittente, mentre ad es. nella TV digitale sulla stessa portante vengono trasmesse (o meglio *multiplate*) più emittenti (§ 10.3.2.1).



11.1.1.2 Collegamenti punto-punto

E' una forma di multiplazione FDM (che sta per *frequency division multiplex)* per mezzo della quale un collegamento tra due località distanti viene *condiviso* per il trasporto di più comunicazioni. Un insieme di N segnali m_i, $i = 1, 2, ..., N$, transita quindi su di uno stesso mezzo trasmissivo, occupando ognuno una differente banda centrata su di una diversa portante f_i, $i = 1, 2, ..., N$, e può essere individualmente demodulato 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.*

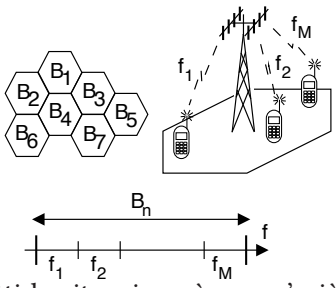


11.1.1.3 Accesso multiplo

E' la tipica strategia di *accesso* alla rete (pag. 828) per le *comunicazioni mobili*, ovvero per la telefonia cellulare e le reti WIFI. Nel primo caso il territorio è suddiviso in *celle*, per ognuna delle quali viene utilizzata una diversa banda (B_n) di frequenze

[1] Ma non sempre questo impedisce la comunicazione, vedi § 16.9.

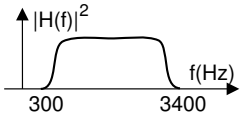
radio, dedicata alla comunicazione tra i terminali ed una unica antenna fissa. All'interno della cella la banda a disposizione è ulteriormente suddivisa tra più canali, ognuno associato ad una diversa portante (f_i), usati a turno dai terminali che desiderano comunicare[2].



Sotto certi aspetti questo caso è in qualche modo antitetico rispetto al § 11.1.1.1, e potrebbe essere indicato come collegamento *multipunto-punto*. In effetti la situazione è un po' più complessa, e gli aspetti qualificanti da un punto di vista sistemistico sono la sincronizzazione tra radiomobili e stazione base, e i protocolli di rete necessari per consentire le fasi di richiesta di accesso, la negoziazione dei parametri di trasmissione, la localizzazione dei radiomobili, e la corretta gestione del cambio di cella, detto *handover*.

11.1.2 Canale telefonico

Le caratteristiche del collegamento offerto dalla comune linea telefonica (§ 24.9.1) rivestono molteplici aspetti. Uno di questi, forse il principale[3], 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 posta pari a 4000 Hz[4]: discutiamo brevemente le origini storiche di tali limitazioni.



L'assenza della regione -300÷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, assieme alla scelta operata nel *vecchio* metodo di multiplazione FDM punto-punto[6], per il quale i singoli canali sono modulati AM-BLU (vedi § 12.1.2), che pure impone di rimuovere le componenti frequenziali più basse.

---

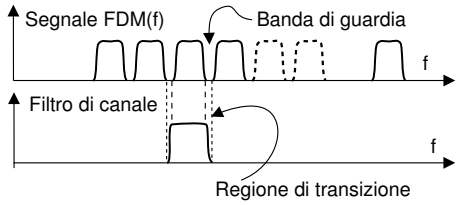
[2] Un minimo di approfondimento (per il GSM) può essere trovato al § 26.1...

[3] Un altro fattore rilevante è la *limitazione della potenza* che è possibile immettere su di un singolo collegamento telefonico e che, associato al precedente, caratterizza il canale telefonico come limitato sia in banda che in potenza, e dunque con capacità (§ 17.3) $C = W \log_2 \left(1 + \frac{\mathcal{P}_s}{N_0 W}\right)$ dipendente solo dal livello di rumore. La limitazione in potenza è motivata storicamente da problemi di *diafonia* (pag. 648) dovuti a fenomeni di induzione elettromagnetica, mentre attualmente è determinata dalla limitata dinamica del segnale che viene campionato e trasmesso in forma numerica (§ 4.3.2).

[4] Questo valore massimo nominale determina che la frequenza di campionamento del PCM telefonico è pari a 2*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, vedi § 10.1.

[5] L'ibrido telefonico è un trasformatore con quattro porte, che realizza la separazione tra le due vie di comunicazione che viaggiano sullo stesso cavo (vedi § 24.9.1). 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 § 24.9.2).

[6] Nel secolo scorso venne definita una vera e propria *gerarchia* di multiplazione, i cui livelli detti di *gruppo*, *super gruppo*, *gruppo master* e *gruppo jumbo* accolgono rispettivamente 12, 60, 600 e 3600 canali voce, per essere trasmessi su doppino, cavo coassiale, o ponte radio. Un approfondimento presso `https://www.vialattea.net/content/883/` e `https://en.wikipedia.org/wiki/L-carrier`.



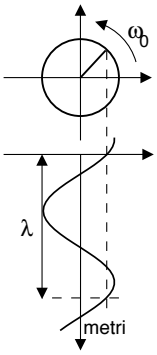
La stessa FDM è anche causa della limitazione per le frequenze da 3400 Hz in poi, dato che è necessario separare i segnali multiplati mediante filtri passa-banda *di canale* che, per essere economicamente realizzabili, devono presentare una regione di transizione di estensione apprezzabile. Tra canali contigui è quindi necessario prevedere un intervallo di frequenze detto *banda di guardia* (pari a 900 Hz), che impone la limitazione a 3400 Hz per la massima frequenza di segnale, in modo da ottenere $300 + (4000 - 3400) = 900$ Hz. In assenza di tale intervallo, all'uscita di un filtro di canale si troverebbe anche parte del segnale presente su di un canale contiguo, producendo interferenza tra comunicazioni diverse.

La limitazione in banda di un canale telefonico tra 300 e 3400 Hz è dunque il motivo per cui la connessione telefonica tra un computer ad un fornitore di connettività Internet *richiedeva* (in tempi pre-ADSL) l'uso di un dispositivo *modem*, che effettua una forma di modulazione sul segnale da trasmettere sul cavo, che arrivava in tale forma fino al provider. Al contrario, nel caso dell'accesso ADSL (vedi § 24.9.4) 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 (§ 19.2.3.1) che è ad uso esclusivo dell'utente.

11.1.3 Antenne e lunghezza d'onda

La trasmissione di un segnale via onda radio necessita di un'antenna di dimensione comparabile alla 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

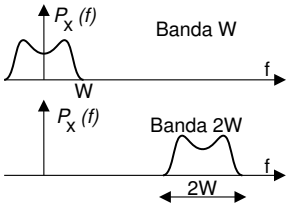
$$\lambda = c \cdot T = \frac{c}{f}$$



Nel caso di segnali modulati il valore di f è quello della portante f_0, in quanto il segnale modulato occupa una banda ristretta attorno ad essa; in figura è rappresentato un vettore rotante a velocità angolare $\omega_0 = 2\pi f_0$ la cui proiezione sulle ascisse produce il valore (del campo elettromagnetico) che *viaggia* a velocità c, e che durante un periodo $T = 1/f_0$ percorre λ metri. Trasmissioni con portanti più elevate necessitano di antenne di dimensioni ridotte; d'altra parte se per assurdo trasmettessimo 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 ![7]

[7] Antenne più corte di λ 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 § 20.5.3 è riportata una tabella dei valori di λ per i diversi servizi di TLC.

11.1.4 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); in tal senso la banda di segnale viene a volte indicata come *banda a frequenze positive*.

11.2 Rappresentazione dei segnali modulati

Terminata la parte introduttiva, affrontiamo gli sviluppi analitici basati sulla possibilità di esprimere un segnale modulato $x(t)$ nella forma

$$x(t) = x_c(t)\cos 2\pi f_0 t - x_s(t)\sin 2\pi f_0 t \tag{11.1}$$

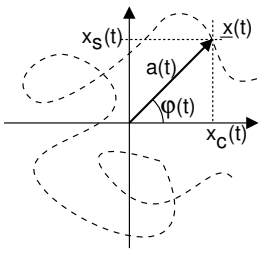
in cui, se f_0 è scelta entro la banda occupata dal segnale, $x_c(t)$ e $x_s(t)$ sono segnali *limitati in banda* con banda contigua all'origine, e le alterazioni prodotte sul segnale modulato da parte del messaggio modulante $m(t)$ possono essere descritte mediante operazioni condotte su $x_c(t)$ ed $x_s(t)$. Ciò significa che $x(t)$ potrà essere sintetizzato, ed il messaggio recuperato, agendo su segnali con banda molto ridotta rispetto alla massima frequenza di $x(t)$. Iniziamo a mostrare come $x_c(t)$ ed $x_s(t)$ siano in realtà la parte reale ed immaginaria di un terzo segnale di banda base, chiamato...

11.2.1 Inviluppo complesso

E' definito come un segnale *complesso* legato a $x_c(t)$ e $x_s(t)$ dalla relazione

$$\underline{x}(t) = x_c(t) + jx_s(t) \tag{11.2}$$

ed è una estensione tempo-variante del concetto di *fasore* (vedi § 2.1.3), che a sua volta consente di rappresentare un segnale del tipo $x(t) = a\cos(\omega_0 t + \varphi)$[8] per mezzo del numero complesso $\underline{x} = ae^{j\varphi}$, mediante la relazione $x(t) = \Re\left\{\underline{x}e^{j\omega_0 t}\right\}$. In modo simile, l'*inviluppo complesso* $\underline{x}(t)$ può essere pensato come un fasore per il quale il modulo a e la fase φ sono funzioni del tempo, ovvero



$$\underline{x}(t) = a(t)\,e^{j\varphi(t)} \tag{11.3}$$

come rappresentato nella figura a fianco assieme ad una sua possibile traiettoria temporale. Ad $\underline{x}(t)$ possiamo quindi associare un segnale *reale*

$$x(t) = \Re\left\{\underline{x}(t)\,e^{j\omega_0 t}\right\} = \Re\left\{a(t)\,e^{j(\omega_0 t+\varphi(t))}\right\} = a(t)\cos(\omega_0 t + \varphi(t)) \tag{11.4}$$

in cui il termine $e^{j\omega_0 t}$ corrisponde ad imprimere ad $\underline{x}(t)$ una rotazione in senso *antiorario*

[8] Per brevità, qui e nel seguito adottiamo a volte la notazione $2\pi f_0 = \omega_0$.

a velocità angolare ω_0. D'altra parte, sviluppando la rappresentazione polare (11.3) come

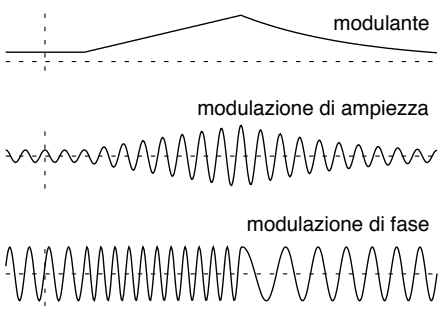
$$\underline{x}(t) = a(t)\, e^{j\varphi(t)} = a(t)\cos\varphi(t) + ja(t)\sin\varphi(t)$$

osserviamo che $x_c(t) = \Re\left\{\underline{x}(t)\right\} = a(t)\cos\varphi(t)$ e $x_s(t) = \Im\left\{\underline{x}(t)\right\} = a(t)\sin\varphi(t)$, permettendoci di dimostrare che la (11.4) è equivalente alla (11.1), in quanto[9]

$$\begin{aligned} x(t) &= \Re\left\{\underline{x}(t)\, e^{j\omega_0 t}\right\} = a(t)\cos(\omega_0 t + \varphi(t)) = \\ &= a(t)\left[\cos\omega_0 t\cos\varphi(t) - \sin\omega_0 t\sin\varphi(t)\right] \\ &= x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t \end{aligned} \tag{11.5}$$

11.2.2 Modulazione di ampiezza, di fase e di frequenza

L'inviluppo complesso è un potente strumento che permette di descrivere il processo di modulazione in modo semplice ed omogeneo. Ad esempio, la moltiplicazione del segnale $a(t)$ di banda base per un coseno a frequenza (*portante*) f_0 (vedi § 3.5.2) $x(t) = a(t)\cos(2\pi f_0 t)$ corrisponde a scrivere l'eq. (11.3) come $\underline{x}(t) = a(t)$, ovvero corrisponde ad un inviluppo complesso $\underline{x}(t) = a(t)$ a fase nulla, e prende il nome di *modulazione di ampiezza*[10] dato che è appunto l'ampiezza della portante a variare in accordo al segnale $a(t)$. Se al contrario consideriamo un inviluppo complesso con modulo costante $\underline{x}(t) = ae^{j\varphi(t)}$ l'andamento della fase $\varphi(t)$ imprime alla portante un diverso tipo di modulazione, detta ora *modulazione di fase*[11] o *angolare*, in quanto il segnale modulante ($\varphi(t)$ in questo caso) altera l'argomento del coseno, ottenendo dalla (11.4) il segnale modulato $x(t) = a\cos(2\pi f_0 t + \varphi(t))$.



Prima di proseguire riflettiamo sull'esempio mostrato in figura, 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 $\underline{x}(t) = \mathrm{e}^{jm(t)}$ la portante modulata angolarmente $x(t) = \cos(2\pi f_0 t + m(t))$ 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 $2\pi f_0 t + \alpha t = 2\pi\left(f_0 + \frac{\alpha}{2\pi}\right)t$ e dunque la frequenza portante *aumenta* di $\frac{\alpha}{2\pi}$.

Per meglio descrivere il caso di modulazione angolare, indichiamo l'argomento del coseno come *fase istantanea*

$$\psi(t) = 2\pi f_0 t + \varphi(t)$$

[9] Si faccia uso della relazione $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$.

[10] Indicata anche come AM (*amplitude modulation*).

[11] Indicata anche come PM (*phase modulation*).

e la sua derivata normalizzata come *frequenza istantanea*

$$f_i(t) = \frac{1}{2\pi}\frac{d}{dt}\psi(t) = f_0 + \frac{1}{2\pi}\frac{d}{dt}\varphi(t) \tag{11.6}$$

In questi termini, la modulazione angolare viene distinta in *modulazione di fase* propriamente detta quando $m(t)$ si limita ad alterare la fase della portante in modo diretto, ovvero

$$\varphi(t) = k_\varphi m(t)$$

mentre viene detta *modulazione di frequenza* quando la fase dipende dall'integrale di $m(t)$, ovvero

$$\varphi(t) = 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau \tag{11.7}$$

dato che in questo caso è la frequenza *istantanea* (11.6) a dipendere direttamente dal segnale modulante: $f_i(t) = f_0 + k_f m(t)$.

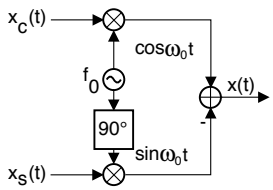
11.2.3 Componenti analogiche di bassa frequenza

Sono anch'esse definite a partire da $a(t)$ e $\varphi(t)$ come

$$x_c(t) = a(t)\cos\varphi(t) \qquad \text{e} \qquad x_s(t) = a(t)\sin\varphi(t) \tag{11.8}$$

e mentre l'eq. (11.2) le identifica con la parte reale ed immaginaria dell'inviluppo complesso $\underline{x}(t) = x_c(t) + jx_s(t)$, l'eq. (11.5) permette loro di descrivere completamente un segnale modulato nella forma $x(t) = x_c(t)\cos 2\pi f_0 t - x_s(t)\sin 2\pi f_0 t$: quest'ultima espressione motiva la scelta dei pedici $_c$ ed $_s$, così come l'appellativo di componente *in fase* (per $x_c(t)$) ed *in quadratura* (per $x_s(t)$) del segnale modulato.

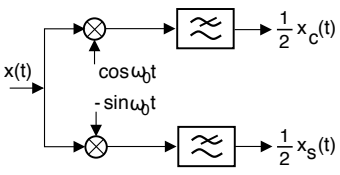
Osserveremo tra breve come, scegliendo per f_0 una frequenza al centro della banda $2W$ del segnale modulato, le componenti analogiche di bassa frequenza $x_c(t)$ e $x_s(t)$ (d'ora in poi *c.a. di b.f.*) risultino essere *limitate in banda*, con banda $2W$ centrata attorno all'origine. D'altra parte, è molto semplice verificare come l'inverso sia vero: il segnale modulato espresso dalla (11.1) può essere infatti ottenuto a partire da $x_c(t)$ e $x_s(t)$ limitate in banda mediante il semplice schema di elaborazione mostrato in figura, detto *modulatore in fase e quadratura*, che rappresenta una via per *sintetizzare* un segnale modulato (in ampiezza, od angolarmente, od entrambe le cose), a partire dalle sue c.a. di b.f., che a loro volta sono ottenibili a partire da $a(t)$ e $\varphi(t)$ in base alle (11.8).



modulatore in fase e quadratura

11.2.4 Demodulazione in fase e quadratura

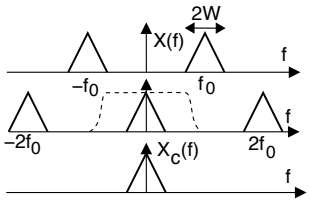
Come il segnale modulato $x(t)$ può essere *sintetizzato* a partire da $x_c(t)$ e $x_s(t)$, così le c.a. di b.f. possono essere recuperate da $x(t)$ adottando lo schema simbolico in figura, in cui il segnale modulato è moltiplicato per due portanti (dette *in fase* ed *in quadratura*), di cui la prima con la medesima frequenza e fase di quella utilizzata dal modu-



latore[12], ovvero pari a $\cos(2\pi f_0 t)$ e per questo detta *coerente*, *sincrona* od *omodina*, mentre la seconda (in quadratura) ha un *anticipo* di fase pari a $\pi/2$, ovvero è pari a $\cos(2\pi f_0 t + \pi/2) = -\sin(2\pi f_0 t)$. Su entrambi i rami è quindi posto un filtro passa basso[13].

Il funzionamento del demodulatore è basato sul fatto che, considerando $x(t)$ espresso nei termini delle sue c.a. di b.f., per il ramo in fase si ottiene[14]:

$$\begin{aligned} x(t)\cos\omega_0 t &= [x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t]\cos\omega_0 t = \\ &= x_c(t)\cos^2\omega_0 t - x_s(t)\sin\omega_0 t\cos\omega_0 t = \\ &= \tfrac{1}{2}x_c(t) + \tfrac{1}{2}x_c(t)\cos 2\omega_0 t - \tfrac{1}{2}x_s(t)\sin 2\omega_0 t \end{aligned}$$



I termini in cui compaiono $\cos 2\omega_0 t$ e $\sin 2\omega_0 t$ sono relativi a componenti di segnale centrate attorno a $2f_0$, che il filtro passa basso (la cui $H(f)$ è tratteggiata in figura) provvede ad eliminare: la banda del filtro deve quindi essere maggiore di W ma inferiore a $2f_0 - W$. Pertanto, non è necessario un filtro rettangolare, e se $f_0 \gg W$ non sussistono particolari problemi realizzativi. Procedendo in maniera simile[15], per il ramo in quadratura si ottiene:

$$\begin{aligned} -x(t)\sin\omega_0 t &= -[x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t]\sin\omega_0 t = \\ &= x_s(t)\sin^2\omega_0 t - x_c(t)\sin\omega_0 t\cos\omega_0 t = \\ &= \frac{1}{2}x_s(t) - \frac{1}{2}x_s(t)\cos 2\omega_0 t - \frac{1}{2}x_c(t)\sin 2\omega_0 t \end{aligned}$$

ed come prima il filtro passa-basso rimuove le componenti a frequenza doppia.

Se i filtri non sono ideali, ma hanno ad esempio una fase lineare (pag. 125), 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, vedi § 13.1.1.1.

Ricostruzione del segnale modulante Una volta che $x_c(t)$ e $x_s(t)$ sono note, queste permettono di risalire alla modulazione di ampiezza e di fase come

$$\begin{cases} a(t) = \left|\underline{x}(t)\right| = \sqrt{x_c^2(t) + x_s^2(t)} \\ \varphi(t) = \arg\left\{\underline{x}(t)\right\} = \arctan 2\,(x_s, x_c) \end{cases} \tag{11.9}$$

in cui si adotta la funzione $\arctan 2\,(x_s, x_c)$, che al contrario di $\arctan \frac{x_s}{x_c}$ tiene conto del segno[16] di x_c ed x_s, e restituisce un angolo compreso nell'intervallo $(-\pi, \pi)$ anziché

[12] Le modalità di sincronizzazione della portante utilizzata al ricevitore rispetto a quella usata in trasmissione sono esposte al § 12.2.1.

[13] Il simbolo $\boxed{\approx}$ rappresenta un filtro passa-basso, poiché viene *cancellata* l'ondina superiore. Nello stesso stile, possono essere indicati un passa-alto $\boxed{\approx}$ ed un passa-banda $\boxed{\approx}$.

[14] Si fa uso delle relazioni $\cos^2\alpha = \frac{1}{2}(1 + \cos 2\alpha)$ e $\sin\alpha\cos\alpha = \frac{1}{2}\sin 2\alpha$

[15] Utilizzando stavolta le relazioni $\sin\alpha\cos\alpha = \frac{1}{2}\sin 2\alpha$ e $\sin^2\alpha = \frac{1}{2}(1 - \cos 2\alpha)$, ed eseguendo il prodotto $-\sin\omega_0 t\,[x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t]$.

[16] Vedi `https://it.wikipedia.org/wiki/Arcotangente2`.

$(-\pi/2, \pi/2)$.

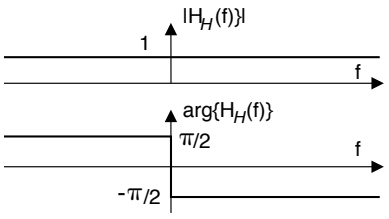
Oltre allo schema circuitale ora discusso esiste anche un approccio analitico che calcola le c.a. di b.f. a partire da $x(t)$ e dalla sua *trasformata di Hilbert* $\widehat{x}(t)$. Definendo quindi il *segnale analitico* $x^+(t)$ (§ 11.2.6) associando ad $x(t)$, è infine possibile esprimere la densità di potenza del segnale modulato $\mathcal{P}_x(f)$ nei termini di quella del suo inviluppo complesso $\mathcal{P}_{\underline{x}}(f)$ (§ 11.2.7). Prendiamo questa strada.

11.2.5 Trasformata di Hilbert

Al contrario di Fourier e di Laplace, quella di *Hilbert* è una trasformata che restituisce nuovamente una funzione del tempo, indicata come

$$\widehat{x}(t) = \mathcal{H}\{x(t)\}$$

ed equivalente al filtraggio di $x(t)$ mediante *un filtro di Hilbert* (§ 11.4.1) la cui risposta in frequenza $H_{\mathcal{H}}(f) = -j \cdot \mathrm{sgn}(f)$ è graficata in figura, e che causa in $X(f)$ una alterazione della fase pari a $\mp\frac{\pi}{2}$ a seconda se $f \gtrless 0$.



Anticipiamo subito (vedi § 11.4.2) che la trasformata di Hilbert di un segnale modulato di cui è noto l'inviluppo complesso $\underline{x}(t)$ risulta pari a

$$\begin{aligned}\hat{x}(t) &= \Im\left\{\underline{x}(t)\, e^{j\omega_0 t}\right\} = a(t)\sin(\omega_0 t + \varphi(t)) = \\ &= x_c(t)\sin\omega_0 t + x_s(t)\cos\omega_0 t\end{aligned} \tag{11.10}$$

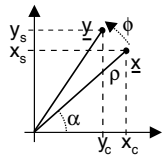
in cui si tiene conto che $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$ e si applicano le (11.8).

Demodulazione delle c.a. di b.f. Affiancando alla (11.10) la relazione (11.1) si imposta il sistema di due equazioni nelle incognite $x_c(t)$ e $x_s(t)$

$$\begin{cases} x(t) = x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t \\ \widehat{x}(t) = x_c(t)\sin\omega_0 t + x_s(t)\cos\omega_0 t \end{cases} \tag{11.11}$$

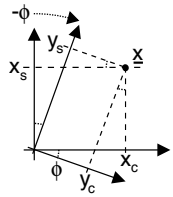
che rappresenta una *rotazione* in senso *orario* del piano dell'inviluppo complesso[17] e

[17] Mostriamo che una matrice di coefficienti della forma $\begin{pmatrix}\cos\phi & -\sin\phi \\ \sin\phi & \cos\phi\end{pmatrix}$ individua una rotazione.



Esprimiamo infatti un numero complesso $\underline{x} = x_c + jx_s$ in forma polare $\underline{x} = \rho e^{j\alpha}$, sussistendo l'uguaglianza $x_c = \rho\cos\alpha$ e $x_s = \rho\sin\alpha$; con riferimento alla figura, immaginiamo ora che $\underline{x}$ ruoti in senso *antiorario* di un angolo (*positivo*) ϕ, ottenendo il nuovo numero complesso $\underline{y} = \underline{x}e^{j\phi} = \rho e^{j\alpha+\phi} = y_c + jy_s$, in cui

$$\begin{cases} y_c = \rho\cos(\alpha+\phi) = \rho(\cos\alpha\cos\phi - \sin\alpha\sin\phi) = x_c\cos\phi - x_s\sin\phi \\ y_s = \rho\sin(\alpha+\phi) = \rho(\sin\alpha\cos\phi + \cos\alpha\sin\phi) = x_c\sin\phi + x_s\cos\phi \end{cases}$$

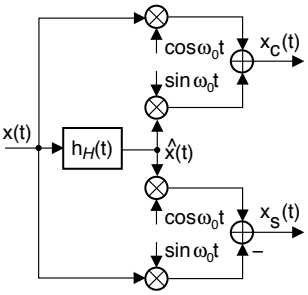


ovvero la matrice dei coefficienti corrisponde a quella preannunciata. Alternativamente, le nuove coordinate y_c , y_s corrispondono a quelle di un vettore fisso, ma riferito ad un sistema di assi ortogonali che ruotano in senso *orario* dello stesso angolo ϕ.

che può essere risolto[18], istante per istante[19], permettendo in definitiva di esprimere le componenti analogiche di bassa frequenza in termini di $x(t)$ e di $\widehat{x}(t)$ come

$$\begin{cases} x_c(t) = x(t)\cos\omega_0 t + \widehat{x}(t)\sin\omega_0 t \\ x_s(t) = -x(t)\sin\omega_0 t + \widehat{x}(t)\cos\omega_0 t \end{cases} \tag{11.12}$$

Alle (11.12) corrisponde lo schema simbolico mostrato a lato, che illustra come le componenti analogiche di bassa frequenza possano essere ottenute direttamente da $x(t)$ grazie all'uso di un *filtro di Hilbert* $h_H(t)$ (§ 11.4.1) per ottenere $\widehat{x}(t)$, e combinando i due segnali per mezzo di oscillatori in quadratura. Una volta determinate $x_c(t)$ e $x_s(t)$ si può procedere come a pag. 348 per ricavare il segnale modulante espresso da $a(t)$ e $\varphi(t)$.



Ora che abbiamo esaminato due diversi metodi per ottenere le c.a. di b.f., affrontiamo il problema di individuare una relazione tra la densità di potenza dell'inviluppo complesso $\mathcal{P}_{\underline{x}}(f)$ e quella $\mathcal{P}_x(f)$ del segnale modulato. A tale scopo, occorre prima definire il...

11.2.6 Segnale analitico

Riprendendo l'analogia introdotta al § 11.2.1 tra inviluppo complesso $\underline{x}(t) = a(t)\,e^{j\varphi(t)}$ e fasore $\underline{x} = ae^{j\varphi}$ osserviamo che per entrambi si può risalire al segnale a cui si riferiscono (una portante, modulata o meno) oltre che mediante la relazione $x(t) = \Re\left\{\underline{x}(t)\,e^{j\omega_0 t}\right\}$, anche come

$$x(t) = \tfrac{1}{2}\left(\underline{x}(t)\,e^{j\omega_0 t} + \underline{x}^*(t)\,e^{-j\omega_0 t}\right) \tag{11.13}$$

in cui vi sono due fasori *coniugati* che ruotano l'uno in senso opposto all'altro (vedi eq. (2.5) a pag. 38), in modo che la loro somma *vettoriale* sia pari[20] a $\Re\left\{\underline{x}(t)\,e^{j\omega_0 t}\right\}$. Proseguendo con l'analogia, come la scomposizione di un coseno

$$x(t) = a\cos(2\pi f_0 t + \varphi) = \tfrac{1}{2}\underline{x}e^{j\omega_0 t} + \tfrac{1}{2}\underline{x}^* e^{-j\omega_0 t}$$

secondo la formula di Eulero[21] dà luogo a due impulsi in frequenza ovvero

---

[18] Verifichiamo che il prodotto tra le matrici dei coefficienti di (11.10) e (11.11) fornisca la matrice identità

$$\begin{pmatrix} \cos & -\sin \\ \sin & \cos \end{pmatrix}\begin{pmatrix} \cos & \sin \\ -\sin & \cos \end{pmatrix} = \begin{pmatrix} \cos^2+\sin^2 & \cos\sin-\cos\sin \\ -\cos\sin+\cos\sin & \sin^2+\cos^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

[19] Dato che i coefficienti $\cos\omega_0 t$, $\sin\omega_0 t$ del sistema (11.11) sono funzione del tempo, le equazioni relative rappresentano una *rotazione oraria* di $\underline{x}(t)$ che "ruota" con velocità angolare ω_0, ossia con un angolo $\omega_0 t$ che *aumenta linearmente* nel tempo. Pertanto le coppie di segnali $(x_c(t), x_s(t))$ e $(x(t), \widehat{x}(t))$ rappresentano entrambe l'evoluzione dell'inviluppo complesso $\underline{x}(t) = a(t)\,e^{j\varphi(t)}$: mentre i segnali di banda base $x_c(t)$ e $x_s(t)$ sono $\Re$ e $\Im$ di $\underline{x}(t)$, i segnali in banda traslata $x(t)$ e $\widehat{x}(t)$ sono $\Re$ e $\Im$ di $\underline{x}(t)\,e^{j\omega_0 t}$, ovvero di $\underline{x}(t)$ *rotante*, vedi le eq. (11.4) e (11.10).

[20] Ricordiamo che la somma di due numeri complessi coniugati è pari al doppio della loro parte reale.

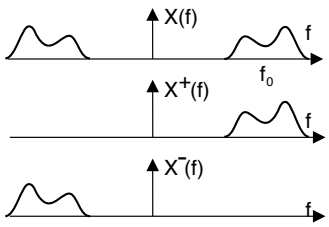
[21] Poniamo qui $\underline{x} = ae^{j\varphi}$

$$X(f) = \tfrac{1}{2}\underline{x}\delta(f - f_0) + \tfrac{1}{2}\underline{x}^*\delta(f + f_0)$$

permettendo di interpretare $\frac{1}{2}\underline{x}e^{j\omega_0 t}$ e $\frac{1}{2}\underline{x}^* e^{-j\omega_0 t}$ nei termini delle componenti a frequenza rispettivamente *positiva* e *negativa* del coseno, così il segnale modulato $x(t) = a(t)\cos(2\pi f_0 t + \varphi(t))$ può considerarsi scomposto nei termini

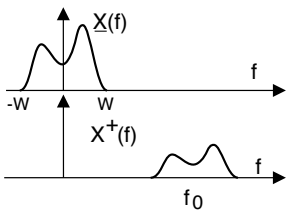
$$x^+(t) = \frac{1}{2}\underline{x}(t)\, e^{j\omega_0 t} \qquad \text{e} \qquad x^-(t) = \frac{1}{2}\underline{x}^*(t)\, e^{-j\omega_0 t} \tag{11.14}$$

dove $x^+(t)$ e $x^-(t)$ individuano rispettivamente le componenti a frequenza *positiva* e *negativa* di $x(t)$, l'una coniugata dell'altra ovvero $x^-(t) = (x^+(t))^*$, in modo da poter scrivere



$$x(t) = x^+(t) + x^-(t) \tag{11.15}$$

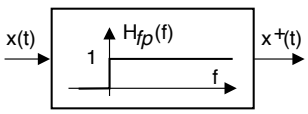
Il segnale *complesso* $x^+(t)$ viene indicato come *segnale analitico*[22] ed in base alla prima delle (11.14) è privo di componenti a frequenza negativa a patto che $\underline{x}(t)$ sia di banda base e con frequenza massima $|W| < f_0$, vedi la figura a lato. In questa ipotesi la sua trasformata vale quindi



$$X^+(f) = \mathcal{F}\left\{x^+(t)\right\} = \frac{1}{2}\underline{X}(f - f_0)$$

pari a zero al di fuori della semiretta $f > 0$.

Incidentalmente, $x^+(t)$ può anche essere pensato come il risultato dell'attraversamento da parte di $x(t)$ di un filtro *ideale* $H_{fp}(f)$[23] con risposta in frequenza a gradino unitario



$$X^+(f) = X(f)\, H_{fp}(f) \tag{11.16}$$

Relazione tra segnale analitico, modulato, e sua trasformata di Hilbert Similmente a $x(t)$ e $\widehat{x}(t)$, anche il segnale analitico $x^+(t)$ è di tipo *passa banda* (benché privo di componenti a frequenza negativa), e si può mostrare[24] che la sua espressione

[22] Vedi ad es. `https://it.wikipedia.org/wiki/Segnale_analitico`

[23] Il pedice $_{fp}$ sta per *frequenze positive*.

[24] L'eguaglianza (11.17) si può dimostrare sia nel dominio del tempo che in quello della frequenza. Partendo dalla prima delle (11.14) si ottiene infatti

$$\begin{aligned} x^+(t) &= \tfrac{1}{2}\underline{x}(t)\, e^{j\omega_0 t} = \tfrac{1}{2}(x_c(t) + jx_s(t))(\cos\omega_0 t + j\sin\omega_0 t) = \\ &= \tfrac{1}{2}\left[(x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t)\right] + j(x_c(t)\sin\omega_0 t + x_s(t)\cos\omega_0 t)] \\ &= \tfrac{1}{2}(x(t) + j\widehat{x}(t)) \end{aligned}$$

Nel dominio della frequenza si applica invece la definizione di filtro di Hilbert (in cui lo sfasamento di $\pm\frac{\pi}{2}$ equivale al prodotto di $X(f)$ per $e^{\pm j\frac{\pi}{2}} = \pm j$) alla trasformata di (11.17), ottenendo

$$X^+(f) = \frac{1}{2}\left(X(f) + j\widehat{X}(f)\right) = \begin{cases} \frac{1}{2}\{X(f) + j[-jX(f)]\} = X(f) & con\ f > 0 \\ \frac{1}{2}\{X(f) + j[jX(f)]\} = 0 & con\ f < 0 \end{cases}$$

nei termini di $x(t)$ e $\widehat{x}(t)$ risulta pari a

$$x^{+}(t) = \frac{1}{2}\left(x(t) + j\widehat{x}(t)\right) \tag{11.17}$$

di cui alla nota[25] si mostra l'equivalenza con (11.16). Infine, con simili passaggi, si ottiene anche

$$x^{-}(t) = \frac{1}{2}\left(x(t) - j\widehat{x}(t)\right) \tag{11.18}$$

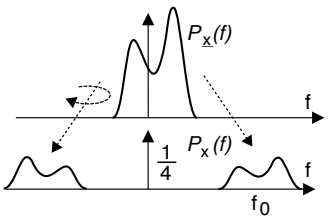
11.2.7 Densità spettrale di segnali passa-banda

Siamo ora in grado di stabilire il legame tra lo spettro dell'inviluppo complesso e quello del segnale modulato. Dalle (11.14) e (11.15) ri-otteniamo la (11.13) ovvero

$$x(t) = x^{+}(t) + x^{-}(t) = \frac{1}{2}\left(\underline{x}(t)\,\mathrm{e}^{j\omega_0 t} + \underline{x}^{*}(t)\,\mathrm{e}^{-j\omega_0 t}\right)$$

la cui trasformata di Fourier, tenendo conto della proprietà di traslazione in frequenza, e che $\mathcal{F}\left\{\underline{x}^{*}(t)\right\} = \underline{X}^{*}(-f)$ (pag. 68), fornisce

$$X(f) = \frac{1}{2}\left(\underline{X}(f - f_0) + \underline{X}^{*}(-f - f_0)\right) \tag{11.19}$$



a cui corrisponde una densità di energia[26]

$$\mathcal{E}_x(f) = \frac{1}{4}\left(\mathcal{E}_{\underline{x}}(f - f_0) + \mathcal{E}_{\underline{x}}(-f - f_0)\right)$$

ovvero una densità di potenza[27]

$$\mathcal{P}_x(f) = \frac{1}{4}\left(\mathcal{P}_{\underline{x}}(f - f_0) + \mathcal{P}_{\underline{x}}(-f - f_0)\right) \tag{11.20}$$

il cui significato è esemplificato alla figura precedente, che raffigura la $\mathcal{P}_{\underline{x}}(f)$ traslare di $\pm f_0$, con una copia *ruotata* per le frequenze negative.

Restringendo ora l'attenzione sul legame tra lo spettro del segnale analitico $x^{+}(t)$ e quello di $\underline{x}(t)$, osserviamo che invertendo la prima delle (11.14) in $\underline{x}(t) = 2x^{+}(t)\,\mathrm{e}^{-j\omega_0 t}$ ed eseguendone la trasformata di Fourier si ottiene

$$\underline{X}(f) = 2X^{+}(f + f_0) \tag{11.21}$$

Osserviamo dunque che in linea di principio $\underline{X}(f)$ *non gode* di simmetria rispetto ad $f = 0$, come peraltro prevedibile visto che $\underline{x}(t)$ è in generale complesso. Per

dato che 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 $f < 0$.

[25] Infatti $H_{fp}(f)$ può essere scritta come $H_{fp}(f) = \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(f) = \frac{1}{2}\left(1 + jH_{\mathcal{H}}(f)\right)$ (vedi eq. (11.25)) , e dunque $H_{fp}(f)\,X(f) = \frac{1}{2}\left(X(f) + j\hat{X}(f)\right)$, da cui la (11.17).

[26] Scriviamo infatti $\mathcal{E}_x(f) = |X(f)|^2 = X(f)\,X^{*}(f)$ da cui otteniamo

$$\begin{aligned}\mathcal{E}_x(f) &= {}^1\!/_4\left(\underline{X}(f - f_0) + \underline{X}^{*}(-f - f_0)\right)\left(\underline{X}^{*}(f - f_0) + \underline{X}(-f - f_0)\right) = \\ &= {}^1\!/_4\left(\underline{X}(f - f_0)\,\underline{X}^{*}(f - f_0) + \underline{X}^{*}(-f - f_0)\,\underline{X}(-f - f_0)\right) = {}^1\!/_4\left(\mathcal{E}_{\underline{x}}(f - f_0) + \mathcal{E}_{\underline{x}}(-f - f_0)\right)\end{aligned}$$

in quanto i prodotti $\underline{X}(f - f_0) \cdot \underline{X}(-f - f_0)$ e $\underline{X}^{*}(-f - f_0) \cdot \underline{X}^{*}(f - f_0)$ sono nulli, dato che in entrambi i casi i fattori risiedono in regioni di frequenza disgiunte.

[27] La (11.20) può essere motivata seguendo le stesse linee guida indicate alla nota 17 a pag. 197.

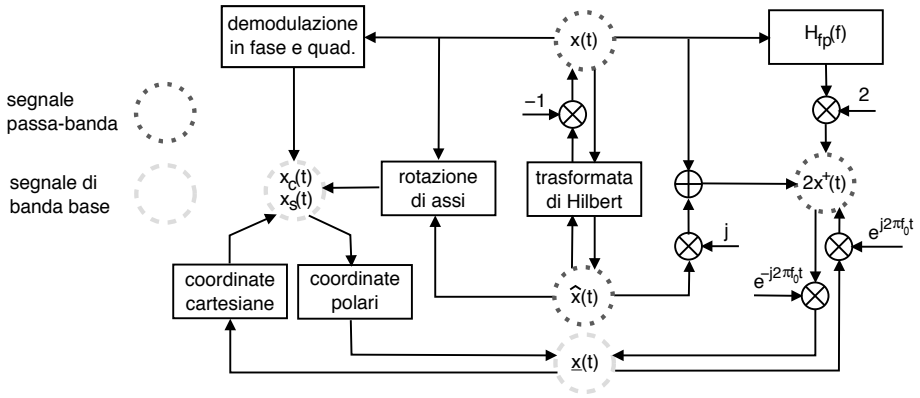


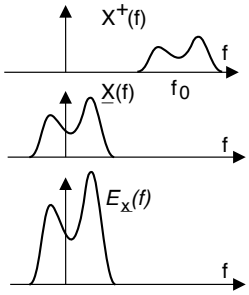
Figura 11.1: Relazioni tra segnale modulato, inviluppo complesso, componenti analogiche e segnale analitico

completare *il giro*, dalla relazione $\mathcal{E}_{\underline{x}}(f) = \left|\underline{X}(f)\right|^2$ otteniamo

$$\mathcal{E}_{\underline{x}}(f) = 4\left|X^+(f+f_0)\right|^2 = 4\mathcal{E}_{x^+}(f+f_0)$$

ed un risultato del tutto simile sussiste anche per segnali di potenza, ovvero

$$\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{x^+}(f+f_0) \qquad (11.22)$$



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.

11.2.8 Schema delle trasformazioni

La figura 11.1 riassume le relazioni esistenti tra le grandezze $x(t)$, $\hat{x}(t)$, e $x^+(t)$, di tipo passa banda, ed $\underline{x}(t)$, $x_c(t)$ e $x_s(t)$, di banda base.

Esempio

- Sia dato il segnale $x(t)$ la cui trasformata $X(f)$ è riportata nella parte superiore di fig. 11.2-a). Derivare l'espressione delle sue componenti analogiche di bassa frequenza, espresse nel dominio della frequenza e del tempo.

  Notiamo che $|X^+(f)| = \frac{k}{2}\text{rect}_{2B}(f-f_0)$, e dunque

  $$\left|\underline{X}(f)\right| = 2\left|X^+(f+f_0)\right| = k\text{rect}_{2B}(f)$$

  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, risulta[28] $\begin{cases} X_c(f) = k\text{rect}_{2B}(f)\, e^{-j2\pi\frac{A}{2\pi B}f} \\ X_s(f) = 0 \end{cases}$, ed effettuando

[28] Approfittiamo dell'occasione per notare che, pur potendo scrivere $\underline{X}(f) = X_c(f) + jX_s(f)$, non

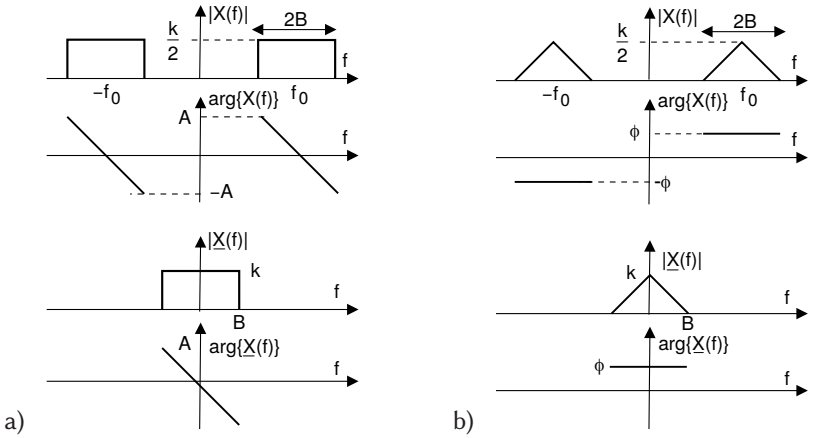


Figura 11.2: Densità spettrali utilizzate negli esempi

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 nella parte superiore di Fig. 11.2-b).

  Eseguendo di nuovo le operazioni di traslazione si ottiene l'inviluppo 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: $\underline{x}(t) = kB\left(\frac{\sin \pi Bt}{\pi Bt}\right)^2 e^{j\phi}$ e dunque

$$\begin{cases} x_c(t) = kB\left(\frac{\sin \pi Bt}{\pi Bt}\right)^2 \cos\phi \\ x_s(t) = kB\left(\frac{\sin \pi Bt}{\pi Bt}\right)^2 \sin\phi \end{cases} \Rightarrow \begin{cases} X_c(f) = k\left(1 - \frac{|f|}{B}\right)\cos\phi \\ X_s(f) = k\left(1 - \frac{|f|}{B}\right)\sin\phi \end{cases}$$

  con $|f| < B$.

11.3 Densità spettrale delle c. analogiche di processi

Quello che ancora manca prima di passare al capitolo successivo è valutare $\mathcal{P}_{x_c}(f)$ e $\mathcal{P}_{x_s}(f)$ nei termini della densità di potenza del processo modulato $\mathcal{P}_x(f)$, estendendo inoltre la trattazione al caso dei processi ergodici. Occorre quindi procedere seguendo le indicazioni del teorema di Wiener, e trasformare le relative funzioni di autocorrelazione $\mathcal{R}_{x_c}(\tau)$ e $\mathcal{R}_{x_s}(\tau)$; una buona dose di calcoli in merito sono svolti al § 11.4.4, arrivando al risultato

$$\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x(\tau)\cos\omega_0\tau + \widehat{\mathcal{R}}_x(\tau)\sin\omega_0\tau \tag{11.23}$$

è assolutamente lecito dire che $\Re\left\{\underline{X}(f)\right\} = X_c(f)$ e $\Im\left\{\underline{X}(f)\right\} = X_s(f)$; 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).

Applicando ora alla (11.23) la formula di Eulero per seno e coseno si ottiene

$$\begin{aligned}\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) &= \mathcal{R}_x(\tau)\frac{e^{j\omega_0\tau}+e^{-j\omega_0\tau}}{2} + \widehat{\mathcal{R}}_x(\tau)\frac{e^{j\omega_0\tau}-e^{-j\omega_0\tau}}{2j} = \\ &= \frac{1}{2}\left[\mathcal{R}_x(\tau) - j\widehat{\mathcal{R}}_x(\tau)\right]e^{j\omega_0\tau} + \frac{1}{2}\left[\mathcal{R}_x(\tau) + j\widehat{\mathcal{R}}_x(\tau)\right]e^{-j\omega_0\tau} = \\ &= \mathcal{R}_x^-(\tau)\,e^{j\omega_0\tau} + \mathcal{R}_x^+(\tau)\,e^{-j\omega_0\tau}\end{aligned}$$

in cui all'ultimo passaggio si è applicata anche a $\mathcal{R}_x(\tau)$ la definizione di segnale analitico eq. (11.17) e (11.18). Non resta quindi che eseguire la trasformata di Fourier, per ottenere

$$\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f) = \mathcal{P}_x^-(f-f_0) + \mathcal{P}_x^+(f+f_0) \tag{11.24}$$

da cui ricaviamo (vedi fig. 11.3) che lo spettro di potenza delle componenti analogiche di un processo si ottiene traslando nell'origine e sovrapponendo le componenti a frequenze positive e negative dello spettro di densità di potenza $\mathcal{P}_x(f)$ del segnale modulato. Come possiamo osservare $\mathcal{P}_{x_c}(f)$ e $\mathcal{P}_{x_s}(f)$ sono entrambe *pari*, in accordo al fatto che $x_c(t)$ ed $x_s(t)$ sono *reali*.

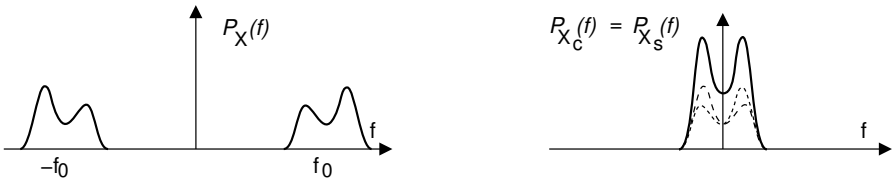


Figura 11.3: Segnale modulato e densità di potenza delle componenti analogiche di b.f.

Rumore bianco passa banda Il risultato mostrato merita un ultimo approfondimento per esaminare il caso in cui il processo $x(t)$ sia di tipo gaussiano, a media nulla, bianco e limitato in banda, ovvero con densità di potenza

$$\mathcal{P}_x(f) = \frac{N_0}{2}\left[rect_{2W}(f-f_0) + rect_{2W}(f+f_0)\right]$$

In tal caso (sempre al § 11.4.4) si trova[29] che $\mathcal{R}_{x_c x_s}(\tau) = 0$ e quindi $x_c(t)$ ed $x_s(t)$ sono incorrelate e, in quanto gaussiane, statisticamente indipendenti. L'applicazione della (11.24) porta dunque a

$$\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f) = \mathcal{P}_x^+(f+f_0) + \mathcal{P}_x^-(f-f_0) = N_0 rect_{2W}(f)$$

e quindi la potenza (e varianza) di entrambe le c.a. di b.f. è pari a quella del segnale modulato, ovvero

$$\mathcal{P}_{x_c} = \int \mathcal{P}_{x_c}(f)\,df = \mathcal{P}_{x_s} = 2N_0W = \mathcal{P}_x$$

come rappresentato in fig. 11.4, mentre l'indipendenza statistica tra le c.a. di b.f. comporta che l'inviluppo complesso $\underline{x}(t) = x_c(t) + jx_s(t)$ ha potenza (e densità di potenza) doppie, ovvero

$$\mathcal{P}_{\underline{x}}(f) = 2N_0 rect_{2W}(f)\,; \qquad \mathcal{P}_{\underline{x}} = 2\mathcal{P}_{x_c} = 2\mathcal{P}_{x_s} = 4N_0W$$

[29] In realtà si ottiene $\mathcal{R}_{x_c x_s}(\tau) = 0$ ogni volta che $\mathcal{P}_x(f)$ ha simmetria *pari* rispetto ad f_0.

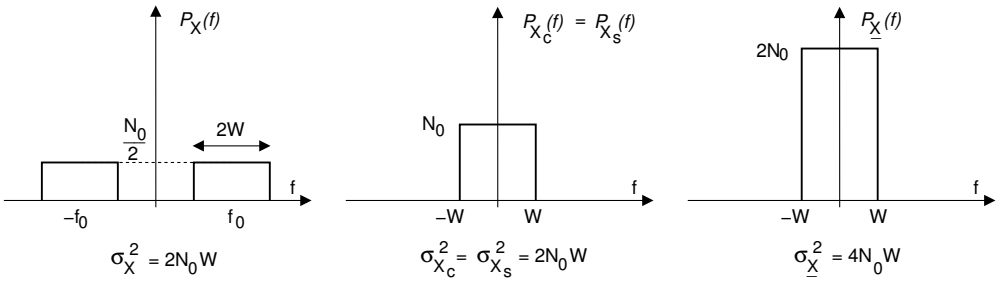


Figura 11.4: Densità di potenza dell'inviluppo complesso per un rumore passa-banda

11.4 Appendici

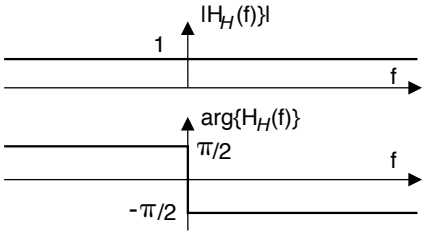
11.4.1 Filtro di Hilbert

Il *filtro di Hilbert* è caratterizzato da una risposta in frequenza descritta analiticamente come

$$H_{\mathcal{H}}(f) = -j \cdot \operatorname{sgn}(f) \tag{11.25}$$

ed il cui andamento di modulo e fase può essere rappresentato come nella figura a a lato. Ricordando infatti che $\pm j = e^{\pm j\frac{\pi}{2}}$ e che $\operatorname{sgn}(f) = \begin{cases} 1 & f > 0 \\ -1 & f < 0 \end{cases}$, otteniamo un andamento *costante* del modulo $|H_{\mathcal{H}}(f)| = 1$, ed un gradino discendente per la fase, ossia

$$\angle H_{\mathcal{H}}(f) = \begin{cases} -\frac{\pi}{2} & f > 0 \\ \frac{\pi}{2} & f < 0 \end{cases}$$



Filtro di Hilbert

Il passaggio di un segnale $x(t)$ attraverso il filtro di Hilbert produce un secondo segnale $\widehat{x}(t)$ detto *trasformata di Hilbert* del primo, indicata come $\widehat{x}(t) = \mathcal{H}\{x(t)\}$, ed il cui andamento in frequenza ha espressione

$$\widehat{X}(f) = \mathcal{F}\{\widehat{x}(t)\} = H_{\mathcal{H}}(f)\, X(f) = -j \cdot \operatorname{sgn}(f) \cdot X(f)$$

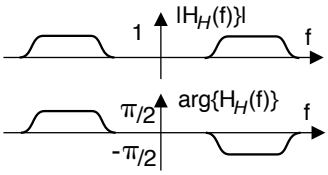
ossia differisce da $X(f)$ per uno sfasamento di $\mp\frac{\pi}{2}$ per frequenze rispettivamente positive o negative. Per la trasformata di Hilbert sussistono le proprietà riportate in nota[30].

[30] Per un approfondimento, vedi ad es. `https://en.wikipedia.org/wiki/Hilbert_transform`, di cui accenniamo brevemente solamente alcuni risultati::

- $\mathcal{H}\{x(t) = x_0\} = 0$: 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 $\widehat{x}(t)$;
- $\mathcal{H}\{\mathcal{H}\{x(t)\}\} = \widehat{\widehat{x}}(t) = -x(t)$: infatti una rotazione di fase pari a π radianti per tutte le frequenze è equivalente ad una inversione di segno;
- $\int_{-\infty}^{\infty} x(t)\,\widehat{x}(t)\,dt = 0$: ortogonalità tra un segnale e la sua trasformata di Hilbert;
- $\mathcal{H}\{x(t) * h(t)\} = \widehat{x}(t) * h(t) = x(t) * \widehat{h}(t)$: la trasformata di Hilbert di una convoluzione (cioè dell'uscita di un filtro) è la convoluzione tra un operando trasformato e l'altro no.

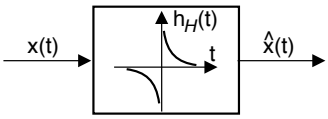
Realizzazione del filtro di Hilbert Sintetizzare un filtro che consegua esattamente la risposta in frequenza descritta dalla (11.25) è un compito pressoché impossibile, a causa della brusca transizione della fase in corrispondenza di $f = 0$.

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.



Risposta impulsiva del filtro di Hilbert Vogliamo dimostrare ora che l'antitrasformata della (11.25) risulta pari a

$$h_{\mathcal{H}}(t) = \mathcal{F}^{-1}\{H_{\mathcal{H}}(f)\} = \frac{1}{\pi t} \tag{11.26}$$



in modo da poter esprimere la trasformata di Hilbert nella forma di un integrale di convoluzione $\widehat{x}(t) = \mathcal{H}\{x(t)\} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau = x(t) * \frac{1}{\pi t}$. Riutilizzando infatti un risultato trovato al § 3.8.6 siamo già arrivati a mostrare che

$$\mathcal{F}^{-1}\left\{-\frac{j}{2\pi f}\right\} = \frac{1}{2}\mathrm{sgn}(t) \tag{11.27}$$

e dato che $H_{\mathcal{H}}(f) = -j \cdot \mathrm{sgn}(f)$, sembra che ci dovrebbe essere un modo semplice di arrivare a *sistemare le cose*. La (11.27), una volta eliminato il termine $1/2$, permette di scrivere $\mathcal{F}\{\mathrm{sgn}(t)\} = -\frac{j}{\pi f}$ e dunque

$$\mathcal{F}\{-j \cdot \mathrm{sgn}(t)\} = -\frac{1}{\pi f} \tag{11.28}$$

Applicando ora alla (11.28) la proprietà di dualità (vedi pag. 65, dove si asserisce che se $G(f) = \mathcal{F}\{g(t)\}$, allora $\mathcal{F}\{G(t)\} = g(-f)$) otteniamo $\mathcal{F}\left\{-\frac{1}{\pi t}\right\} = -j \cdot \mathrm{sgn}(-f) = j \cdot \mathrm{sgn}(f)$, e dunque arriviamo al risultato anticipato

$$\mathcal{F}^{-1}\{H_{\mathcal{H}}(f)\} = \mathcal{F}^{-1}\{-j \cdot \mathrm{sgn}(f)\} = \frac{1}{\pi t}$$

11.4.2 Trasformata di Hilbert di un segnale modulato

Si intende dimostrare che se $x_c(t)$ ed $x_s(t)$ sono limitate in banda $\pm W$ con $W < f_0$, allora risulta

$$\begin{cases} \mathcal{H}\{x_c(t)\cos\omega_0 t\} = x_c(t)\sin\omega_0 t \\ \mathcal{H}\{x_s(t)\sin\omega_0 t\} = -x_s(t)\cos\omega_0 t \end{cases} \tag{11.29}$$

e dunque dalla (11.1) si ottiene

$$\hat{x}(t) = \mathcal{H}\{x_c(t)\cos 2\pi f_0 t - x_s(t)\sin 2\pi f_0 t\} = x_c(t)\sin\omega_0 t + x_s(t)\cos\omega_0 t$$

e quindi $\hat{x}(t) = \Im\left\{\underline{x}(t)\, e^{j\omega_0 t}\right\}$ come espresso dall'eq. (11.10). Limitiamoci a dimostrare la prima delle (11.29), ovvero che

$$\mathcal{H}\{x_c(t)\cos 2\pi f_0 t\} = x_c(t)\sin 2\pi f_0 t \tag{11.30}$$

Iniziamo con il considerare che $\mathcal{F}$-trasformando l'argomento di (11.30) possiamo evidenziarne le componenti a frequenza positiva e negativa $X_c(f-f_0)$ e $X_c(f+f_0)$

$$x_c(t)\cos 2\pi f_0 t = \frac{x_c(t)}{2}\left(e^{j2\pi f_0 t}+e^{-j2\pi f_0 t}\right) \overset{\mathcal{F}}{\Rightarrow} \frac{1}{2}\left[X_c(f-f_0)+X_c(f+f_0)\right] \qquad (11.31)$$

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

$$\frac{1}{2}\left[X_c(f-f_0)+X_c(f+f_0)\right] \overset{\mathcal{H}}{\Rightarrow} \frac{1}{2}\left[X_c(f-f_0)\,e^{-j\frac{\pi}{2}}+X_c(f+f_0)\,e^{j\frac{\pi}{2}}\right]$$

e quindi $\mathcal{F}$-antitrasformando questa espressione si ottiene la $\mathcal{H}$-trasformata del segnale (11.30)

$$\frac{1}{2}\left[X_c(f-f_0)\,e^{-j\frac{\pi}{2}}+X_c(f+f_0)\,e^{j\frac{\pi}{2}}\right] \overset{\mathcal{F}^{-1}}{\Rightarrow} \frac{x_c(t)}{2}\left(e^{j2\pi f_0 t}e^{-j\frac{\pi}{2}}+e^{-j2\pi f_0 t}e^{j\frac{\pi}{2}}\right)$$

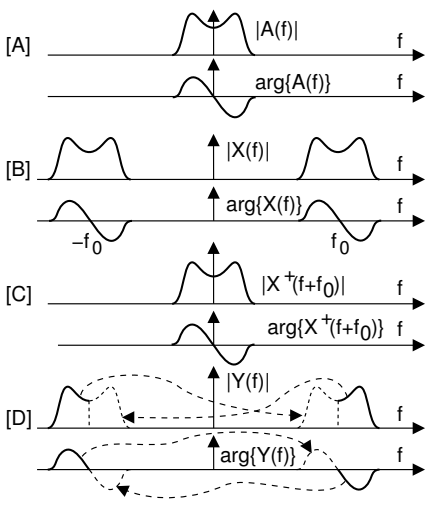
risultato che, anche se non ancora nella forma anticipata, poteva comunque essere ottenuto anche direttamente a partire dal secondo membro di (11.31), invocando subito la limitazione ad una semibanda di $x_c(t)\,e^{\pm j2\pi f_0 t}$. Per ottenere la (11.30) è ora sufficiente moltiplicare e dividere per $j=e^{j\frac{\pi}{2}}$, 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 $e^{j\frac{\pi}{2}}=j$, e $e^{j\pi}=-1$.

11.4.3 Trasmissione a banda laterale unica

Con riferimento alla figura che segue, consideriamo un segnale $a(t)$ reale e limitato in banda, con $A(f)=A^*(-f)$ (grafico [A]). In virtù delle proprietà di simmetria coniugata per segnali reali la conoscenza del solo contenuto a frequenze positive $f>0$, ovvero di $A^+(f)=A(f)\,\mathrm{rect}_W\left(f-\frac{W}{2}\right)$, è sufficiente a definire $a(t)$ in modo completo. Se consideriamo ora 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: $X^+(f_0+\alpha)=\{X^+(f_0-\alpha)\}^*$.



Questo risultato mostra come sia *teoricamente* possibile (con una fotocopiatrice ed un paio di forbici!) 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 (frecce tratteggiate) spostando le copie duplicate di $Y^+(f)$ e $Y^-(f)$ come indicato dalle frecce.

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, e si trova al § 12.1.2.

11.4.4 Processo passa banda

Svolgiamo ora l'approfondimento dei passaggi che portano ai risultati discussi al § 11.3, ovvero l'espressione della densità di potenza $\mathcal{P}_{x_c}(f)$ e $\mathcal{P}_{x_s}(f)$ in funzione di $\mathcal{P}_x(f)$ eq. (11.24), svolgendo i calcoli in modo da tenere in conto anche il caso dei processi aleatori: pertanto, occorrerà prima ottenere un risultato relativo alle rispettive funzioni di autocorrelazione $\mathcal{R}_{x_c}(\tau)$ e $\mathcal{R}_{x_s}(\tau)$, e quindi effettuare la trasformata di Fourier come prescritto dal teorema di Wiener.

Osserviamo innanzitutto che quando un processo aleatorio presenta una $\mathcal{P}_x(f)$ limitata in banda attorno ad f_0, la relativa funzione di autocorrelazione $\mathcal{R}_x(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_x(f)\}$ può essere espressa nei termini delle componenti analogiche di bassa frequenza della funzione di autocorrelazione stessa:

$$\mathcal{R}_x(\tau) = \mathcal{R}_c(\tau)\cos\omega_0\tau - \mathcal{R}_s(\tau)\sin\omega_0\tau$$

D'altra parte, una qualunque realizzazione di un processo $x(t)$ limitato in banda attorno ad f_0 ammette la rappresentazione $x(t) = x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t$, ma data la natura aleatoria di $x(t)$, gli stessi $x_c(t)$ ed $x_s(t)$ sono realizzazioni di processi, in generale statisticamente *dipendenti*, 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 un processo stazionario ed ergodico: come già osservato al § 7.5.3, $x(t)$ è solamente *ciclostazionario*[31].

Come prima cosa, proviamo a calcolare la funzione di autocorrelazione dell'inviluppo complesso di una generica realizzazione, che per l'ergodicità corrisponde al relativo momento misto:

$$\begin{aligned}
\mathcal{R}_{\underline{x}}(\tau) &= E\left\{\underline{x}^*(\tau)\,\underline{x}(t+\tau)\right\} = \\
&= E\left\{\left[x_c(\tau) - jx_s(\tau)\right]\left[x_c(t+\tau) + jx_s(t+\tau)\right]\right\} = \\
&= E\left\{x_c(\tau)\,x_c(t+\tau) + x_s(\tau)\,x_s(t+\tau) + \right. \\
&\quad \left. + j\left[x_c(\tau)\,x_s(t+\tau) - x_s(\tau)\,x_c(t+\tau)\right]\right\} = \\
&= \mathcal{R}_{x_c}(\tau) + \mathcal{R}_{x_s}(\tau) + j\left[\mathcal{R}_{x_c x_s}(\tau) - \mathcal{R}_{x_s x_c}(\tau)\right]
\end{aligned} \tag{11.32}$$

Queste quattro quantità sono calcolate al § 11.4.5, e nel caso in cui $x_c(t)$ e $x_s(t)$ siano

[31] Come illustrato al § 6.3.7, il processo risultante diviene ergodico qualora al coseno sia aggiunta una fase aleatoria uniformemente distribuita.

stazionari ed ergodici, il risultato finale fornisce le espressioni

$$\begin{cases} \mathcal{R}_{x_c}(\tau) & = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x(\tau)\cos\omega_0\tau + \widehat{\mathcal{R}}_x(\tau)\sin\omega_0\tau \\ \mathcal{R}_{x_c x_s}(\tau) & = -\mathcal{R}_{x_s x_c}(\tau) = \widehat{\mathcal{R}}_x(\tau)\cos\omega_0\tau - \mathcal{R}_x(\tau)\sin\omega_0\tau \end{cases} \tag{11.33}$$

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, sostituendo (11.33) in (11.32), risulti

$$\mathcal{R}_{\underline{x}}(\tau) = 2\left[\mathcal{R}_{x_c}(\tau) + j\mathcal{R}_{x_c x_s}(\tau)\right] \tag{11.34}$$

e pertanto otteniamo

$$\mathcal{P}_{\underline{x}}(f) = \mathcal{F}\left\{\mathcal{R}_{\underline{x}}(\tau)\right\} = 2\left[\mathcal{P}_{x_c}(f) + j\mathcal{P}_{x_c x_s}(f)\right] \tag{11.35}$$

in cui $\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f)$ sono reali pari in quanto $x_c(t)$ ed $x_s(t)$ sono reali. Prima di giungere alle conclusioni espresse al § 11.4.4.1 ed anticipate al § 11.3, prendiamoci *il lusso* di sviluppare una serie di considerazioni basate sui risultati fin qui ottenuti:

1. la (11.35) sembra indicare 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 dispari[32]: pertanto $\mathcal{P}_{x_c x_s}(f) = \mathcal{F}\left\{\mathcal{R}_{x_c x_s}(\tau)\right\}$ è completamente immaginario, e dunque $\mathcal{P}_{\underline{x}}(f)$ è reale;

2. se risulta $\mathcal{R}_{x_c x_s}(\tau) = 0$ per ogni τ, allora la potenza mutua $\mathcal{P}_{x_c x_s}(f)$ si annulla, e la (11.35) fornisce $\mathcal{P}_{\underline{x}}(f) = 2\mathcal{P}_{x_c}(f)$ reale pari; la presenza di $\mathcal{P}_{x_c x_s}(f) \neq 0$ può invece rendere $\mathcal{P}_{\underline{x}}(f)$ asimmetrico, permettendo di ottenere ancora $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_x^+(f + f_0)$ come espresso dalla (11.22);

3. invertendo i due punti precedenti osserviamo che, se $\mathcal{P}_x(f)$ ha simmetria pari rispetto ad f_0, allora $\mathcal{P}_{\underline{x}}(f)$ è pari, e quindi deve risultare $\mathcal{R}_{x_c x_s}(\tau) = 0$, ovvero le c.a. di b.f. $x_c(t)$ ed $x_s(t)$ risultano mutuamente incorrelate; **se** inoltre queste sono congiuntamente gaussiane, *allora risultano anche statisticamente indipendenti;*

4. se $x_c(t)$ e $x_s(t)$ sono a media nulla, la potenza $\mathcal{P}_{x_c}$ (uguale a $\mathcal{P}_{x_s}$ in virtù della prima delle (11.33)) si calcola come $\mathcal{R}_{x_c}(0) = \mathcal{R}_{x_s}(0)$. Dato che $\mathcal{R}_{x_c x_s}(\tau) = -\mathcal{R}_{x_s x_c}(\tau)$ è dispari (punto 1), deve risultare che $\mathcal{R}_{x_c x_s}(0) = 0$; in questo caso la (11.34) fornisce $\mathcal{R}_{\underline{x}}(0) = 2\mathcal{R}_{x_c}(0) = 2\mathcal{R}_{x_s}(0)$, e dunque si ottiene

   $$\mathcal{P}_{\underline{x}} = \sigma^2_{\underline{x}} = \mathcal{R}_{\underline{x}}(0) = 2\mathcal{R}_{x_c}(0) = 2\mathcal{R}_{x_s}(0) = 2\sigma^2_{x_c} = 2\sigma^2_{x_s} = 2\mathcal{P}_{x_c} = 2\mathcal{P}_{x_s}$$

   In definitiva, le componenti analogiche di bassa frequenza hanno entrambe potenza pari a *metà* di quella dell'inviluppo complesso;

5. l'eq. (11.22) asserisce che $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{x^+}(f + f_0)$, ed in modo simile si può trovare che $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{x^-}(f - f_0)$, e quindi $\mathcal{P}_{x^+} = \mathcal{P}_{x^-} = \frac{1}{4}\mathcal{P}_{\underline{x}}$. Dato poi che $\mathcal{P}_x = \mathcal{P}_{x^+} + \mathcal{P}_{x^-}$ in quanto $x^+(t)$ e $x^-(t)$ sono ortogonali perché definiti su bande

[32] Infatti (eq. (11.33)) $\mathcal{R}_{x_c x_s}(\tau) = \widehat{\mathcal{R}}_x(\tau)\cos\omega_0\tau - \mathcal{R}_x(\tau)\sin\omega_0\tau$, 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.

disgiunte, in base al punto 4 si ottiene

$$\mathcal{P}_x = \mathcal{P}_{x^+} + \mathcal{P}_{x^-} = \frac{1}{4}\left[\mathcal{P}_{\underline{x}} + \mathcal{P}_{\underline{x}}\right] = \frac{1}{2}\mathcal{P}_{\underline{x}} = \mathcal{P}_{x_c} = \mathcal{P}_{x_s}$$

e dunque $x_c(t)$ e $x_s(t)$ hanno (ciascuno) potenza pari a quella di $x(t)$, ovvero $\mathcal{P}_{x_c} = \mathcal{P}_{x_s} = \mathcal{P}_x$;

6. se consideriamo $\widehat{x}(t)$ l'uscita del filtro di Hilbert per il quale risulta $|H_H(f)|^2 = 1$, si ottiene che $\mathcal{P}_{\hat{x}}(f) = |H_H(f)|^2 \mathcal{P}_x(f) = \mathcal{P}_x(f)$, e dunque antitrasformando $\mathcal{R}_{\widehat{x}}(\tau) = \mathcal{R}_x(\tau)$;
7. è possibile mostrare che, esprimendo 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
$$\begin{cases} \mathcal{R}_c(\tau) = \mathcal{R}_{x_c}(\tau) \\ \mathcal{R}_s(\tau) = -\mathcal{R}_{x_c x_s}(\tau) \end{cases}$$
8. la prima delle (11.33) ci dice che $\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f)$ in quanto $\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau)$, e che risulta $\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x(\tau)\cos\omega_0\tau + \widehat{\mathcal{R}}_x(\tau)\sin\omega_0\tau$; applicando ora la formula di Eulero per seno e coseno si ottiene
$$\begin{aligned} \mathcal{R}_{x_c}(\tau) &= \mathcal{R}_{x_s}(\tau) = \\ &= \mathcal{R}_x(\tau)\tfrac{e^{j\omega_0\tau}+e^{-j\omega_0\tau}}{2} + \widehat{\mathcal{R}}_x(\tau)\tfrac{e^{j\omega_0\tau}-e^{-j\omega_0\tau}}{2j} \\ &= \tfrac{1}{2}\left[\mathcal{R}_x(\tau) - j\widehat{\mathcal{R}}_x(\tau)\right]e^{j\omega_0\tau} + \tfrac{1}{2}\left[\mathcal{R}_x(\tau) + j\widehat{\mathcal{R}}_x(\tau)\right]e^{-j\omega_0\tau} \\ &= \mathcal{R}_x^-(\tau)\,e^{j\omega_0\tau} + \mathcal{R}_x^+(\tau)\,e^{-j\omega_0\tau} \end{aligned}$$
infatti i termini tra parentesi quadre corrispondono alla definizione di componenti a frequenze positive e negative ottenute tramite trasformata di Hilbert, per come espressa dalla (11.17).

11.4.4.1 Conclusioni

Al punto 8) del precedente elenco abbiamo mostrato che

$$\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x^-(\tau)\,e^{j\omega_0\tau} + \mathcal{R}_x^+(\tau)\,e^{-j\omega_0\tau}$$

e quindi

$$\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f) = \mathcal{P}_x^-(f - f_0) + \mathcal{P}_x^+(f + f_0)$$

dunque lo spettro di densità di potenza delle componenti analogiche di un processo si ottiene traslando nell'origine e sovrapponendo (vedi fig. 11.5) le componenti a frequenze positive e negative dello spettro di densità di potenza $\mathcal{P}_x(f)$ del segnale modulato. Dunque $\mathcal{P}_{x_c}(f)$ e $\mathcal{P}_{x_s}(f)$ sono entrambe pari, come deve essere per $x_c(t)$ ed $x_s(t)$ reali.

11.4.4.2 Processo gaussiano bianco limitato in banda

Se $x(t)$ è un processo gaussiano stazionario ergodico, bianco ed a media nulla, con densità spettrale $\mathcal{P}_x(f) = \frac{N_0}{2}$ limitata in banda $\pm W$ attorno ad f_0, allora (vedi fig. 11.6)

- le relative c.a. di b.f. $x_c(t)$ e $x_s(t)$ sono processi congiuntamente gaussiani, stazionari, ergodici, statisticamente indipendenti ed a media nulla, con potenza $\mathcal{P}_{x_c} = \mathcal{P}_{x_s} = \mathcal{P}_x = 2N_0W$, pari alle varianze $\sigma_x^2 = \sigma_{x_c}^2 = \sigma_{x_s}^2$. Le rispettive densità

di potenza valgono

$$\mathcal{P}_{x_c}(f) = \mathcal{P}_{x_s}(f) = \mathcal{P}_x^+(f+f_0) + \mathcal{P}_x^-(f-f_0) = N_0 rect_{2W}(f) \tag{11.36}$$

- il suo inviluppo complesso $\underline{x}(t)$ relativo ad f_0 ha potenza (e densità di potenza) doppie

$$\mathcal{P}_{\underline{x}} = 2\mathcal{P}_{x_c} = 4N_0W; \qquad \mathcal{P}_{\underline{x}}(f) = 2N_0 rect_{2W}(f) \tag{11.37}$$

Infatti la simmetria pari di $\mathcal{P}_x(f)$ attorno ad f_0 rende $x_c(t)$ e $x_s(t)$ incorrelate, come mostrato al punto 3) di pag. 360: pertanto $\mathcal{P}_{x_c x_s}(f) = \mathcal{F}\left\{\mathcal{R}_{x_c x_s}(\tau)\right\} = 0$ (punto 2) e dunque la (11.35) si semplifica nella (11.37).

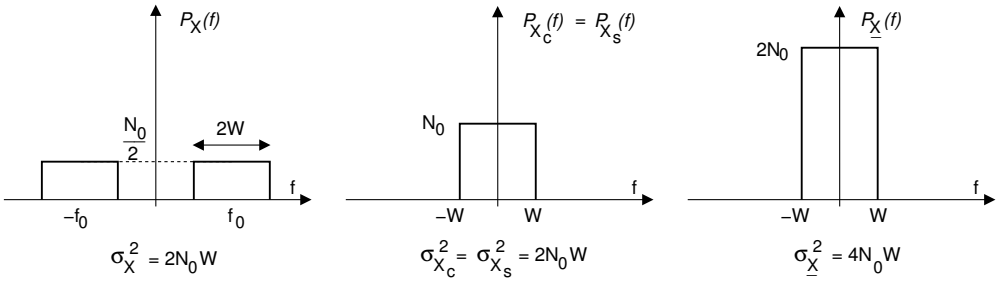


Figura 11.6: Densità di potenza delle c.a. di b.f. per un processo gaussiano bianco e limitato e limitato in banda

11.4.5 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 (pag. 349) $x_c(t) = x(t)\cos\omega_0 t + \widehat{x}(t)\sin\omega_0 t$, iniziamo a svolgere i calcoli per $\mathcal{R}_{x_c}(\tau)$:

$$\begin{aligned}
\mathcal{R}_{x_c}(\tau) &= E\{x_c(\tau)\, x_c(t+\tau)\} = \\
&= E\{[x(t)\cos\omega_0 t + \widehat{x}(t)\sin\omega_0 t]\cdot \\
&\cdot\ [x(t+\tau)\cos\omega_0(t+\tau) + \widehat{x}(t+\tau)\sin\omega_0(t+\tau)]\} = \\
&= E\{x(t)\, x(t+\tau)\}\cdot\cos\omega_0 t\cdot\cos\omega_0(t+\tau) + \\
&+ E\{x(t)\,\widehat{x}(t+\tau)\}\cdot\cos\omega_0 t\cdot\sin\omega_0(t+\tau) + \\
&+ E\{\widehat{x}(t)\, x(t+\tau)\}\cdot\sin\omega_0 t\cdot\cos\omega_0(t+\tau) + \\
&+ E\{\widehat{x}(t)\,\widehat{x}(t+\tau)\}\cdot\sin\omega_0 t\cdot\sin\omega_0(t+\tau)
\end{aligned}$$

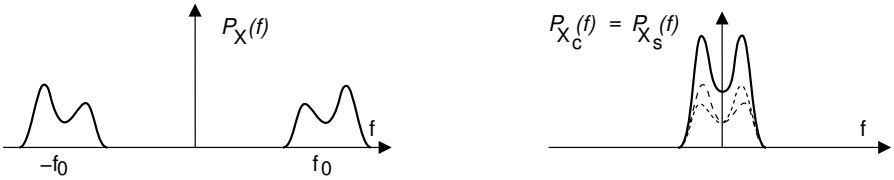


Figura 11.5: Processo passa banda e densità di potenza delle c.a. di b.f.

Valutiamo quindi singolarmente i quattro valori attesi, procedendo con il calcolo di medie temporali in virtù dell'ergodicità, e indicando con $\overline{x\,(t)}$ la media temporale di $x\,(t)$, ossia $\overline{x\,(t)} = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} x\,(t)\,dt$:

$$\begin{aligned}
E\{x\,(t)\,x\,(t+\tau)\} &= \overline{x\,(t)\,x\,(t+\tau)} = \mathcal{R}_x\,(\tau)\\
E\{x\,(t)\,\widehat{x}\,(t+\tau)\} &= \overline{x\,(t)\,\widehat{x}\,(t+\tau)} = \mathcal{R}_{x\widehat{x}}\,(\tau) = x\,(-\tau) * \widehat{x}\,(\tau) =\\
&= x\,(-\tau) * x\,(\tau) * \frac{1}{\pi\tau} = \mathcal{R}_x\,(\tau) * \frac{1}{\pi\tau} = \widehat{\mathcal{R}}_x\,(\tau)\\
E\{\widehat{x}\,(t)\,x\,(t+\tau)\} &= \overline{\widehat{x}\,(t)\,x\,(t+\tau)} = \mathcal{R}_{\widehat{x}x}\,(\tau) = \widehat{x}\,(-\tau) * x\,(\tau) =\\
&= x\,(-\tau) * \left(-\frac{1}{\pi\tau}\right) * x\,(\tau) = x\,(-\tau) * x\,(\tau) * \left(-\frac{1}{\pi\tau}\right) =\\
&= \mathcal{R}_x\,(\tau) * \left(-\frac{1}{\pi\tau}\right) = -\widehat{\mathcal{R}}_x\,(\tau)\\
E\{\widehat{x}\,(t)\,\widehat{x}\,(t+\tau)\} &= \overline{\widehat{x}\,(t)\,\widehat{x}\,(t+\tau)} = \mathcal{R}_{\widehat{x}\widehat{x}}\,(\tau) = \widehat{x}\,(-\tau) * \widehat{x}\,(\tau) =\\
&= x\,(-\tau) * \left(-\frac{1}{\pi\tau}\right) * x\,(\tau) * \frac{1}{\pi\tau} =\\
&= x\,(-\tau) * x\,(\tau) * \left(-\frac{1}{\pi\tau}\right) * \frac{1}{\pi\tau} = -\widehat{\widehat{\mathcal{R}}}_x\,(\tau) = \mathcal{R}_x\,(\tau)
\end{aligned}$$

Sostituendo le relazioni ora trovate nella espressione di $\mathcal{R}_{x_c}\,(\tau)$, si ottiene

$$\begin{aligned}
\mathcal{R}_{x_c}\,(\tau) &= \mathcal{R}_x\,(\tau)\cdot\cos\omega_0 t\cdot\cos\omega_0\,(t+\tau) + \widehat{\mathcal{R}}_x\,(\tau)\cdot\cos\omega_0 t\cdot\sin\omega_0\,(t+\tau) +\\
&- \widehat{\mathcal{R}}_x\,(\tau)\cdot\sin\omega_0 t\cdot\cos\omega_0\,(t+\tau) + \mathcal{R}_x\,(\tau)\cdot\sin\omega_0 t\cdot\sin\omega_0\,(t+\tau) =\\
&= \tfrac{1}{2}\mathcal{R}_x\,(\tau)\,[\cos\omega_0\,(-\tau) + \cos\omega_0\,(2t+\tau)] +\\
&+ \tfrac{1}{2}\widehat{\mathcal{R}}_x\,(\tau)\,[\sin\omega_0\,(\tau) + \sin\omega_0\,(2t+\tau)] +\\
&- \tfrac{1}{2}\widehat{\mathcal{R}}_x\,(\tau)\,[\sin\omega_0\,(-\tau) + \sin\omega_0\,(2t+\tau)] +\\
&+ \tfrac{1}{2}\mathcal{R}_x\,(\tau)\,[\cos\omega_0\,(-\tau) - \cos\omega_0\,(2t+\tau)] =\\
&= \mathcal{R}_x\,(\tau)\cdot\cos\omega_0\tau + \widehat{\mathcal{R}}_x\,(\tau)\cdot\sin\omega_0\tau
\end{aligned}$$

che costituisce il risultato anticipato alla (11.33). Per l'espansione dei termini trigonometrici, si è fatto uso delle relazioni

$$\begin{aligned}
\cos\alpha\cdot\cos\beta &= \tfrac{1}{2}[\cos(\alpha-\beta) + \cos(\alpha+\beta)]\\
\sin\alpha\cdot\sin\beta &= \tfrac{1}{2}[\cos(\alpha-\beta) - \cos(\alpha+\beta)]\\
\sin\alpha\cdot\cos\beta &= \tfrac{1}{2}[\sin(\alpha-\beta) + \sin(\alpha+\beta)]
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{R}_{x_c x_s}(\tau) &= E\{x_c(\tau)\, x_s(t+\tau)\} = \\
&= E\{[x(t)\cos\omega_0 t + \widehat{x}(t)\sin\omega_0 t]\cdot \\
&\cdot\ [\widehat{x}(t+\tau)\cos\omega_0(t+\tau) - x(t+\tau)\sin\omega_0(t+\tau)] = \\
&= E\{x(t)\,\widehat{x}(t+\tau)\}\cdot\cos\omega_0 t\cdot\cos\omega_0(t+\tau) + \\
&- E\{x(t)\,x(t+\tau)\}\cdot\cos\omega_0 t\cdot\sin\omega_0(t+\tau) + \\
&+ E\{\widehat{x}(t)\,\widehat{x}(t+\tau)\}\cdot\sin\omega_0 t\cdot\cos\omega_0(t+\tau) + \\
&- E\{\widehat{x}(t)\,x(t+\tau)\}\cdot\sin\omega_0 t\cdot\sin\omega_0(t+\tau)
\end{aligned}$$

I valori attesi che vediamo comparire sono stati già ottenuti e quindi possiamo scrivere direttamente lo sviluppo dei calcoli, in cui si applicano nuovamente le identità trigonometriche note:

$$\begin{aligned}
\mathcal{R}_{x_c x_s}(\tau) &= \widehat{\mathcal{R}}_x(\tau)\cdot\cos\omega_0 t\cdot\cos\omega_0(t+\tau) - \mathcal{R}_x(\tau)\cdot\cos\omega_0 t\cdot\sin\omega_0(t+\tau) + \\
&+ \mathcal{R}_x(\tau)\cdot\sin\omega_0 t\cdot\cos\omega_0(t+\tau) - \widehat{\mathcal{R}}_x(\tau)\cdot\sin\omega_0 t\cdot\sin\omega_0(t+\tau) = \\
&= \tfrac{1}{2}\widehat{\mathcal{R}}_x(\tau)\,[\cos\omega_0(-\tau) + \cos\omega_0(2t+\tau)] + \\
&- \tfrac{1}{2}\mathcal{R}_x(\tau)\,[\sin\omega_0(\tau) + \sin\omega_0(2t+\tau)] + \\
&- \tfrac{1}{2}\mathcal{R}_x(\tau)\,[\sin\omega_0(-\tau) + \sin\omega_0(2t+\tau)] + \\
&- \tfrac{1}{2}\widehat{\mathcal{R}}_x(\tau)\,[\cos\omega_0(-\tau) - \cos\omega_0(2t+\tau)] = \\
&= -\mathcal{R}_x(\tau)\cdot\sin\omega_0(2t+\tau) + \widehat{\mathcal{R}}_x(\tau)\cdot\cos\omega_0(2t+\tau)
\end{aligned}$$

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 sensatamente essere posto a zero, e dunque ottenere il risultato previsto alla (11.33).

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 12

Modulazione (e ritorno) di segnali analogici

Analizziamo le tecniche adottate per modulare (in ampiezza o angolarmente) una portante sinusoidale con un segnale informativo di natura analogica[1], studiando allo stesso tempo le caratteristiche spettrali del segnale ottenuto. Sono quindi discussi i diversi approcci di demodulazione, i circuiti che li realizzano, e l'influenza di eventuali inaccuratezze, mentre al capitolo successivo sono gli discussi gli approcci puramente *numerici*. Lo studio di come le diverse scelte condizionino le prestazioni del ricevitore nel caso di un canale distorcente ed in presenza di rumore al lato ricevente viene affrontato rispettivamente ai capp. 13 e 14.

12.1 Modulazione di ampiezza - *AM*

Al § 11.2 si è mostrato come un segnale modulato $x(t)$ può essere rappresentato nei termini delle sue *componenti analogiche di bassa frequenza* $x_c(t)$ e $x_s(t)$: quando queste *non sono* scelte in modo indipendente[2], possiamo individuare le seguenti classi di segnali *modulati in ampiezza*:

- *banda laterale doppia*: così chiamata in quanto $\mathcal{P}_x(f)$ è simmetrico rispetto ad f_0, conseguenza dell'essere $x_s(t)$ nulla. Si tratta del caso introdotto al § 3.5.2, ora indicato con gli acronimi BLD o DSB (*double side band*);
- *banda laterale unica* (BLU o SSB - *single side band*): qui sono presenti sia $x_c(t)$ che $x_s(t)$, con il vincolo $x_s(t) = \widehat{x}_c(t)$. Ciò fa sì che (come vedremo) la densità $\mathcal{P}_x(f)$ del segnale modulato giaccia tutta all'*esterno* (od all'interno) di $\pm f_0$;
- *banda laterale ridotta* (BLR o VSB - *vestigial side band*[3]) è una via di mezzo tra i due casi precedenti, in quanto $\mathcal{P}_x(f)$ non è simmetrica rispetto ad f_0, e pur giacendo su entrambi i lati, occupa una banda minore del caso BLD.

[1] Per i segnali numerici si usano tecniche peculiari, esposte al capitolo 16.

[2] Qualora $x_c(t)$ e $x_s(t)$ siano due segnali indipendenti, la forma di modulazione di ampiezza risultante viene detta segnale QAM (*quadrature amplitude modulation*), vedi § 16.3.

[3] Come sarà più chiaro nel seguito, l'acronimo VSB evoca il fatto che, anziché sopprimere completamente una delle due bande laterali, se ne mantengono *delle vestigia*.

Per completare la classificazione, per ognuna delle possibilità illustrate 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 segnale modulante $m(t)$, ma solamente alla portante, e quindi priva di contenuto informativo ai fini della trasmissione. I tre sottocasi citati sono indicati come:

- portante intera (PI o LC - *large carrier*);
- portante soppressa (PS o SC - *suppressed carrier*);
- portante parzialmente soppressa (PPS).

12.1.1 Banda laterale doppia - *BLD*

Come anticipato, questo è il caso in cui l'inviluppo complesso $\underline{x}(t)$ del segnale modulato presenta *una sola* componente analogica di bassa frequenza, che *per convenzione*[4] è posta pari a $x_c(t)$, la cui dipendenza dal segnale modulante $m(t)$ è espressa nella forma generale $x_c(t) = a_p + k_a m(t)$, e quindi

$$x_{BLD}(t) = \left(a_p + k_a m(t)\right) \cos \omega_0 t \tag{12.1}$$

Pertanto l'inviluppo complesso *è reale* e vale $\underline{x}(t) = a_p + k_a m(t)$; considerando poi $m(t)$ a media nulla, la relativa densità di potenza ha valore

$$\mathcal{P}_{\underline{x}}(f) = a_p^2 \delta(f) + k_a^2 \mathcal{P}_m(f)$$

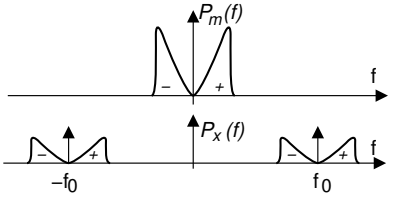
Ricordando ora che $\mathcal{P}_x(f) = \frac{1}{4}\left(\mathcal{P}_{\underline{x}}(f - f_0) + \mathcal{P}_{\underline{x}}(-f - f_0)\right)$ (eq. (11.20)) e che per $\underline{x}(t)$ reale la relativa densità di potenza è pari ovvero $\mathcal{P}_{\underline{x}}(f) = \mathcal{P}_{\underline{x}}(-f)$, la densità di potenza del segnale modulato risulta pari a

$$\mathcal{P}_x(f) = \frac{a_p^2}{4}\left[\delta(f - f_0) + \delta(f + f_0)\right] + \frac{k_a^2}{4}\left[\mathcal{P}_m(f - f_0) + \mathcal{P}_m(f + f_0)\right] \tag{12.2}$$

La potenza *totale* di $x(t)$ vale perciò

$$\mathcal{P}_x = \int \mathcal{P}_x(f)\, df = \frac{a_p^2}{2} + \frac{k_a^2}{2}\mathcal{P}_m \tag{12.3}$$

mentre la corrispondente densità spettrale è raffigurata a lato, dove si è posto $k_a = 1$.



[4] Considerando che la portante del segnale *ricevuto* può avere una fase arbitraria, e che con una traslazione temporale ci si può sempre ricondurre ad usare una funzione $\cos \omega_0 t$, tale convenzione individua il caso più generale di un segnale modulato del tipo $x(t) = a(t)\cos(\omega_o t + \varphi)$ con φ costante. Infatti, introducendo un ritardo $\tau = \frac{\varphi}{2\pi f_o}$ si ottiene $x(t - \tau) = a(t - \tau)\cos(2\pi f_0 (t - \tau) + \varphi) = a(t - \tau)\cos(2\pi f_0 t)$.

D'altra parte, risultando $a(t)\cos(\omega_o t + \varphi) = a(t)(\cos(\omega_o t)\cos\varphi - \sin(\omega_o t)\sin\varphi)$ si ottiene che la presenza di una fase incognita φ determina la ricezione di un segnale modulato le cui c.a. di b.f. risultano pari a $x_c(t) = a(t)\cos\varphi$ e $x_s(t) = a(t)\sin\varphi$, e che quindi variano *in simultanea*. Pertanto, in base ai risultati del § 13.1.2.4, il segnale modulato equivale a quello in cui è presente la sola componente in fase $x_c(t)$, ma al quale un errore nella fase di demodulazione imprime una rotazione di angolo φ al piano dell'inviluppo complesso.

12.1.1.1 Portante soppressa - *PS*

Osserviamo che nell'espressione (12.3) della $\mathcal{P}_x$ di un segnale AM-BLD il termine $\frac{a_p^2}{2}$ rappresenta la potenza della portante *non modulata*[5], concentrata per metà ad f_0 e per metà a $-f_0$. Evidentemente, ponendo $a_p = 0$ nella (12.1) tale componente *svanisce*, dando luogo al sottocaso di *portante soppressa*, a cui corrisponde una densità di potenza pari a

$$\mathcal{P}_x(f) = \frac{k_a^2}{4}\left[\mathcal{P}_m(f - f_0) + \mathcal{P}_m(f + f_0)\right]$$

La demodulazione di AM-BLD-PS si effettua in modo coerente (§ 12.2.1), dopo aver ricostruito la portante per quadratura (§ 12.2.2.1) o mediante un *Costas Loop*[6], oppure mediante demodulatore ad inviluppo (§ 12.2.5), dopo aver elaborato la portante ricostruita come spiegato al § 12.1.1.3.

12.1.1.2 Portante intera - *PI*

Questo caso si verifica qualora si ponga $a_p \neq 0$, con un valore scelto in modo che $x_c(t)$ *sia sempre positiva*, e ciò accade se $a_p \geq k_a \cdot \max\{|m(t)|\}$, in modo che risulti sempre (vedi fig. 12.1)

$$x_c(t) = a_p + k_a m(t) \geq 0 \quad \text{per} \quad \forall t$$

e quindi in modo che la portante modulata *non inverta mai la fase*, come invece accade per i casi di portante soppressa (o parzialmente).

Un modo equivalente di esprimere questa condizione è

$$a_p^2 \geq k_a^2 m^2(t) \quad \text{per} \quad \forall t$$

ed indicando con P_I^{Max} il *massimo* valore di $m^2(t)$ ([7]), essa può essere soddisfatta qualora $\left(\frac{a_p}{k_a}\right)^2 > P_I^{Max}$, permettendo così di dimensionare l'uno rispetto all'altro[8].

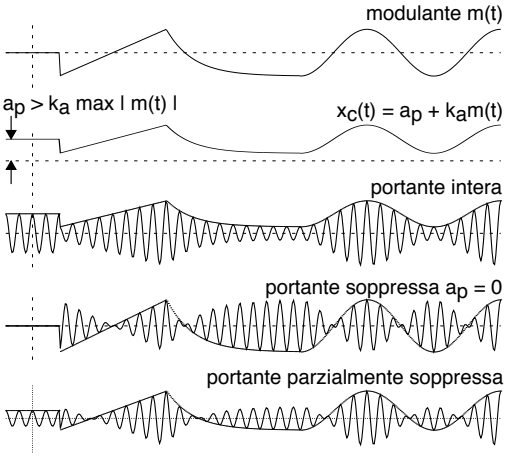


Figura 12.1: Modulazione di ampiezza BLD

Indice di modulazione per portante intera Il rapporto

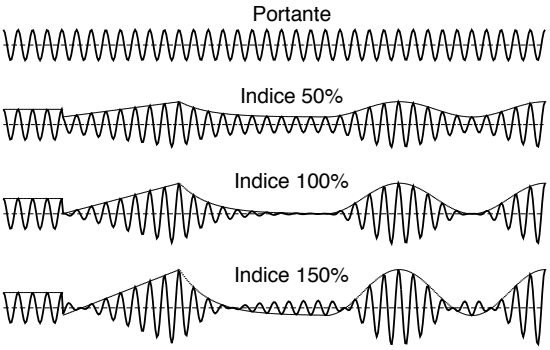
$$I_a = \frac{k_a \cdot \max\{|m(t)|\}}{a_p} \tag{12.4}$$

[5] Cioè che *non dipende* dal messaggio modulante $m(t)$.

[6] Vedi nota 18 a pag. 373.

[7] Il segnale $\mathcal{P}_I(t) = m^2(t)$ può essere indicato come *potenza istantanea* di $m(t)$, e P_I^{Max} indicato come la sua *potenza di picco*.

[8] Ad 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 § 6.2.3) 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...



prende il nome di *indice di modulazione* e nel caso di portante intera assume valori compresi tra zero ed uno, o tra 0 e 100 in termini percentuali. Un indice I_a del 100% corrisponde allo sfruttamento di *tutta* la dinamica della portante, fatto rilevante ai fini delle considerazioni svolte al § 12.1.1.4.

Per $k_a = 0$ si ottiene $I_a = 0$ ed assenza di modulazione; a valori k_a crescenti corrisponde un aumento di I_a e dunque una maggiore variazione dell'ampiezza della portante, finché per $I_a > 100\%$ si verifica una *sovramodulazione* e non ci troviamo più in condizioni di portante *intera* bensì *parzialmente soppressa* (§ 12.1.1.3), come mostrato alla figura precedente.

La ragione principale della modulazione a portante intera è che in tal caso il componente di ricezione può fare a meno di conoscere il valore di f_0 ed utilizzare il semplice *demodulatore di inviluppo* descritto al § 12.2.5.

12.1.1.3 Portante parzialmente soppressa - *PPS*

Se il valore a_p nella (12.1) è inferiore a quello necessario ad ottenere una portante intera, ma non è nullo, si ottiene il caso della portante *parzialmente soppressa*, che permette di risparmiare potenza (vedi § 12.1.1.4). Il residuo di portante presente può essere usato per la sua *ri-generazione* al lato ricevente mediante l'uso di un PLL (§ 12.2.2.2), in modo da *sommarla* al segnale ricevuto, ri-producendo così il termine $a_p \cos \omega_0 t$ e riconducendosi al caso di portante intera.

12.1.1.4 Efficienza energetica per portante intera e *PPS*

Nell'espressione (12.3) della potenza totale $\mathcal{P}_x = \frac{1}{2}\left(a_p^2 + k_a^2 \mathcal{P}_m\right)$ per un generico segnale AM-BLD notiamo che solo $\frac{1}{2}k_a^2\mathcal{P}_m = \mathcal{P}_u$ esprime un segnale *utile*, mentre $\frac{1}{2}a_p^2$ rappresenta la potenza spesa per la portante senza trasportare informazione. Pertanto si definisce una *efficienza energetica*

$$\eta = \frac{\mathcal{P}_u}{\mathcal{P}_x} = \frac{\frac{1}{2}k_a^2\mathcal{P}_m}{\frac{1}{2}\left(a_p^2 + k_a^2\mathcal{P}_m\right)} = \frac{1}{1 + \frac{a_p^2}{k_a^2\mathcal{P}_m}}$$

il cui valore indica la frazione di potenza trasmessa che è utile ai fini della ricostruzione del messaggio.

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!

12.1.2 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

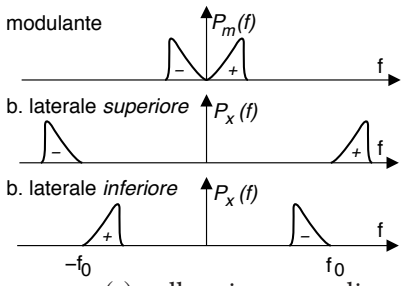
una banda *uguale* a quella di $m(t)$. Tale risultato è ottenuto realizzando un segnale modulato $x\,(t)$ le cui componenti analogiche $x_c(t)$ ed $x_s(t)$ sono *dipendenti* tra loro, ed in particolare imponendo che $x_c(t) = m(t)$ e $x_s(t) = \widehat{m}(t)$: infatti in tal modo si ottiene

$$\begin{aligned} x_{BLU}\,(t) &= m\,(t)\cos\omega_0 t - \widehat{m}\,(t)\sin\omega_0 t = \\ &= m\,(t)\,\frac{e^{j\omega_0 t}+e^{-j\omega_0 t}}{2} - \widehat{m}\,(t)\,\frac{e^{j\omega_0 t}-e^{-j\omega_0 t}}{2j} = \\ &= e^{j\omega_0 t}\frac{1}{2}\left[m\,(t)+j\widehat{m}\,(t)\right] + e^{-j\omega_0 t}\frac{1}{2}\left[m\,(t)-j\widehat{m}\,(t)\right] \end{aligned} \tag{12.5}$$

Ricordando ora che $\frac{1}{2}\left[m\,(t) \pm j\widehat{m}\,(t)\right] = m^{\pm}\,(t)$ (vedi eq. (11.17) e (11.18)) corrisponde al contenuto a frequenze positive (e negative) di $m\,(t)$, allora (assumendo $x(t)$ di energia) la trasformata di Fourier di ambo i membri di (12.5) fornisce

$$\begin{aligned} X_{BLU}\,(f) &= \delta\,(f-f_0) * M^{+}\,(f) + \delta\,(f+f_0) * M^{-}\,(f) = \\ &= M^{+}\,(f-f_0) + M^{-}\,(f+f_0) \end{aligned} \tag{12.6}$$

e quindi il segnale modulato AM-BLU è ottenuto a partire dai contenuti a frequenze positive e negative di $m\,(t)$, traslati *ai lati* della portante f_0, come mostrato alla riga centrale della figura a lato, in cui il segnale modulato BLU risulta *(*nel dominio della frequenza*) esterno* ad f_0, circostanza indicata come segnale BLU in *banda laterale superiore*. Il caso opposto (*banda laterale inferiore*, riga in basso della figura) si ottiene cambiando segno a $x_s\,(t)$ nella prima eguaglianza di (12.5); scriviamo dunque l'espressione generale come



$$x_{BLU}\,(t) = \frac{k_a}{\sqrt{2}}\,m\,(t)\cos\omega_0 t \mp \frac{k_a}{\sqrt{2}}\,\widehat{m}\,(t)\sin\omega_0 t \tag{12.7}$$

con − e + rispettivamente per ottenere un segnale BLU con banda superiore o inferiore.

Dopo aver notato che stiamo trattando di un caso a *portante soppressa*, osserviamo che il segnale modulato BLU (12.7) ha una potenza (vedi § 12.4.5)

$$\mathcal{P}_x = 2\cdot\left(\tfrac{k_a^2}{2}\cdot\mathcal{P}_m\cdot\tfrac{1}{2}\right) = \tfrac{k_a^2}{2}\mathcal{P}_m$$

eguale a quella di un segnale AM-BLD per il quale $x_c(t) = k_a m\,(t)$ e $x_s(t) = 0$.

Qualora si consideri infine un segnale modulante $m\,(t)$ realizzazione di un processo ergodico, al § 12.4.4 si dimostra il risultato del tutto simile alla (12.6), ovvero

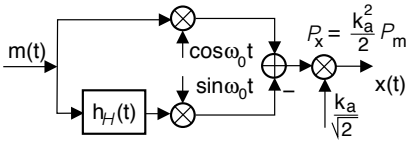
$$\mathcal{P}_x\,(f) = \mathcal{P}_{m^+}\,(f-f_0) + \mathcal{P}_{m^-}\,(f+f_0) \tag{12.8}$$

Il vantaggio di questa tecnica di modulazione è subito evidente: consente di risparmiare banda, permettendo la trasmissione di più messaggi in divisione di frequenza FDM, vedi pag. 343.

12.1.2.1 Generazione di segnali *BLU*

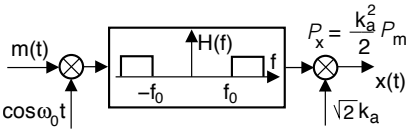
Un segnale BLU può essere ottenuto mediante due possibili tecniche analogiche, ed una terza che è percorribile nel caso di segnali campionati:

- la prima (modulatore di *Hartley*) consiste nell'uso di un *filtro di Hilbert* per calcolare $\widehat{m}(t)$, da usare assieme ad $m(t)$ in un modulatore in fase ed in quadratura, introdotto al § 11.2.3. E' 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, vedi § 11.4.1;



modulatore di Hartley

- nella seconda viene prima generato un segnale BLD, che viene poi filtrato passa-banda in modo da eliminare una delle due bande laterali. Qualora sia necessario trasmettere componenti spettrali di $m\,(t)$ prossime a frequenza zero si determina un problema simile a quello del caso precedente, complicando la realizzazione del filtro.



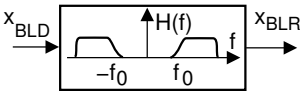
filtraggio passa banda

- la tecnica numerica nota come modulatore di *Weaver* si avvale della possibilità di eseguire prodotti complessi su valori del segnale modulante campionato. Volendo ottenere un segnale BLU con banda laterale *superiore*, il segnale di banda base $m(t)$ (con banda W) viene moltiplicato per l'esponenziale complesso $e^{-j2\pi\frac{W}{2}t}$ in modo da centrare $m^{+}(t)$ attorno alla frequenza zero, ed eliminando $m^{-}(t)$ (ora centrato a frequenza $-W$) dal segnale *complesso* ottenuto, mediante un filtro *passa basso* con banda $\frac{W}{2}$. Il risultato viene quindi moltiplicato per una singola portante a frequenza $f_0 + \frac{W}{2}$, ottenendo il risultato desiderato.

La trasmissione FDM di segnali BLU è stata lungamente usata per multiplare in forma analogica svariati canali telefonici (vedi § 11.1.1.2). Pertanto, la limitazione sulla minima frequenza di un canale telefonico a 300 Hz è motivata anche dalla necessità di effettuare modulazioni BLU.

12.1.3 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 televisivo analogico[9] (vedi § 25.1). Si ricorre allora alla modulazione a *banda laterale ridotta* (BLR), ottenuta facendo transitare un segnale modulato BLD attraverso un 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.



[9] Nel caso ad esempio di ampie zone di immagine uniformi ed a luminosità costante, il segnale è praticamente costante.

	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	$\frac{k_a}{\sqrt{2}}m(t)\cos(\omega_0 t) - \frac{k_a}{\sqrt{2}}\widehat{m}(t)\sin(\omega_0 t)$	$\frac{k_a^2}{2}\mathcal{P}_m$	$\sqrt{\frac{2\mathcal{P}_x}{\mathcal{P}_m}}$
BLD-PI	$[a_p + k_a m(t)]\cos(\omega_0 t)$ con $a_p \geq k_a \cdot \max\{\lvert m(t)\rvert\}$	$\frac{a_p^2}{2} + \frac{k_a^2}{2}\mathcal{P}_m$	$\sqrt{\frac{2\mathcal{P}_x - a_p^2}{\mathcal{P}_m}}$

Tabella 12.1: Espressione dei segnali AM e relativa potenza

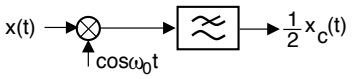
12.1.4 Potenza di un segnale AM

La tabella 12.1 riporta uno schema riassuntivo dell'espressione del segnale per i diversi tipi di modulazione di ampiezza, assieme ai valori k_a ed a_p tali da determinare uno specifico valore per la potenza totale $\mathcal{P}_x$, di cui ai §§ 12.4.5 e 12.4.5.1 si discute lo schema di calcolo.

12.2 Demodulazione di ampiezza

Il segnale informativo $m(t)$ può essere recuperato a partire da quello modulato $x(t)$ mediante il processo di demodulazione, che nel caso AM può avvenire mediante diverse tecniche, denominate *omodina*, di *inviluppo*, *in fase e quadratura*, *eterodina*; ognuna di esse ha il suo campo di applicazione, assieme a pregi e difetti.

12.2.1 Demodulazione coerente o omodina



Si tratta del circuito già noto (vedi § 11.2.4) di estrazione della componente in fase $x_c(t)$ mediante moltiplicazione o *mixing*[10] 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 avere la stessa fase e la stessa frequenza della portante ricevuta[11], condizione indicata anche con il nome di demodulazione *omodina, sincrona, coerente*, a *conversione diretta*, o zero-IF[12]. Il metodo è applicabile a tutti i tipi di modulazione di ampiezza, in quanto per tutti la componente in fase è direttamente legata al messaggio $m(t)$; nella pratica, nei casi di BLD-PI ed in quelli ad esso riconducibili, può essere invece preferibile adottare il demodulatore di inviluppo (§ 12.2.5).

[10] Il dispositivo fisico che effettua la moltiplicazione per una portante viene indicato in letteratura con il termine di *mixer*, il cui significato letterale è *mescolatore*. Dato che lo stesso termine è usato anche per indicare un circuito od apparato in grado di realizzare la *somma* di più segnali, come ad esempio avviene per il mixer *audio* di un sistema di amplificazione sonora, per distinguere i due casi si può parlare di mixer *additivo* oppure *moltiplicativo*, come nel nostro caso. In appendice 12.4.1 sono illustrate due tecniche di realizzazione del mixer.

[11] Dato che un qualunque canale presenta un ritardo di propagazione τ, la portante del segnale ricevuto sarà nella forma $\cos 2\pi f_0 (t-\tau) = \cos(2\pi f_0 t - 2\pi f_0 \tau) = \cos(2\pi f_0 t - \varphi)$, ovvero sarà sempre presente una fase $\varphi = 2\pi f_0 \tau$ *incognita*. Nel caso poi di un collegamento radiomobile, può anche essere presente un errore di frequenza, dovuto all'*effetto doppler*., vedi § 20.4.6.

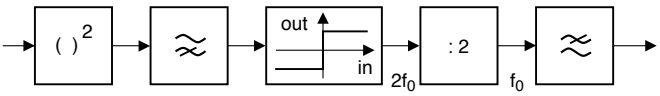
[12] Le ultime due definizioni sono orientate a differenziarsi dal metodo di demodulazione *eterodina*, che in realtà si è affermato *prima* della praticabilità di quello omodina, per i motivi esposti al § 12.2.7.

12.2.2 Sincronizzazione di portante

Individua il compito di generare presso il demodulatore una *copia* della portante quanto più possibile *coerente* con la fase di quella ricevuta. Descriviamo due dei metodi utilizzati a questo scopo, mentre un terzo attuabile con tecniche totalmente digitali è descritto al § 12.4.3.

12.2.2.1 Metodo della quadratura

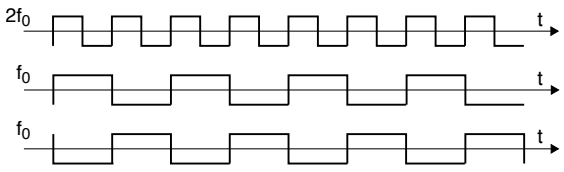
Anche se nel segnale ricevuto non vi è traccia della portante, come per BLD-PS, la portante di demodulazione può essere comunque ottenuta mediante lo schema simbolico rappresentato in figura, che come prima cosa eleva al quadrato il segnale modulato ricevuto $x(t) = m(t)\cos(\omega_0 t + \varphi)$, producendo



$$\frac{1}{2} m^2(t) \left[1 + \cos(2\omega_0 t + 2\varphi)\right]$$

Il termine di banda base $\frac{1}{2}m^2(t)$ viene quindi rimosso dal filtro passa alto, mentre il termine $\cos(2\omega_0 t + 2\varphi)$ è convertito in un'onda quadra a frequenza $2f_0$ mediante il dispositivo non lineare *squadratore*[13], che produce in uscita la funzione *segno* di ciò che si presenta in ingresso. A sua volta l'onda quadra attraversa un divisore di frequenza[14], ottenendo così una nuova onda quadra, ma a frequenza f_0; come noto (§ 2.5.3) l'onda quadra contiene anche tutte le armoniche dispari, che sono rimosse dal filtro passa basso di uscita, ottenendo in definitiva la portante desiderata.

Qualora il divisore sia implementato mediante un multivibratore bistabile[15] che commuta sul *fronte di salita* dell'ingresso, il metodo è affetto da una *ambiguità di segno*, che corrisponde ad un eventuale errore di fase pari a π, come mostrato in figura.



12.2.2.2 Phase Locked Loop o *PLL*

Una seconda tecnica (nota come *circuito ad aggancio di fase*) adotta invece un approccio a *controreazione,* e si basa sull'utilizzo di un dispositivo chiamato *oscillatore controllato in tensione* (VOLTAGE CONTROLLED OSCILLATOR o VCO[16]) il quale genera una sinusoide

$$y(t) = \sin\left(\omega_0 t + 2\pi k_f \int_{-\infty}^{t} \varepsilon(\tau)\, d\tau\right)$$

la cui fase varia nel tempo in proporzione all'integrale del segnale di ingresso $\varepsilon(\tau)$[17].

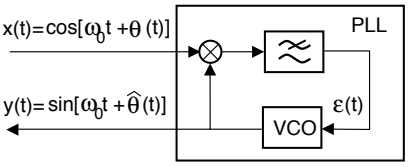
[13] Realizzato mediante un amplificatore ad elevato guadagno, portato a lavorare in *saturazione*.

[14] Vedi ad es. `https://it.wikipedia.org/wiki/Divisore_di_frequenza`

[15] Vedi ad es. `https://it.wikipedia.org/wiki/Multivibratore`

[16] Vedi ad es. `https://it.wikipedia.org/wiki/Oscillatore_controllato_in_tensione`

[17] Se ad esempio $\varepsilon(\tau) = \Delta/k_f$ ossia *è costante*, si ottiene $y(t) = \sin(2\pi f_0 t + 2\pi\Delta t) = \sin[2\pi(f_0 + \Delta)t]$, ovvero la frequenza si è alterata di una quantità pari a Δ. Infatti, il VCO realizza il processo di *modulazione di frequenza*, vedi eq. (11.7) a pag. 347.



Lo schema a lato illustra come il ruolo del VCO sia quello di generare una portante sfasata di $\pi/2$ rispetto a quella del segnale $x\,(t)$ in arrivo, mentre a quest'ultimo è richiesto di contenere almeno *un residuo* di portante[18]. In uscita dal VCO è pertanto presente il segnale $y\,(t) = \sin(\omega_0 t + \widehat{\theta}\,(t))$ in cui

$$\widehat{\theta}\,(t) = 2\pi k_f \int_{-\infty}^{t} \varepsilon\,(\tau)\,d\tau \tag{12.9}$$

rappresenta la *stima* della fase $\theta\,(t)$ del segnale in ingresso, valutata all'istante t. Eseguendo ora il prodotto tra $y\,(t)$ ed il segnale ricevuto[19] $x\,(t) = \cos\,(\omega_0 t + \theta\,(t))$ si ottiene[20]

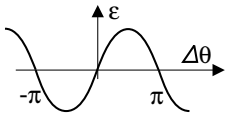
$$\tfrac{1}{2}\sin\left[2\omega_0 t + \theta\,(t) + \widehat{\theta}\,(t)\right] + \tfrac{1}{2}\sin\left[\theta\,(t) - \widehat{\theta}\,(t)\right]$$

il cui primo termine è centrato a frequenza doppia ($2\omega_0$) e viene eliminato dal filtro passa basso (detto anche *filtro di loop*), alla uscita del quale troviamo dunque

$$\varepsilon(t) = \tfrac{1}{2}\sin\left[\theta\,(t) - \widehat{\theta}\,(t)\right] = \tfrac{1}{2}\sin(\Delta\theta\,(t))$$

dove $\Delta\theta\,(t) = \theta\,(t) - \widehat{\theta}\,(t)$ rappresenta l'errore di fase che desideriamo annullare, ed $\varepsilon(t)$ è la grandezza in ingresso al VCO.

Pensiamo ora al caso in cui la $\theta\,(t)$ presente nel segnale di ingresso sia *costante*: nel momento in cui $\Delta\theta = 0$, si ottiene che anche $\varepsilon = 0$, ed il VCO *non altera* la fase (esatta) della portante generata. Se invece $\Delta\theta \gtrless 0$ (e $|\Delta\theta| < \pi$)[21], allora $\varepsilon \gtrless 0$, e dunque (vedi figura a lato) il VCO è portato ad aumentare (diminuire) la fase della propria portante, riducendo di conseguenza l'errore di fase[22]. 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 $\widehat{\theta}\,(t)$ e l'integrale di $\varepsilon\,(t)$ che compare nella (12.9)[23].



[18] Un diverso circuito controreazionato in grado di operare anche per segnali a *portante soppressa* prende il nome di *Costas loop*, vedi `https://en.wikipedia.org/wiki/Costas_loop`, mentre al § 16.11.1 si discute di una realizzazione relativa ad una trasmissione a *spettro espanso*.

[19] Trascuriamo la presenza di eventuali modulazioni, il cui effetto si intende *mediato* dalla caratteristica passa-basso del PLL, dovuta sia all'integratore presente nel VCO, che al filtro di loop.

[20] Utilizziamo qui la relazione $\cos\alpha \sin\beta = \frac{1}{2}\,[\sin\,(\alpha+\beta) + \sin\,(\alpha-\beta)]$.

[21] La grandezza di controllo $\varepsilon(t)$ proporzionale a $\sin(\Delta\theta)$ si azzera 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 $\widehat{\theta}\,(t)$ che causa un ulteriore aumento o diminuzione di $\Delta\theta$, finché questo non raggiunge il valore 0 o 2π, corrispondenti a condizioni di stabilità. In altre parole, se $|\Delta\theta| < \pi$ si determina un transitorio alla fine del quale $\varepsilon \to 0$, mentre se $\pi < |\Delta\theta| < 3\pi$ il transitorio converge verso $\varepsilon \to 2\pi$, e così via.

[22] Notiamo che un moltiplicatore, seguito da un filtro passabasso, esegue il calcolo dell'intercorrelazione tra gli ingressi del moltiplicatore (vedi § 7.5.4), che nel nostro caso è una sinusoide.

[23] Inoltre, le prestazioni del PLL dipendono fortemente anche dalla banda e dall'ordine del *filtro di loop*, che limita la velocità di variazione di $\varepsilon\,(t)$ e l'estensione dell'intervallo di aggancio. Lo studio teorico si basa sull'uso della trasformata di Laplace e sulla approssimazione $\sin(\Delta\theta) \simeq \Delta\theta$, in quanto così il PLL

Al § 12.4.3 viene illustrato come utilizzare il PLL allo scopo di generare una portante *stabile* di modulazione a frequenza qualsiasi, a partire da un oscillatore al quarzo.

12.2.3 Errori di fase e di frequenza

Cosa accade se la sincronizzazione di portante non è *perfetta*? Qualora tra la portante del demodulatore (§ 12.2.1) e quella del segnale in arrivo $x(t)$ siano presenti errori di fase θ e/o di frequenza Δf, ovvero risulti $x(t) = \cos(2\pi(f_o + \Delta f)t + \theta)$, il risultato della demodulazione (non più coerente) risulta pari a[24]:

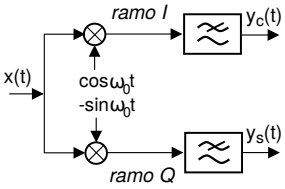
$$\begin{aligned} y(t) &= x_c(t)\cos\omega_0 t\cos[(\omega_o + \Delta\omega)t + \theta] \\ &= \tfrac{1}{2}x_c(t)[\cos(\Delta\omega t + \theta) + \cos((2\omega_0 + \Delta\omega)t + \theta)] \end{aligned}$$

Mentre il termine a frequenza (circa) doppia viene eliminato come di consueto dall'apposito filtro, sul segnale demodulato $y(t)$ si manifestano ora le seguenti distorsioni:

- in assenza di errori di frequenza ($\Delta\omega = 0$) si ottiene $y(t) = \frac{1}{2}x_c(t)\cos\theta \leq \frac{1}{2}x_c(t)$ cioè una attenuazione, che può annullare $y(t)$ se $\theta = \pm\frac{\pi}{2}$, mentre per $\theta = \pi$ si ottiene una inversione di segno di $x_c(t)$;
- qualora $\Delta\omega \neq 0$ si ottiene $y(t) = \frac{1}{2}x_c(t)\cos(\Delta\omega t)$ 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.

12.2.3.1 Demodulazione I e Q in presenza di errore di fase

Poniamoci ora nel caso in cui nel segnale modulato siano presenti entrambe le c.a. di b.f., ovvero $x(t) = x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t$, e si desideri demodularle entrambe. Si ricorre allora al demodulatore *in fase e quadratura* (§ 11.2.4), che prevede due rami (detti anche I e Q) con portanti di demodulazione, appunto, in quadratura.



Applicando i risultati del § 11.2.4, e con riferimento alla notazione adottata nella figura che segue, in condizioni di *coerenza* si ottiene $y_c(t) = \frac{1}{2}x_c(t)$ e $y_s(t) = \frac{1}{2}x_s(t)$. Se viceversa il segnale ricevuto presenta una fase incognita θ, e dunque $x(t) = x_c(t)\cos(\omega_0 t + \theta) - x_s(t)\sin(\omega_0 t + \theta)$, si ottiene invece[25]

$$\begin{aligned} y_c(t) &= 1/2\,(x_c(t)\cos\theta - x_s(t)\sin\theta) \\ y_s(t) &= 1/2\,(x_c(t)\sin\theta + x_s(t)\cos\theta) \end{aligned} \tag{12.10}$$

---

può essere studiato come un sistema di controllo *linearizzato*, sommariamente descritto al § 12.3.2.1. Per approfondimenti, vedi `http://it.wikipedia.org/wiki/Phase-locked_loop`.

[24] Si applichi $\cos\alpha\cos\beta = \frac{1}{2}[\cos(\alpha+\beta) + \cos(\alpha-\beta)]$.

[25] Per il ramo in fase risulta

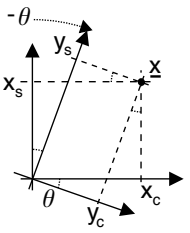
$$\begin{aligned} y_c(t) &= (x_c(t)\cos(\omega_0 t + \theta) - x_s(t)\sin(\omega_0 t + \theta))\cdot\cos\omega_0 t = \\ &= x_c(t)\cos(\omega_0 t + \theta)\cos\omega_0 t - x_s(t)\sin(\omega_0 t + \theta)\cos\omega_0 t = \\ &= \tfrac{1}{2}x_c(t)[\cos(2\omega_0 t + \theta) + \cos\theta] - \tfrac{1}{2}x_s(t)[\sin(2\omega_0 t + \theta) - \sin(-\theta)] \end{aligned}$$

mentre svolgendo simili sviluppi per il ramo in quadratura, si giunge a

$$y_s(t) = \tfrac{1}{2}x_c(t)[\sin\theta - \sin(2\omega_0 t + \theta)] + \tfrac{1}{2}x_s(t)[\cos(\theta) - \cos(2\omega_0 t + \theta)]$$

Anche qui i filtri passabasso eliminano le componenti centrate a $2f_0$, permettendo di ottenere la (12.10).

Ovviamente, per $\theta = 0$ le (12.10) si riducono al caso noto, mentre *curiosamente* per uno sfasamento $\theta = \frac{\pi}{2}$ le due c.a. di b.f. (a parte un segno) si invertono di ruolo. Un ragionamento più approfondito è fornito a pag. 349, e dimostra che θ rappresenta l'angolo di cui ruota il piano dell'inviluppo complesso tra $\underline{x}(t)$ e $\underline{y}(t)$. Ad ogni modo il sistema (12.10) è perfettamente invertibile, qualora θ sia noto.



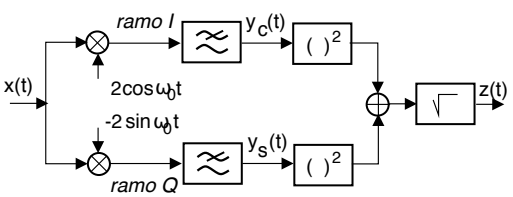
12.2.4 Demodulazione incoerente

Si tratta di uno schema utile nella *fase di ricerca* della regione di frequenza in cui è presente un segnale[26], ovvero quando si desidera verificare la presenza *o meno* di un segnale ad una determinata frequenza. In tale schema la coerenza di fase tra la portante ricevuta e quella di demodulazione viene deliberatamente *trascurata*, adottando una architettura che utilizza anche il ramo *in quadratura*.

Se consideriamo un segnale AM-BLD-PS ricevuto in presenza di una fase θ incognita rispetto alla portante del ramo I del demodulatore, ovvero $x(t) = m(t)\cos(\omega_0 t + \theta)$, il relativo inviluppo complesso rispetto ad f_0 (e $\theta = 0$) risulta pari a

$$\underline{x}(t) = m(t)\, e^{j\theta} = m(t)\cos\theta + jm(t)\sin\theta$$

le cui parti reale ed immaginaria corrispondono all'uscita dei filtri passa-basso posti sui rami del demodulatore I-Q mostrato in figura, ossia $y_c(t) = m(t)\cos\theta$ e $y_s(t) = m(t)\sin\theta$, come si ottiene (a parte un fattore $1/2$) dalle (12.10) avendo posto $x_c(t) = m(t)$ e $x_s(t) = 0$. Dunque il segnale $z(t)$ di uscita corrisponde a



$$z(t) = \sqrt{y_c^2(t) + y_s^2(t)} = |m(t)|\sqrt{\cos^2\theta + \sin^2\theta} = |m(t)|$$

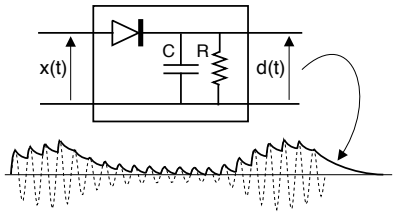
Pertanto, nonostante l'ignoranza della fase θ, siamo ancora in grado di individuare la *presenza* di un segnale modulante. L'operazione di modulo impedisce l'uso dello schema per demodulare generici segnali BLD-PS, mentre il caso PI sarebbe perfettamente demodulabile, ma per quello è più che sufficiente il demodulatore *di inviluppo* discusso al § seguente. Infine, al § 14.4.2 si illustra come usare il demodulatore incoerente per decidere per la presenza o meno di una sinusoide a cui è sovrapposto un rumore gaussiano, e viene valutata la relativa probabilità di errore.

[26] La ricerca dell'emittente può essere l'azione banale di sintonizzare a mano la propria radio sul programma preferito, oppure (come si dice, in modalità *ricerca automatica*), mediante un circuito del tipo di cui stiamo discutendo, con il quale vengono *provate* diverse portanti di demodulazione, finché non si riscontra un segnale in uscita.

In generale, la ricezione della comunicazione vera e propria viene preceduta da una fase di *acquisizione della portante*, svolta ad esempio come qui accennato, dopodiché la sincronizzazione è mantenuta mediante interventi automatici (ad es. via PLL), necessari qualora si tratti di dover compensare le variazioni di frequenza dovute ad esempio al movimento reciproco di trasmettitore e ricevitore (*effetto doppler*), come per il caso delle comunicazioni con mezzi mobili, vedi § 20.4.6.

12.2.5 Demodulatore di inviluppo per AM-BLD-PI

Si tratta del semplice circuito *non lineare* riportato in figura[27]. 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 (pag. 22) $\tau = RC$, abbastanza grande rispetto al periodo della portante $\frac{1}{f_0}$, 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, il valore di τ deve essere scelto né troppo piccolo né troppo grande, per evitare una eccessiva *seghettatura*, ed al contempo riuscire ad inseguire anche le variazioni più rapide del messaggio[28].



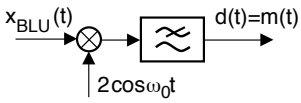
La semplicità del circuito è tale da farlo usare nel maggior numero di casi possibili, anche se il suo uso prevalente è per la demodulazione di segnali a *portante intera*. D'altra parte, la contemporanea presenza di altri segnali modulati con portante diversa da quella del segnale desiderato rendono obbligatoria l'adozione di ulteriori provvedimenti, come discusso nel § 12.2.7 relativo alla demodulazione *eterodina*.

12.2.6 Demodulazione per segnali a banda laterale unica e ridotta

Nel caso di segnali BLU (§ 12.1.2)

$$x_{BLU}(t) = m(t)\cos\omega_0 t - \widehat{m}(t)\sin\omega_0 t$$

il segnale modulante $m(t)$ può essere riottenuto a partire da $x(t)$ utilizzando il demodulatore omodina mostrato in figura, dato che la componente in fase $x_c(t)$ dell'inviluppo complesso corrisponde proprio pari ad $m(t)$. Occorre però prestare 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 $x_s(t)$, come mostrato al § 12.2.3.1 in uscita dal demodulatore si ottiene (nel caso di banda laterale superiore):



$$d(t) = m(t)\cos(\Delta\omega t + \theta) - \widehat{m}(t)\sin(\Delta\omega t + \theta)$$

Pertanto la modulazione BLU è più sensibile di quella BLD agli errori della portante di demodulazione, dato che anche un semplice errore di fase θ produce non solo un affievolimento, ma un vero fenomeno di *interferenza* tra $m(t)$ e $\widehat{m}(t)$. Per evitare che ciò accada, nella trasmissione blu è spesso presente una portante parzialmente soppressa, in modo da agevolare il funzionamento delle tecniche di recupero portante.

[27] Il simbolo ⊣▷⊢ rappresenta un *diodo*, costituito da un bipolo di materiale semiconduttore drogato, che ha la particolarità di condurre in un solo verso (quello della freccia).

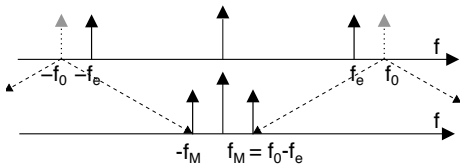
[28] Presso `http://it.wikipedia.org/wiki/Rivelatore_d'inviluppo` qualche linea guida di progetto.

Anche nel caso BLR è possibile ricorrere ad un demodulatore di tipo omodina, evitando i problemi di sincronizzazione di fase illustrati, purché il filtro $H(f)$ usato in trasmissione per rimuovere parte di una banda laterale presenti alcune condizioni di simmetria attorno a f_0[29].

Ampiezza dei segnali BLU Le variazioni di ampiezza dei segnali AM-BLU sono ben maggiori che nel caso BLD, a causa del brusco troncamento spettrale causato dal filtro di Hilbert, e dalla distorsione di fase non lineare associata, e di ciò va tenuto conto per evitare fenomeni di saturazione e distorsione non lineare (§ 8.3), anche adottando adeguate contromisure[30].

12.2.7 Demodulatore eterodina

Individua la tecnica di utilizzare una frequenza di demodulazione *differente* da quella della portante[31], e fu inventata per rendere udibili i segnali in *codice Morse*[32] trasmessi via radio in forma di una portante intermittente f_0: dato che $\cos\alpha\cos\beta = {}^1\!/_2\,(\cos(\alpha+\beta)+\cos(\alpha-\beta))$, scegliendo la frequenza di eterodina f_e di poco inferiore a quella della portante f_0 il termine a frequenza $f_0 - f_e$ cade infatti nella banda udibile. Da un punto di vista grafico il risultato equivale a *sommare e sottrarre* la frequenza eterodina alle frequenze dell'altro segnale, traslando così f_0 in $f_0 \pm f_e$: indichiamo di qui in poi la differenza $f_0 - f_e$ con il termine di *media frequenza* f_M, detta anche frequenza *intermedia* o IF.



Piccola storia della radio Con lo sviluppo della trasmissione radio di segnali modulati, le diverse emittenti eseguivano trasmissioni AM-BLD-PI ognuna su di una portante differente; sebbene fosse possibile *sintonizzare* ciascuna emittente con un demodulatore omodina centrato sulla relativa portante, i dispositivi del tempo soffrivano di fenomeni di *deriva*, e non essendo ancora stato inventato il PLL, la portante di demodulazione *slittava*. Inoltre la tecnica omodina soffriva anche del problema del *rientro* della portante di demodulazione sull'altro ingresso del mixer, comportando l'insorgenza di una *componente continua* in uscita dal mixer stesso, in grado di mandare *in saturazione* il successivo stadio di amplificazione[33].

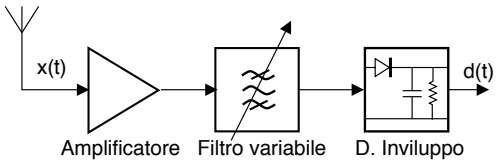
[29] Si può dimostrare che per l'inviluppo complesso $\underline{H}(f)$ di $H(f)$ deve risultare: $\underline{H}(f) + \underline{H}^*(-f) = cost$ perché in tal modo il residuo di banda parzialmente soppressa si combina esattamente con ciò *che manca* alla banda laterale *non* soppressa.

[30] Vedi ad es. `https://en.wikipedia.org/wiki/Amplitude-companded_single-sideband_modulation`

[31] Per la storia in maggior dettaglio, vedi ad es. `https://en.wikipedia.org/wiki/Heterodyne`

[32] In origine il segnale telegrafico (lett. *scrittura a distanza*) era trasmesso via cavo; per approfondimenti vedi `https://en.wikipedia.org/wiki/Morse_code` e `https://en.wikipedia.org/wiki/Wireless_telegraphy`

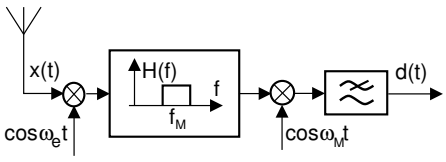
[33] A causa di fenomeni di induzione elettromagnetica che si manifestano tra conduttori presenti all'interno del circuito di demodulazione (che perdipiù aumentano con il valore delle frequenze in gioco) la portante di demodulazione omodina può appunto *rientrare* nella via percorsa dal segnale modulato. Se in ingresso al mixer è presente, oltre al segnale da demodulare *a portante soppressa*, anche un termine



Si provò quindi ad adottare una modulazione a portante intera in modo da poter adottare un demodulatore ad inviluppo, ma in tal caso l'emittente desiderata doveva essere prima *selezionata* anteponendo al demodulatore un filtro passa banda *variabile* centrato sulla portante dell'emittente desiderata (vedi figura), filtro di difficile realizzazione all'aumentare della frequenza[34].

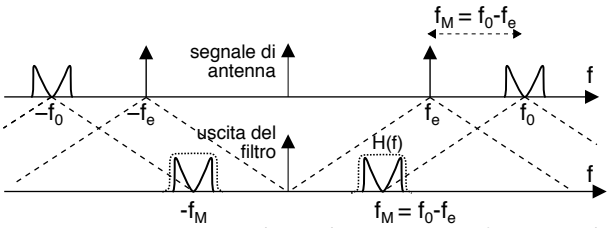
12.2.7.1 Supereterodina

La serie di considerazioni sopra svolte portò alla scelta di adottare in modo sistematico la tecnica di demodulazione eterodina, detta *super-* qualora si scelga una frequenza intermedia f_M *più elevata* di quelle dello spettro udibile[35], dando luogo allo schema di ricevitore che potremmo definire *in due passi* mostrato a lato: volendo sintonizzare l'emittente con portante f_0 il segnale ricevuto viene innanzitutto moltiplicato per una portante *eterodina* $f_e = f_0 - f_M$, in modo che lo spettro dell'emittente centrata su f_0 sia traslato alla *frequenza intermedia* $f_M = f_0 - f_e$. A quel punto un filtro passa banda *fisso* centrato su f_M permette di isolare l'emittente desiderata, che viene successivamente portata in banda base ad opera dello stadio di demodulazione *omodina* operante a frequenza f_M.



ricevitore supereterodina

Da un punto di vista grafico accade quanto mostrato in figura, con la frequenza f_e che viene *sommata e sottratta* a tutte le frequenze in ingresso al primo mixer, portando in f_M l'emittente centrata su di una f_0 distante da f_e di una quantità pari alla loro differenza $f_0 - f_e$.



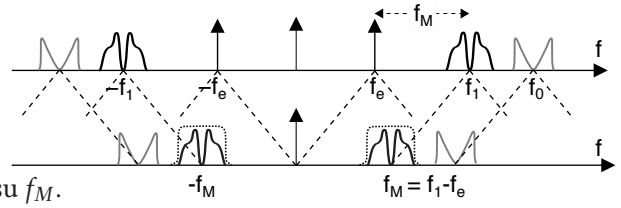
La sintonia di una diversa emittente avviene variando esclusivamente f_e, e quindi

alla stessa frequenza portante, si verifica un fenomeno noto come *self-mixing* dovuto all'eguaglianza $\cos^2 \alpha = 1/2\,(1 + \cos 2\alpha)$ che determina la comparsa di un termine *in continua*, e che non può essere eliminato mediante filtraggio passa alto qualora il segnale modulato presenti componenti energetiche prossime a frequenza zero. Lo stadio di amplificazione successivo mantiene un funzionamento lineare solo per valori di ingresso compresi in uno specifico intervallo, come discusso al § 8.3, mentre il valore medio ora presente può portare il segnale di ingresso *al difuori* di tale dinamica.

[34] Le 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.

[35] Il prefisso *super* venne scelto come contrazione di *supersonic heterodyne*, e dato che agli inizi del '900 di certo non esistevano aerei *supersonici*, indicava il concetto di *sopra i suoni*, in contrapposizione al suo uso originario di traslare il segnale *radiotelegrafico* in banda audio.

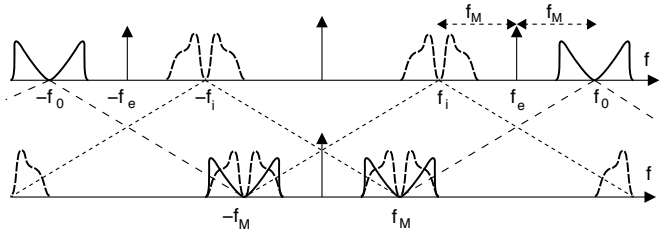
volendo ricevere ad esempio quella centrata in f_1, si imposta $f_e = f_1 - f_M$ come mostrato in questa seconda figura, in modo che ora sia la seconda emittente *a cadere* dentro il filtro centrato su f_M.



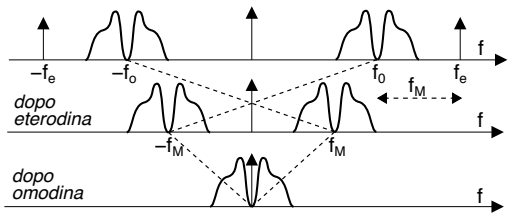
12.2.7.2 Frequenza immagine

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 $f_0 = f_e + f_M$, anche quella a portante $f_i = f_e - f_M$, 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 $f_0 = f_e + f_M$, sia 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 le trasmissioni centrate su portanti a frequenze minori di f_e.



Scelta della frequenza di eterodina Le trasmissioni *broadcast* AM[36] adottano portanti nella regione di frequenze detta delle *onde medie* (540-1600 KHz) con modulazione AM-BLD-PI ed utilizzano un ricevitore per il quale si sceglie una f_e *maggiore* della frequenza f_0 da sintonizzare anziché *minore* come prima illustrato, con il risultato che ora la frequenza immagine f_i è quella che si trova *al disopra* della f_0, come mostrato nello schema a lato. Per queste trasmissioni si è scelto di utilizzare una frequenza intermedia f_M pari a 455 KHz, quindi volendo ad esempio sintonizzare una emittente con $f_0 = 600$ KHz occorre una $f_e = f_M + f_0 = 1055$ KHz, ma allo stesso tempo anche l'emittente relativa alla portante $f_i = f_e + f_M = 1510$ KHz viene traslata nella banda del filtro a frequenza intermedia. Pertanto, prima del mixer operante ad f_e va posto un filtro che lasci passare solo le emittenti centrate a portanti inferiori ad f_e, reintroducendo l'esigenza di un filtro variabile, ma meno complesso di quello di pag. 378 dato che questo non ha lo scopo di filtrare una sola emittente, ma l'intera banda.



[36]Vedi ad es. `https://en.wikipedia.org/wiki/AM_broadcasting`

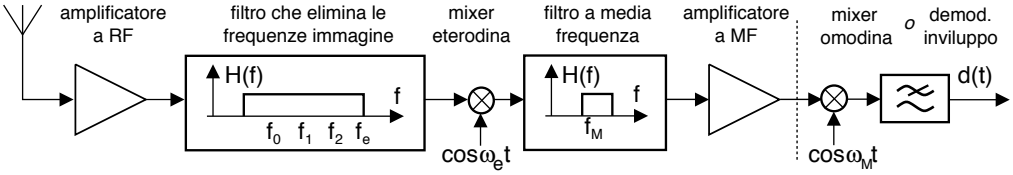


Figura 12.2: Schema di un ricevitore supereterodina con f_e maggiore della frequenza sintonizzata

La scelta $f_M = 455KHz$, inferiore alla minima frequenza di 510 KHz, permette di utilizzare per la media frequenza una regione dello spettro libera da altre trasmissioni[37], che altrimenti potrebbero essere amplificate dagli stadi ad alto guadagno posti dopo il filtro MF. La scelta di $f_e > f_0$ permette poi di posizionare il filtro *passa banda* che elimina le frequenze immagine *al disotto* della f_e, rendendo più semplice la sua realizzazione. La figura 12.2 mostra lo schema generale[38] (compresi gli stadi di amplificazione) per un ricevitore supereterodina con $f_e > f_0$. Riassumiamo i vantaggi ottenuti:

- la sintonia avviene mediante la variazione di f_e, ed il resto non cambia;
- la separazione tra f_0 ed f_M scongiura il rischio di instabilità che si potrebbe verificare se il segnale uscente dal filtro di media frequenza, amplificato, fosse ri-captato dallo stadio di ingresso, mentre ora invece l'amplificazione può aver luogo proprio nello stadio a media frequenza;
- il ridotto valore di f_M rispetto alla banda di frequenze di cui si opera la sintonia permette la realizzazione di un filtro passa banda a media frequenza di ridotta complessità e migliore selettività;
- per un segnale a portante intera lo stadio omodina è sostituito da uno ad inviluppo, senza necessità di generare f_0.

Conversione di frequenza multipla Notiamo che lo stadio di eterodina può essere ulteriormente ripartito in due conversioni di frequenza successive (vedi ad es. la fig. 25.6 a pag. 869), di cui la seconda conversione opera la sintonia, mentre la prima ha il solo scopo di traslare la banda di interesse in una regione centrata su di una frequenza inferiore, in cui il mezzo trasmissivo (ad es. un cavo coassiale) presenta minore attenuazione. Inoltre, la tecnica di mixing eterodina viene utilizzata anche negli apparati ripetitori[39], in cui la frequenza di trasmissione deve differire da quella di ricezione per evitare fenomeni di auto-interferenza.

[37] Il valore della frequenza intermedia utilizzata per le diverse bande in cui operano sistemi di radio diffusione è determinato in seno ad enti di standardizzazione, e le autorità di concessione della licenza di trasmissione evitano di assegnare alle emittenti frequenze nella stessa banda in cui è prevista l'uso di una frequenza intermedia, allo scopo di impedire *interferenze* nella medesima banda da parte di una diversa trasmissione. Oltre alla MF a 455 KHz del broadcast AM, abbiamo ad esempio valori di media frequenza pari a 10.7 MHz per il broadcast FM, 38.9 MHz per la televisione, 70 MHz per trasmissioni a microonde terrestri e satellitari.

[38] Nel caso di trasmissione a portante intera lo stadio eterodina finale viene rimpiazzato da un demodulatore inviluppo, oppure ancora da un demodulatore in fase e quadratura per gli usi più generali.

[39] Vedi ad es. il caso di un *trasponder* satellitare, § 25.3.3

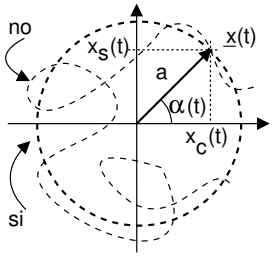
Realizzazione numerica Dopo lo stadio di eterodina, il segnale centrato a media frequenza presenta un valore di frequenza massima W assai ridotto rispetto alla sua versione modulata, permettendo di attuare su di esso le tecniche di (sotto)campionamento (§ 4.8) ed operare le restati operazioni, come la demodulazione in fase e quadratura, in via completamente numerica. Dato che attualmente tutti i ricevitori operano in questo modo, la questione verrà approfondita in una futura edizione.

12.3 Modulazione angolare

In questo caso l'informazione contenuta nel messaggio $m(t)$ è impressa sulla portante modificandone la fase $\alpha(t)$, ottenendo un segnale modulato

$$x(t) = a\cos(2\pi f_0 t + \alpha(t)) \tag{12.11}$$

il cui inviluppo complesso (vedi eq. (11.3)) vale $\underline{x}(t) = ae^{j\alpha(t)} = x_c(t) + jx_s(t)$, dove $x_c(t) = a\cos\alpha(t)$ e $x_s(t) = a\sin\alpha(t)$.



Notiamo subito che a differenza della modulazione di ampiezza, il modulo di $\underline{x}(t)$ è rigorosamente *costante*, e la sua fase $\alpha(t)$ può evolvere nel tempo unicamente su di una circonferenza di raggio a. Si è già mostrato al § 11.2.2 come il legame tra messaggio $m(t)$ e fase dell'inviluppo complesso $\alpha(t)$ possa essere descritto come modulazione di *fase* (o PM, *phase modulation*) qualora risulti $\alpha(t) = k_\phi m(t)$, oppure nei termini di una modulazione di *frequenza* (o FM) qualora si scelga $\alpha(t) = 2\pi k_f \int_{-\infty}^{t} m(\tau)\,d\tau$, dove k_ϕ e k_f sono coefficienti di proporzionalità che *dosano* l'intensità della modulazione introdotta.

La relazione che lega la *frequenza istantanea* $f_i(t)$ ad $m(t)$ dipende dal legame tra $\alpha(t)$ ed $m(t)$; ricordando la definizione (eq. (11.6)) di $f_i(t)$ come la derivata della *fase istantanea* $\psi(t) = 2\pi f_0 t + \alpha(t)$, ovvero

$$f_i(t) = \frac{1}{2\pi}\frac{d}{dt}\psi(t) = f_0 + \frac{1}{2\pi}\frac{d}{dt}\alpha(t) \tag{12.12}$$

si ottiene la tabella a lato che riassume la dipendenza della fase $\alpha(t)$ e della frequenza istantanea $f_i(t)$ da $m(t)$ per entrambi i tipi di modulazione angolare. Le due alternative PM e FM sono quindi esaminate assieme, in quanto intercambiabili qualora si effettui

	$\alpha(t)$	$f_i(t)$
PM	$k_\phi m(t)$	$f_0 + \frac{k_\phi}{2\pi}\frac{d}{dt}m(t)$
FM	$2\pi k_f \int_{-\infty}^{t} m(\tau)\,d\tau$	$f_0 + k_f m(t)$

Tabella 12.2: Legame tra segnale modulante $m(t)$, fase modulata $\alpha(t)$, e frequenza istantanea $f_i(t)$

- una PM con $m(t)$ pari all'integrale del messaggio informativo *oppure*
- una FM con $m(t)$ pari alla derivata del messaggio informativo.

Prima di affrontare gli aspetti della generazione, ricezione, e determinazione della densità di potenza di un segnale modulato angolarmente, analizziamo due sue peculiarità.

Non linearità Una caratteristica *fondamentale* della modulazione angolare è che il segnale modulato $x(t)$ dipende da quello modulante $m(t)$ in modo fortemente *non lineare,* e pertanto lo spettro di densità di potenza $\mathcal{P}_x(f)$ di (12.11) non può essere calcolato allo stesso modo del caso AM. Infatti, l'inviluppo complesso di un segnale modulato angolarmente può essere espresso[40] come:

$$\underline{x}(t) = ae^{j\alpha(t)} = a\left[1 + j\alpha(t) - \frac{\alpha^2(t)}{2} - j\frac{\alpha^3(t)}{3!} + ...\right] \tag{12.13}$$

da cui risulta evidente che, anche se $\mathcal{P}_\alpha(f)$ può essere espressa in funzione di $\mathcal{P}_m(f)$ in base alle relazione di tab. 12.2, nulla può essere detto in generale per $\mathcal{P}_{\underline{x}}(f)$, e dunque per $\mathcal{P}_x(f) = \frac{1}{4}\mathcal{P}_{\underline{x}}(f - f_0) + \frac{1}{4}\mathcal{P}_{\underline{x}}(-f - f_0)$ (eq. (11.20)). 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 pari* alla somma dei risultati individuali: $FM\left\{m_1(t) + m_2(t)\right\} \neq FM\{m_1(t)\} + FM\{m_2(t)\}$.

Ampiezza costante La circostanza che $\underline{x}(t) = ae^{j\alpha(t)}$ presenti un modulo costante pari ad a, indipendentemente dall'ampiezza del segnale modulante, è particolarmente utile qualora per $m(t)$ siano previste forti variazioni di dinamica, come ad es. nel caso del segnale FDM (pag. 343) utilizzato per trasmettere più canali telefonici[41]. 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 f_n) può variare di molto. Il segnale complessivo $y(t)$ viene dunque applicato all'ingresso di un modulatore FM e trasmesso a piena potenza, senza subire distorsioni di non linearità (vedi § 8.3 e 13.3.3).

12.3.1 Generazione di un segnale a modulazione angolare

Il metodo *più diretto* di generare un segnale FM è quello di utilizzare un VCO (introdotto al § 12.2.2.2), 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. D'altra parte, per effettuare una modulazione PM per la quale $\alpha(t) = k_\phi m(t)$ si può usare un modulatore FM a partire da una fase modulante $\alpha(t) = 2\pi k_f \int_{-\infty}^{t} m'(\tau)d\tau$, ponendo $m'(t) = \frac{1}{2\pi}\frac{k_\phi}{k_f}\frac{d}{dt}m(t)$. Un terzo metodo di modulazione è illustrato per un caso particolare, al § 12.4.6.

Entrambi i segnali FM e PM possono infine essere ottenuti mediante il modulatore in fase e quadratura (pag. 347) alimentato dalle c.a. di b.f. $x_c(t) = \cos\alpha(t)$ e $x_s(t) = \sin\alpha(t)$, come effettivamente accade in diversi casi di modulazione numerica, vedi il cap. 16.

[40] Si fa qui uso della espansione in serie di potenze dell'esponenziale: $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + ...$

[41] Un altro caso di multiplex FDM è quello del downlink di un trasponder DVB-S, introdotto al § 25.3

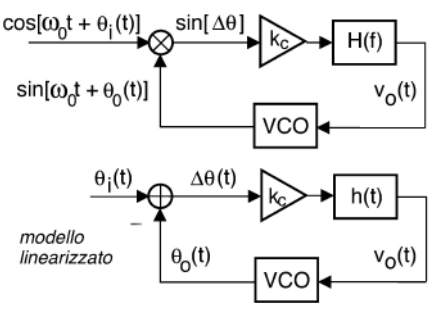
12.3.2 Ricezione di un segnale a modulazione angolare

Di base, si può utilizzare un demodulatore *coerente* in fase e quadratura (§ 11.2.4) per ottenere le c.a. di b.f. $x_c(t)$ ed $x_s(t)$ a partire dal segnale modulato, e da queste ricavare la fase modulata $\alpha(t)$ applicando la seconda eq. (11.9), ovvero $\alpha(t) = \arctan 2(x_s, x_c)$ ([42]), ottenendo infine $m(t)$ invertendo le relazioni di tab. 12.2. Una soluzione del genere è tuttavia possibile solo nell'ambito di una implementazione *numerica*, a causa della difficoltà realizzativa di un dispositivo circuitale che presenti esattamente la relazione non lineare di tipo arcotangente. Illustriamo quindi i due metodi più comunemente usati nel mondo *analogico*.

12.3.2.1 Ricevitore a PLL

Al § 12.2.2.2 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, realizzando al contempo la funzione desiderata.

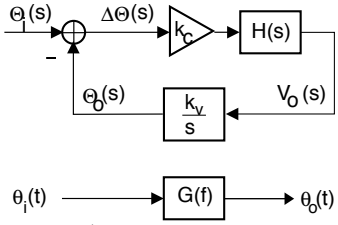
La figura a lato riporta lo schema generale di un PLL, in cui il VCO genera un segnale pari a $\sin(\omega_0 t + \theta_o(t))$, con $\theta_o(t) = k_v \int_{-\infty}^{t} v_o(\tau) d\tau$, mentre il segnale ricevuto ha la forma $x(t) = \cos(\omega_0 t + \theta_i(t))$. Lo schema può essere analizzato con i metodi dei controlli automatici, in quanto rappresenta un sistema che tenta di mantenere nullo l'errore $\sin \Delta\theta$, con $\Delta\theta(t) = \theta_i(t) - \theta_o(t)$ (vedi § 12.2.2.2); tale analisi si basa sulla *linearizzazione* $\sin \Delta\theta \simeq \Delta\theta$, valida per $\Delta\theta$ piccolo. In tal caso l'analisi di Laplace[43] permette (vedi fig. sotto) di scrivere la relazione



$$\Theta_o(s) = \frac{k_c k_v H(s)}{s + k_c k_v H(s)} \Theta_i(s) \tag{12.14}$$

che consente di esprimere $\theta_o(t)$ (fase del VCO) come la versione *filtrata* della fase della portante modulata $\theta_i(t)$, da parte della risposta in frequenza *ad anello chiuso*

$$G(f) = \left. \frac{k_c k_v H(s)}{s + k_c k_v H(s)} \right|_{s=j2\pi f}$$



Ricordando che la fase $\theta_o(t)$ del VCO corrisponde a $\theta_o(t) = k_v \int_{-\infty}^{t} v_o(\tau) d\tau$, possiamo constatare come il segnale $v_o(t)$ al suo ingresso (ovvero l'uscita del filtro di loop $H(s)$) corrisponda alla ricostruzione del messaggio modulante $m(t)$ nel caso di modulazione FM! Pertanto, il segnale $v_o(t)$ realizza la funzione di demodulazione di frequenza.

[42] Ricordiamo che arctan 2 restituisce un angolo compreso nell'intervallo $(-\pi, \pi)$ anziché $(-\pi/2, \pi/2)$.

[43] La (12.14) dà luogo ad una funzione di trasferimento ad anello chiuso il cui ordine dipende da come è realizzato il filtro passa basso. Per approfondimenti, vedi `https://it.wikipedia.org/wiki/Phase-locked_loop`.

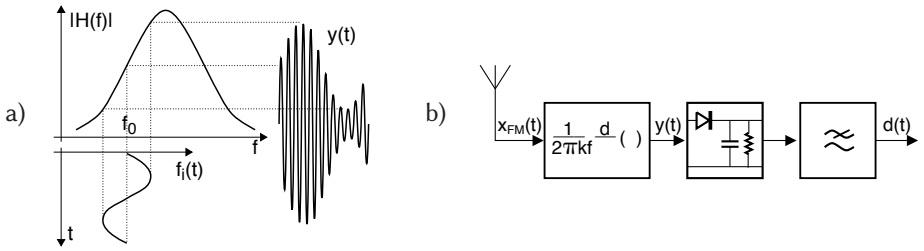


Figura 12.3: a) - conversione FM-AM; b) - schema del demodulatore a discriminatore

12.3.2.2 Ricevitore a discriminatore

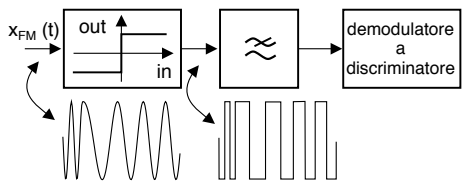
Questa seconda architettura di demodulatore di frequenza si basa su di un fenomeno detto *conversione FM-AM*, di cui in figura 12.3-a) è mostrato il principio di funzionamento più semplice, e noto come *rivelatore a pendenza* (SLOPE DETECTOR).

Un circuito risonante *accordato* ad un frequenza maggiore di f_0 realizza una risposta in frequenza $H(f)$ il cui modulo aumenta in maniera pressoché lineare nella banda di segnale, simulando così l'*effetto di una derivata* (vedi § 3.6). La figura mostra come, al variare della frequenza istantanea $f_i(t) = f_o + k_f m(t)$, l'ampiezza del segnale (passabanda) $y(t)$ uscente dal derivatore *vari* in misura del valore (tempo variante) di $|H(f)|_{f=f_i(t)}$, ottenendo così un segnale modulato *in ampiezz*a dalle stesse variazioni di $f_i(t)$, ovvero di $m(t)$. La figura 12.3-b) mostra quindi uno schema che utilizza il fenomeno descritto per ricostruire il segnale modulante $m(t)$ a partire da $y(t)$ mediante un semplice demodulatore di inviluppo (§ 12.2.5). Svolgendo infatti i passaggi, il segnale uscente dal derivatore risulta pari a[44]

$$\begin{aligned} y(t) &= \frac{1}{2\pi k_f}\frac{d}{dt} a\cos\left(2\pi f_0 t + 2\pi k_f \int_{-\infty}^{t} m(\tau)d\tau\right) = \\ &= \frac{1}{2\pi k_f}\left(2\pi f_0 + 2\pi k_f m(t)\right) a\sin\left(2\pi f_0 t + 2\pi k_f \int_{-\infty}^{t} m(\tau)d\tau\right) \end{aligned}$$

che corrisponde ad un segnale modulato sia angolarmente che in ampiezza, ed in particolare la cui ampiezza risulta $a(t) = a\left(f_0/k_f + m(t)\right)$. Pertanto con una scelta opportuna[45] di f_0/k_f la modulazione di ampiezza è riconducibile al caso BLD-PI (§ 12.1.1.2), e quindi il messaggio $m(t)$ può essere recuperato mediante un demodulatore d'inviluppo (§ 12.2.5).

Il risultato ottenuto è valido purché il segnale modulato $x(t)$ sia esso stesso privo di variazioni di ampiezza: per questo motivo il derivatore è spesso preceduto da un blocco *squadratore*, che produce una versione, appunto, "squadrata" del segnale ricevuto e quindi priva di modulazione di ampiezza. Essendo



[44] La derivata di $\cos[\alpha(t)]$ è pari a $-\sin[\alpha(t)]\cdot\alpha'(t)$, ma il segno $-$ è ininfluente ai fini dell'elaborazione successiva.

[45] L'utilizzo del demodulatore inviluppo è possibile solo nel caso di una modulazione a portante intera, ovvero per cui $\frac{f_0}{k_f} + m(t) > 0$ per $\forall t$, e dunque è necessario che risulti $k_f < \frac{f_0}{\max_t\{|m(t)|\}}$.

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 mediante un filtro passa basso posto a valle dello squadratore.

12.3.3 Densità spettrale di segnali a modulazione angolare

Come già osservato a pag. 382, la relazione (12.13) che esprime l'inviluppo complesso di un segnale modulato angolarmente nei termini di una serie di potenze

$$\underline{x}(t) = ae^{j\alpha(t)} = a\sum_{n=0}^{\infty}\frac{[j\alpha(t)]^n}{n!} \tag{12.15}$$

non può essere utilizzata in modo diretto per ottenere quella dello spettro di densità di potenza $\mathcal{P}_x(f)$ del segnale modulato in funzione di una generica fase modulante $\alpha(t)$; ciononostante, la (12.15) costituisce comunque un punto di partenza per analizzare altri aspetti della situazione.

Osserviamo innanzitutto che, essendo $|\underline{x}(t)| = a$, la sua potenza totale ha sempre valore $\mathcal{P}_{\underline{x}} = a^2$, indipendentemente da $\alpha(t)$, e dunque[46] $\mathcal{P}_x = \frac{a^2}{2}$. Inoltre, la presenza nella (12.15) di potenze di $\alpha(t)$ di qualunque ordine sembrerebbe indicare che $\mathcal{P}_{\underline{x}}(f)$ abbia una banda infinita: in realtà la presenza dei fattoriali a denominatore fa sì che la serie possa essere troncata ad un certo ordine $N < \infty$, e dunque $\underline{x}(t)$ sia da considerare limitato in banda.

Per speculare sull'influenza di $\alpha(t)$ sul segnale modulato, notiamo che quanto più $|\alpha(t)|$ è piccolo, tanto prima la (12.15) può essere troncata con errori trascurabili; se poi $\alpha(t)$ si mantiene sempre *molto piccolo*, ci si può limitare al solo primo termine ($n = 1$), dando così luogo ad un comportamento *lineare*, dato che in tal caso si ottiene $\underline{x}(t) = a(1 + j\alpha(t))$. Se viceversa $\alpha(t)$ assume valori *molto elevati*, e quindi (12.15) comprende parecchi termini, subentra un secondo aspetto peculiare dell'FM, indicato come *conversione ampiezza → 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 direttamente proporzionali alle ampiezze di $m(t)$, e quindi *l'andamento* della densità di potenza $\mathcal{P}_x(f)$ in funzione di f *riflette* quello (funzione di m) della densità di probabilità di $p_M(m)$ che descrive le ampiezze di $m(t)$, come torneremo ad approfondire al § 12.3.3.3. Infine, a valori intermedi della dinamica di $\alpha(t)$ corrisponde una $\mathcal{P}_x(f)$ che sarà una *via di mezzo* tra i due casi estremi discussi, e che pertanto possono essere pensati come *casi limite* tra cui porre la densità di potenza effettiva.

Dato che la natura non lineare della modulazione angolare rende necessario studiare ogni caso individualmente, il calcolo di $\mathcal{P}_x(f)$ viene svolto nel seguito per il caso *particolare* di un segnale $m(t)$ sinusoidale, considerando le due possibilità estreme

[46] Da un lato, $\frac{a^2}{2}$ è banalmente la potenza della portante di ampiezza a. Da una altro punto vista, lo stesso risultato si ottiene a partire dalla $\mathcal{P}_x(f) = \frac{1}{4}\left(\mathcal{P}_{\underline{x}}(f - f_0) + \mathcal{P}_{\underline{x}}(-f - f_0)\right)$ (eq. 11.20), da cui mediante integrazione in frequenza otteniamo $\mathcal{P}_x = \frac{1}{4}2\mathcal{P}_{\underline{x}} = \frac{a^2}{2}$.

di $\alpha(t)$ molto piccolo o molto grande, ed i risultati vengono quindi estrapolati per approssimare altre situazioni.

12.3.3.1 Segnale modulante sinusoidale

Per questo calcolo esprimiamo il segnale modulante come $m(t) = \cos(2\pi wt)$, con w che indica la frequenza di modulazione. L'espressione della fase modulante $\alpha(t)$ e della relativa frequenza istantanea $f_i(t) = f_0 + \frac{1}{2\pi}\frac{d}{dt}\alpha(t)$ per il segnale modulato $x(t) = a\cos(2\pi f_0 t + \alpha(t))$ risulta allora quella riportata alla tabella seguente, per i casi di modulazione di fase e di frequenza, assieme all'espressione della massima deviazione di fase $\Delta\alpha = \max\{|\alpha(t)|\}$ e di frequenza $\Delta f = \max\{|f_i(t) - f_0|\}$.

	$\alpha(t)$	$f_i(t)$	$\Delta\alpha$	Δf
PM	$k_\phi m(t) = k_\phi \cos(2\pi wt)$	$f_0 - wk_\phi \sin(2\pi wt)$	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

Osserviamo che in entrambi i casi sia $\alpha(t)$ che $f_i(t)$ variano sinusoidalmente con frequenza 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'inviluppo complesso come[47]

$$\underline{x}(t) = ae^{j\alpha(t)} = ae^{j\beta\sin(2\pi wt)}$$

Notiamo ora che $\underline{x}(t)$ è periodico di periodo $\frac{1}{w}$, e dunque per esso vale lo sviluppo in serie di Fourier (§ 2.2) $\underline{x}(t) = a\sum_{n=-\infty}^{\infty} X_n e^{j2\pi nwt}$, i cui coefficienti sono definiti come

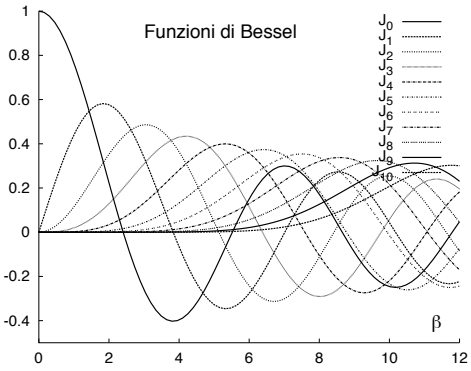
$$X_n = w\int_{-\frac{1}{2w}}^{\frac{1}{2w}} e^{j\beta\sin(2\pi wt)} e^{-j2\pi nwt} dt = \mathcal{J}_n(\beta) \tag{12.16}$$

ovvero sono pari[48] alle *funzioni di Bessel* del *primo tipo*, ordine n ed argomento β. Queste hanno l'andamento mostrato alla figura 12.4, in cui sono riportate anche le proprietà che le caratterizzano. I valori di X_n si ottengono quindi tracciando una linea verticale nel diagramma di figura in corrispondenza del valore adottato per β, e individuando il valore di ciascuna $\mathcal{J}_n$ per quel β.

Osserviamo ora che in presenza di un valore di β elevato, in base all'ultima proprietà mostrata in fig. 12.4 ovvero che $\mathcal{J}_n(\beta) \simeq 0$ con $n > \beta$ se $\beta \gg 1$, le funzioni di Bessel di ordine $n > \beta$ sono 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

[47] Si è sostituito *cos* con *sin* nel caso PM per omogeneità di formulazione, senza alterare la sostanza delle cose.

[48] Le *funzioni di Bessel* del primo tipo, ordine n ed argomento β sono definite come $\mathcal{J}_n(\beta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j(\beta\sin x - nx)} dx$, riconducibili alla (12.16) mediante un cambio di variabile.



Proprietà delle funzioni di Bessel

- $\mathcal{J}_n(\beta)$ è reale con $\forall n, \beta$
- $\mathcal{J}_n(\beta) = \mathcal{J}_{-n}(\beta)$ con n pari
- $\mathcal{J}_n(\beta) = -\mathcal{J}_{-n}(\beta)$ con n dispari
- $\sum_{n=-\infty}^{+\infty} \mathcal{J}_n^2(\beta) = 1$
- $\mathcal{J}_n(\beta) \simeq 0$ con $n > \beta$ se $\beta \gg 1$

Figura 12.4: Andamento delle funzioni di Bessel del primo tipo e relative proprietà

$$\underline{x}_{FM}(t) \simeq a \sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta)\, \mathrm{e}^{j2\pi n w t} \quad \stackrel{\mathcal{F}}{\Longrightarrow} \quad \underline{X}_{FM}(f) \simeq a \sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta)\, \delta(f - nw)$$

Pertanto il segnale modulato $x(t) = \Re\left\{\underline{x}(t)\mathrm{e}^{j\omega_0 t}\right\}$ risulta pari a[49]

$$x(t) \simeq a \sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta) \cos 2\pi(f_0 + nw)t \tag{12.17}$$

ossia è costituito da $2\beta + 1$ cosinusoidi a frequenza $f_0 \pm nw$ centrate *attorno* ad f_0, e dunque il relativo spettro di densità di potenza ha espressione[50]

$$\mathcal{P}_x(f) \simeq \frac{a^2}{4} \sum_{n=-\beta}^{\beta} |\mathcal{J}_n(\beta)|^2 \left[\delta(f - f_0 - nw) + \delta(f + f_0 + nw)\right] \tag{12.18}$$

ed è formato da impulsi centrati a frequenza $f = \pm f_0 \pm nw$.

Osserviamo che in base alla proprietà $\sum_{n=-\infty}^{+\infty} \mathcal{J}_n^2(\beta) = 1$, estendendo la somma in (12.18) per $-\infty > n > \infty$ ed integrando su f si ottiene un risultato già noto, ovvero la potenza totale $\mathcal{P}_x$ eguaglia quella della portante non modulata, pari a $\frac{a^2}{2}$, indipendentemente dall'indice di modulazione β.

Modulazione a basso indice Come anticipato al § 12.3.3 e mostrato in fig. 12.4, qualora $\beta \ll 1$ le funzioni di Bessel $\mathcal{J}_n(\beta)$ con $n > 1$ presentano valori che possono essere trascurati. Pertanto in tal caso $x(t)$ occupa una banda $B \simeq 2w$, in modo del tutto simile all'AM-BLD.

[49] Infatti

$$\underline{x}(t)\mathrm{e}^{j\omega_0 t} = \sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta)\, \mathrm{e}^{j2\pi n w t} \mathrm{e}^{j2\pi f_0 t} = \sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta)\, \mathrm{e}^{j2\pi(nw+f_0)t}$$

la cui parte reale è appunto pari a $\sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta) \cos 2\pi(nw + f_0)t$

[50] Infatti ad ogni termine $\mathcal{J}_n(\beta) \cos 2\pi(f_0 + nw)t$ della (12.17) corrisponde una densità di potenza

$$\mathcal{P}(f) = \frac{\mathcal{J}_n^2(\beta)}{4}\left[\delta(f - f_0 - nw) + \delta(f + f_0 + nw)\right]$$

e la potenza della somma è pari alla somma delle potenze, in virtù della ortogonalità tra cosinusoidi.

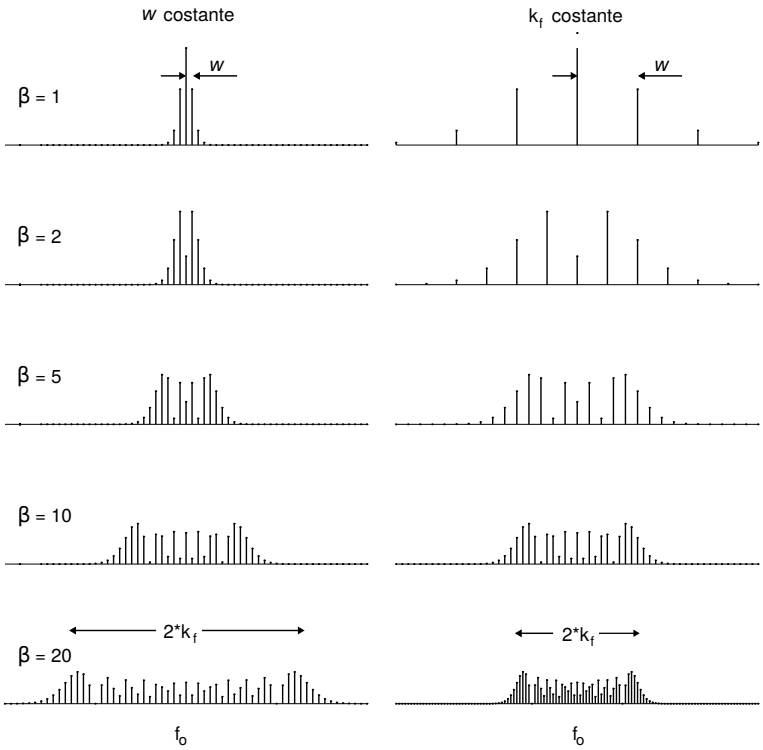


Figura 12.5: Spettro di ampiezza $|X(f)|$ per segnale FM a modulazione sinusoidale. Solo la riga $\beta = 10$ mantiene la stessa scala di frequenze a sin. come a ds.

Modulazione ad alto indice All'aumentare di $\beta = \frac{k_f}{w}$ nella (12.18) aumenta il numero dei termini rilevanti, e la fig. 12.5 mostra l'aspetto di $|X(f)|$ per $f > 0$ ([51]) calcolato per valori di β crescenti, mantenendo fisso w (a sinistra) oppure k_f (a destra). Osserviamo che

- mantenendo w fisso ed aumentando k_f, ovvero l'ampiezza di $m(t)$, il numero di righe spettrali a frequenza $f_0 \pm nw$ aumenta, occupando una banda crescente, che per β molto grande si estende da $f_0 - \beta w$ a $f_0 + \beta w$, dato che $\mathcal{J}_n(\beta) \simeq 0$ per $n > \beta$;
- mantenendo k_f fisso e diminuendo w, ossia la frequenza modulante, la banda occupata tende a ridursi, mentre le nuove righe spettrali a frequenza $f_0 \pm nw$ si infittiscono. Per $w \to 0$ si ha $\beta \to \infty$ mentre la spaziatura tra le righe spettrali tende ad annullarsi, producendo una densità spettrale praticamente continua.

Notiamo che in entrambi i casi all'aumentare di β la banda occupata a frequenze positive è bene approssimata dal valore $B = 2\beta w = 2\frac{k_f}{w} \cdot w = 2k_f$, ossia pari al doppio della massima deviazione di frequenza *istantanea* Δf, vedi la tabella a pag. 386.

[51] Ovvero, la fig. 12.5 mostra $\left|X^+(f)\right| = \frac{1}{2}\underline{X}(f - f_0) = \frac{a}{2}\sum_{n=-\beta}^{\beta} |\mathcal{J}_n(\beta)|\,\delta(f - f_0 - nw)$

12.3.3.2 Regola di Carson

Come appena discusso la modulazione FM da parte di un tono sinusoidale a frequenza w produce un segnale modulato $x(t)$ la cui banda varia tra $2w$ e $2k_f$ nei casi di indice di modulazione β rispettivamente basso od alto. Una formula in grado di esprimere questo tipo di relazione è

$$B_C \simeq 2\left(k_f + w\right) = 2w\left(\beta + 1\right) \tag{12.19}$$

nota come *regola di Carson*[52], in grado di tener conto di entrambi i fattori che concorrono alla determinazione della banda, e che fornisce i valori esatti[53] sia per $\beta \ll 1$, che per $\beta \to \infty$, in entrambi i casi in cui sia k_f ad aumentare, o w a diminuire.

Sebbene questo risultato si riferisca al caso di $m(t) = \cos(2\pi wt)$, la (12.19) viene spesso adottato come una buona approssimazione anche per altri tipi di segnali modulanti, come illustrato appresso.

Modulazione multitono Si riferisce ad un segnale FM per il quale $m(t)$ è la combinazione di più sinusoidi a frequenza w_i ed ampiezza k_f^i. In tal caso la trattazione matematica si complica, e perviene al risultato che nel segnale modulato $x(t)$ compaiono, oltre a componenti spettrali già analizzate e relative a ciascuna w_i, centrate in f_0 e spaziate da multipli di w_i, anche componenti spaziate a frequenze somma e differenza delle combinazioni dei multipli delle w_i. In questo caso l'occupazione di banda è approssimata riscrivendo la (12.19) come $B_C = 2w_M\left(\beta_M + 1\right)$ in cui $w_M = \max_i\{w_i\}$ è la più grande delle frequenze modulanti e $\beta_M = {}^{\Delta f_M}/_{w_M}$ è l'indice di modulazione *equivalente* per questo caso, avendo definito $\Delta f_M = \sum_i k_f^i$ come la massima deviazione della frequenza istantanea Δf.

Modulazione per segnali qualsiasi Nel caso di un segnale modulante generico, limitato in banda tra $-W$ e W, e che produce da una deviazione massima della frequenza istantanea $\Delta f = k_f \cdot \max\{|m(t)|\}$, l'occupazione di banda è approssimata riscrivendo nuovamente la regola di Carson come $B_C \simeq 2W\left(\beta_s + 1\right)$ con $\beta_s = \frac{\Delta f}{W}$.

Allargamento spettrale L'applicazione della regola di Carson mostra che la banda occupata dal segnale modulato può risultare $\beta_s + 1$ volte più estesa di quella W del segnale modulante, un comportamento del tutto nuovo rispetto a quanto avviene nel caso dell'AM. Nonostante possa sembrare un aspetto negativo, al § 14.4.2 si mostra come una maggiore occupazione di banda consenta di migliorare l'SNR dopo demodulazione, superando in tal modo le prestazioni ottenibili nel caso AM. Al contrario, se $\beta_s \ll 1$ il comportamento si avvicina molto a quello lineare (vedi appendice 12.4.6).

[52] J. R. Carson fu uno dei primi a studiare le tecniche di modulazione negli anni '20, vedi ad es. `http://en.wikipedia.org/wiki/John_Renshaw_Carson`

[53] Nel caso di modulazione sinusoidale ad alto indice la (12.19) esprime la banda entro cui è contenuto il 98% della potenza del segnale modulato. Per indici $2 < \beta < 10$ ne fornisce invece una stima *per difetto*, ed una approssimazione più corretta è $B_C \simeq 2w\left(\beta + 2\right)$.

12.2.3.3 Densità spettrale per FM ad alto indice

Riprendiamo il ragionamento iniziato al § 12.3.3 e relativo a come la densità di probabilità $p_M(m)$ del segnale modulante si rifletta sulla densità di potenza $\mathcal{P}_x(f)$ del corrispondente segnale FM nel caso di modulazione ad alto indice, ovvero qualora $\beta_s \gg 1$[54]. In questo caso, la frazione di potenza totale $\mathcal{P}_x = \frac{a^2}{2}$ del segnale FM che si distribuisce tra le frequenze f_1 ed f_2 è pari alla frazione di tempo per cui la frequenza istantanea $f_i(t) = f_0 + k_f m(t)$ permane nello stesso intervallo, ovvero pari alla frazione di tempo per cui il segnale modulante $m(t)$ assume valori compresi tra $m_1 = \frac{f_1 - f_0}{k_f}$ e $m_2 = \frac{f_2 - f_0}{k_f}$. Tale frazione è proprio pari alla probabilità di trovare $m_1 \leq m(t) \leq m_2$, ovvero $Prob\{m_1 \leq m(t) \leq m_2\} = \int_{m_1}^{m_2} p_M(m)\, dm$, dove $p_M(m)$ è la densità di probabilità che descrive il processo modulante.

Si può affermare dunque che qualora si generi un segnale FM ad alto indice a 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. Tale conclusione mantiene validità purché $\beta \gg 1$; nel caso contrario, sono validi i ragionamenti sviluppati al § 12.3.3.5.

Esempio Un processo $m(t)$ limitato in banda $\pm W$ e con con d.d.p. uniforme $p_M(m) = \frac{1}{\Delta_M}\text{rect}_{\Delta_M}(m)$ modula ad alto indice una portante a frequenza f_0 ed ampiezza a, con un coefficiente di modulazione k_f. Determinare la $\mathcal{P}_x(f)$ del segnale modulato.

Notiamo 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 ottiene[55]:

$$\mathcal{P}_x(f) = \frac{a^2}{4\Delta_M k_f}\left[\text{rect}_{\Delta_M k_f}(f - f_0) + \text{rect}_{\Delta_M k_f}(f + f_0)\right]$$

Esempio 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. 163), e dunque per $\beta \gg 1$ risulterà $\mathcal{P}_{\underline{x}}(f) = \frac{a^2}{1-(f/k_f)^2}$, con l'andamento rappresentato dalla fig. 6.4 a pag. 163, ovvero il quadrato dell'andamento a cui tendono (per $\beta \to \infty$) i grafici in basso

[54] Sinceramente non ho afferrato appieno il motivo di questa limitazione. Quel che ho trovato esprime che "ciò equivale ad avere $\Delta f \gg W$, e quindi comporta il rispetto di una condizione detta *approssimazione quasi stazionaria*" e fa riferimento ad uno studio di H.E. Rowe del 1965. La tesi è che in tal caso la frequenza istantanea varia *lentamente* rispetto al periodo $1/W$ della massima frequenza modulante, e dunque il segnale modulato osservato per un breve periodo è approssimato ad una sinusoide non modulata ed a frequenza costante pari a $f_i(t) = f_0 + k_f\gamma$ in cui γ è il valore di $m(t)$ praticamente costante nel periodo di osservazione. A me sembra che perché ciò avvenga, sia sufficiente che $f_0 \gg \Delta f$. Ma forse lo capirò con il tempo.

[55] Volendo 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 12.3.3.4) 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 le particolarità di $p_M(m)$ uniforme, in questo caso è da preferire il primo risultato.

di fig. 12.5. Pertanto le righe spettrali, addensandosi, tendono a disporsi in accordo all'andamento della densità di probabilità $p_M(m)$ del processo armonico.

12.3.3.4 Indice di modulazione per processi

Nel caso in cui non sia nota la d.d.p. del processo modulante, ma solo la sua potenza, oppure qualora non sussista la condizione di alto indice $\beta_s \gg 1$, oppure ancora non sia possibile definire il valore della massima deviazione di frequenza $\Delta f = k_f \cdot \max\{|m(t)|\}$ come ad esempio per $m(t)$ gaussiano, allora l'occupazione di banda può ancora essere approssimata mediante un'ultima variazione della regola di Carson, che viene ora applicata dopo aver definito un indice di modulazione β_p *per processi* come

$$\beta_p = \begin{cases} \sigma_\alpha & \text{(PM)} \\ \frac{\sigma_f}{W} & \text{(FM)} \end{cases}$$

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 istantanea[56], e $\sigma_\alpha = k_\phi\sqrt{\mathcal{P}_m}$ è la deviazione standard della fase modulante[57]. L'applicazione della regola di Carson con il nuovo valore di β_p fornisce 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 25.2).

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.

12.3.3.5 Densità spettrale per FM a basso indice

In questo caso si suppone l'indice di modulazione β piccolo a sufficienza, in modo che lo sviluppo in serie dell'inviluppo complesso del segnale modulato possa essere arrestato ai primi termini. Sotto opportune ipotesi, si può mostrare che vale il risultato

$$\mathcal{P}_{\underline{x}}(f) \simeq a^2 \mathrm{e}^{-\sigma_\alpha^2}\left[\delta(f) + \mathcal{P}_\alpha(f) + \frac{1}{2}\mathcal{P}_\alpha(f) * \mathcal{P}_\alpha(f) + \frac{1}{3!}\mathcal{P}_\alpha(f) * \mathcal{P}_\alpha(f) * \mathcal{P}_\alpha(f) + \cdots\right]$$

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	$k_\phi^2 \mathcal{P}_m(f)$	$k_\phi^2 P_m$
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 $\mathrm{e}^{-\sigma_\alpha^2}$ diminuisce, riducendo la concentrazione di potenza a frequenza portante. Dato che risulta comunque $P_{\underline{x}} = a^2$, la potenza residua si distribuisce sugli altri termini,

[56] Infatti, 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.

[57] Come sopra, partendo dalla relazione $\alpha(t) = k_\phi m(t)$.

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 12.4.6 è illustrata una tecnica di modulazione per segnali FM modulati a basso indice.

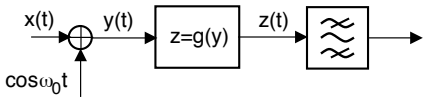
12.4 Appendici

12.4.1 Mixer mediante non linearità

Illustriamo un modo di realizzare il dispositivo che effettua la funzione di moltiplicazione tra un segnale modulante ed una portante. Sebbene esistano schemi circuitali capaci di realizzare esattamente il prodotto tra due segnali analogici[58], l'approccio che segue è molto semplice, e fa uso di un sommatore, un oscillatore, un dispositivo non lineare, ed un filtro passa-banda, come mostrato in figura.

Il dispositivo non lineare ha una caratteristica ingresso-uscita del tipo

$$z = a_1 y + a_2 y^2 + a_3 y^3 + \cdots$$



e quando in ingresso viene applicata la somma di due segnali $x(t) + \cos\omega_0 t$, produce

$$z(t) = a_1 (x(t) + \cos\omega_0 t) + a_2(x^2(t) + \cos^2\omega_0 t + 2x(t)\cos\omega_0 t) + a_3(\cdots) + \ldots$$

da cui, osservando che i termini $\cos^n \omega_0 t$ sono relativi a componenti centrate a frequenza nf_0[59], il filtro passa banda può estrarre il termine $x(t)\cos\omega_0 t$ a cui siamo interessati.

Da un punto di vista circuitale[60] il dispositivo non lineare può essere costituito da un semplice diodo, per il quale la corrente che lo attraversa è espressa in funzione della tensione V ai suoi capi in base all'espressione $I = I_s\left(e^{V/\alpha} - 1\right)$ in cui I_s ed α sono delle costanti; per piccoli valori di V la relativa espansione in serie di Taylor permette infatti di scrivere $e^{V/\alpha} - 1 \simeq {}^{V}/_{\alpha} + \frac{({}^{V}/_{\alpha})^2}{2}$ e dunque, ponendo $V = x(t) + \cos\omega_0 t$ si ottiene il risultato anticipato.

12.4.2 Mixer a commutazione

Non è strettamente necessario disporre di un oscillatore sinusoidale per realizzare il prodotto di un segnale con una portante: è sufficiente un'onda quadra ed un filtro! Infatti, un qualunque segnale periodico

$$y(t) = g(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT)$$

[58] Vedi ad es. `https://en.wikipedia.org/wiki/Ring_modulation` e `https://en.wikipedia.org/wiki/Gilbert_cell`

[59] Infatti la formula di *de Moivre* asserisce che $(\cos\alpha + j\sin\alpha)^n = \cos n\alpha + j\sin n\alpha$, come confermato anche dalla formula di Eulero $(\cos\alpha + j\sin\alpha)^n = \left(e^{j\alpha}\right)^n = e^{jn\alpha} = \cos n\alpha + j\sin n\alpha$.

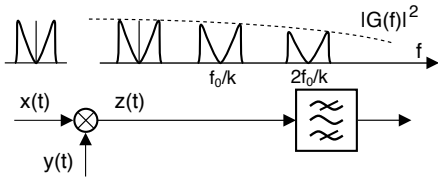
[60] Per un esempio di progettazione elettronica che realizza le funzioni descritte, si veda ad esempio `https://digilander.libero.it/ingcasanof/quinta/misure/modulatore_am_con_fet/modulatore_am%20con%20fet.htm`

di periodo $T = k/f_0$ (con k intero) possiede una densità di potenza[61]

$$\mathcal{P}_y(f) = \frac{|G(f)|^2}{T} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{k} f_0\right) \tag{12.20}$$

Il prodotto di tale segnale per $x(t)$ produce un segnale $z(t)$ con densità di potenza[62]

$$\mathcal{P}_z(f) = \mathcal{P}_x(f) * \mathcal{P}_y(f) = \frac{|G(f)|^2}{T} \sum_{n=-\infty}^{\infty} \mathcal{P}_x\left(f - \frac{n}{k} f_0\right) \tag{12.21}$$



Pertanto, il desiderato spettro di potenza si ottiene inserendo dopo il moltiplicatore un filtro passa banda centrato su una delle armoniche a frequenza $\frac{n}{k} f_0$ di $y(t)$, ovvero su una delle repliche spettrali che compongono $\mathcal{P}_z(f)$. L'inviluppo mostrato in figura è relativo ad una scelta per $g(t)$ del tipo $g(t) = rect_\tau(t)$ con τ sufficientemente minore di k/f_0 e scelto in modo opportuno[63], in modo che se la banda di $x(t)$ è sufficientemente ridotta rispetto a f_0/k l'entità della distorsione lineare di ampiezza può essere considerata trascurabile. Lo stesso dispositivo può essere usato anche per i moltiplicatori di demodulazione: in tal caso, il filtro da usare sarà un passa basso.

Dato che un punto vista circuitale il prodotto per un'onda quadra è assimilabile ad un interruttore che si apre e si chiude, un dispositivo del genere viene detto *switching mixer*.

12.4.3 Sintesi di frequenza con PLL ed oscillatore a cristallo

Questa sottosezione illustra una tecnica per generare una portante di modulazione (oppure eterodina) che sia *stabile* in frequenza. Dato che tutti questi elementi sono ormai integrati in un unico *chip*, possiamo anche visitare quelli a catalogo di un produttore *a caso*[64]*!* Essenzialmente il circuito PLL illustrato al § 12.2.2.2 non è in grado di generare una portante *di modulazione* stabile a sufficienza, in quanto ottenuta a partire da oscillatori realizzati mediante circuiti di tipo analogico[65], la cui frequenza dipende anche dalla precisione dei valori dei componenti utilizzati e dalla temperatura di lavoro, oltre a presentare una variabilità che aumenta al crescere della frequenza di

[61] La (12.20) si ottiene applicando l'espressione per i coefficienti Y_n dello sviluppo in serie di $y(t)$ in funzione dei campioni della trasformata di un suo periodo $G(f)|_{f=n/T}$ data dalla (3.4) ovvero $Y_n = \frac{1}{T} G\left(\frac{n}{T}\right)$, a quella $\mathcal{P}_y(f) = \sum_{n=-\infty}^{\infty} |Y_n|^2 \delta(f - n/T)$ fornita dalla (7.11) per la densità di potenza di un segnale periodico.

[62] Infatti in base alla § 7.5.3 ad un prodotto tra processi statisticamente indipendenti corrisponde la convoluzione in frequenza delle relative densità spettrali, e la (12.21) è conseguenza della convoluzione con gli impulsi presenti nella (12.20).

[63] Dato che $G(f)$ si annulla per $f = m/\tau$, se scegliessimo $\tau = 1/hf_0$ avremmo $G(f) = 0$ per $f = mhf_0$ impedendo il funzionamento del circuito per qualche armonica di f_0/k. In particolare, scegliendo $h = 1$ lo schema sarebbe del tutto inutilizzabile!

[64] Vedi ad es. `https://www.analog.com/en/product-category/fractional-n-pll.html`

[65] Vedi ad es. `https://it.wikipedia.org/wiki/Oscillatore`

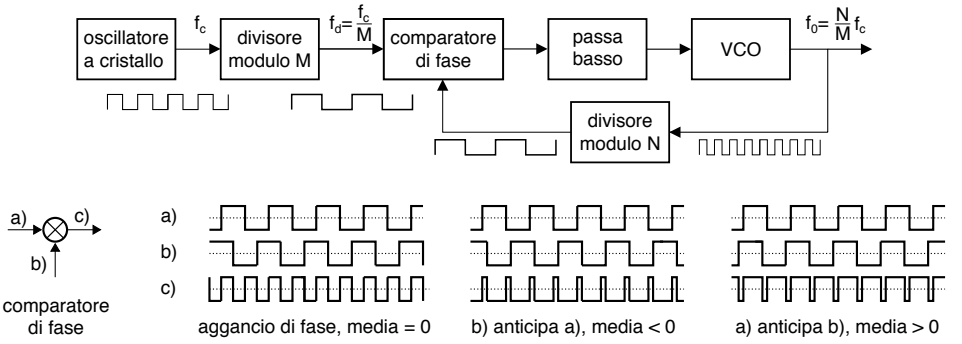


Figura 12.6: Sintesi di frequenza con PLL ed oscillatore a cristallo (sopra), e funzionamento del comparatore di fase per forme d'onda rettangolari (sotto)

oscillazione. Al contrario, gli oscillatori basati su di un cristallo[66] sono molto più stabili e precisi, tipicamente dell'ordine di ±20 PPM[67], anche se il loro costo, disponibilità e fragilità peggiora all'aumentare della frequenza, arrivando in pratica a qualche decina di MHz.

Per generare portanti più elevate si utilizza la frequenza dell'oscillatore a cristallo come un *riferimento* a cui far *agganciare* un circuito PLL modificato come in figura 12.6, in cui le forme d'onda sono di tipo rettangolare in modo da poterne ottenere *di nuove* con periodo multiplo di quello di partenza (e dunque frequenza pari ad un suo sottomultiplo) mediante l'uso di un circuito divisore modulo N[68]. In particolare, il moltiplicatore è ora detto *comparatore di fase*[69] ed opera (per forme d'onda bipolari) come mostrato nella parte inferiore di fig. 12.6, producendo un'onda bipolare a *valor medio* positivo o negativo a seconda se l'ingresso a sinistra sia in anticipo od in ritardo rispetto a quello proveniente dal basso, valor medio che è proporzionale allo slittamento in eccesso rispetto alla condizione di uno sfasamento pari ad un quarto di periodo.

La frequenza f_0 generata dal VCO si aggancia quindi ad un valore N volte maggiore di quello f_d in ingresso al comparatore, in virtù del divisore per N posto a valle del VCO; essendo il divisore di tipo programmabile, la frequenza del segnale prodotto dal vco può essere modificata variando il valore di N. Allo stesso tempo, anche la frequenza f_c prodotta dall'oscillatore a cristallo viene divisa per un diverso numero M, in modo che la frequenza di uscita risulti pari a $f_0 = \frac{N}{M} f_c$. In tal modo il circuito può essere programmato per generare frequenze $\frac{N}{M}$ volte maggiori di quella prodotta dal cristallo, mantenendo la sua stessa precisione: ad esempio un cristallo con precisione di 20 PPM e frequenza $f_c = 10$ MHz, dopo aver scelto $M = 1$ ed $N = 20$, permette di ottenere $f_0 = 200$ MHz ±4 KHz.

[66] Vedi ad es. `https://it.wikipedia.org/wiki/Oscillatore_al_cristallo`

[67] Abbreviazione di *parti per milione*: 10 PPM equivalgono a 10 cicli ogni 10^6, ovvero un valore compreso tra 999.990 ed 1.000.010 per una frequenza *nominale* di 1 MHz.

[68] Vedi ad es. `http://studenti.fisica.unifi.it/~carla/appunti/2008-9/slides-7.pdf`

[69] Vedi ad es. `https://it.wikipedia.org/wiki/Comparatore_di_fase`

Moltiplicatore di frequenza Il medesimo schema può essere impiegato per produrre *più segnali* di clock tutti multipli di una comune velocità di partenza, ad esempio nell'ambito di reti logiche complesse, o all'interno di microprocessori.

12.4.3.1 Sintesi digitale diretta

Si tratta della possibilità di generare le portanti di modulazione e demodulazione per via completamente numerica, nella forma di una sequenza costituita da valori dei campioni della forma d'onda sinusoidale letti da una memoria che viene indirizzata da un contatore ciclico. L'argomento sarà approfondito in una prossima edizione.

12.4.4 Densità di potenza per segnali AM a banda laterale unica

Affrontiamo il problema di dimostrare che per un segnale AM-BLU a banda laterale *superiore* (§ 12.1.2) si ottiene

$$\mathcal{P}_x(f) = \mathcal{P}_{m^+}(f - f_0) + \mathcal{P}_{m^-}(f + f_0)$$

come asserito all'eq. (12.8). A tale proposito osserviamo che la (11.20) stabilisce

$$\mathcal{P}_x(f) = \tfrac{1}{4}\left(\mathcal{P}_{\underline{x}}(f - f_0) + \mathcal{P}_{\underline{x}}(-f - f_0)\right) \tag{12.22}$$

in cui $\mathcal{P}_{\underline{x}}(f) = \mathcal{F}\left\{\mathcal{R}_{\underline{x}}(\tau)\right\}$, e dato che nel nostro caso $\underline{x}(t) = m(t) + j\widehat{m}(t) = 2m^+(t)$ (vedi eq. (11.17)), otteniamo

$$\mathcal{R}_{\underline{x}}(\tau) = \left\langle \underline{x}(t),\, \underline{x}(t+\tau)\right\rangle = \left\langle 2m^+(t),\, 2m^+(t+\tau)\right\rangle = 4\mathcal{R}_{m^+}(\tau) \tag{12.23}$$

in cui si è fatto uso della notazione di prodotto scalare $\langle\, .\, ,\, .\,\rangle$ per generalizzare il calcolo dell'autocorrelazione sia al caso di segnale di potenza, sia a quello di un processo. Eseguendo ora la trasformata di (12.23) otteniamo

$$\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{m^+}(f) = 4\mathcal{P}_m(f)\left|H_{fp}(f)\right|^2$$

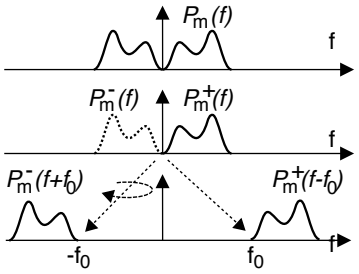
in cui $H_{fp}(f)$ è definito al § 11.2.6, e pertanto $\mathcal{P}_{\underline{x}}(f)$ esiste solamente sul semiasse delle frequenze positive. Dunque la (12.22) si riscrive considerando che per $f > 0$ si ha

$$\mathcal{P}_x(f)|_{f>0} = \frac{1}{4}\mathcal{P}_{\underline{x}}(f - f_0) = \mathcal{P}_{m^+}(f - f_0)$$

mentre per $f < 0$ risulta

$$\mathcal{P}_x(f)|_{f<0} = \mathcal{P}_{m^+}(-f - f_0) = \mathcal{P}_{m^-}(f + f_0)$$

come mostrato in figura.



12.4.5 Calcolo della potenza di un segnale AM BLU

Come anticipato in fondo al § 12.1.2, mostriamo che se

$$x_{BLU}(t) = \frac{k_a}{\sqrt{2}}\left(m(t)\cos\omega_0 t - \widehat{m}(t)\sin\omega_0 t\right) \tag{12.24}$$

allora $\mathcal{P}_x = \frac{k_a^2}{2}\mathcal{P}_m$. Possiamo innanzitutto scrivere che

$$\mathcal{P}_x = \mathcal{P}_{x^+} + \mathcal{P}_{x^-} = 2\mathcal{P}_{x^+} \tag{12.25}$$

in quanto le componenti a frequenza positiva e negativa di $x(t)$ sono ortogonali [70], 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'inviluppo 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 (12.25), fornisce $\mathcal{P}_x = 2\mathcal{P}_{x^+} = \frac{1}{2}\mathcal{P}_{\underline{x}}$.

Come sappiamo $\mathcal{P}_{\underline{x}} = \mathcal{R}_{\underline{x}}(0)$ in cui, nell'ipotesi di processo ergodico, $\mathcal{R}_{\underline{x}}(0)$ è l'autocorrelazione di un qualunque membro, ad es. proprio di (12.24), e dunque essendo in tal caso $\underline{x}(t) = \frac{k_a}{\sqrt{2}}[m(t) + j\widehat{m}(t)]$, si ottiene

$$\mathcal{P}_x = \frac{1}{2}\mathcal{P}_{\underline{x}} = \frac{1}{2}\mathcal{R}_{\underline{x}}(0) = \frac{1}{2}\left(\frac{k_a}{\sqrt{2}}\right)^2 [\mathcal{R}_{mm}(0) + \mathcal{R}_{\widehat{m}\widehat{m}}(0) + 2j\mathcal{R}_{m\widehat{m}}(0)]$$

Osserviamo ora che $\mathcal{R}_{m\widehat{m}}(0) = \int_{-\infty}^{\infty} m(t)\,\widehat{m}(t)\,dt = 0$ in quanto $m(t)$ ed $\widehat{m}(t)$ sono ortogonali; inoltre, $\mathcal{R}_{mm}(0) = \mathcal{P}_m = \mathcal{R}_{\widehat{m}\widehat{m}}(0)$ (non dimostrato). Pertanto si ottiene

$$\mathcal{P}_x = \frac{1}{2}\frac{k_a^2}{2}[\mathcal{P}_m + \mathcal{P}_m] = \frac{1}{4}k_a^2 \cdot 2\mathcal{P}_m = \frac{k_a^2}{2}\mathcal{P}_m$$

12.4.5.1 Calcolo della potenza di segnali BLD-PI, PS, PPS

La tabella al § 12.1.4 è calcolata adottando lo stesso procedimento sopra esposto al § 12.4.5, in cui ora

$$\mathcal{P}_{\underline{x}} = \mathcal{P}_{x_c=} \begin{cases} k_a^2\mathcal{P}_m & \text{(BLD-PS)} \\ a_p^2 + k_a^2\mathcal{P}_m & \text{(BLD-PI, BLD-PPS)} \end{cases}$$

12.4.6 Modulazione FM a basso indice

Riprendiamo qui il caso in cui $\beta \ll 1$ e di conseguenza $\Delta\alpha \ll 1$, consentendo quindi di arrestare al 1ř ordine lo sviluppo in serie di potenze (eq. (12.15)) di $\underline{x}(t)$. Se il segnale modulante è cosinusoidale, il segnale FM risulta

$$x_{FM}(t) = a\cos\left(\omega_0 t + 2\pi k_f \int_{-\infty}^{t} \cos(2\pi w\tau)\, d\tau\right) = a\cos(\omega_0 t + \beta\sin(2\pi w\tau))$$

Ricordando che $\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$, $x_{FM}(t)$ può essere riscritto come

$$x_{FM}(t) = a\cos\omega_0 t\cos(\beta\sin 2\pi wt) - a\sin\omega_0 t\sin(\beta\sin 2\pi wt)$$

che, se $\beta \ll 1$, diviene

$$x_{FM}(t) = a\cos\omega_0 t - \beta a\sin\omega_0 t\sin 2\pi wt \tag{12.26}$$

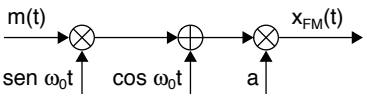
che confrontiamo con l'espressione

$$x_{AM}(t) = a_p\cos\omega_0 t + k_a\cos\omega_0 t\cos 2\pi wt \tag{12.27}$$

che si otterrebbe per modulazione a portante intera, o ridotta, dello stesso $m(t)$.

Il confronto tra (12.26) e (12.27) rivela che mentre nell'AM il segnale modulante moltiplica una portante *in fase* a quella (più o meno) intera, nell'FM a basso indice $m(t)$

[70] Infatti risulta $\int_{-\infty}^{\infty} X^+(f)\,X^-(f)\,df = 0$. dato che i due termini non si sovrappongono in frequenza.



opera su di una portante in *quadratura*. Questa considerazione è alla base dello *schema di modulazione* per segnali FM a basso indice mostrato a lato e realizzabile *sommando* alla portante un segnale AM-BLD, modulato su di una portante in quadratura.

D'altra parte, uno schema di modulazione del genere produce anche una modulazione AM parassita: questa può essere eliminata in ricezione dall'azione congiunta di uno squadratore e di un filtro passa basso, come discusso in fondo al § 12.3.2.2.

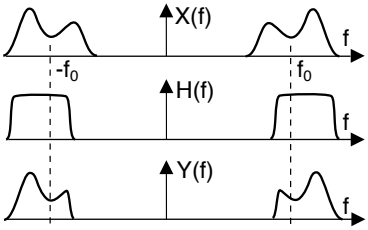
Capitolo 13

Distorsione per segnali modulati

Forti della descrizione analitica dei segnali modulati (cap. 11), e dello studio delle possibili tecniche di mo-demodulazione (cap. 12), occupiamoci di investigare l'influenza che i fenomeni di distorsione *lineare* e *non* (cap. 8) possono avere sui segnali di natura passa-banda, o modulati. A tal fine, viene dapprima esaminato come le operazioni di filtraggio possano essere descritte anche in termini di inviluppo complesso, e quindi sono determinate le condizioni *favorevoli* che un canale di comunicazione passa banda dovrebbe avere, così come viene approfondita la possibilità di *semplificare* l'analisi degli effetti per casi particolari di segnale modulato. I risultati sono quindi applicati alle tecniche di modulazione affrontate al capitolo 12, in modo da analizzare gli effetti della distorsione lineare nei diversi casi. Infine, viene mostrato come la modulazione angolare *non risenta* di fenomeni di distorsione *non* lineare.

13.1 Filtraggio passa banda

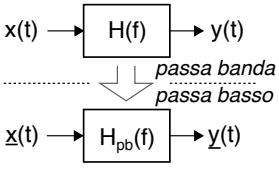
Qualora un filtro presenti una risposta in frequenza $H(f)$ di tipo *passa banda* come raffigurato a lato, la corrispondente risposta impulsiva $h(t)$ può essere considerata descritta nei termini delle componenti in fase ed in quadratura



$$h(t) = h_c(t)\cos\omega_0 t - h_s(t)\sin\omega_0 t \qquad (13.1)$$

riferite ad una frequenza f_0, ovvero nei termini del relativo inviluppo complesso $\underline{h}(t) = h_c(t) + jh_s(t)$ in modo del tutto simile (§ 11.2.1) a quanto avviene per il segnale modulato $x(t)$ in ingresso.

Ciò consente di ri-definire l'operazione di filtraggio di $x(t)$ attraverso $H(f)$ nei termini del filtraggio del relativo inviluppo complesso $\underline{x}(t)$ da parte di filtro $H_{pb}(f)$ passa basso detto *equivalente di banda base* e descritto da una risposta impulsiva *complessa*



$$h_{pb}(t) = \tfrac{1}{2}\underline{h}(t) \qquad (13.2)$$

Qualora infatti sia per il segnale modulato $x(t)$ che per la risposta impulsiva $h(t)$

valgano le condizioni di limitazione in banda, l'inviluppo complesso dell'uscita da $H(f)$ passa banda può essere calcolato a partire da $\underline{x}(t)$ e $\underline{h}(t)$ (valutati rispetto alla medesima frequenza f_0) come[1]

$$\underline{y}(t) = \tfrac{1}{2}\underline{x}(t) * \underline{h}(t) \tag{13.3}$$

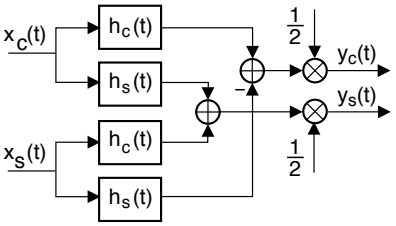
da cui si ottiene l'espressione di $y_c(t)$ ed $y_s(t)$ in funzione delle c.a. di b.f. di $x(t)$ e di quelle del filtro[2]:

$$\begin{aligned} \underline{y} &= \tfrac{1}{2}\underline{x} * \underline{h} = \tfrac{1}{2}\left[x_c + jx_s\right] * \left[h_c + jh_s\right] = \\ &= \tfrac{1}{2}\left[x_c * h_c - x_s * h_s\right] + j\tfrac{1}{2}\left[x_s * h_c + x_c * h_s\right] \end{aligned} \tag{13.4}$$

Pertanto le componenti reale e immaginaria di $\underline{y}(t)$ possono essere ottenute mediante 4 filtri *di banda base* operanti su $x_c(t)$ e $x_s(t)$, dato che dalla (13.4) otteniamo

$$\begin{cases} y_c(t) = \tfrac{1}{2}\left[x_c(t) * h_c(t) - x_s(t) * h_s(t)\right] \\ y_s(t) = \tfrac{1}{2}\left[x_s(t) * h_c(t) + x_c(t) * h_s(t)\right] \end{cases} \tag{13.5}$$

a cui corrisponde lo schema simbolico mostrato sotto.



Il risultato (13.5) ha una doppia valenza, sia positiva che negativa. Da un lato asserisce che si può *intenzionalmente* eseguire un filtraggio di tipo passa banda su di un segnale modulato senza che sia necessario realizzare il filtro *per davvero*, operando invece sulle relative c.a. di b.f. del segnale, e del filtro. D'altro canto se l'effetto filtrante è causato dal canale di comunicazione ed *è già avvenuto*, l'effetto prodotto sulle c.a. di b.f. $y_c(t)$ e $y_s(t)$ ottenute mediante demodulazione in fase e quadratura (§ 11.2.4) del segnale modulato prende il nome di...

13.1.1 Intermodulazione tra componenti analogiche

Le (13.5) mostrano come sia $y_c(t)$ che $y_s(t)$ dipendano in generale da entrambe le componenti $x_c(t)$ e $x_s(t)$, in un modo *apparentemente* ineliminabile. Infatti, le informazioni contenute in $x_c(t)$ ed $x_s(t)$ sono ora mescolate in modo da non poter essere recuperate mediante una operazione di equalizzazione (§§ 15.3 e 18.4) attuata separatamente su ciascuna delle c.a. di b.f. $y_c(t)$ e $y_s(t)$. A meno che...

[1] Per 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

$$x^+(t) * h^+(t) = \left[x * h_{fp}\right] * \left[h * h_{fp}\right] = \left[x * h\right] * \left[h_{fp} * h_{fp}\right] = y * h_{fp} = y^+(t)$$

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:

$$\begin{aligned} \tfrac{1}{2}\underline{x}(t) * \underline{h}(t) &= \tfrac{1}{2}\left[2x^+(t)\,\mathrm{e}^{-j\omega_0 t}\right] * \left[2h^+(t)\,\mathrm{e}^{-j\omega_0 t}\right] = \\ &= 2\int_{-\infty}^{\infty} x^+(\tau)\,\mathrm{e}^{-j\omega_0\tau} h^+(t-\tau)\,\mathrm{e}^{-j\omega_0(t-\tau)}\,d\tau = \\ &= 2\mathrm{e}^{-j\omega_0 t}\int_{-\infty}^{\infty} x^+(\tau)\,h^+(t-\tau)\,d\tau = 2\mathrm{e}^{-j\omega_0 t} y^+(t) = \underline{y}(t) \end{aligned}$$

[2] Tralasciamo di indicare la dipendenza da t per semplicità di notazione.

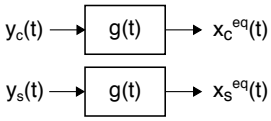
13.1.1.1 Equalizzazione in fase e quadratura

Consideriamo il caso in cui $\underline{h}(t)$ presenti *una sola* delle due c.a. di b.f.[3], ovvero $\underline{h}(t)$ sia solo reale o solo immaginario: se ad esempio risulta $\underline{h}(t) = h_c(t)$, le (13.5) divengono

$$\begin{cases} y_c(t) = \frac{1}{2} x_c(t) * h_c(t) \\ y_s(t) = \frac{1}{2} x_s(t) * h_c(t) \end{cases} \tag{13.6}$$

e quindi $y_c(t)$ e $y_s(t)$ risultano affette unicamente da distorsione lineare (§ 8.2).

Ciò consente di ri-ottenere le componenti *trasmesse* $x_c(t)$ e $x_s(t)$ a partire da quelle *ricevute* $y_c(t)$ e $y_s(t)$ mediante un procedimento di *equalizzazione*[4] svolto su entrambi i rami in modo indipendente mediante due identici filtri con risposta impulsiva $g(t)$ *reale* e tale che $g(t) * h_c(t) = 2\delta(t-\tau)$.



equalizzazione in fase e quadratura

Così facendo vengono ripristinate le condizioni di *canale perfetto* (pag. 231), cioè

$$\begin{cases} x_c^{eq}(t) = y_c(t) * g(t) = \frac{1}{2} x_c(t) * h_c(t) * g(t) = x_c(t-\tau) \\ x_s^{eq}(t) = y_s(t) * g(t) = \frac{1}{2} x_s(t) * h_c(t) * g(t) = x_s(t-\tau) \end{cases}$$

13.1.1.2 Equalizzazione complessa

Qualora nella (13.5) siano presenti entrambe $h_c(t)$ e $h_s(t)$ ed entrambe $x_c(t)$ e $x_s(t)$ non è possibile equalizzare le c.a. di b.f. in modo indipendente, e apparentemente la distorsione lineare può essere rimossa solo operando sul segnale modulato, con tutte le difficoltà legate alle elevate frequenze in gioco. Se invece rimuoviamo il vincolo di operare mediante un filtro *fisicamente realizzabile*, ossia con risposta impulsiva $g(t)$ *reale*, scopriamo che l'equalizzazione *può ancora* essere svolta *in banda base* ricorrendo ad una convoluzione *complessa*

$$\underline{x}^{eq}(t) = \underline{y}(t) * \underline{g}(t) \tag{13.7}$$

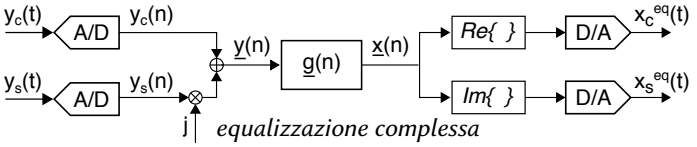
in cui $\underline{g}(t)$ è definita in modo che

$$\underline{h}^{eq}(t) = \underline{g}(t) * \underline{h}(t) = 2\delta(t-\tau) \tag{13.8}$$

ovvero in modo da rendere *perfetto* l'inviluppo complesso della risposta impulsiva del canale, dato che combinando le (13.3), (13.7) e (13.8) si ottiene

$$\underline{x}^{eq}(t) = \underline{y}(t) * \underline{g}(t) = \frac{1}{2} \underline{x}(t) * \underline{h}(t) * \underline{g}(t) = \underline{x}(t) * \delta(t-\tau) = \underline{x}(t-\tau)$$

L'equalizzazione complessa si rende possibile realizzando un filtro numerico (§ 5.3) con risposta impulsiva $\underline{g}(n)$ *complessa* ottenuta a partire dai campioni di $\underline{g}(t)$, ed operante sui campioni delle c.a. di b.f. come mostrato in figura.

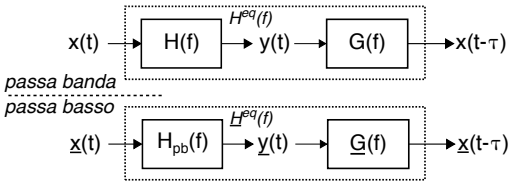


equalizzazione complessa

[3] Una considerazione del tutto simile può essere svolta qualora sia l'inviluppo complesso *del segnale modulato* $\underline{x}(t)$ ad essere solo reale od immaginario, ma viene rimandata al § 13.2.

[4] Come già evidenziato al § 18.4 è preferibile realizzare l'equalizzazione operando sulle c.a. di b.f. $x_c(t)$ e $x_s(t)$ *da trasmettere*, in modo da evitare di rendere colorato il rumore in ingresso al ricevitore. Nel caso in cui $h_c(t)$ non sia nota, occorre che presso il ricevitore venga effettuata una sua stima, vedi § 18.4.

13.1.1.3 Canale equalizzato

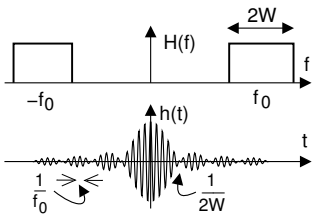


Effettuare l'equalizzazione direttamente a radio frequenza porta ad un risultato lievemente diverso da quando la si realizzi con l'equivalente di banda base: infatti anche se in entrambi i casi il risultato è sempre quello di ottenere un canale equalizzato *perfetto*, nel primo caso tale proprietà compete al canale passa banda ovvero $h^{eq}(t) = \delta(t-\tau)$, mentre nel secondo è l'equivalente *passa basso* ad essere perfetto, ovvero $\underline{h}^{eq}(t) = \delta(t-\tau)$. Prendiamo spunto da questa osservazione per analizzare i casi in cui *non si verifica* distorsione lineare per un segnale modulato.

13.1.2 Assenza di distorsione lineare nel filtraggio passa banda

Prendiamo in esame i requisiti affinché un canale non produca effetti di distorsione lineare su di un segnale modulato che lo attraversa, aggiungendo dettagli un po' per volta, per arrivare ad un risultato più generale.

13.1.2.1 Canale passa banda ideale



E' descritto da una risposta in frequenza $H(f)$ nulla ovunque, tranne che negli intervalli di frequenze $f_0 - W \leq |f| \leq f_0 + W$ in cui ha valore unitario, ovvero

$$H(f) = rect_{2W}(f - f_0) + rect_{2W}(f + f_0)$$

da cui antitrasformando si ottiene facilmente

$$\begin{aligned} h(t) &= 2W\,\mathrm{sinc}(2Wt)\left(e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}\right) = \\ &= 4W\,\mathrm{sinc}(2Wt)\cos 2\pi f_0 t \end{aligned} \tag{13.9}$$

Confrontando la (13.9) con la (13.1), riconosciamo che

$$h_c(t) = 4W\,\mathrm{sinc}(2Wt) \quad \text{e} \quad h_s(t) = 0$$

e dunque $\underline{h}(t) = h_c(t) + jh_s(t)$ è reale, ed $\underline{H}(f) = 2rect_{2W}(f)$. Un segnale modulato in transito non subisce nessuna intermodulazione, ne distorsione lineare, e neanche ritardo.

A questo caso particolare aggiungiamo le considerazioni relative ad un canale perfetto (ovvero con modulo costante e fase lineare) svolte a pag. 231, ove si afferma che in tal caso la forma d'onda del segnale in transito non subisce alterazione se non per un ritardo temporale: verifichiamo come questa ipotesi si rifletta sulle c.a. di b.f. di un segnale modulato. Come anticipato al § 13.1.1.3, ad essere perfetta può essere sia la risposta in frequenza $H(f)$ del canale passa banda, sia quella $\underline{H}(f)$ del suo equivalente passa basso (13.2). Iniziamo dal secondo.

13.1.2.2 Canale equivalente passa basso perfetto

L'inviluppo complesso in questo caso ha trasformata

$$\underline{H}(f) = 2\mathrm{rect}_{2W}(f)\,\mathrm{e}^{-j2\pi f\tau}$$

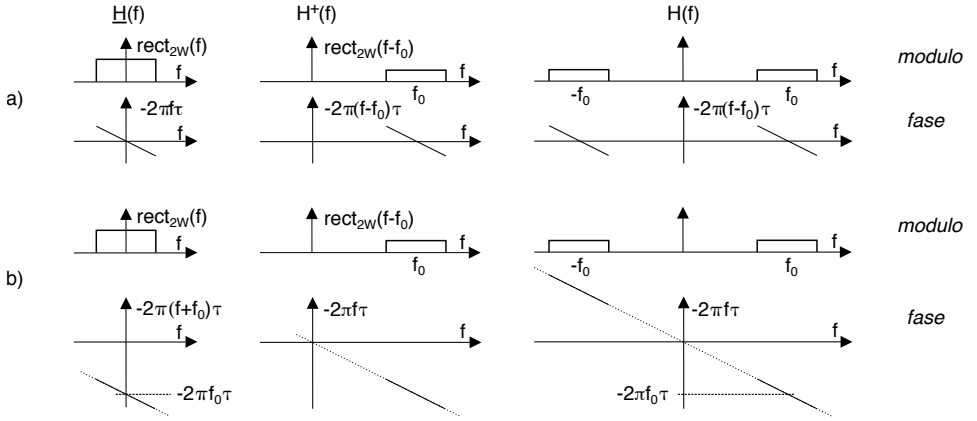


Figura 13.1: Condizioni di assenza di distorsione lineare per segnali modulati, nei casi a) - canale equivalente passa basso perfetto, b) - canale passa banda perfetto

a cui è associata una $\underline{h}(t) = 4W\text{sinc}(2W(t-\tau))$. Ricordando (eq. (11.19)) che $H^+(f) = \frac{1}{2}\underline{H}(f-f_0)$ e che $H^-(f) = \frac{1}{2}\underline{H}^*(-f-f_0)$, per il canale passa banda si ottiene

$$H(f) = \text{rect}_{2W}(f-f_0)\,\text{e}^{-j2\pi(f-f_0)\tau} + \text{rect}_{2W}(f+f_0)\,\text{e}^{-j2\pi(f+f_0)\tau} \tag{13.10}$$

mostrata in figura 13.1-a), a cui corrisponde[5] una risposta impulsiva

$$h(t) = 4W\text{sinc}(2W(t-\tau))\cos\omega_0 t \tag{13.11}$$

e dunque anche in questo caso $\underline{h}(t) = h_c(t) = 4W\text{sinc}(2W(t-\tau))$ è solamente reale. Un segnale modulato $x(t)$ che attraversi il canale produce in uscita[6]

$$y(t) = x_c(t-\tau)\cos\omega_0 t - x_s(t-\tau)\sin\cos\omega_0 t \tag{13.12}$$

manifestando un ritardo τ a carico solamente delle c.a. di b.f., e non della portante. Indicando $\varphi(f) = -2\pi f\tau$ come fase $\varphi_{pb}(f)$ di $\underline{H}(f)$, il rapporto $-\frac{1}{2\pi}\frac{d}{df}\varphi(f) = \tau$ è stato indicato al § 8.2.2 come ritardo *di gruppo* $\tau_g(f)$, in questo caso costante in quanto $\underline{H}(f)$ ha fase lineare.

13.1.2.3 Canale passa banda perfetto

In questo caso la risposta in frequenza assume la forma (vedi fig. 13.1-b)

$$H(f) = [\text{rect}_{2W}(f-f_0) + \text{rect}_{2W}(f+f_0)]\,\text{e}^{-j2\pi f\tau} \tag{13.13}$$

---

[5] Espandiamo la (13.10) come

$$H(f) = \text{rect}_{2W}(f-f_0)\,\text{e}^{-j2\pi f\tau}\text{e}^{j2\pi f_0\tau} + \text{rect}_{2W}(f+f_0)\,\text{e}^{-j2\pi f\tau}\text{e}^{-j2\pi f_0\tau}$$

da cui antitrasformando si ottiene

$$h(t) = 2W\text{sinc}(2W(t-\tau))\,\text{e}^{j2\pi f_0(t-\tau)}\text{e}^{j2\pi f_0\tau} + 2W\text{sinc}(2W(t-\tau))\,\text{e}^{-j2\pi f_0(t-\tau)}\text{e}^{-j2\pi f_0\tau} =$$
$$= 2W\text{sinc}(2W(t-\tau))\left(\text{e}^{j2\pi f_0 t} + \text{e}^{-j2\pi f_0 t}\right) = 4W\text{sinc}(2W(t-\tau))\cos\omega_0 t$$

[6] La (13.12) si ottiene considerando W abbastanza elevato da poter assimilare $2W\text{sinc}(2Wt) \to \delta(t)$ ossia ad un impulso, in modo che la (13.3) produca $\underline{y}(t) = \frac{1}{2}\underline{x}(t) * 2\delta(t-\tau) = x_c(t-\tau) + jx_s(t-\tau)$.

a cui corrisponde[7] una risposta impulsiva

$$h(t) = 4W\text{sinc}(2W(t-\tau))\cos\omega_0(t-\tau) \tag{13.14}$$

ed un segnale analitico $H^+(f) = \text{rect}_{2W}(f-f_0)\,e^{-j2\pi f\tau}$, e quindi un inviluppo complesso

$$\underline{H}(f) = 2H^+(f+f_0) = 2\text{rect}_{2W}(f)\,e^{-j2\pi(f+f_0)\tau}$$

da cui, essendo $e^{-j2\pi(f+f_0)\tau} = e^{-j2\pi f\tau}e^{-j2\pi f_0\tau} = e^{-j2\pi f\tau}(\cos\omega_0\tau - j\sin\omega_0\tau)$, si ha

$$\underline{h}(t) = 4W\text{sinc}(2W(t-\tau))(\cos\omega_0\tau - j\sin\omega_0\tau)$$

e pertanto

$$h_c(t) = 4W\text{sinc}(2W(t-\tau))\cos\omega_0\tau \quad \text{e} \quad h_s(t) = 4W\text{sinc}(2W(t-\tau))\sin\omega_0\tau$$

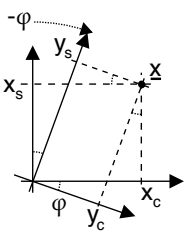
Ponendo ora $\phi = \omega_0\tau$ e applicando le (13.5), in base all'osservazione alla nota 6 in uscita dal filtro troviamo componenti analogiche pari a

$$\begin{cases} y_c(t) = \frac{1}{2}[x_c(t-\tau)\cos\phi - x_s(t-\tau)\sin\phi] \\ y_s(t) = \frac{1}{2}[x_c(t-\tau)\sin\phi + x_s(t-\tau)\cos\phi] \end{cases} \tag{13.15}$$

Il risultato (13.15) mostra che, oltre al ritardo $\tau = \tau_g$ già trovato per il caso al § 13.1.2.2, si osserva anche una *rotazione* del piano dell'inviluppo complesso, potendo esprimere le (13.15) come

$$\begin{bmatrix} y_c(t) \\ y_s(t) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}\begin{bmatrix} x_c(t-\tau) \\ x_s(t-\tau) \end{bmatrix} \tag{13.16}$$

in cui la matrice dei coefficienti è costante, e corrisponde ad una rotazione *antioraria*[8] di $\underline{y}(t)$ rispetto a $\underline{x}(t)$, equivalente alla rotazione *oraria* degli assi mostrata in figura, di un angolo $\phi = \omega_0\tau = 2\pi f_0\tau$ pari alla fase di $H(f)$ calcolata per $f = f_0$ come mostrato in fig. 13.1-b). Indicando quest'ultima come $\varphi(f)$, osserviamo che al § 8.2.2 abbiamo definito $\frac{-\phi}{2\pi f_0} = \frac{-\varphi(f_0)}{2\pi f_0} = \tau$ come *ritardo di fase* $\tau_f(f)$ calcolato per $f = f_0$, ovvero τ rappresenta anche il ritardo della portante, cioè



$$y(t) = x(t-\tau) = x_c(t-\tau)\cos\omega_0(t-\tau) - x_s(t-\tau)\sin\omega_0(t-\tau)$$

come previsto, vista l'ipotesi di partenza (13.13), nonché per il risultato che per un canale perfetto si ha $\tau_g = \tau_f$. L'effetto della rotazione ϕ può quindi essere annullato se la fase della portante di demodulazione produce una rotazione opposta, oppure digitalmente per eseguire la rotazione inversa descritta a pagina 349 operando sui campioni delle c.a di b.f..

Infine, notiamo che le (13.16) possono essere riscritte in forma *compatta* come

[7] Il termine tra parentesi quadre in (13.13) ha anti-trasformata

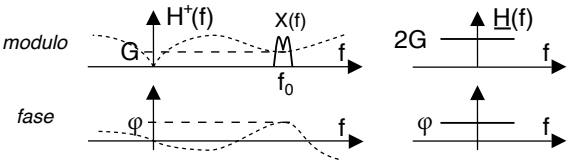
$$\mathcal{F}^{-1}\{\text{rect}_{2W}(f-f_0) + \text{rect}_{2W}(f+f_0)\} = 2W\text{sinc}(2Wt)\left(e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}\right) = 2W\text{sinc}(2Wt)\cos\omega_0 t$$

ma il termine $e^{-j2\pi f\tau}$ presente in (13.13) produce un ritardo nell'antitrasformata, dunque il risultato (13.14)

[8] L'analisi che interpreta la trasformazione legata ad un sistema lineare con matrice dei coefficienti pari a $\begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$ come una *rotazione* è stata svolta alla nota 17 di pag. 349.

$\underline{y}(t) = \underline{x}(t-\tau)\, e^{j\phi}$.

13.1.2.4 Segnale a banda stretta



L'ultimo caso in cui non si determina distorsione lineare sul segnale modulato $x(t)$ si verifica quando il segnale occupa una banda B *molto piccola*[9] rispetto alla frequenza portante f_0, in modo che la risposta in frequenza del canale $H(f)$ possa ritenersi approssimativamente *costante*[10] nella banda del segnale ovvero per $f \simeq f_0$, cioè

$$H(f) \simeq H(f_0) = G\mathrm{e}^{j\phi \mathrm{sgn}(f)}$$

e dunque $\underline{H}(f) = 2H^{+}(f+f_0) = 2Ge^{j\phi}$, a cui corrisponde un inviluppo complesso

$$\underline{h}(t) = 2Ge^{j\phi}\delta(t) = 2G(\cos\phi + j\sin\phi)\,\delta(t)$$

In uscita da $H(f)$ si osserva pertanto (eq. 13.3)

$$\begin{aligned}\underline{y}(t) &= \tfrac{1}{2}\underline{x}(t) * \underline{h}(t) = (x_c(t) + jx_s(t)) * G(\cos\phi + j\sin\phi)\,\delta(t) = \\ &= G\left[(x_c(t)\cos\phi - x_s(t)\sin\phi) + j(x_c(t)\sin\phi + x_s(t)\cos\phi)\right]\end{aligned}$$

che può essere scritta in forma matriciale come

$$\begin{bmatrix} y_c(t) \\ y_s(t) \end{bmatrix} = G\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}\begin{bmatrix} x_c(t) \\ x_s(t) \end{bmatrix}$$

in cui compare la stessa matrice dei coefficienti del caso precedente, che di nuovo rappresenta la rotazione *oraria* del piano dell'inviluppo complesso di un angolo pari alla fase $\varphi(f)$ di $H(f)$ valutata per $f = f_0$, a cui corrisponde il segnale ricevuto

$$y(t) = x_c(t)\cos\omega_0(t-\tau) - x_s(t)\sin\omega_0(t-\tau)$$

Riassumendo Esprimendo la risposta in frequenza del canale nella forma $H(f) = |H(f)|\, e^{j\varphi(f)}$ e la trasformata del corrispondente inviluppo complesso come $\underline{H}(f) = \left|\underline{H}(f)\right| e^{j\varphi_{pb}(f)}$ in cui $\varphi_{pb}(f) = \varphi(f+f_0)$, abbiamo mostrato che

- un canale *passabanda ideale* non altera il segnale modulato in transito;
- un canale *equivalente passabasso perfetto* introduce nelle c.a. di b.f. un ritardo $\tau_g = -\frac{1}{2\pi}\frac{d}{df}\varphi_{pb}(f)\Big|_{f=0}$ che dipende *dalla pendenza* della fase di $\underline{H}(f)$ nell'origine, ovvero di quella $\varphi(f)$ del passa banda per $f = f_0$;
- un canale *passabanda perfetto* oltre al ritardo τ_g delle c.a. di b.f. introduce anche una rotazione del piano di $\underline{y}(t)$ di un angolo $\phi = \omega_0\tau_f$ che dipende *dal valore* della fase di $\underline{H}(f)$ nell'origine in quanto $\tau_f = -\frac{\varphi(f_0)}{2\pi f_0} = -\frac{\varphi_{pb}(0)}{2\pi f_0}$;

[9] Condizione indicata anche come *piccola banda frazionale*, definita come $B/f_0 \ll 1$.

[10] Detta anche condizione per un *fading piatto* nel caso di un collegamento radio, vedi pag. 682, mentre dal punto di vista circuitale ciò corrisponde a realizzare le condizioni di adattamento di impedenza (vedi § 18.1.1.4) 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.

- un segnale modulato *a banda stretta* subisce la rotazione suddetta, ma non il ritardo delle c.a di b.f.

L'approccio unitario agli ultimi tre aspetti delineati è fornito al § seguente, che illustra il *caso particolare* di distorsione lineare dovuta alla sola risposta di fase.

13.1.3 Ritardo di fase, di gruppo, e distorsione di tempo di transito

Al § 8.2.2 si è affermato che l'attraversamento da parte di un segnale modulato in ampiezza $x(t) = a(t)\cos(2\pi f_0 t)$ *a banda stretta* di un canale che (nella banda di segnale) presenta una risposta in frequenza $H(f) = e^{j\varphi(f)}$ (ovvero modulo costante e fase *generica*) produce in uscita

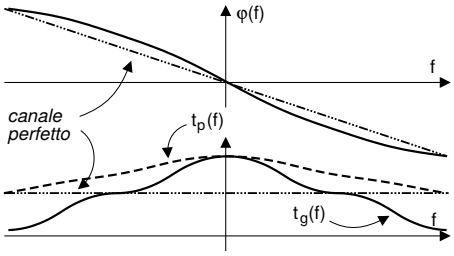
$$y(t) = a\left(t - \tau_g(f_0)\right)\cos\left(2\pi f_0\left(t - \tau_f(f_0)\right)\right) \tag{13.17}$$

in cui

$$\tau_f(f) = -\frac{\varphi(f)}{2\pi f} \qquad e \qquad \tau_g(f) = -\frac{1}{2\pi}\frac{d}{df}\varphi(f)$$

sono rispettivamente indicati come ritardo *di fase* della *portante* $\tau_f(f)$ (eq. 8.7) e ritardo *di gruppo* $\tau_g(f)$ (eq. 8.8) del segnale *modulato*, ovvero il ritardo con cui il *gruppo di frequenze* presenti in $a(t)$ si presenta in uscita. Tale risultato viene dimostrato in appendice 13.4.1 come l'esito di una approssimazione al primo ordine dello sviluppo in serie di potenze di $\varphi(f)$.

Nel caso di canale perfetto si ottiene $\varphi(f) = -2\pi f\tau$, 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 $\varphi(f)$ viaggia *parallela* alla risposta in fase del canale perfetto, mentre risulta $t_p(f) = \tau$ quando $\varphi(f)$ *la interseca*.



Qualora il segnale in transito *non sia* a banda stretta, può essere considerato come la sovrapposizione di tante componenti $x_i(t) = a_i(t)\cos(2\pi f_i t)$, tutte a banda stretta, centrate su portanti f_i tra loro contigue: pertanto l'inviluppo di ampiezza $a_i(t)$ di ognuna di esse si presenta in uscita con un diverso ritardo $\tau_i = t_g(f_i)$. L'effetto della fase non lineare $\varphi(f)$ è dunque espresso nei termini della sua derivata normalizzata $t_g(f) = -\frac{1}{2\pi}\frac{d}{df}\varphi(f)$, misurabile strumentalmente mediante segnali a banda stretta, ed utilizzata per descrivere l'entità della *distorsione di fase*, per questo detta anche *distorsione di tempo di transito*, come definito al § 8.2.2.

13.1.4 Assenza di intermodulazione tra componenti analogiche

Come ultimo aspetto delle possibili tipologie di canale che possono tornare utili, ripartiamo dal sistema (13.5) e dall'osservazione che qualora l'inviluppo complesso $\underline{h}(t)$ associato alla risposta impulsiva $h(t)$ del canale passa banda $H(f)$ (considerato

per $|f - f_0| \leq W$ ovvero entro la banda di segnale) sia completamente reale, ovvero

$$\underline{h}(t) = h_0(t) \tag{13.18}$$

le (13.5) si riducono alle (13.6) e si può procedere ad una equalizzazione *in fase e quadratura*, ovvero attuata con filtri *fisicamente realizzabili*. In realtà la (13.18) è una condizione solamente *sufficiente* ad ottenere assenza di intermodulazione, dato che in presenza di un ritardo τ sulla portante (come mostrato al § 13.1.2.3) $\underline{h}(t)$ diviene

$$\underline{h}(t) = h_0(t)(\cos\phi - j\sin\phi) = h_0(t)\,\mathrm{e}^{-j\phi} \tag{13.19}$$

in cui $\phi = \omega_0\tau$. Mostriamo come la (13.19) si riflette su $\underline{H}(f) = \left|\underline{H}(f)\right| \mathrm{e}^{j\varphi_{pb}(f)}$.

La condizione (13.18) $\underline{h}(t) = h_0(t)$ reale implica (pag. 65) una $\underline{H}(f) = H_0(f)$ a *simmetria coniugata* rispetto a $f = 0$, ovvero modulo $\left|\underline{H}(f)\right| = |H_0(f)|$ *pari*, e fase $\varphi_{pb}(f) = \varphi(f + f_0)$ *dispari*; d'altra parte dalla (13.19) si ottiene

$$\underline{H}(f) = H_0(f)\,\mathrm{e}^{-j\phi} = |H_0(f)|\,\mathrm{e}^{j\varphi_{pb}(f)}\mathrm{e}^{-j\phi} = |H_0(f)|\,\mathrm{e}^{j(\varphi_{pb}(f)-\phi)}$$

Pertanto

> *Un canale $H(f)$ non provoca interferenza intersimbolica nei confronti di un segnale modulato se e solo se il corrispondente equivalente passa basso $\underline{H}(f)$ presenta (nella banda di segnale) modulo pari e fase dispari, a meno di una costante ϕ.*

Essendo $\phi = \omega_0\tau$, ciò corrisponde a scomporre $H^+(f)$ nella cascata di $2H_0(f - f_0)$ e di un canale perfetto con $h(t) = \delta(t - \tau)$ in cui $\tau = \frac{-\phi}{\omega_0} = \frac{-\varphi(f)}{2\pi f_0} = \tau_f(f_0)$ è il ritardo di fase della portante, come mostrato in fig. 13.2.

Osservazione Qualora ϕ sia pari ad un multiplo di $\frac{\pi}{2}$ l'evoluzione di $\underline{h}(t)$ si sviluppa lungo gli assi che definiscono il piano dell'inviluppo complesso. Infatti in base alla (13.19) per $\phi = 0$ si ottiene $\underline{h}(t) = h_o(t)$ ovvero $h_c(t) = h_0(t)$, per $\phi = \frac{\pi}{2}$ si ha $\underline{h}(t) = -jh_o(t) = jh_s(t)$, e per $\phi = \pi$ o $\phi = \frac{3\pi}{2}$ si ottiene rispettivamente $\underline{h}(t) = -h_o(t) = h_c(t)$ e $\underline{h}(t) = jh_o(t) = jh_s(t)$.

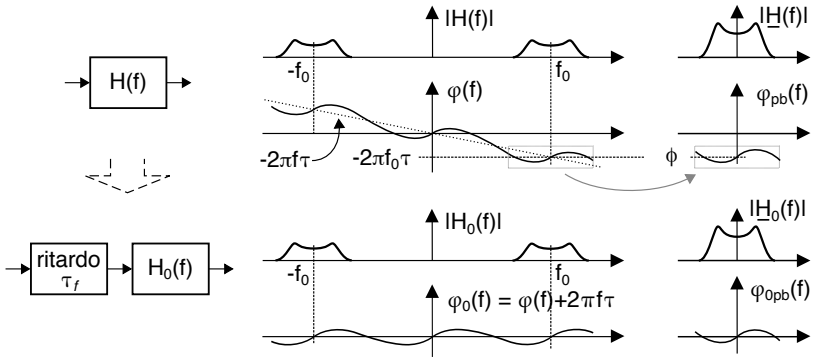


Figura 13.2: Condizione per assenza di intermodulazione nelle componenti analogiche di bassa frequenza

13.2 Distorsione lineare per segnali modulati

Dopo aver analizzato le particolarità del canale per quanto riguarda la trasmissione di generici segnali passa banda, passiamo ad analizzare gli effetti della *distorsione lineare* sui tipi di modulazione descritti al cap. 12, e che estendono le conseguenze già note dal cap. 8.

13.2.1 Modulazione di ampiezza

Distinguiamo tra i tre casi seguenti:

AM-BLD-PS o Banda Laterale Doppia, Portante Soppressa In questo caso $x_s(t) = 0$, e quindi il sistema (13.5) si riduce a $\begin{cases} y_c(t) = \frac{1}{2} x_c(t) * h_c(t) \\ y_s(t) = \frac{1}{2} x_c(t) * h_s(t) \end{cases}$ e non si verifica intermodulazione. In assenza di errore di fase sulla portante è quindi sufficiente un demodulatore omodina in quanto l'informazione si trova sulla componente $y_c(t) = x_c(t) * h_c(t)$, e l'effetto del canale può essere equalizzato con un filtro fisicamente realizzabile. Se invece è presente un errore di fase ϕ (vedi § 12.2.3.1) adottando un demodulatore in fase e quadratura otteniamo

$$\begin{cases} y_c(t) = \frac{1}{2} \left[x_c(t) * h_c(t) \cos\phi - x_s(t) * h_s(t) \sin\phi \right] \\ y_s(t) = \frac{1}{2} \left[x_c(t) * h_s(t) \sin\phi + x_c(t) * h_s(t) \cos\phi \right] \end{cases}$$

ovvero $\underline{y}(t) = \frac{1}{2} \underline{x}(t) * \underline{h}(t) \cdot \mathrm{e}^{-j\phi}$, reversibile numerizzando le c.a. di b.f. e mettendo assieme la procedura di equalizzazione complessa (§ 13.1.1.2) con quella di inversione della rotazione (§ 11.2.5).

AM-BLD-PI o Banda Laterale Doppia, Portante Intera In questo caso il demodulatore è sempre di tipo *ad inviluppo*, tipicamente operante sul segnale a media frequenza prodotto da uno stadio eterodina, la fase della cui portante è indifferente a quella del segnale modulato.

Il problema principale si presenta quando $|H(f)|$ è molto ridotto in corrispondenza della portante ovvero per $f = f_0$, tipicamente a causa di una attenuazione *selettiva* come esemplificato all'esempio di pag. 238, poi ripreso al § 20.3.3. In tal caso infatti l'indice di modulazione I_a (eq. 12.4) supera il 100%, ed il segnale diviene *sovramodulato*, rendendo impraticabile la demodulazione di inviluppo.

AM-BLU o Banda Laterale Unica In questo caso il segnale modulato contiene ambedue le c.a. di b.f., e dunque in presenza di distorsione lineare insorge intermodulazione (eq. 13.5), particolarmente deleteria se non equalizzata, operazione da svolgere per via numerica sui campioni dell'inviluppo complesso ricevuto $\underline{y}(t)$ (§ 13.1.1.2)

13.2.2 Modulazione di Frequenza

Quando un segnale modulato angolarmente attraversa un canale affetto da distorsione lineare (di modulo, di fase, od entrambe), si verificano[11] due fenomeni di *conversione*,

[11] Vedi *E. Bedrosian*, Distortion and Crosstalk of Linearly Filtered, Angle-Modulated Signals, presso `https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19660014309.pdf`

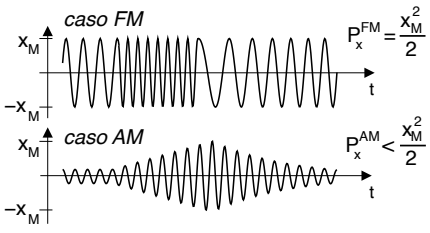
indicati come conversione PM-AM e PM-PM. La prima consiste in una modulazione di ampiezza *sovrapposta*, come avviene per il caso discusso al § 12.3.2.2 relativo alla demodulazione FM mediante discriminatore; la seconda invece si sostanzia in una *alterazione* della modulazione di fase. In entrambi i casi l'effetto della distorsione *dipende* dal messaggio modulante, e dunque non può essere considerato di natura *additiva*. Anche se la modulazione AM *parassita* può essere rimossa da un limitatore in ricezione (vedi § 12.3.2.2), ciò non è possibile per l'errore introdotto nella fase modulante; quest'ultima presenta inoltre anche termini *non lineari* ovvero legati alle sue potenze, e dunque non eliminabili mediante equalizzazione dopo demodulazione.

13.3 Distorsione non lineare di segnali modulati

Già al § 8.3 si è analizzato il fenomeno della distorsione *di non linearità*, focalizzando l'attenzione sui segnali di banda base, e sulla generazione di componenti spettrali a frequenze assenti dal segnale di ingresso, eliminabili mediante filtraggio. Nel caso di segnali modulati una non linearità produce conseguenze peculiari, affrontate in questa sede.

13.3.1 Limitazione di potenza per modulazione AM

Una delle prime conseguenze della non linearità degli apparati sui segnali modulati differenzia i casi FM ed AM per quanto riguarda la massima potenza a cui è possibile trasmettere. Nel primo caso infatti il segnale mantiene sempre la stessa ampiezza, che può essere posta pari a x_M ossia al massimo valore di ingresso al dispositivo non lineare (tipicamente, un amplificatore) prima che inizi a manifestarsi il fenomeno della sataurazione (vedi fig. a pag. 241). Pertanto la potenza del segnale trasmesso è pari a $\mathcal{P}_x^{FM} = {}^{x_M^2}/_2$, condizione detta anche di *piena potenza*.



Al contrario, nel caso della trasmissione di un segnale AM l'inviluppo di ampiezza del segnale è fortemente variabile nel tempo: per evitare di operare in regione non lineare, questa volta x_M è il valore *massimo* dell'inviluppo di ampiezza, e la potenza di trasmissione del segnale deve essere minore di quella *piena*, una sorta di *arretramento* indicato anche come *back-off*. Pertanto a differenza del caso FM, la trasmissione AM deve operare con una potenza $\mathcal{P}_x^{AM} < \mathcal{P}_x^{FM}$ minore della massima consentita dall'amplificatore, mentre per trasmettere la stessa potenza è necessario ricorrere ad un amplificatore sovradimensionato.

13.3.2 Distorsione di terza armonica

Riprendiamo il discorso iniziato al § 8.3.2, in cui si è analizzato l'effetto di una non linearità su di un processo gaussiano stazionario a media nulla, con potenza $\mathcal{P}_x$ e densità spettrale $\mathcal{P}_x(f)$, giungendo a dimostrare la comparsa dei termini di secondo e

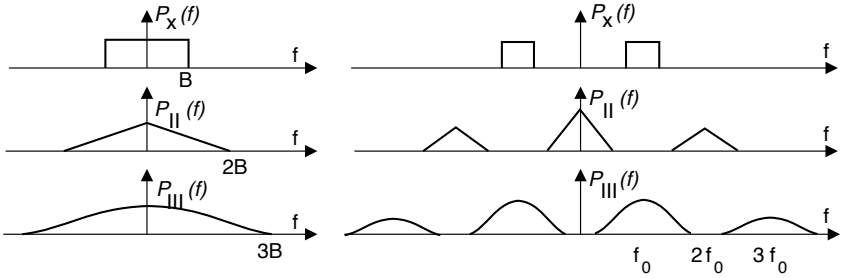


Figura 13.3: Densità spettrale di segnali affetti da distorsioni non lineari; a sinistra per banda base, a destra per segnale modulato

terzo ordine $\mathcal{P}_{II}(f)$ e $\mathcal{P}_{III}(f)$ espressi come

$$\mathcal{P}_{II}(f) = G^2 2\alpha^2 \cdot \mathcal{P}_x(f) * \mathcal{P}_x(f); \quad \mathcal{P}_{III}(f) = G^2 6\beta^2 \cdot \mathcal{P}_x(f) * \mathcal{P}_x(f) * \mathcal{P}_x(f) \quad (13.20)$$

Dato che ora il processo è un segnale modulato con portante f_0, eseguendo la costruzione grafica (§ 3.4.3) per le convoluzioni presenti nella (13.20) osserviamo (fig. 13.3) che il termine $\mathcal{P}_{II}(f)$ occupa regioni di frequenza (con banda doppia rispetto a $\mathcal{P}_x(f)$) centrate ad $f = 0$ e $f = 2f_0$, disgiunte dalla banda di $\mathcal{P}_x(f)$: pertanto $\mathcal{P}_{II}(f)$ può *essere non considerato* fonte di disturbo - se non a danno di eventuali altre trasmissioni a frequenza $2f_0$. Infine, la convoluzione tra $\mathcal{P}_I(f)$ e $\mathcal{P}_{II}(f)$ fornisce una $\mathcal{P}_{III}(f)$ costituita anch'essa da due componenti, di cui una centrata ad $f = 3f_0$ e che, come per $\mathcal{P}_{II}(f)$, non produce disturbo se non ad altre trasmissioni; mentre una parte più consistente di $\mathcal{P}_{III}(f)$ è centrata sulla stessa portante f_0 di $x(t)$, e dunque costituisce effettivamente fonte di disturbo, come si dice, *in banda*. In definitiva, vi sono almeno tre buone ragioni per *tenere d'occhio* il valore di β (vedi (8.10) per la sua definizione), 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.

13.3.3 Insensibilità della modulazione angolare alle non linearità

Di nuovo, la modulazione FM si dimostra più tollerante della AM rispetto alle non-linearità, al punto che la distorsione *in banda* discussa al § precedente si dimostra *non essere* un disturbo. Infatti, applicando la notazione introdotta con (8.10) ad un segnale modulato angolarmente, ovvero del tipo $x(t) = \cos[\omega_0 t + \varphi(t)]$, l'attraversamento di un dispositivo con caratteristica ingresso-uscita approssimata come $y(t) = x(t) + \alpha x^2(t) + \beta x^3(t)$ produce una uscita (vedi eq. (8.11))

$$y(t) = \frac{\alpha}{2}a^2(t) + \left(1 + \frac{3}{4}\beta\right)\cos[\omega_0 t + \varphi(t)] + \frac{\alpha}{2}\cos[2\omega_0 t + 2\varphi(t)] + \frac{\beta}{4}\cos[3\omega_0 t + 3\varphi(t)]$$

Dopo che i termini a frequenza $2\omega_0$ e $3\omega_0$, nonché la costante additiva, sono eliminati mediante un filtro passa-banda centrato in $f = f_0 = \frac{\omega_0}{2\pi}$, rimane solamente il termine $z(t) = \left(1 + \frac{3}{4}\beta\right)\cos[\omega_0 t + \varphi(t)]$. Pertanto, la modulazione di fase $\varphi(t)$ *è esattamente*

la stessa di quella impressa dal modulatore, e quindi i fenomeni non lineari *non hanno conseguenze sulla modulazione angolare!* Tranne, ovviamente, che per le interferenze causate alle trasmissioni su portanti a frequenza doppia e tripla di f_0.

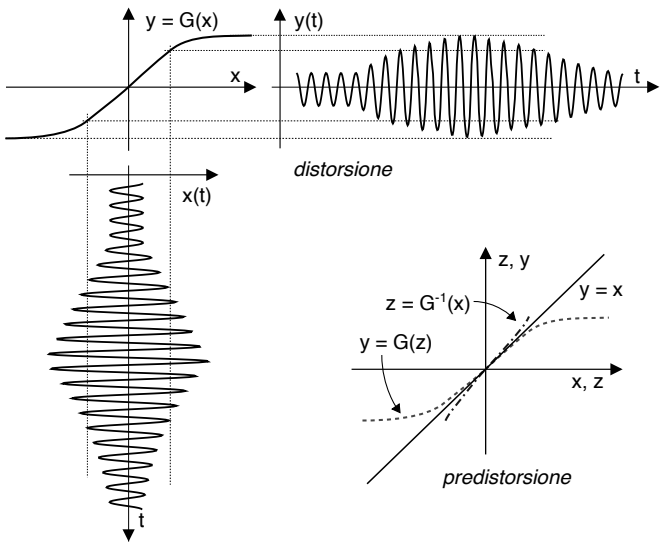
L'insensibilità dei segnali modulati angolarmente nei riguardi delle non linearità è stata ad esempio sfruttata nei collegamenti in ponte radio progettati per trasmettere un segnale FDM in modulazione di frequenza (§ 11.1.1.2), adottando un basso indice di modulazione (risparmiando banda) e trasmettendo a piena potenza (vedi § 8.3). In questo modo, la potenza del segnale trasmesso *non dipende* dal numero di canali contemporaneamente attivi.

13.3.4 Predistorsione

A differenza dei segnali modulati angolarmente, quelli a modulazione di ampiezza - qualora non limitati in potenza come descritto al § 13.3.1 - subiscono pienamente l'effetto di distorsione non lineare, che si ripercuote sulla ampiezza della portante, ed in egual misura sul segnale demodulato.

Nel caso in cui i parametri che caratterizzano la nonlinearità[12] $y = G(x)$ siano noti, un rimedio che viene tentato è quello di far passare il segnale modulato (prima della sua trasmissione) attraverso un nuovo elemento non lineare appositamente realizzato in modo che effettui una trasformazione *inversa* ovvero $z = G^{-1}(x)$ con G^{-1} tale che

$$G(z) = G\left(G^{-1}(x)\right) = x$$



fornendo quindi in ingresso all'elemento non lineare G il segnale *predistorto* $z = G^{-1}(x)$ anziché quello originale x, neutralizzando così il fenomeno di non linearità.

13.4 Appendice

13.4.1 Derivazione del tempo di ritardo di gruppo

Dimostriamo qui il risultato (13.17). Svolgiamo i calcoli rappresentando sia l'ingresso $x(t) = a(t)\cos(2\pi f_0 t)$ *a banda stretta* che la risposta impulsiva $h(t)$ del canale nei termini dei corrispondenti inviluppi complessi, in modo da poter scrivere (vedi eq. (13.3))

$$\underline{Y}(f) = \tfrac{1}{2}\underline{X}(f)\,\underline{H}(f) \tag{13.21}$$

[12] Ovvero i coefficienti dello sviluppo in serie della caratteristica ingresso-uscita, vedi nota 16 a pag. 242.

Lo scopo è dimostrare che, se il canale è affetto dalla sola distorsione di fase, ovvero descritto da una risposta in frequenza $H(f) = 1 \cdot e^{j\varphi(f)}$, il segnale in uscita avrà la forma $y(t) \simeq a\left(t - \tau_g(f_0)\right) \cos\left(2\pi f_0 \left(t - \tau_f(f_0)\right)\right)$.

Per quanto riguarda l'ingresso, ad esso corrisponde

$$\underline{X}(f) = X_c(f) = A(f) \tag{13.22}$$

mentre per quanto riguarda il canale risulta

$$\underline{H}(f) = 2H^+(f + f_0) = 2e^{j\varphi(f+f_0)} = 2e^{j\varphi_{pb}(f)} \tag{13.23}$$

dove il pedice $_{pb}$ simboleggia che ci riferiamo alla fase dell'equivalente *passa-basso* del canale, ovvero che $\varphi_{pb}(f) = \varphi(f + f_0)$ è la fase di $\underline{H}(f)$, e non di $H(f)$. Sviluppando ora $\varphi_{pb}(f)$ in serie di Maclaurin, e troncando la stessa al primo termine, per f prossimo a zero si ottiene[13]

$$\varphi_{pb}(f) \simeq -2\pi \left(f_0 \tau_f(f_0) + f \tau_g(f_0)\right) \tag{13.24}$$

e quindi sostituendo (13.22) e (13.23) in (13.21), ed utilizzando (13.24) otteniamo

$$\begin{aligned} \underline{Y}(f) &= \tfrac{1}{2}\underline{X}(f)\,\underline{H}(f) = A(f)\, e^{j\varphi_{pb}(f)} \simeq A(f)\, e^{-j2\pi\left(f_0\tau_f(f_0) + f\tau_g(f_0)\right)} \\ &= e^{-j2\pi f_0 \tau_f(f_0)} \cdot A(f)\, e^{-j2\pi f \tau_g(f_0)} \end{aligned}$$

da cui, ricordando la proprietà di traslazione temporale, otteniamo l'antitrasformata

$$\underline{y}(t) = e^{-j2\pi f_0 \tau_f(f_0)} \cdot a\left(t - \tau_g(f_0)\right)$$

a cui corrisponde[14] il segnale modulato $y(t) \simeq a\left(t - \tau_g(f_0)\right) \cos\left(2\pi f_0 \left(t - \tau_f(f_0)\right)\right)$.

Teniamo ora a precisare che il risultato mostrato *perde validità* sia qualora la risposta in frequenza non abbia modulo costante, sia nel caso in cui la *derivata della fase* $\frac{d\varphi(f)}{df}$ non possa essere considerata sufficientemente costante nell'intervallo di frequenze occupato da $A(f)$: nel secondo caso infatti nello sviluppo di Maclaurin di cui alla nota 13 occorre tener conto anche dei termini legati alle derivate successive, la cui importanza relativa aumenta con le *potenze* di f, ovvero tanto più ci si discosta dalla frequenza centrale f_0 del gruppo (di frequenze).

---

[13] Infatti, ricordando le definizioni (§ 8.2.2) $t_p(f) = -\frac{\varphi(f)}{2\pi f}$ per il ritardo *della portante* e $t_g(f) = -\frac{1}{2\pi}\frac{d}{df}\varphi(f)$ per il ritardo *di gruppo*, sussistono i passaggi

$$\begin{aligned} \varphi_{pb}(f) &\simeq \varphi_{pb}(0) + f \cdot \left.\frac{d\varphi_{pb}(f)}{df}\right|_{f=0} = \varphi(f_0) + f \cdot \left.\frac{d\varphi(f)}{df}\right|_{f=f_0} \\ &= 2\pi\left(f_0 \frac{\varphi(f_0)}{2\pi f_0} + f\frac{1}{2\pi}\left.\frac{d\varphi(f)}{df}\right|_{f=f_0}\right) = -2\pi\left(f_0 t_p(f_0) + f t_g(f_0)\right) \end{aligned}$$

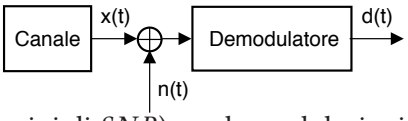
[14] E' sufficiente applicare le definizione

$$\begin{aligned} y(t) &= \Re\left\{\underline{y}(t)\, e^{j\omega_0 t}\right\} = \Re\left\{e^{-j2\pi f_0 \tau_f(f_0)} \cdot a\left(t - \tau_g(f_0)\right) \cdot e^{j\omega_0 t}\right\} = \\ &= \Re\left\{a\left(t - \tau_g(f_0)\right) \cdot e^{j\left(2\pi f_0 t - 2\pi f_0 \tau_f(f_0)\right)}\right\} = a\left(t - \tau_g(f_0)\right) \cos\left(2\pi f_0 \left(t - \tau_f(f_0)\right)\right) \end{aligned}$$

Capitolo 14

Prestazioni delle trasmissioni modulate

Chiudiamo con questo capitolo la parte relativa alla modulazione di segnali analogici, occupandoci di stabilire l'effetto che il rumore $n(t)$ presente all'uscita del canale produce sul risultato del processo di demodulazione $d(t)$.



Dopo aver applicato i risultati del § 11.3 alla caratterizzazione del rumore dopo demodulazione, viene definito un *indice di qualità* del collegamento rispetto al quale confrontare le prestazioni (in termini di SNR) per le modulazioni AM e FM, scoprendo come nel secondo caso la qualità possa essere migliorata a spese dell'occupazione di banda. Infine si affronta l'analisi della detezione di sinusoide nel rumore, specializzando i risultati sulla decisione statistica (§ 6.6.1) al caso dei processi circolari.

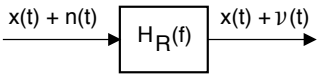
14.1 Il rumore additivo nei segnali modulati

Consideriamo un segnale modulato $x(t)$ ed affetto da un rumore additivo gaussiano bianco a media nulla $n(t)$ (§ 8.4.2) con densità di potenza

$$\mathcal{P}_n(f) = \frac{N_0}{2}$$

la cui occupazione spettrale è considerata costante a tutte le frequenze di interesse.

Filtro di ricezione Prima ancora di essere demodulato il segnale ricevuto viene fatto passare attraverso un *filtro di ricezione* passa-banda $H_R(f)$ centrato sulle frequenze del segnale, in modo da limitare la banda del rumore ricevuto e conseguentemente ridurre la potenza del rumore in ingresso al demodulatore. La risposta in frequenza $H_R(f)$ del filtro ha *modulo costante* nella banda del segnale, e tende a zero al di fuori di essa, in modo che il segnale utile $x(t)$ transita inalterato, mentre il rumore $n(t)$ diviene *limitato in banda*, producendo l'uscita $\nu(t)$.



14.1.1 Rapporto segnale-rumore

La qualità di ricezione (in funzione della frequenza) dipende dalla densità di potenza $\mathcal{P}_x(f)$ del segnale modulato e da quella $\mathcal{P}_\nu(f)$ del rumore filtrato, in base al 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. 12), mentre per quanto riguarda il rumore, dopo il filtraggio $\mathcal{P}_\nu(f)$ risulta pari a

$$\mathcal{P}_\nu(f) = \mathcal{P}_n(f)\,|H_R(f)|^2 = \frac{N_0}{2}\,|H_R(f)|^2 \tag{14.1}$$

D'altra parte, ha senso valutare l'SNR *complessivo*, ovvero il rapporto tra la potenza di segnale e quella di rumore *totali*: la prima risulta allora pari a

$$\mathcal{P}_x = \int_{-\infty}^{\infty} \mathcal{P}_x(f)\,df$$

mentre per quella di rumore dalla (14.1) si ottiene

$$\mathcal{P}_\nu = \frac{N_0}{2}\int_{-\infty}^{\infty} |H_R(f)|^2\,df$$

valutando cioè la potenza di rumore che attraversa il filtro di ricezione $H_R(f)$.

14.1.2 Banda di rumore

Definiamo questo concetto con l'aiuto della fig. 14.1, dove in alto è rappresentata la densità spettrale del segnale modulato, che occupa una banda[1] B_{RF}. Nel caso fosse possibile adottare come filtro di ricezione un passa banda *ideale* (§ 13.1.2.1) si otterrebbe $\mathcal{P}_\nu = N_0 B_{RF}$; invece $H_R(f)$ presenta una regione di transizione (vedi fig. 14.1-b)) che ne *accresce* la banda ad un valore $B_\nu > B_{RF}$. La potenza totale del rumore uscente da $H_R(f)$ risulta pertanto pari a

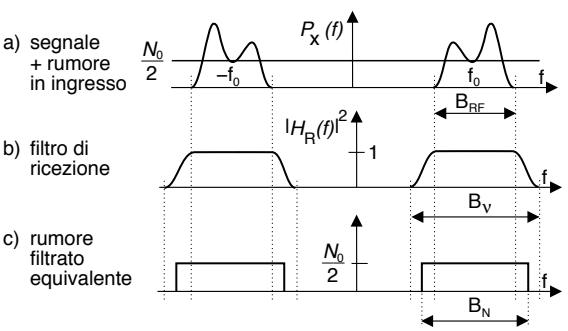


Figura 14.1: Densità spettrale al ricevitore, filtro di ricezione e rumore bianco passa banda equivalente

$$\begin{aligned}\mathcal{P}_\nu &= \frac{N_0}{2}\cdot\int_{-\infty}^{\infty}|H_R(f)|^2\,df = \\ &= N_0\int_0^{\infty}|H_R(f)|^2\,df = N_0 B_N\,|H_R(f_0)|^2\end{aligned}$$

[1] Il pedice $_{RF}$ sta per *radio frequency* ed indica l'occupazione di banda a frequenze positive di un segnale modulato.

Il termine $B_{RF} \leq B_N \leq B_\upsilon$ 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 *equivalente* che lascia passare la stessa potenza di rumore, come rappresentato in fig. 14.1-c).

Dato che l'effettiva banda B_υ del filtro di ricezione dipende da complessità e costo del filtro, e dunque può essere pensata come *negoziabile* in sede progettuale, a volte si procede assumendo $B_N = B_{RF}$ ovvero come nel caso ideale, con l'accortezza che in tal caso i valori di SNR calcolati al § 14.2 saranno pari al *massimo* possibile, a cui *defalcare* successivamente le penalizzazioni legate alla effettiva implementazione.

14.1.3 Demodulazione del processo di rumore

Il rumore $\nu(t)$ in uscita dal filtro di ricezione $H_R(f)$ è un processo ergodico bianco a media nulla di tipo *passa-banda*, e può quindi essere descritto nei termini delle sue componenti analogiche di bassa frequenza:

$$\nu(t) = \nu_c(t) \cos \omega_0 t - \nu_s(t) \sin \omega_0 t \tag{14.2}$$

Allo scopo di valutare la densità di potenza $\mathcal{P}_{\nu_c,\nu_s}(f)$ delle c.a. di b.f. di $\underline{\nu}(t)$, facciamo riferimento alla figura 14.2 che a sinistra mostra la densità di potenza $\mathcal{P}_n(f) = \frac{N_0}{2}$ del rumore $n(t)$ in ingresso ad un filtro di ricezione $H_R(f)$ ideale e con risposta in frequenza unitaria $|H_R(f_0)|^2 = 1$; pertanto risulta

$$\mathcal{P}_\nu(f) = \frac{N_0}{2}\mathrm{rect}_{B_N}(f - f_0) + \frac{N_0}{2}\mathrm{rect}_{B_N}(f + f_0)$$

e quindi dato che $\mathcal{P}_\nu^+(f) = \mathcal{P}_\nu^-(-f)$ la (11.36) di pag. 362 fornisce

$$\mathcal{P}_{\nu_c}(f) = \mathcal{P}_{\nu_s}(f) = \mathcal{P}_\nu^+(f + f_0) + \mathcal{P}_\nu^-(f - f_0) = N_0 \mathrm{rect}_{B_N}(f)$$

Come discusso al § 11.4.4.2, $\nu_c(t)$ e $\nu_s(t)$ sono due processi congiuntamente gaussiani, ergodici, a media nulla ed incorrelati e pertanto statisticamente indipendenti in quanto gaussiani. Abbiamo inoltre verificato come presentino anche uguale varianza (e

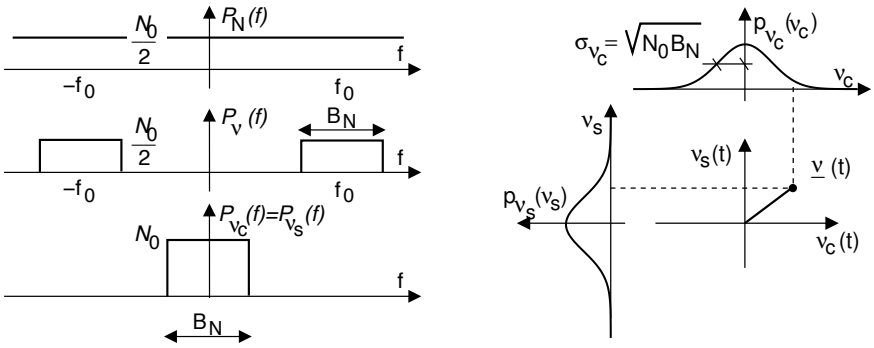


Figura 14.2: Densità spettrale e d.d.p. delle c.a. di b.f. del rumore demodulato

potenza), a sua volta uguale alla potenza del rumore filtrato $\mathcal{P}_\nu$, ovvero

$$\sigma^2_{\nu_c} = \sigma^2_{\nu_s} = \mathcal{P}_\nu = N_0 B_N$$

Il lato destro di fig. 14.2 rappresenta come nelle condizioni descritte la posizione di $\underline{\nu}(t)$ nel piano dell'inviluppo complesso sia una v.a. bidimensionale a componenti gaussiane indipendenti e con identica d.d.p.; osserviamo inoltre che nel caso in cui la banda di $\nu(t)$ sia *stretta rispetto a* f_0, l'inviluppo complesso $\underline{\nu}(t) = \nu_c(t) + j\nu_s(t)$ evolve *lentamente* rispetto alla velocità di rotazione di $\underline{\nu}(t)\, e^{j\omega_0 t}$.

In definitiva quindi, operando una demodulazione coerente in fase ed in quadratura (§ 12.2.3.1) del segnale ricevuto, nelle componenti analogiche risultanti saranno presenti i termini additivi $\nu_c(t)$ e $\nu_s(t)$, entrambi di potenza $\mathcal{P}_\nu = N_0 B_N$.

14.1.4 SNR di sistema

La quantità

$$SNR_0 = \frac{\mathcal{P}_x}{WN_0} = SNR_0 \tag{14.3}$$

individua il parametro *di sistema* (o di *riferimento*) rispetto al quale confrontare l'SNR ottenuto per i tipi di modulazione discussi al cap. 12. La (14.3) è definita a partire dai valori delle *condizioni operative*, ovvero la potenza ricevuta $\mathcal{P}_x$, il livello di rumore $N_0/2$, e la massima frequenza W del segnale modulante; viceversa non dipende dai parametri *di trasmissione*, come l'indice di modulazione. In pratica SNR_0 corrisponde all'SNR *in assenza di modulazione*, ovvero ciò che si otterrebbe ricevendo direttamente il segnale di banda base con potenza $\mathcal{P}_x$ in presenza di rumore additivo $\mathcal{P}_N(f) = N_0/2$ attraverso un filtro passa basso con banda $2W$.

Notiamo infine (e questo è valido anche per i casi che seguono) che SNR_0 può riferirsi indifferentemente sia alle potenze di segnale che a quelle disponibili (vedi § 18.1.1.3), in quanto

$$SNR_0 = \frac{\mathcal{P}_x}{WN_0} = \frac{\mathcal{P}_x}{WN_0}\frac{4R_g}{4R_g} = \frac{W_{d_x}}{W_{d_N}}$$

14.2 Prestazioni delle trasmissioni AM

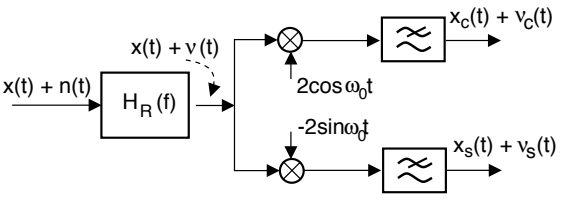
Esprimiamo innanzitutto il segnale modulato nei termini delle sue componenti analogiche

$$x_{AM}(t) = x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t$$

a cui sommare il rumore *filtrato* $\nu(t)$ (eq. (14.2)).

All'uscita da un demodulatore coerente in fase e quadratura (§ 12.2.3.1) si osserva quindi un segnale $d(t)$ espresso dalle relative c.a di b.f.

$$\begin{cases} d_c(t) = x_c(t) + \nu_c(t) \\ d_s(t) = x_s(t) + \nu_s(t) \end{cases}$$



	$x_c(t)$	$\mathcal{P}_x$	$\mathcal{P}_{x_c}$	B_N	$\mathcal{P}_{\nu_c}$	SNR
BLD-PS	$m(t)$	$\frac{1}{2}\mathcal{P}_m$	$\mathcal{P}_m = 2\mathcal{P}_x$	$2W$	$2WN_0$	$\frac{\mathcal{P}_x}{WN_0} = SNR_0$
BLU-PS	$\frac{1}{\sqrt{2}}m(t)$	$\frac{1}{2}\mathcal{P}_m$	$\frac{1}{2}\mathcal{P}_m = \mathcal{P}_x$	W	WN_0	SNR_0
BLD-PI	$\sqrt{\eta}\left(a_p + m(t)\right)$	$\frac{1}{2}\mathcal{P}_m$	$\eta\mathcal{P}_m = 2\eta\mathcal{P}_x$	$2W$	$2WN_0$	$\eta \cdot SNR_0$

Tabella 14.1: Potenza di segnale e di rumore dopo demodulazione AM

mentre tra la potenza del segnale ricevuto $x(t)$ e quella delle sue c.a. di b.f. sussiste[2] la relazione

$$\mathcal{P}_x = \frac{1}{2}\mathcal{P}_{x_c} + \frac{1}{2}\mathcal{P}_{x_s} \tag{14.4}$$

14.2.1 Potenza di segnale e di rumore dopo demodulazione ed SNR

Nel caso di modulazione AM il segnale modulante viene ricavato a partire dalla sola componente in fase $d_c(t) = x_c(t) + \nu_c(t)$, i cui termini identifichiamo come componenti di segnale e di rumore, ottenendo così l'SNR *dopo demodulazione*

$$SNR_d = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} \tag{14.5}$$

Il valore della (14.5) per il caso di modulazione BLD-PS, BLU-PS e BLD-PI (cap. 12) è calcolato ai §§ seguenti a partire dall'espressione

$$x_{BLD}(t) = x_c(t)\cos\omega_0 t = \left(a_p + m(t)\right)\cos\omega_0 t$$

a parità di SNR_0, ossia considerando fissi W (di $m(t)$), N_0 (del rumore) e la $\mathcal{P}_x$ ricevuta, ed i risultati riportati in tab.[3] 14.1 assieme alle grandezze che concorrono al calcolo.

La banda di rumore (§ 14.1.2) indicata in 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 otteniamo sono *i migliori* possibili, dato che se $B_N > B_{RF}$, l'SNR risulterà peggiore.

14.2.1.1 Modulazione BLD-PS

La prima riga di tab. 14.1 riassume come per $x_{AM}(t) = m(t)\cos\omega_0 t$ si ottenga $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$ ovvero $\mathcal{P}_m = 2\mathcal{P}_x$, e dato che $\mathcal{P}_{x_c} = \mathcal{P}_m$, a numeratore della (14.5) possiamo scrivere $\mathcal{P}_{x_c} = \mathcal{P}_m = 2\mathcal{P}_x$. Per quanto riguarda il denominatore, nel caso BLD la banda di $x(t)$ è pari a $2W$ ovvero al doppio di quella di $m(t)$, e quindi con una densità

[2] Infatti i segnali $x_c(t)\cos\omega_0 t$ e $x_s(t)\sin\omega_0 t$ risultano *ortogonali*, e le potenze si sommano. Volendo sviluppare i calcoli, possiamo valutare $\mathcal{P}_x$ come

$$\begin{aligned}\mathcal{P}_x &= E\left\{(x_{AM}(t))^2\right\} = E\left\{(x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t)^2\right\} = \\ &= E\left\{(x_c(t)\cos\omega_0 t)^2\right\} + E\left\{(x_s(t)\sin\omega_0 t)^2\right\} - 2E\left\{x_c(t)\,x_s(t)\cos\omega_0 t\sin\omega_0 t\right\}\end{aligned}$$

Possiamo ora aggiungere ad entrambe le portanti una fase aleatoria uniforme in modo da renderle anch'esse processi, indipendenti da $x_c(t)$ ed $x_s(t)$. Al § 7.5.3 si è mostrato che il prodotto di processi indipendenti ed a media nulla ha potenza pari al prodotto delle potenze, e dunque i primi due termini sono rispettivamente pari a $\frac{1}{2}\mathcal{P}_{x_c}$ e $\frac{1}{2}\mathcal{P}_{x_v}$. Per quanto riguarda il terzo termine, esso rappresenta il valore atteso del prodotto di processi indipendenti ed a media nulla, e dunque è nullo. Infine, sviluppando i calcoli a partire dalle medie temporali anziché di insieme si perviene al medesimo risultato.

[3] La tabella estende quella al § 12.1.4, rispetto alla quale si considera il termine k_a ora inglobato in $m(t)$.

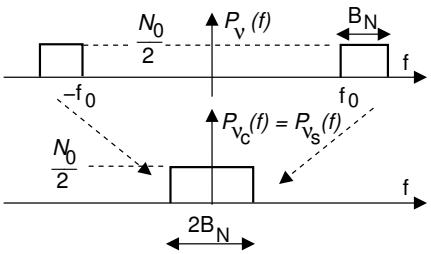
$\mathcal{P}_{\nu_c}(f) = N_0$ (vedi fig. 14.2) la potenza di rumore vale $\mathcal{P}_{\nu_c} = 2WN_0$, e dunque

$$SNR_{BLD} = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} = \frac{2\mathcal{P}_x}{2WN_0} = \frac{\mathcal{P}_x}{WN_0} = SNR_0$$

ovvero le prestazioni dopo demodulazione sono esattamente pari all'SNR_0 di riferimento definito al § 14.1.4: dunque la modulazione BLD-PS *non altera* il rapporto SNR_0 di banda base, ovvero è come se il processo di modulazione fosse *trasparente*.

14.2.1.2 Modulazione BLU-PS

In questo caso si ha $x_{AM}(t) = \frac{1}{\sqrt{2}} m(t) \cos\omega_0 t - \frac{1}{\sqrt{2}} \hat{m}(t) \sin\omega_0 t$ (vedi § 12.1.4 e 12.4.5) da cui si può ottenere[4] $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$ come per il caso BLD-PS, e dato che ora risulta $\mathcal{P}_{x_c} = E\left\{\left(\frac{1}{\sqrt{2}} m(t)\right)^2\right\} = \frac{1}{2}\mathcal{P}_m$, al numeratore di (14.5) possiamo scrivere $\mathcal{P}_{x_c} = \mathcal{P}_x$, *la metà* del caso precedente.



Per quanto riguarda invece la componente di rumore, alla figura a lato si mostra come anche la densità di potenza $\mathcal{P}_\nu(f)$ del rumore che attraversa $H_R(f)$ occupa una banda a sua volta dimezzata, e quindi dopo demodulazione la densità di potenza $\mathcal{P}_{\nu_c}(f)$ occupa una banda $\pm B_N$ come nel caso AM-BLD, ma possiede una valore $N_0/2$ *uguale* a quello della $\mathcal{P}_\nu(f)$ in ingresso, e non *doppio* come al § 14.1.3.

Pertanto la potenza $\mathcal{P}_{\nu_c}$ del rumore demodulato sul ramo in fase (con un filtro $H_R(f)$ ideale ed a banda minima $B_N = W$) è pari a $2W \cdot N_0/2 = WN_0$, permettendo di scrivere

$$SNR_{BLU} = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} = \frac{\mathcal{P}_x}{WN_0} = SNR_0$$

e cioè si ottengono prestazioni *identiche* a quelle del caso AM-BLD, ma utilizzando solo *metà* della banda altrimenti necessaria.

14.2.1.3 Modulazione BLD-PI

In questo caso il segnale ricevuto ha espressione

$$x_{PI}(t) = \sqrt{\eta}\left(a_p + m(t)\right)\cos\omega_0 t$$

dove $\eta = \frac{\mathcal{P}_m}{a_p^2 + \mathcal{P}_m}$ è pari *all'efficienza* della BLD-PI introdotta al § 12.1.1.4, in modo da poter scrivere che la potenza del segnale ricevuto vale[5] $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$, uguale ai due casi

[4] Riprendendo l'approccio adottato alla nota 2, consideriamo le portanti in fase e quadratura come realizzazioni di un processo armonico con potenza 1/2, moltiplicate per un processo statisticamente indipendente $m(t)/\sqrt{2}$ con potenza $\mathcal{P}_m/2$. La potenza di ciascuna c.a. di b.f. è il prodotto di queste due, e dunque partendo dalla (14.4)

$$\mathcal{P}_x = \frac{1}{2}\mathcal{P}_{x_c} + \frac{1}{2}\mathcal{P}_{x_s} = \frac{1}{2}\cdot 2\cdot \mathcal{P}_{x_c} = 1\cdot\frac{1}{2}\cdot\mathcal{P}_m = \frac{1}{2}\cdot\mathcal{P}_m$$

[5] Infatti, considerando nuovamente la portante in fase come un processo armonico indipendente da $m(t)$ possiamo scrivere $\mathcal{P}_x = \eta\left(a_p^2 + \mathcal{P}_m\right)\cdot 1/2 = 1/2\cdot\mathcal{P}_m$ dato che $E\left\{(a_p + m(t))^2\right\} = a_p^2 + \mathcal{P}_m$, in quanto

precedenti.

Per valutare l'SNR_d, a numeratore della (14.5) non consideriamo l'intera potenza $\mathcal{P}_{x_c}$ di $x_c(t) = \sqrt{\eta}\left(a_p + m(t)\right)$, ma solo quella della sua componente *utile* $u(t) = \sqrt{\eta} m(t)$, che ha potenza $\mathcal{P}_u = \eta\mathcal{P}_m = 2\eta\mathcal{P}_x$, mentre la potenza ηa_p^2 si riferisce invece alla portante non modulata, e non trasporta informazione. Dato che per quanto riguarda il rumore demodulato siamo nella stessa condizione del caso AM-BLD ovvero $\mathcal{P}_{\nu_c} = 2WN_0$, possiamo scrivere

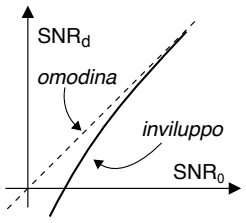
$$SNR_{PI} = \frac{\mathcal{P}_u}{\mathcal{P}_{\nu_c}} = \frac{2\eta\mathcal{P}_x}{2WN_0} = \eta\frac{\mathcal{P}_x}{WN_0} = \eta \cdot SNR_0$$

e dunque constatiamo che la presenza della portante comporta una riduzione di prestazioni in misura esattamente pari all'efficienza $\eta = \frac{\mathcal{P}_m}{a_p^2+\mathcal{P}_m}$.

L'analisi esposta si riferisce però ad una demodulazione IQ coerente, mentre per il caso BLD-PI si usa il demodulatore *di inviluppo* (§ 12.2.5), che fornisce come risultato il modulo dell'inviluppo complesso ovvero

$$d(t) = \left|\underline{x}(t) + \underline{\nu}(t)\right| = \sqrt{\left[\sqrt{\eta}\left(a_p + m(t)\right) + \nu_c(t)\right]^2 + \nu_s^2(t)}$$

Finché $\left|\underline{\nu}(t)\right|$ è piccolo e trascurabile rispetto ad a_p, ci si ritrova approssimativamente nel caso precedente; al contrario per bassi valori di SNR_0 la potenza utile $\mathcal{P}_u$ diviene una frazione di $\mathcal{P}_x$ ancora più piccola di quanto non sia $\mathcal{P}_u = \eta\mathcal{P}_m = 2\eta\mathcal{P}_x$, dando luogo ad un SNR peggiore del caso di demodulazione in fase e quadratura, come illustrato in figura.



14.3 Prestazioni della modulazione di frequenza

Occupiamoci ora della valutazione dell'SNR dopo demodulazione per il caso di una trasmissione FM (§ 12.3), analizzando come esso dipenda dalle condizioni di ricezione (potenza ricevuta $\mathcal{P}_x$, densità di potenza del rumore $N_0/2$ e banda del ricevitore B_N) e dai parametri di trasmissione (indice di modulazione β e banda del segnale modulante W). Anticipiamo che la natura *non lineare* della modulazione FM porterà a sviluppi del tutto diversi dal caso dell'AM, infatti troveremo che se la potenza del rumore in ingresso al ricevitore non è eccessiva

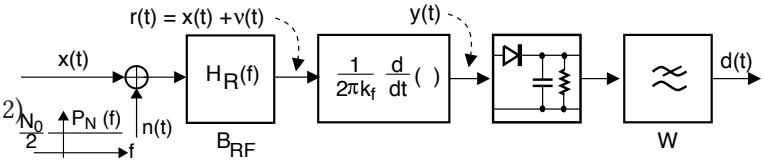
- quando la potenza del segnale ricevuto $\mathcal{P}_x$ *aumenta*, quella del segnale demodulato resta *costante*, mentre invece *diminuisce* quella del rumore dopo demodulazione;
- l'SNR dopo demodulazione *migliora* all'aumentare della banda occupata.

Per arrivare a questi risultati non banali, valutiamo innanzitutto ciò che accade nella ricezione di una portante non modulata, e quindi analizziamo 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.

$E\left\{a_p \cdot m(t)\right\} = 0$ qualora $m(t)$ sia a media nulla.

14.3.1 Rumore dopo demodulazione FM

L'analisi viene svolta considerando un demodulatore *a discriminatore* (§ 12.3.2.2) alla cui uscita $r(t)$ del filtro di ricezione è presente una portante *non modulata* $x(t)$ di ampiezza[6] $A = \sqrt{2\mathcal{P}_x}$, oltre che un rumore gaussiano bianco limitato in banda $\nu(t)$, ovvero



Demodulatore FM a discriminatore

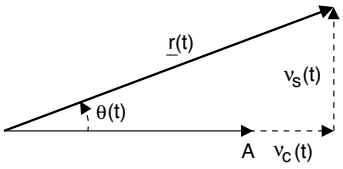
$$r(t) = A\cos\omega_0 t + \nu_c(t)\cos\omega_0 t - \nu_s(t)\sin\omega_0 t$$

La banda del filtro $H_R(f)$ (e dunque di $\nu(t)$) deve essere sufficiente a far passare le frequenze che *sarebbero* presenti se la portate *fosse* modulata, e che nel caso FM può essere stimata applicando la *regola di Carson* (eq. 12.19), ossia $B_{RF} = B_C \simeq 2W(\beta + 1)$.

In presenza di una portante non modulata, le componenti analogiche di bassa frequenza dell'inviluppo complesso $\underline{r}(t)$ del segnale ricevuto

$$\underline{r}(t) = r_c(t) + jr_s(t) \quad \text{sono espresse come} \quad \begin{cases} r_c(t) = A + \nu_c(t) \\ r_s(t) = \nu_s(t) \end{cases} \tag{14.6}$$

di cui a fianco è rappresentata la costruzione vettoriale: $r_c(t)$ è la somma tra l'ampiezza A della portate ed una v.a. $\nu_c(t)$ gaussiana a media nulla e deviazione standard $\sigma = \sqrt{N_0 B_N} \geq \sqrt{N_0 B_{RF}}$, mentre $r_s(t)$ consiste in un'altra v.a. $\nu_s(t)$ della stessa natura di $\nu_c(t)$ ma ad essa incorrelata[7].



Ricordando ora che nel caso FM il *segnale informativo* è legato alla *derivata* della fase $\theta(t)$ di $\underline{r}(t)$, esprimiamo $r(t)$ mettendo $\theta(t)$ in evidenza

$$r(t) = \Re\left\{\underline{r}(t)\,e^{j\omega_0 t}\right\} = \Re\left\{\left|\underline{r}(t)\right| e^{j\theta(t)} e^{j\omega_0 t}\right\} = \left|\underline{r}(t)\right|\cos\left(\omega_0 t + \theta(t)\right)$$

in cui possiamo considerare il termine $\left|\underline{r}(t)\right|$ *rimosso* dal limitatore (vedi § 12.3.2.2) che usualmente è anteposto al discriminatore. Il segnale $y(t)$ in uscita dal derivatore è quindi descritto (a parte il segno) come

$$\begin{aligned} y(t) = \frac{1}{2\pi k_f}\frac{d}{dt}r(t) \Rightarrow \frac{1}{2\pi k_f}\frac{d}{dt}\cos\left(\omega_0 t + \theta(t)\right) = \\ = \left(\frac{f_0}{k_f} + \frac{1}{2\pi k_f}\frac{d}{dt}\theta(t)\right)\sin\left(\omega_0 t + \theta(t)\right) \end{aligned}$$

e viene a sua volta elaborato da parte del demodulatore di inviluppo come fosse un segnale BLD-PI (§ 12.1.1.2), fornendo in definitiva un segnale *demodulato* dovuto al solo rumore

[6] Con questa posizione, la potenza della portante risulta $\frac{\left(\sqrt{2\mathcal{P}_x}\right)^2}{2} = \frac{2\mathcal{P}_x}{2} = \mathcal{P}_x$.

[7] Si veda il § 14.4.1 per una analisi più approfondita degli aspetti statistici della questione, che portano a definire $\rho = \left|\underline{r}(t)\right|$ una v.a. di RICE.

$$d(t) = \frac{1}{2\pi k_f}\frac{d}{dt}\theta(t) = \nu_d(t) \tag{14.7}$$

14.3.2 Caso di basso rumore

Con riferimento all'ultima figura, osserviamo che qualora $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma^2_{\nu_c} = \sigma^2_{\nu_s} = N_0 B_N$ i valori di $\nu_c(t)$ e $\nu_s(t)$ risultano *piccoli* rispetto ad A, e l'inviluppo complesso ricevuto $\underline{r}(t)$ rimane *prossimo* a quello della portante non modulata, dato che in questo caso $\underline{\nu}(t)$ ha modulo *abbastanza* più piccolo di A. L'angolo $\theta(t)$ che compare nella (14.7) può dunque essere approssimato come

$$\theta(t) = \arctan\frac{\nu_s(t)}{A+\nu_c(t)} \simeq \arctan\frac{\nu_s(t)}{A} \simeq \frac{\nu_s(t)}{A}$$

la cui densità spettrale di potenza vale

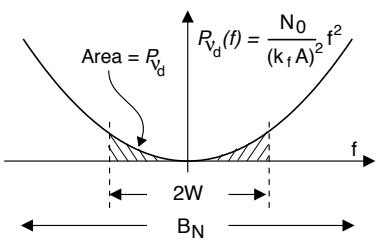
$$\mathcal{P}_\theta(f) = \frac{1}{A^2}\mathcal{P}_{\nu_s}(f) = \frac{N_0}{A^2} \tag{14.8}$$

in quanto $\mathcal{P}_{\nu_s}(f) = N_0$ come discusso al § 14.1.3. Ricordiamo ora (vedi § 3.6) che l'operazione di derivata svolta dal discriminatore equivale a moltiplicare lo spettro di ampiezza del segnale in ingresso per $j2\pi f$, ovvero moltiplicare la sua densità di potenza per $(2\pi f)^2$: applichiamo questo risultato per ottenere la densità di potenza di $\nu_d(t)$ (14.7) a partire dalla (14.8), in modo che la densità di potenza del rumore *demodulato* $\nu_d(t)$ risulti

$$\mathcal{P}_{\nu_d}(f) = \frac{1}{(2\pi k_f)^2}(2\pi f)^2\mathcal{P}_\theta(f) = \left(\frac{f}{k_f}\right)^2\frac{N_0}{A^2} = \frac{N_0}{(k_f A)^2}f^2$$

e quindi la relativa potenza totale $\mathcal{P}_{\nu_d} = \sigma^2_{\nu_d}$ si calcola come

$$\mathcal{P}_{\nu_d} = \int_{-\infty}^{\infty}\mathcal{P}_{\nu_d}(f)\,df = 2\int_0^W \frac{N_0}{(k_f A)^2}f^2 df = \\ = 2\frac{N_0}{(k_f A)^2}\cdot\frac{f^3}{3}\Big|_0^W = \frac{2}{3}\frac{N_0}{(k_f A)^2}W^3 \tag{14.9}$$

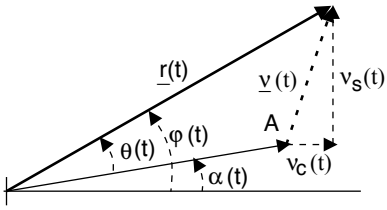


in cui W è la banda del segnale *modulante* (se ci fosse), ed il rumore è limitato in tale banda in virtù del filtro passa basso posto a valle del discriminatore[8].

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 demodulato ha un andamento *parabolico*.

Segnale presente Continuando nell'ipotesi di basso rumore ovvero $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma^2_{\nu_c} = \sigma^2_{\nu_s} = N_0 B_N$, possiamo osservare che (vedi fig. a lato) la presenza di una fase

[8] Si noti che le potenze $\sigma^2_{\nu_c}$ e $\sigma^2_{\nu_s}$ delle c.a. di b.f. del rumore in ingresso al discriminatore sono invece relative alla banda B_N, $\geq$ di quella B_{RF} del segnale *modulato*.



modulante $\alpha(t)$ nel segnale

$$x_{FM}(t) = A\cos(2\pi f_0 t + \alpha(t))$$

comporta che la fase $\varphi(t)$ dell'inviluppo complesso del segnale ricevuto $r(t)$ è costituita dalla somma tra $\alpha(t)$ e l'angolo $\theta(t)$ dovuto al rumore sovrapposto alla portante di ampiezza A, cioè $\varphi(t) = \alpha(t) + \theta(t)$. Pertanto l'uscita (14.7) del discriminatore diviene

$$d(t) = \frac{1}{2\pi k_f}\frac{d}{dt}\alpha(t) + \frac{1}{2\pi k_f}\frac{d}{dt}\theta(t) = s_d(t) + \nu_d(t)$$

ed il rapporto tra le potenze dei due termini definisce l'SNR dopo demodulazione come $SNR_d = \frac{\mathcal{P}_{s_d}}{\mathcal{P}_{\nu_d}}$, dove quindi $\mathcal{P}_{s_d}$ è la potenza di segnale *utile* demodulato $s_d(t) = \frac{1}{2\pi k_f}\frac{d}{dt}\alpha(t)$, e $\mathcal{P}_{\nu_d}$ è la potenza del rumore demodulato calcolata alla (14.9).

Ricordando (§ 11.2.2) che $\alpha(t) = 2\pi k_f \int_{-\infty}^{t} m(\tau)\,d\tau$, per la potenza di $s_d(t)$ si ottiene[9] $\mathcal{P}_{s_d} = \mathcal{P}_m = \int_{-W}^{W} \mathcal{P}_m(f)\,df$, e quindi

$$SNR_d = \frac{\mathcal{P}_{s_d}}{\mathcal{P}_{\nu_d}} = \frac{\mathcal{P}_m}{\frac{2}{3}\frac{N_0}{(k_f A)^2}W^3} = 3\frac{\mathcal{P}_m k_f^2}{W^2 N_0 W}\frac{A^2}{2} = 3\frac{\sigma_{f_d}^2}{W^2}\frac{\mathcal{P}_x}{N_0 W} = 3\beta^2 SNR_0$$

avendo sostituito $\mathcal{P}_m k_f^2$ con $\sigma_{f_d}^2$ (vedi sotto), $\frac{A^2}{2}$ con la potenza della portante ricevuta $\mathcal{P}_x$, $\frac{\sigma_{f_d}}{W}$ con l'indice di modulazione β (§ 12.3.3.4), e $\frac{\mathcal{P}_x}{N_0 W}$ con l'SNR *di sistema* SNR_0 (§ 14.1.4). Il risultato ottenuto conferma la seconda affermazione di inizio sezione: si ha un *miglioramento* rispetto all'SNR_0 (e dunque rispetto all'AM) *tanto maggiore* quanto maggiore *è la banda occupata* dal segnale modulato $B_{RF} \simeq 2W(\beta + 1)$ (eq. (12.19) a pag. 389), ovvero quanto più è grande l'indice di modulazione β.

Discussione dei passaggi 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 $f_i(t)$ (§ 12.3) rispetto a quella della portante f_0. Ricordiamo quindi che $f_i(t) = \frac{1}{2\pi}\frac{d}{dt}\psi(t)$ in cui $\psi(t)$ è la fase istantanea $\psi(t) = 2\pi f_0 t + \alpha(t)$, e dato che per l'FM $\alpha(t) = 2\pi k_f \int_{-\infty}^{t} m(\tau)\,d\tau$ si ottiene

$$f_d(t) = \frac{1}{2\pi}\frac{d}{dt}\left(2\pi f_0 t + 2\pi k_f \int_{-\infty}^{t} m(\tau)\,d\tau\right) - f_0 = f_0 + k_f m(t) - f_0 = k_f m(t)$$

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 proporzionale alla larghezza di banda del segnale modulato[10]. D'altra parte, questo risultato è un aspetto della conversione AM-FM che avviene per alto indice di modulazione, come

[9] Dato che gli operatori di derivata ed integrale si annullano, ovvero $\mathcal{P}_{s_d} = Pot\left\{\frac{1}{2\pi k_f}\frac{d}{dt}\alpha(t)\right\} = Pot\left\{\frac{1}{2\pi k_f}\frac{d}{dt}2\pi k_f \int_{-\infty}^{t} m(\tau)\,d\tau\right\} = Pot\{m(t)\} = \mathcal{P}_m$. In definitiva, abbiamo semplicemente demodulato!

[10] Infatti, il rapporto $\frac{\sigma_{f_d}}{W}$ definito al § 12.3.3.4 come indice di modulazione β_p, rappresenta appunto una misura del rapporto tra l'occupazione di banda *efficace* del segnale modulato, e la massima frequenza W presente nel segnale modulante.

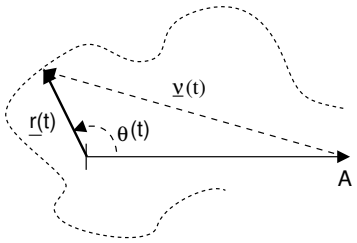
descritto al § 12.3.3.3.

Discussione del risultato Notiamo innanzitutto che se $\beta < \frac{1}{\sqrt{3}} \simeq 0,58$ il valore di $SNR = 3\beta^2 SNR_0$ non aumenta affatto, anzi le prestazioni peggiorano. Ma con bassi indici di modulazione abbiamo già visto (§ 12.3.3.1) che l'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 $\beta > \frac{1}{\sqrt{3}}$: se ad esempio $\beta = 5$ si ottiene $3\beta^2 = 75$ 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[11] avviene *a spese dell'occupazione di banda*, e pertanto costituisce una manifestazione del *compromesso banda-potenza*, vedi pagg. 461 e 565.

Verrebbe ora quasi il desiderio di aumentare indefinitamente β (nei limiti della banda disponibile) per migliorare a piacere l'SNR. Peccato non sia possibile, dato che ad un certo punto l'analisi effettuata *perde validità*: infatti, aumentando β anche la banda di rumore del ricevitore deve crescere, essendo aumentata la banda del segnale modulato. Pertanto le condizioni $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma^2_{\nu_c} = \sigma^2_{\nu_s} = N_0 B_N$ non sono più verificate, con le conseguenze illustrate di seguito.

14.3.3 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*, ed all'aumentare del rumore l'SNR degrada molto rapidamente. Per indagarne le cause facciamo riferimento allo schema a lato, che mostra l'inviluppo complesso della portante non modulata A, del rumore in ingresso $\underline{\nu}(t)$, e del segnale ricevuto $\underline{r}(t)$, notando 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 alla figura 14.3-a. 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 crepitìo indistinto. Si è trovato che questo effetto si manifesta a partire da un SNR_0 *di sistema* inferiore[12] a 10-25 dB, e per valori SNR_0 minori di tale valore l'effetto aumenta molto rapidamente, cosicché si parla di *effetto soglia*. Le curve di 14.3-b riportano un

[11] Miglioramento che può essere sfruttato quando ad esempio il collegamento è di tipo *punto-punto*, come nel caso di un ponte radio con antenne direttive od una comunicazione satellitare, in modo da contenere la potenza irradiata entro il *cono di emissione* e non invadere lo spettro radio riservato ad altre trasmissioni.

[12] L'effetto soglia interviene *prima* per i valori di β più elevati, vedi fig. 14.3-b.

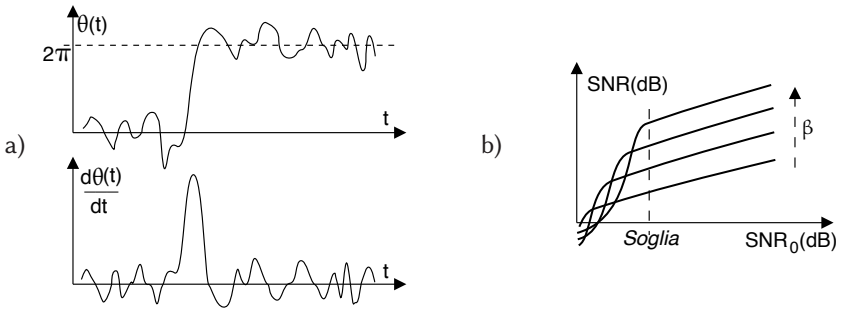


Figura 14.3: a) - Rumore impulsivo dopo demodulazione FM; b) - effetto soglia

tipico andamento dell'SNR dopo demodulazione, con l'indice β che svolge il ruolo di parametro, e possiamo osservare come con un SNR_0 inferiore alla soglia le prestazioni degradino rapidamente. Si è trovato che demodulando con un PLL, anziché con un discriminatore, il valore di soglia si riduce di circa 3 dB.

Nella pratica comune il segnale di rumore può essere costituito da una *interferenza* dovuta ad una emittente adiacente (ossia con una portante prossima a quella della emittente sintonizzata) che *sovramodula*, ovvero adotta un indice di modulazione troppo elevato, ed invade la banda delle emittenti contigue.

Esercizio Sia dato un trasmettitore FM con potenza trasmessa 1 Watt e segnale modulante $m(t)$ con banda $\pm W = \pm 10$ MHz. Un collegamento con attenuazione disponibile A_d = 100 dB lo interfaccia ad un ricevitore con temperatura di sistema T_{ei} = 2900 ^{o}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 § 18.2, dove è mostrato che $T_{e_i} = T_0 (F-1) + T_g = T_0 F$ 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) Qui è utilizzata la relazione $W_R = W_T G_d$ dal § 19.1, in modo da scrivere $SNR = 3\beta^2 SNR_0 = 3\beta^2 \frac{W_R}{N_0 W} = 3\beta^2 \frac{W_T G_d}{N_0 W}$; 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 $\beta_{min} = \sqrt{\frac{SNR \cdot N_0 W}{3 \cdot W_T G_d}} = \sqrt{\frac{10^4 \cdot 4 \cdot 10^{-20} \cdot 10^7}{3 \cdot 10^{-10}}} = 3.65$. Applicando la regola di Carson per la banda: $B_{RF} \simeq 2W \cdot (\beta + 1) = 2 \cdot 10^7 \cdot 4.65 = 9.3 \cdot 10^7 = 93$ MHz.

3) La validità dei risultati 2) dipende dal verificarsi delle condizioni di basso rumore, ovvero deve risultare $W_R \gg \sigma^2_{\nu_c} = \sigma^2_{\nu_s} = N_0 B_N = N_0 B_{RF} = 4 \cdot 10^{-20} \cdot 9.3 \cdot 10^7 = 3.72 \cdot 10^{-12}$ Watt, ma poiché $W_R = \frac{W_T}{A_d} = \frac{1}{10^{10}} = 10^{-10}$, si ha $\frac{W_R}{\sigma^2_{\nu_c}} = \frac{10^{-10}}{3.72 \cdot 10^{-12}} = 26$, che soddisfa abbastanza bene l'esigenza di *basso rumore*. Per trovare β_{Max} partiamo dal vincolo che debba risultare $W_R \geq 10 \cdot \sigma^2_{\nu_c} = 10 \cdot N_0 \cdot B_{RF} = 10 \cdot N_0 \cdot 2W \cdot (\beta_{Max} + 1)$, da cui otteniamo $\beta_{Max} = \frac{W_R}{10 \cdot N_0 \cdot 2W} - 1 = \frac{10^{-10}}{8 \cdot 10^{-12}} - 1 = 12.5 - 1 = 11.5$, al quale corrisponde una banda $B_{RF} = 2W \cdot (\beta_{Max} + 1) = 2 \cdot 10^7 \cdot 12.5 = 250$ MHZ, ed un guadagno di $SNR = 10 \lg_{10} 3\beta^2_{Max} \simeq 26$ dB, mentre con β_{min} nominale si sarebbe ottenuto $10 \lg_{10} \left(3 \cdot 3.65^2\right) = 16$ dB.

4) Il concetto di margine è introdotto al § 19.1; 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': $SNR = 10^4 = 3\beta'^2 \frac{W_T G_d G'_d}{N_0 W} = 3\beta^2 \frac{W_T G_d}{N_0 W} \frac{\beta'^2}{\beta^2} G'_d$; 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.

14.3.4 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 una densità spettrale $\mathcal{P}_m(f)$ a sua volta bianca, l'$SNR(f)$ alle frequenze più elevate sarebbe molto peggiore del suo valore per frequenze inferiori. Nella pratica, si possono verificare (ad esempio) i seguenti problemi:

- Nelle trasmissioni FDM-FM (vedi § 11.1.1.2), 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;
- nell'FM *broadcast* (vedi § 25.2), 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 quanto il suo ruolo è quello di *enfatizzare* le frequenze più elevate. In tal modo anche $m(t)$ presenta uno spettro *parabolico* e l'SNR sarà lo stesso a tutte le frequenze! L'alterazione introdotta su $m(t)$ viene quindi rimossa mediante una rete di *de-enfasi* posta in ricezione (praticamente un integratore, ovvero un passa-basso) tale da ripristinare l'originale sagoma spettrale del segnale, rendendo la densità di potenza del rumore costante in frequenza.

Con un po' di riflessione, ci si accorge che l'uso di una coppia enfasi-deenfasi equivale ad effettuare una trasmissione a modulazione di fase (vedi pag. 382). 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.

14.4 Detezione di sinusoide nel rumore

Concludiamo questo capitolo con un argomento diverso dai precedenti: anziché calcolare l'SNR dopo democulazione, affrontiamo il problema di decidere se nelle vicinanze di una determinata frequenza f_0 sia presente o meno un segnale *a banda stretta*, ad esempio per effettuare una operazione di sintonizzazione automatica. A questo scopo torniamo ad occuparci della demodulazione *incoerente in fase e quadratura* introdotta al § 12.2.4, ora applicata al problema di rilevare la presenza (o meno) di una sinusoide $s(t)$ *immersa nel rumore* entro una banda B_N, affrontato mediante il formalismo della *verifica di ipotesi* (§ 6.6.1) basata sul confronto tra il valore di una variabile di osservazione ρ, che rappresenta il modulo dell'inviluppo complesso ricevuto, ed una soglia di decisione λ, da posizionare a seconda del criterio adottato. Lo scopo è quello di arrivare ad una espressione per la d.d.p. di ρ a partire dalle uscite del demodulatore in fase e quadratura, secondo lo schema di fig. 14.4. Al § 16.6 verrà adottato uno schema simile, applicato al caso della trasmissione numerica.

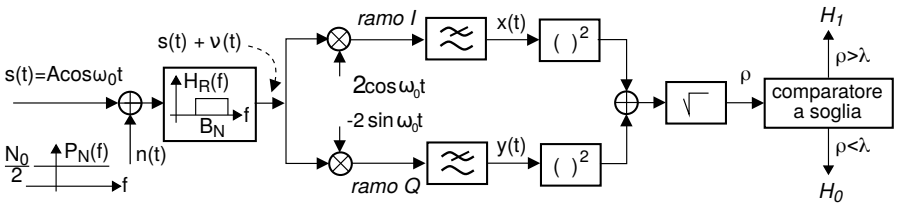


Figura 14.4: Detezione incoerente di sinusoide immersa nel rumore

14.4.1 Descrizione statistica del modulo dell'inviluppo complesso

Negli sviluppi che seguono scegliamo di indicare le uscite in fase e quadratura del demodulatore IQ rispettivamente come $x(t)$ e $y(t)$. Se in ingresso è presente il solo rumore $n(t)$, $x(t)$ e $y(t)$ corrispondono alle c.a. di b.f. $\nu_c(t)$ e $\nu_s(t)$ della sua versione filtrata[13]; se invece in ingresso è presente anche $s(t) = A\cos\omega_0 t$, nell'uscita $x(t)$ del ramo in fase troviamo anche la *componente in fase* di $s(t)$, pari ad A, che diventa dunque il valor medio della v.a. estratta da $x(t)$. La fig. 14.5-a) rappresenta la d.d.p. delle v.a. x ed y estratte dai processi $x(t)$ e $y(t)$, mostrando anche le *curve di livello* (vedi § 6.5) della gaussiana bidimensionale risultante.

Consideriamo ora che la sinusoide $s(t)$, quando presente, può in realtà avere una fase φ qualsiasi, e dunque come discusso al § 13.1.2.3 il piano (x, y) ruota dello stesso angolo φ, causando la traslazione della d.d.p. bidimensionale della stessa quantità, come illustrato in fig. 14.5-b). Dunque la differenza tra quando $s(t)$ è presente o meno consiste nello *scostamento dall'origine* del valor atteso della gaussiana, in qualsiasi direzione, e per questo in ingresso al comparatore di fig. 14.4 viene posta la grandezza $\rho = \sqrt{x^2 + y^2}$, che è a sua volta una v.a., ed il cui valore viene rapportato alla soglia λ.

[13] Ovvero (§ 14.1.3) $x(t)$ e $y(t)$ sono processi congiuntamente gaussiani ed incorrelati con media nulla e varianza $\sigma^2 = N_0 B_N$.

Allo scopo di valutare la d.d.p. della v.a. ρ ovvero del modulo dell'inviluppo complesso demodulato $\underline{z} = x + jy$ e dunque poter individuare λ secondo il criterio di massima verosimiglianza (§ 6.6.2.1), applichiamo i risultati ottenuti al § 6.4.2 sulle trasformazioni di v.a. , in modo da passare dalla rappresentazione cartesiana $\underline{z} = x + jy$ a quella polare $z = \rho e^{j\varphi}$, come rappresentato in fig. 14.5-c). Definiamo dunque la trasformazione in oggetto, assieme alle rispettive funzioni inverse, come

$$\begin{cases} \rho = \sqrt{x^2 + y^2} \\ \varphi = \arctan \frac{y}{x} \end{cases} \qquad \begin{cases} x = \rho \cos \varphi \\ y = \rho \sin \varphi \end{cases} \tag{14.10}$$

e mostriamo che, nei due casi di segnale assente o presente, la v.a. ρ assume rispettivamente la d.d.p. di *Rayleigh* oppure quella di *Rice*.

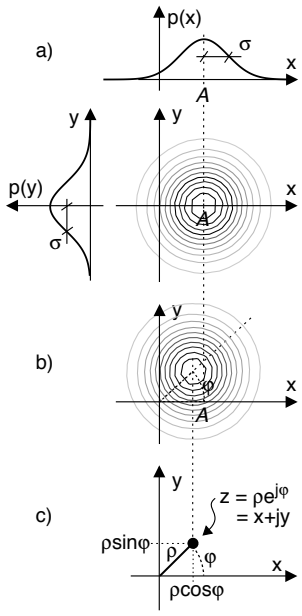


Figura 14.5: v.a. gaussiana bidimensionale in coordinate polari

Variabile aleatoria di Rayleigh In *assenza* di segnale, x ed y sono due v.a. gaussiane indipendenti, a media nulla e uguale varianza σ^2, la cui d.d.p. congiunta si ottiene[14] come prodotto delle d.d.p. marginali, e vale

$$p_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{14.11}$$

La $p_{P,\Phi}(\rho, \varphi)$ viene quindi calcolata come prescritto dalla (6.25) di pag. 166, valutando[15] le espressioni per $p_{X,Y}(x(\rho, \varphi), y(\rho, \varphi))$ e $|\mathbf{J}(x, y/\rho, \varphi)|$, e ottenendo così

$$p_{P,\Phi}(\rho, \varphi) = \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \quad \text{con} \quad \begin{cases} 0 & < \rho < & \infty \\ -\pi & < \varphi < & \pi \end{cases}$$

Le d.d.p. marginali $p_P(\rho)$ e $p_\Phi(\varphi)$ si ottengono quindi saturando[16] la d.d.p. congiunta rispetto all'altra variabile, in modo da ottenere

$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \text{ con } \rho \geq 0; \quad p_\Phi(\varphi) = \frac{1}{2\pi} \text{ con } -\pi < \varphi \leq \pi \tag{14.12}$$

---

[14] Vedi anche il § 6.5.1. Basta moltiplicare: $p_X(x)\, p_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right)$

[15] Il calcolo dei due termini si esegue come

$$p_{X,Y}(x(\rho, \varphi), y(\rho, \varphi)) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\rho^2\left(\cos^2\varphi + \sin^2\varphi\right)}{2\sigma^2}\right) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

$$|\mathbf{J}(x, y/\rho, \varphi)| = \left|\begin{bmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \varphi} \end{bmatrix}\right| = \left|\begin{bmatrix} \cos\varphi & -\rho\sin\varphi \\ \sin\varphi & \rho\cos\varphi \end{bmatrix}\right| = \rho\left(\cos^2\varphi + \sin^2\varphi\right) = \rho$$

[16] Svolgiamo il calcolo solo per la prima relazione:

$$p_P(\rho) = \int_{-\pi}^{\pi} p_{P,\Phi}(\rho, \varphi)\, d\varphi = \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \cdot \int_{-\pi}^{\pi} d\varphi = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

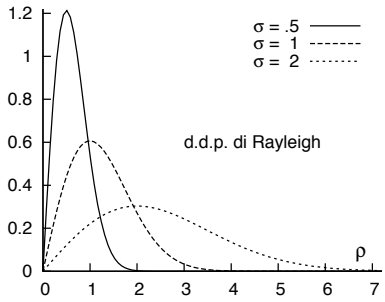


Figura 14.6: Densità di probabilità di Rayleigh

L'espressione di $p_P(\rho)$ in (14.12) prende nome di d.d.p. di Rayleigh, graficata in fig. 14.6, mentre il valor medio e la varianza della v.a. ρ valgono rispettivamente

$$m_P = \sigma\sqrt{\frac{\pi}{2}} \quad \text{e} \quad \sigma_P^2 = \sigma^2\left(2 - \frac{\pi}{2}\right) \tag{14.13}$$

E' inoltre possibile mostrare[17] che per la v.a. di Rayleigh vale la proprietà

$$Pr\{\rho > \lambda\} = \int_\lambda^\infty p_P(\rho)\, d\rho = \exp\left(-\frac{\lambda^2}{2\sigma^2}\right) \tag{14.14}$$

Il valore (14.14) può rappresentare la probabilità di *mancare un bersaglio* per una distanza superiore a λ, considerando gli errori di puntamento orizzontale e verticale entrambi gaussiani, indipendenti, a media nulla ed uguale varianza.

Variabile aleatoria di Rice Consideriamo ora il caso in cui il tono $s(t)$ *sia presente*, e senza perdita di generalità assumiamo che abbia fase $\varphi = 0$ in modo che la trasformazione (14.10) possa ancora essere applicata considerando, al posto di x, una v.a. x', sempre gaussiana con varianza σ^2, ma ora con media pari ad A, ovvero la componente in fase di $s(t)$. In questo caso il prodotto tra le d.d.p. marginali si scrive come[18]

$$p_{X',Y}(x', y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x'-A)^2 + y^2}{2\sigma^2}\right) \tag{14.15}$$

e l'operazione di cambio di variabile porta[19] alla d.d.p. $p_P(\rho)$ detta di Rice, che ha espressione

---

[17] Dato che $\frac{d}{d\rho}\exp\left(-\frac{\rho^2}{2\sigma^2}\right) = -\frac{\rho}{\sigma^2}\exp\left(-\frac{\rho^2}{2\sigma^2}\right)$, si ottiene

$$\int_\lambda^\infty \frac{\rho}{\sigma^2}e^{-\frac{\rho^2}{2\sigma^2}}d\rho = -\int_\lambda^\infty -\frac{\rho}{\sigma^2}e^{-\frac{\rho^2}{2\sigma^2}}d\rho = -\left[e^{-\frac{\rho^2}{2\sigma^2}}\right]_\lambda^\infty = e^{-\frac{\lambda^2}{2\sigma^2}}$$

[18] Infatti in questo caso risulta

$$p_{X'}(x)\, p_Y(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x'-A)^2}{2\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right)$$

[19] Sostituendo nell'esponente della (14.15) $x' = \rho\cos\varphi$ 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

$$\begin{aligned} p_{P,\Phi}(\rho,\varphi) &= p_{X',Y}\left(x'(\rho,\varphi), y = y(\rho,\varphi)\right)\left|\mathbf{J}\left(x', y/\rho,\varphi\right)\right| \\ &= \frac{\rho}{2\pi\sigma^2}\exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right)\exp\left(\frac{\rho A\cos\varphi}{\sigma^2}\right) \end{aligned}$$

A questo punto la saturazione della d.d.p. congiunta, operata eseguendo $p_P(\rho) = \int_{-\pi}^{\pi} p_{P,\Phi}(\rho,\varphi)\, d\varphi$, determina il risultato (14.16).

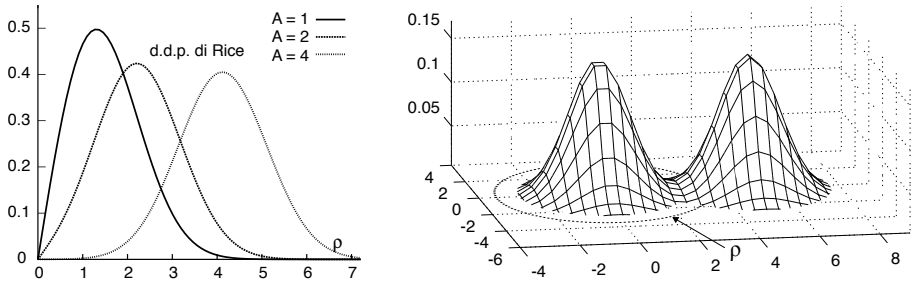


Figura 14.7: Densità di probabilità di Rice con $\sigma = 1$ (a sin) e coppia di gaussiane bidimensionali a varianza unitaria, la prima a media nulla, la seconda centrata in (5,0) (a ds)

$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{\rho A}{\sigma^2}\right) \text{ per } \rho \geq 0 \tag{14.16}$$

dove $I_0(z) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{e}^{z\cos\varphi} d\varphi$ è la funzione *modificata* di Bessel del primo tipo ed ordine zero[20], la cui espressione non ne permette il calcolo in forma chiusa, ma che può essere approssimata come $I_0(z) \sim \mathrm{e}^{\frac{z^2}{4}}$ per $z \ll 1$, e come $I_0(z) \sim \frac{\mathrm{e}^z}{\sqrt{2\pi z}}$ per $z \gg 1$.

Nella parte sinistra di fig. 14.7 è mostrato l'andamento di $p_P(\rho)$ con $\sigma = 1$ e tre diversi valori di A, in modo da poterlo confrontare con quello della seconda curva per la d.d.p. di Rayleigh alla fig. 14.6, ottenuta 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. Nella parte a destra di fig. 14.7 sono invece raffigurate le gaussiane bidimensionali che danno luogo alle distribuzioni di Rayleigh e di Rice.

14.4.2 Detezione incoerente di sinusoide nel rumore

Come fatto osservare nella discussione di fig. 14.5, se il segnale $s(t)$ si presenta con una fase $\varphi \neq 0$ ovvero $s(t) = A\cos(\omega_0 t + \varphi)$, il piano dell'inviluppo complesso ruota dello stesso angolo, ed è per questo motivo che abbiamo scelto il *modulo* ρ dell'inviluppo complesso come grandezza su cui operare la decisione, di cui abbiamo trovato la d.d.p. per i casi di segnale assente e presente.

Compromesso tra banda di ricerca e probabilità di detezione Prima di procedere osserviamo che qualora la frequenza di $s(t)$ fosse pari a $f = f_0 + \Delta f$, l'inviluppo complesso $\underline{z} = x + jy$ *ruoterebbe* con velocità angolare $2\pi\Delta f$, ma il suo modulo ρ resterebbe costante e pari ad A, dando luogo anche in questo caso alla d.d.p. di Rice. Ciò consente l'adozione dello schema di fig. 14.4 per la ricerca di una sinusoide che cade entro *tutta* la B_N del filtro di ingresso; d'altra parte, all'aumentare di B_N aumenta anche la potenza σ^2 del rumore, causando come vedremo tra breve un peggioramento delle prestazioni del decisore.

[20] Anche nella figura a pag. 387 si parla di funzioni di Bessel $J_n(x)$, ma queste *modificate* sono in relazione a quelle, come $I_n(x) = j^{-n} J_n(jx)$ - vedi
`https://it.wikipedia.org/wiki/Armoniche_cilindriche`.

Definizione del problema Analizziamo i risultati fin qui ottenuti nell'ottica della decisione di ipotesi statistica (§ 6.6.1), allo scopo di definire il criterio con cui scegliere la soglia di decisione λ da utilizzare nello schema di fig. 14.4.

Il caso di segnale assente (in cui la v.a. di osservazione ρ ha d.d.p. di Rayleigh) viene quindi indicato come *ipotesi* H_0, mentre quello in cui $s(t)$ è presente *ipotesi* H_1 (e ρ ha d.d.p. di Rice). In entrambi i casi la dinamica dei valori di ρ è direttamente legata (attraverso le (14.12) e (14.16)) alla potenza di rumore in ingresso $\sigma^2 = N_0 B_N$, pari a quella delle c.a. di b.f. $\nu_c(t)$ e $\nu_s(t)$, mentre per quanto riguarda il valor medio di ρ, nell'ipotesi H_0 si ha $m_\rho = \sigma\sqrt{\frac{\pi}{2}}$ (eq. (14.12)), e per H_1 risulta $m_\rho \to A$ quando $A \gg \sigma$.

Decisione di massima verosimiglianza In figura 14.8 oltre alle d.d.p. condizionate alle ipotesi $p(\rho/H_0)$ e $p(\rho/H_1)$ e calcolate per $\sigma = 1$ ed $A = 4$, viene mostrato anche il valore λ_{ML}[21] per cui esse si intersecano ovvero $p(\rho/H_0)\big|_{\rho=\lambda_{ML}} = p(\rho/H_1)\big|_{\rho=\lambda_{ML}}$, e la regola di decisione $\frac{p(\rho/H_1)}{p(\rho/H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1$ che ne consegue corrisponde al criterio di massima verosimiglianza (§ 6.6.2.1), attuato nella forma $\frac{\rho}{\lambda_{ML}} \underset{H_0}{\overset{H_1}{\gtrless}} 1$ ovvero $\rho \underset{H_0}{\overset{H_1}{\gtrless}} \lambda_{ML}$.

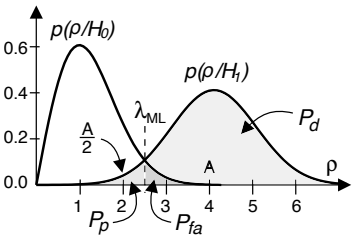


Figura 14.8: Posizione della soglia per il decisore di massima verosimiglianza

Probabilità di errore A seguito della decisione si possono verificare i due tipi di evento di errore

- *falso allarme* con probabilità $Pr(e/H_0) = \int_\lambda^\infty p(\rho/H_0)\, d\rho = P_{fa}$
- *perdita* con probabilità $Pr(e/H_1) = \int_0^\lambda p(\rho/H_1)\, d\rho = P_p$

rispettivamente pari alle aree colorate in celeste e giallo di fig. 14.8. Osserviamo quindi che la scelta $\lambda = \lambda_{ML}$ risulta *ottima* qualora non sussistano *costi* per i due tipi errori P_{fa} e P_p (vedi sotto), e le probabilità *a priori* di H_0 ed H_1 siano uguali. Infatti in tal caso la probabilità di errore complessiva

$$P_e = Pr(H_0)\, Pr(e/H_0) + Pr(H_1)\, Pr(e/H_1) = \tfrac{1}{2} P_{fa} + \tfrac{1}{2} P_p \qquad (14.17)$$

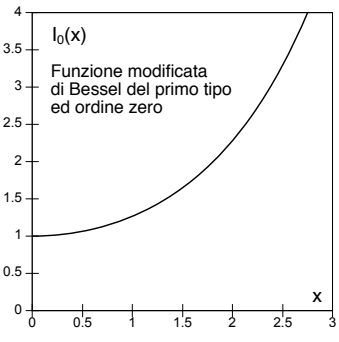
risulta *minima*, dato che spostando λ a destra o sinistra rispetto a λ_{ML}, una delle due aree aumenta più di quanto non diminuisca l'altra.

Calcolo della soglia Tutto bello, ma volendo ottenere il valore di λ_{ML}, come si fa? La condizione che per $\rho = \lambda_{ML}$ risulti $p(\rho/H_0) = p(\rho/H_1)$ comporta l'uguaglianza tra le d.d.p. di Rayleigh (14.12) e di Rice (14.16), ovvero

$$\frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{\rho A}{\sigma^2}\right)$$

da cui con alcuni passaggi si ottiene

$$\exp\left(\frac{A^2}{2\sigma^2}\right) = I_0\left(\frac{\rho A}{\sigma^2}\right)$$



[21] Il valore di λ_{ML} va calcolato per via numerica una volta noti σ ed A.

Una volta noti i valori di A e σ il primo membro è un numero, e dunque per via numerica si trova il valore di $\lambda_{ML} = \rho$ che rende il secondo membro pari al primo.

Decisione Bayesiana Qualora si conoscano i valori di $Pr(H_0)$ e $Pr(H_1)$ e questi siano diversi da 1/2, ponendo $\lambda = \lambda_{ML}$ la (14.17) non è più minimizzata. In tal caso la soglia ottima viene invece stabilita secondo il criterio di *massima probabilità a posteriori* o MAP, vedi § 17.1.2, ovvero scegliendo l'ipotesi H_i la cui probabilità *a posteriori* $p(H_i/\rho)$ è massima. Applicando il teorema di Bayes (§ 6.1.4) si ottiene $p(H_i/\rho) = \frac{p(\rho/H_i)Pr(H_i)}{p(\rho)}$ e dunque la regola di decisione diviene

$$\frac{p(H_1/\rho)}{p(H_0/\rho)} = \frac{p(\rho/H_1)\,Pr(H_1)}{p(\rho/H_0)\,Pr(H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \quad \text{ovvero} \quad \frac{p(\rho/H_1)}{p(\rho/H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{Pr(H_0)}{Pr(H_1)} \tag{14.18}$$

che nel caso di ipotesi equiprobabili $Pr(H_0) = Pr(H_1)$ degenera nel criterio di ML.

Costo delle decisioni Allarghiamo ora il campo di applicazione della decisione statistica a situazioni in cui può essere associato un differente *costo* ai due tipi di errore, così come si può associare un *guadagno* all'evento di decisione corretta (o *detezione*) la cui probabilità $P_d = \int_\lambda^\infty p(\rho/H_1)\,d\rho$ è misurata dall'area *verde* in fig. 14.8. Ad esempio, nell'ambito del *telerilevamento* si tenta di massimizzare la probabilità di detezione a spese di quella di falso allarme[22], mentre in *campo medico* si tende a preferire un falso allarme, piuttosto che trascurare l'importanza di un sintomo o referto. In questi casi nella 14.18 compare un altro termine[23] che tiene conto dei costi associati alla decisione, in modo da preferire uno dei due tipi di errore rispetto all'altro.

Criterio di Neyman-Pearson In alcuni casi la probabilità a priori $Pr(H_1)$ che il segnale sia presente non è nota in quanto l'evento è di natura *sporadica*, e noi lì, in attesa che si verifichi. Un possibile approccio è allora quello di fissare la P_{fa} massima tollerata, e quindi tentare di massimizzare la prob. di detezione P_d, come avviene adottando il criterio di *Neyman-Pearson*[24], sulla cui descrizione non ci addentriamo.

Decisore per SNR elevato Torniamo ad investigare sulla applicazione del criterio di massima verosimiglianza, la cui soglia di decisione λ_{ML} può essere fissata una volta nota l'ampiezza A della sinusoide e la deviazione standard σ del rumore; a volte però tali grandezze non sono note, se non *a grandi linee!*

In particolare, qualora sia noto solamente che $A/\sigma \gg 1$ e dunque in presenza di una ampiezza $A \gtrsim 10\sigma$ ben maggiore della dinamica del rumore[25], notiamo che all'aumentare di $\frac{A}{\sigma}$ le curve di fig. 14.8 si allontanano ma non cambiano larghezza, ed il valore di λ_{ML} si avvicina (da destra) ad $\frac{A}{2}$. Ponendo quindi $\lambda = \frac{A}{2}$ e sostituendo le espressioni di Rayleigh (14.12) e di Rice (14.16) per le d.d.p. condizionate in quella

[22] A meno che decidere per H_1 non possa provocare *danni collaterali* documentabili dai media.

[23] Vedi ad es. `http://webuser.unicas.it/tortorella/TTII/PDF2003/decisione_bayes.pdf`

[24] Vedi ad es. `https://en.wikipedia.org/wiki/Neyman-Pearson_lemma`

[25] Un modo di ricondursi a questo caso è quello di diminuire la banda del filtro di ingresso, riducendo così $\sigma^2 = N_0 B_N$. In questo modo però, come osservato a pag. 14.4.2, si riduce l'intervallo di frequenza Δf che può essere analizzato.

(14.17) della P_e, otteniamo

$$P_e = \frac{1}{2}\int_{\frac{A}{2}}^{\infty} \frac{\rho}{\sigma^2}\exp\left(-\frac{\rho^2}{2\sigma^2}\right)d\rho + \frac{1}{2}\int_{0}^{\frac{A}{2}} \frac{\rho}{\sigma^2}\exp\left(-\frac{\rho^2+A^2}{2\sigma^2}\right)I_0\left(\frac{\rho A}{\sigma^2}\right)d\rho \qquad (14.19)$$

Per ciò che riguarda il primo termine, applicando il risultato (14.14) 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 14.8 tracciata per $A = 4$, o le considerazioni riportate al § 14.5.1), e quindi può essere trascurato, fornendo in definitiva

$$P_e \simeq \frac{1}{2}\exp\left(-\frac{A^2}{8\sigma^2}\right) \qquad (14.20)$$

per $\frac{A}{\sigma} \gg 1$. Ricordando ora che $\frac{A^2}{2}$ rappresenta la potenza della sinusoide, e che σ^2 è la potenza del rumore, il risultato trovato ha una immediata interpretazione in termini di $SNR = \frac{A^2/2}{\sigma^2}$:

$$P_e \simeq \frac{1}{2}\exp\left(-\frac{SNR}{4}\right) \qquad (14.21)$$

14.5 Appendice

14.5.1 Approssimazione della d.d.p. di Rice per SNR elevato

Come già osservato a pag. 429, la funzione modificata di Bessel può essere approssimata come $I_0\left({}^{\rho A}/_{\sigma^2}\right) \sim {}^{\exp\left(\frac{\rho A}{\sigma^2}\right)}/_{\sqrt{2\pi\frac{\rho A}{\sigma^2}}}$ per $\frac{\rho A}{\sigma^2} \gg 1$, e quindi in tal caso la funzione *integranda* che compare al secondo termine di (14.19) diviene

$$\begin{aligned} p_{RICE}(\rho) &\simeq \frac{\rho}{\sigma^2}e^{-\frac{\rho^2+A^2}{2\sigma^2}}I_0\left(\frac{\rho A}{\sigma^2}\right) = \frac{\rho}{\sigma^2}e^{-\frac{\rho^2}{2\sigma^2}}e^{-\frac{A^2}{2\sigma^2}}e^{\frac{\rho A}{\sigma^2}}\frac{\sigma}{\sqrt{2\pi\rho A}} = \\ &= \sqrt{\frac{\rho}{2\pi\sigma^2 A}}\exp\left(-\frac{(\rho-A)^2}{2\sigma^2}\right) < \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\rho-A)^2}{2\sigma^2}\right) \end{aligned} \qquad (14.22)$$

in cui l'ultimo passaggio tiene conto che nelle ipotesi poste risulta anche $\frac{A}{\sigma} \gg 1$, permettendo di scrivere $\rho \simeq A + \varepsilon$ con $\varepsilon \ll A$, e dunque $\sqrt{\frac{\rho}{2\pi\sigma^2 A}} \simeq \sqrt{\frac{A+\varepsilon}{2\pi\sigma^2 A}} < \sqrt{\frac{A}{2\pi\sigma^2 A}} = \frac{1}{\sqrt{2\pi}\sigma}$.

Dato che la (14.22) è a tutti gli effetti la d.d.p. di una gaussiana con media A e varianza σ^2, l'integrale a secondo membro di (14.19) risulta inferiore a

$$\int_{-\infty}^{\frac{A}{2}} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\rho-A)^2}{2\sigma^2}\right)d\rho = \frac{1}{2}erfc\left\{\frac{{}^{A}/_{2}}{\sqrt{2}\sigma}\right\} \qquad (14.23)$$

Considerando di nuovo il verificarsi di ${}^{A}/_{\sigma} \gg 1$, anche per l'argomento dell'*erfc* risulta $z = \frac{{}^{A}/_{2}}{\sqrt{2}\sigma} \gg 1$, ed in tal caso vale l'approssimazione[26] $erfc(z) \simeq \frac{1}{z\sqrt{\pi}}e^{-z^2}$. Sostituendo questa in (14.23) e quindi nella (14.19), il secondo membro di (14.19) si approssima

[26] Vedi `https://it.wikipedia.org/wiki/Funzione_degli_errori`

come

$$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{\sqrt{\pi}}\cdot\frac{\sqrt{2}\sigma}{A/2}\cdot\exp\left(-\frac{A^2}{4\cdot 2\cdot\sigma^2}\right)=\frac{\sigma}{\sqrt{2\pi}A}\exp\left(-\frac{A^2}{8\sigma^2}\right)$$

che, essendo per ipotesi $\frac{A}{\sigma}\gg 1$, risulta trascurabile rispetto al primo termine $\frac{1}{2}\exp\left(-\frac{A^2}{8\sigma^2}\right)$ della (14.19).

Parte III

Trasmissione dei Segnali

Prefazione alla terza parte

Dopo che nelle prime due parti del testo sono state gettate le basi della teoria dei segnali, intesi questi ultimi come le entità apportatrici di informazione, e dopo aver analizzato le loro caratteristiche nel tempo, in frequenza e dal punto di vista energetico, gli aspetti probabilistici applicati ai loro valori (singoli o multipli), il campionamento, la densità spettrale ed il filtraggio, il processo di modulazione, la teoria dell'informazione e della codifica di sorgente, dopo tutto questo, la terza parte si occupa di studiare gli aspetti legati alla necessità di *trasmettere a distanza* le informazioni impresse sul segnale.

Il capitolo 15 affronta la *trasmissione dati in banda base*, in cui l'informazione simbolica codificata da *cifre binarie* (bit) viene impressa su di un segnale *analogico* per il solo scopo di poterlo trasmettere. Il *segnale dati* è quindi definito come un'onda PAM di cui valutiamo la densità spettrale, ed individuiamo le *condizioni di Nyquist* per l'assenza di ISI. Si passa poi a determinare la *probabilità di errore* sul simbolo e sul bit per una trasmissione multilivello a coseno rialzato, ed a descrivere le modalità per attuare il controllo di errore, come FEC e codifica di canale. Il capitolo termina affrontando l'argomento della *acquisizione della temporizzazione*, e quello del *ricevitore ottimo*, che suddivide l'impulso a coseno rialzato tra trasmettitore e ricevitore.

Le possibilità offerte dalla combinazione delle tecniche di modulazione con quelle di trasmissione numerica sono affrontate al cap. 16, in cui si descrive come i metodi a portante singola costituiscano una vera e propria *famiglia* di tecniche. A queste si aggiunge la tecnica OFDM a portante multipla, che viene analizzata in dettaglio nei suoi diversi aspetti: architetturali, di prestazione, di equalizzazione, di distribuzione ottima della potenza, di sincronizzazione e di approccio all'acceso multiplo. Infine si affrontano i sistemi a *spettro espanso*, incluse le relative problematiche di sincronizzazione, per assolvere alle comuni necessità quotidiane, dal modem ADSL alla televisione, dalla telefonia cellulare al WiFi.

I tempi sono a questo punto maturi per tornare ad affrontare (al cap. 17) gli aspetti della teoria dell'informazione legati alla *capacità di canale* ed al *controllo di errore:* un susseguirsi di definizioni e risultati teorici porta a stabilire le limitazioni intrinseche del processo comunicativo, al cui raggiungimento tendono le tecniche di codifica,

affrontate in un crescendo di approcci via via più sofisticati fino ai metodi iterativi come *turbo* e LDPC adottati dai sistemi più recenti.

Il capitolo 18 sviluppa tre argomenti che, anche se per così dire *marginali* rispetto alla trasmissione dei segnali, sono profondamente legati agli aspetti *fisici* che la caratterizzano: innanzi tutto viene svolta una analisi degli effetti che i *circuiti elettrici* producono sui segnali in transito; quindi, viene descritta la modalità con cui tenere conto del *rumore* introdotto dagli apparati, compreso il caso dei collegamenti sviluppati mediante una catena di ripetitori; infine, si affronta in modo abbastanza approfondito il tema della *equalizzazione*, ovvero come ovviare agli effetti di distorsione lineare eventualmente introdotta da un canale.

La caratterizzazione dei mezzi fisici di trasmissione di natura *cablata*, ovvero il cavo in rame e la fibra ottica, viene svolta al cap. 19, ponendola nel contesto del *bilancio di collegamento*, ovvero della valutazione della potenza necessaria a coprire una determinata distanza[27]. Il caso del canale radio viene quindi trattato al cap. 20, in cui viene approfondito lo studio delle particolarità che rendono le trasmissioni mobili una sorta di palestra, in cui le basi teoriche discusse nella prima parte dispiegano tutta la loro versatilità, consentendo di ottenere un modello concettuale dei fenomeni, e volgerli a vantaggio dell'esigenza trasmissiva.

L'ultimo capitolo (21) di questa parte dedica una sessantina di nuove pagine alla trattazione dei sistemi multi-antenna (MIMO), la cui realizzazione si è resa possibile grazie ai progressi tecnologici intervenuti nel frattempo, che consentono ai moderni sistemi radio una estrema flessibilità di utilizzo, a tutto vantaggio delle prestazioni delle attuali tecniche di trasmissione, siano esse di tipo punto-punto, di accesso multiplo, o broadcast, raggiungendo il triplice traguardo di prestazioni migliori, un utilizzo più efficiente della risorsa radio, ed una estrema flessibilità nel suo impiego.

[27] O equivalentemente, della distanza che è possibile coprire conoscendo la potenza trasmessa, e quella che è necessario ricevere.

Capitolo 15

Trasmissione dati in banda base

Dopo che un segnale *analogico* è divenuto sequenza *numerica* mediante campionamento (cap. 4), e la sequenza convertita in bit dalla quantizzazione (vedi § 4.3.1.1), non esistono più differenze rispetto ai dati nativamente numerici, come i documenti su di un computer. In questo capitolo si affrontano gli aspetti legati alla trasmissione di *un segnale analogico* di *banda base* realizzato a partire da *dati numerici* (di qualunque natura), discutendo le alternative per realizzare il segnale in accordo a requisiti sia nel tempo che in frequenza, ed arrivando ad una espressione per la probabilità di errore in presenza di rumore additivo. Sono inoltre introdotte le problematiche della equalizzazione, del controllo di errore e della sincronizzazione.

15.1 Trasmissione su canale numerico

Al primo capitolo (§ 1.2.2) abbiamo illustrato 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*), che viaggia su *canale analogico*. Se quest'ultimo presenta una risposta in frequenza di tipo *passa-banda*[1], anche il segnale dati dovrà presentare le medesime caratteristiche frequenziali, ed al capitolo 16 sono 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 segnale dati viene detto *di banda base*, ed il *modem* che lo genera è indicato come *codificatore di linea*.

15.1.1 Trasmissione numerica di banda base

La figura 15.1 rappresenta lo schema generale di un canale numerico, con evidenziati i principali elementi che lo compongono, la funzione dei quali viene ora brevemente illustrata. L'ingresso al canale è descritto, nella sua forma più generale, come un flusso (o sequenza) di *cifre binarie* con valore logico di 0 ed 1, indicate come bit (*binary digit*), che pervengono ad una velocità (binaria) di f_b bit/secondo.

[1] Ovvero lascia passare solo frequenze comprese in un intervallo che non comprende l'origine.

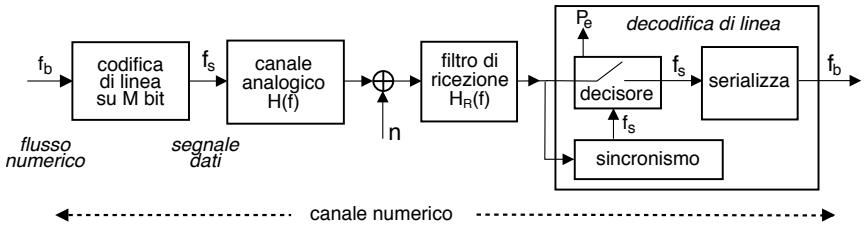
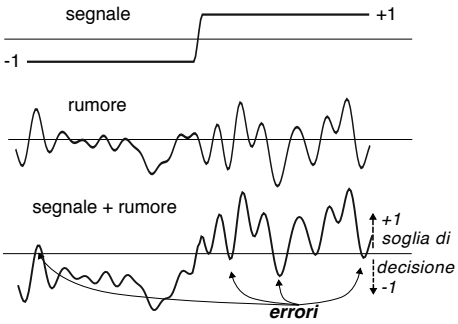


Figura 15.1: Elementi costitutivi di un canale numerico

Il primo elemento del canale numerico prende il nome di *codificatore di linea* (§ 15.1.2) e produce un *segnale dati* analogico che, ogni T_s secondi, trasporta un *simbolo*, che a sua volta rappresenta *uno o più bit*. Pertanto per la *frequenza di simbolo*[2] $f_s = 1/T_s$ si può scrivere $f_s \leq f_b$, con il segno di uguale nel caso di simboli *binari*. Al contrario, se ogni simbolo rappresenta $M > 1$ bit, come vedremo al § 15.1.2.4 si ottiene un *risparmio di banda* per il segnale uscente dal codificatore, in modo da adattare lo stesso alla *banda passante* che caratterizza la risposta in frequenza $H(f)$ del canale[3]. Per qualunque scelta di M, ad ogni possibile simbolo è associato ad un diverso *livello* del segnale analogico trasmesso.

L'elemento che svolge le funzioni inverse a quelle finora descritte è il *decodificatore di linea* posto al lato ricevente, che *ricostruisce* la sequenza dei simboli trasmessi a partire dal segnale analogico ricevuto, innanzitutto *campionandolo* [4](con ritmo f_s), e quindi *decidendo* quale simbolo sia stato trasmesso all'istante $t = nT_s$, in base al confronto tra il valore campionato, ed alcune *soglie di decisione*.

La presenza di un processo di *rumore additivo* $n(t)$ all'uscita del canale fa si che il segnale preso in esame dal decisore possa superare la 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 § 15.4.3, 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, come evidente dalla figura a lato per una trasmissione *a due livelli*. Fortunatamente la potenza di rumore può essere resa



[2]Indichiamo T_s come *periodo di simbolo*, mentre il suo inverso $f_s = 1/T_s$ è detto *frequenza di simbolo*, *baud-rate* o *frequenza di segnalazione*, e si misura in simboli/secondo, unità di misura indicata anche come *baud*, in memoria di Émile Baudot, vedi `http://it.wikipedia.org/wiki/Codice_Baudot`.

[3]Se 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 in*terferenza tra simboli* (vedi § 15.1.2.2).

[4]Sembra giusto sottolineare che questo *campionamento* non ha lo scopo discusso al cap. 4, ma si tratta piuttosto qualcosa di più simile al filtro adattato (§ 7.6), che *decide* in base al superamento di una soglia. D'altra parte, mentre la decisione operata dal quantizzatore introduce un errore, quella del ricevitore numerico discrimina tra informazioni *già discrete*.

minima progettando adeguatamente il filtro $H_R(f)$ di ricezione (vedi § 14.1.1), che può anche realizzare un *filtro adattato* (§ 7.6), oppure ancora un dispositivo di *equalizzazione* (§ 18.4).

Il confronto con la soglia di decisione non avviene *di continuo* come in figura, bensì agli istanti $t = nT_s$ corrispondenti a quelli in cui sono codificati i simboli. Nello schema di fig. 15.1 è infatti presente anche un imprescindibile dispositivo di *sincronizzazione*, che osservando il segnale ricevuto[5] genera un *segnale di orologio* (CLOCK) che consente al decisore di operare *al passo* con il ritmo $f_s = 1/T_s$: al § 15.7 sono descritte alcune tecniche per affrontare questo aspetto.

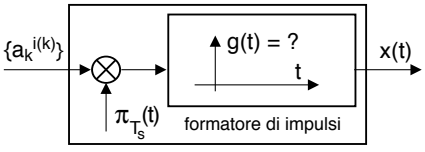
Infine, nel caso in cui ogni simbolo codifichi più di un bit (vedi la trasmissione *multilivello* al § 15.1.2.4), è necessaria la presenza di un *serializzatore* che provvede ad emettere uno dopo l'altro i bit corrispondenti a ciascun simbolo ricevuto.

15.1.2 Codifica di linea e segnale dati

Poniamoci nel caso più generale in cui una *sorgente discreta* con alfabeto composto da L simboli[6] emetta la sequenza $\{a_k^{i(k)}\}$, dove il pedice $_k$ individua l'istante temporale $t = kT_s$ mentre l'indice $i\,(k)$ rappresenta l'identità (con i da 1 a L) del simbolo emesso in tale istante. Come anticipato i simboli sono prodotti con frequenza $f_s = 1/T_s$, ed il codificatore di linea fa corrispondere alla sequenza $\{a_k^{i(k)}\}$ un *segnale dati* $x\,(t)$ espresso come

$$x\,(t) = g\,(t) * \sum_k a_k^{i(k)} \cdot \delta\,(t - kT_s) = \sum_k a_k^{i(k)} \cdot g\,(t - kT_s) \tag{15.1}$$

a cui corrisponde lo schema simbolico mostrato a lato, in cui $\pi_{T_s}\,(t) = \sum_k \delta\,(t - kT_s)$ è un treno di impulsi (§ 3.7) con periodo $T_s = 1/f_s$, schema del tutto analogo al SAMPLE&HOLD introdotto al § 4.2.4, tranne che ora $g\,(t)$ è generico e prende il nome di *impulso dati*.



Generazione del SEGNALE DATI

Dato che i simboli $a_k^{i(k)}$ sono entità casuali, il segnale $x\,(t)$ risulta essere un processo aleatorio, e la valutazione della relativa densità di potenza $\mathcal{P}_x\,(f)$ coinvolge il calcolo della autocorrelazione $\mathcal{R}_x\,(\tau)$ e l'applicazione del teorema di Wiener (§ 7.2.1). Tale sviluppo è svolto al § 7.7.4, dove l'ipotesi aggiuntiva di simboli a_k statisticamente indipendenti ed a media nulla porta per $\mathcal{P}_x\,(f)$ all'espressione

$$\mathcal{P}_x\,(f) = \sigma_A^2 \cdot \frac{\mathcal{E}_g\,(f)}{T_s} \tag{15.2}$$

in cui $\mathcal{E}_g\,(f)$ è lo spettro di densità di energia di $g\,(t)$ che determina la *sagomatura* della densità di potenza $\mathcal{P}_x\,(f)$, mentre σ_A^2 è la varianza dei valori a_k[7] che ne determina

[5] Oppure mediante una seconda linea di trasmissione.

[6] Se L risulta essere una potenza di due ovvero $L = 2^M$, ogni diverso valore rappresenta un gruppo di $M = \log_2 L$ cifre binarie (*bit*), e la trasmissione convoglia un messaggio numerico con *frequenza binaria* pari a $f_b \left[\frac{bit}{secondo}\right] = M \left[\frac{bit}{simbolo}\right] \cdot f_s \left[\frac{simboli}{secondo}\right]$.

[7] La modalità generale di calcolo per σ_A^2 viene descritta alla nota 68 di pag. 224.

il *fattore di scala.*

Onda PAM Il segnale dati (15.1) in alcuni contesti è indicato come *onda PAM*, acronimo di *Pulse Amplitude Modulation,* ad indicare la variazione (o modulazione) che i valori a_k determinano sugli impulsi $g(t - mT)$ che si susseguono sull'asse temporale $\sum_m g(t - mT)$[8]. Un ulteriore punto di vista è illustrato al § 24.9.5.

15.1.2.1 Segnale dati binario e onda rettangolare

Partiamo dalla (15.2) per ottenere l'espressione della densità di potenza relativa ad una trasmissione *binaria*, ossia con $f_s = f_b$ ed $a_k = \{1, -1\}$, con valori a_k *equiprobabili* in modo che la sequenza $\{a_k\}$ sia a media nulla, ed adottando un impulso rettangolare $g(t) = \text{rect}_\tau(t - \tau/2)$ con $\tau \leq T_b$ in modo da mantenere la sua durata inferiore al periodo di simbolo. In tal caso il segnale dati $x(t)$ assume l'aspetto mostrato in fig. 15.2 e per esso si ottiene[9] $\sigma_A^2 = 1$ e $\mathcal{E}_G(f) = \tau^2 \text{sinc}^2(f\tau)$, dunque la (15.2) diviene

$$\mathcal{P}_x(f) = \sigma_A^2 \cdot \frac{\mathcal{E}_G(f)}{T_b} = \frac{\tau^2}{T_b} \cdot \text{sinc}^2(f\tau)$$

con l'aspetto[10] mostrato sulla destra di fig. 15.2 ed espresso in deciBel (§ 8.1), per i casi $\tau = T_b$ e $\tau = T_b/2$, indicati nell'ordine come *Not-Return-to-Zero* (NRZ) e RZ (vedi § 15.2.1). Osserviamo che il lobo principale di $\mathcal{P}_x(f)$ è delimitato in $|f| < \frac{1}{\tau}$, ovvero $|f| < f_b$ qualora $\tau = T_b$, mentre nel caso RZ il lobo principale *si dilata* fino a $|f| < 2f_b$ e *si abbassa* di 6 dB[11]. L'occupazione di banda complessiva sarà infine approssimata ad un valore pari ad alcuni multipli dell'ampiezza del primo lobo[12].

15.1.2.2 Distorsione lineare e interferenza intersimbolica

Qualora il canale attraversato dal segnale dati (15.1) non sia un *canale perfetto* (§ 8) ovvero presenti una risposta impulsiva $h(t) \neq a\delta(t - \tau)$, $x(t)$ subisce *distorsione*

[8] Infatti, se i valori a_k fossero tutti uguali, il segnale $\sum_m g(t - mT)$ sarebbe semplicemente periodico, come descritto a pag. 81.

[9] Svolgendo i conti si ha

$$\sigma_A^2 = E\{a_k^2\} - (E\{a_k\})^2 = E\{a_k^2\} = \sum p_i \left(a_k^{(i)}\right)^2 = \tfrac{1}{2} \cdot 1^2 + \tfrac{1}{2} \cdot (-1)^2 = 2 \cdot \tfrac{1}{2} \cdot 1 = 1$$

essendo $E\{a_k\} = 0$, mentre per quanto riguarda $\mathcal{E}_G(f)$ si ottiene

$$\mathcal{E}_G(f) = |G(f)|^2 = |\mathcal{F}\{g(t)\}|^2 = \left|\tau \text{sinc}(f\tau)\, e^{-j2\pi f\tau/2}\right|^2 = \tau^2 \text{sinc}^2(f\tau)$$

[10] Estendiamo il risultato al caso noto di segnale periodico. Ponendo $a_k = (-1)^k$ si genera un'onda rettangolare, il cui spettro (mancando la componente aleatoria) *è a righe*, con lo stesso inviluppo di tipo $\text{sinc}^2(fT_b)$.

[11] Infatti il fattore $\frac{\tau^2}{T_b}$ passa da T_b (NRZ) a $\frac{T_b}{4}$ (RZ), pari ad una riduzione di 6 dB.

[12]

f_b	apparato	T_b	$10/T_b$
$2.4 \cdot 10^3$	Modem (anni '80)	$4.2 \cdot 10^{-3}$	$24\ KHz$
$28.8 \cdot 10^3$	Modem (anni '90)	$3.5 \cdot 10^{-5}$	$288\ KHz$
$10 \cdot 10^6$	Thin Ethernet (anni '90)	10^{-7}	$100\ MHz$
$100 \cdot 10^6$	Fast Ethernet	10^{-8}	$1\ GHz$

Nella tabella a fianco è riportata l'occupazione di banda necessaria a contenere 10 lobi di un $\text{sinc}(fT_b) = 1/T_b \mathcal{F}\{\text{rect}_{T_b}(t)\}$, ovvero relativa ad una trasmissione binaria a velocità $f_b = 1/T_b$ per alcuni casi tipici *del passato*: osserviamo che un'onda rettangolare può *andar bene* a basse velocità di trasmissione, infatti già per 10 Msimboli/sec, velocità di una LAN, occorrono 100 MHz di banda.

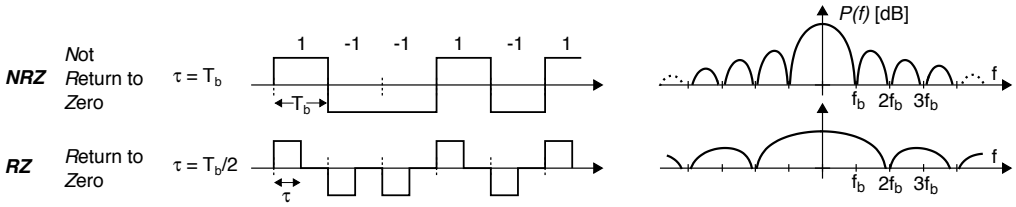
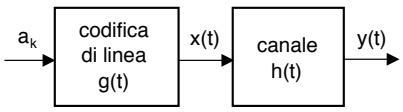


Figura 15.2: Segnale dati binario e relativa densità di potenza espressa in dB/Hz

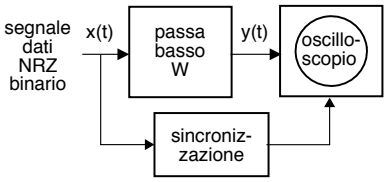
lineare (§ 8.2), ed in uscita dal canale si presenta[13] il nuovo segnale dati

$$y(t) = \sum_k a_k \cdot \tilde{g}(t - kT_s) \quad \text{in cui} \quad \tilde{g}(t) = g(t) * h(t)$$



L'effetto della convoluzione tra $g(t)$ ed $h(t)$ è quello di *disperdere* nel tempo la forma d'onda $g(t)$, che anche se delimitata entro un periodo di simbolo come per il caso dell'onda rettangolare, arriva ad *invadere* gli intervalli temporali riservati ai simboli adiacenti, dando luogo al fenomeno della *interferenza intersimbolica* indicata anche come ISI ovvero *Inter Symbolic Interference*.

15.1.2.3 Diagramma ad occhio



Per valutare sperimentalmente l'effetto della distorsione lineare sul segnale dati $x(t)$ facciamo riferimento ad uno schema di misura rappresentato a lato, che prevede di limitare la banda di $x(t)$ ad una frequenza W e quindi visualizzare il segnale distorto $y(t)$ ponendolo in ingresso ad un *oscilloscopio*[14], la cui *base dei tempi* è sincronizzata con quella di $x(t)$, in modo da visualizzare sullo schermo la sovrapposizione di forme d'onda corrispondenti a coppie di simboli binari.

L'esito di tale operazione viene raffigurato in fig. 15.3, la cui parte superiore mostra il segnale dati $x(t)$ originario realizzato in versione NRZ mediante un impulso $g(t) = rect_{T_b}(t)$ e valori a_k binari pari a 11110101000, a fianco della sua versione filtrata a $W = 2f_b$ (al centro) e $W = f_b/2$ (a destra), ossia limitato in una banda W pari rispettivamente al *doppio* ed alla *metà* della *larghezza* del primo lobo di $\mathcal{P}_x(f)$ (vedi fig. 15.2). Come evidente la limitazione di banda causa una alterazione della forma d'onda, ed il campionamento di $x(t)$ agli istanti di simbolo $t = kT_s$ produce valori diversi da quelli originari.

La riga inferiore di fig. 15.3 mostra invece la visualizzazione dei casi descritti da parte dell'oscilloscopio, e la disposizione risultante dei tracciati è detta *diagramma ad occhio* (traduzione di EYE DIAGRAM), termine che deriva dall'aspetto che assume il

[13] Possiamo infatti sviluppare le seguente uguaglianze

$$\begin{aligned} y(t) &= [\textstyle\sum_k a_k \cdot g(t - kT_s)] * h(t) = [g(t) * \sum_k a_k \cdot \delta(t - kT_s)] * h(t) = \\ &= g(t) * h(t) * \textstyle\sum_k a_k \cdot \delta(t - kT_s) = \tilde{g}(t) * \sum_k a_k \cdot \delta(t - kT_s) = \\ &= \textstyle\sum_k a_k \cdot \tilde{g}(t - kT_s) \end{aligned}$$

[14] Vedi `https://it.wikipedia.org/wiki/Oscilloscopio`

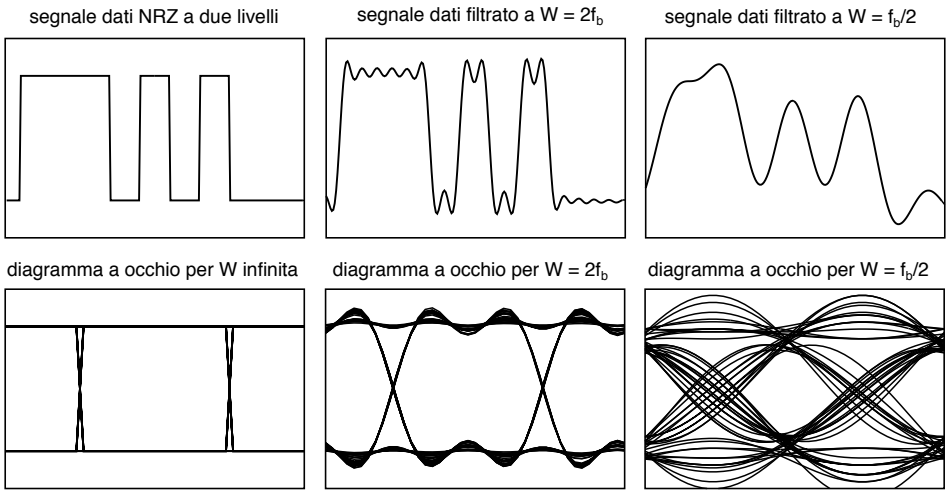


Figura 15.3: Segnale dati a banda infinita e limitata, e relativo diagramma ad occhio

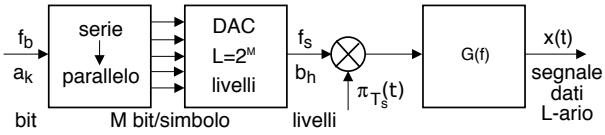
disegno che si forma, analogia che apparirà più evidente in seguito all'adozione di un impulso $g(t)$ limitato in banda (figura 15.8), ed in presenza di rumore (figura 15.10).

Riservandoci di riprendere l'argomento nel seguito, osserviamo che il problema *non* si presenta se

- la frequenza di simbolo è molto inferiore alla banda del canale, *ovvero*
- la risposta impulsiva $h(t)$ del canale ha una duratamolto inferiore a T_s.

15.1.2.4 Trasmissione multilivello

Nel caso in cui la banda a disposizione per la trasmissione *sia scarsa*, una soluzione di semplice attuazione è quella di ricorrere ad una trasmissione non più *binaria*, ma che impieghi simboli ad L valori, detti *livelli*[15]. A tale scopo, M simboli binari della sequenza originaria a_k sono raggruppati assieme, ed emessi *in contemporanea* da un dispositivo *serie/parallelo*[16] come una unica parola binaria di M bit, posta quindi in ingresso ad un converitore D/A (pag. 97) che produce in uscita uno tra $L = 2^M$ possibili valori, ampiezze, o livelli, che rappresentano i valori dei simboli per la nuova sequenza b_h, con $h = {}^{k}/_{M}$. Dato che occorrono $T_s = MT_b$ secondi per accumulare M bit, i simboli b_h della nuova sequenza



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

[16] Si tratta di un componente di elettronica digitale noto come *registro a scorrimento* (vedi `https://it.wikipedia.org/wiki/Registro_a_scorrimento`), costituito da M celle di memoria di un bit, ciascuna delle quali (con frequenza f_b) *copia* il contenuto della precedente, mentre la prima è caricata (*in serie*) con un nuovo bit; al termine di M cicli i bit vengono letti tutti assieme, appunto, *in parallelo*.

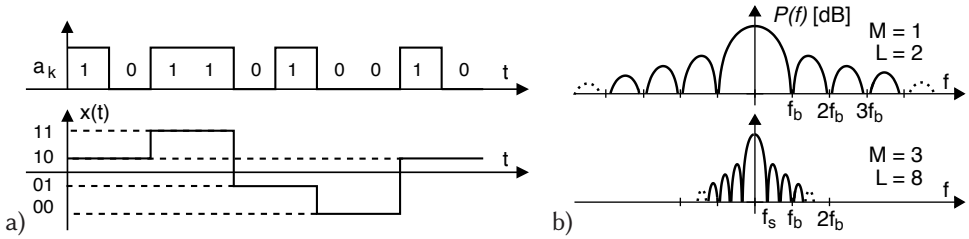


Figura 15.4: a) - codifica a quattro livelli di una sequenza binaria; b) - densità spettrale per un segnale binario ed uno ad otto livelli a parità di f_b

sono prodotti ad una velocità indicata come frequenza *di simbolo*, pari a

$$f_s = f_b/M = 1/MT_b = 1/T_s$$

ed utilizzati per produrre un segnale dati $x(t) = \sum_h b_h \cdot g(t - hT_s)$ ad L livelli e frequenza di simbolo f_s[17].

Il risultato finale della codifica multilivello è esemplificato in figura 15.4-a), che mostra un segnale dati binario ad onda rettangolare NRZ assieme al corrispondente segnale dati $x(t)$ ottenuto per $L = 4$. All'aumento del periodo di simbolo T_s corrisponde quindi l'aumento della durata di $g(t)$, ovvero una *contrazione* della $G(f)$ che compare nella (15.2), determinando quindi la *riduzione* della banda occupata da $x(t)$, come mostrato in fig. 15.4-b), relativa all'uso di $M = 3$ bit/simbolo ($L = 8$). 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.

Sembrerebbe tutto risolto, se non che al § 15.4 si mostra come, a meno di non aumentare la potenza del segnale dati, in presenza di rumore la codifica multilivello causi un *peggioramento* della probabilità di errore del decisore, in quanto a parità di dinamica complessiva del segnale i valori dei livelli risultano ora *ravvicinati*. Questo fenomeno è rappresentato in figura 15.5, in cui a sinistra si mostra un segnale dati ad 8 livelli, al centro il diagramma ad occhio corrispondente, ed a destra lo stesso diagramma, per un segnale filtrato a meno della sua banda.

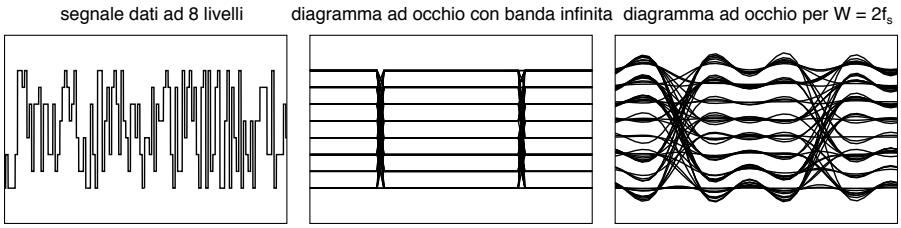


Figura 15.5: Segnale dati ad 8 livelli e diagramma ad occhio per banda infinita o limitata

[17] In ricezione si effettua il procedimento inverso, ripristinando la codifica binaria originaria di M bit a cui il codificatore ha associato il valore L-ario ricevuto, e quindi *serializzando* gli M bit, in modo da ri-ottenere la sequenza binaria di partenza. Vedi anche fig.15.11.

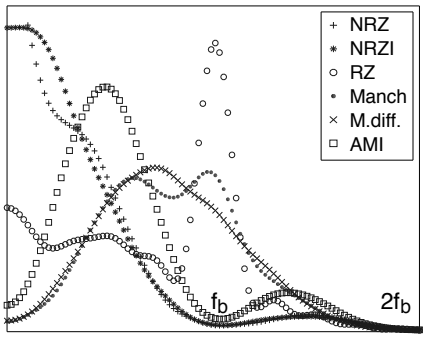
15.2 Scelta dell'impulso dati

Fino ad ora si è ragionato sulla base di un segnale dati generato secondo lo schema del § 15.1.2, in cui l'impulso $g(t)$ è definito come $g(t) = rect_\tau(t)$, mentre nei sistemi di comunicazione questa è solamente una di diverse alternative, i cui criteri di scelta andiamo ora ad illustrare.

15.2.1 Codici di linea a banda infinita

Come anticipato al § 15.1.2 il grafico dello spettro di densità di potenza $\mathcal{P}_X(f) = \sigma_A^2 \frac{\mathcal{E}_g(f)}{T_s}$ di un segnale dati dipende direttamente[18] 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 $g(t) = rect_\tau(t)$ si ottiene che $\mathcal{P}_X(f)$ ha andamento di tipo $sinc^2(f\tau)$ (vedi eq. (3.28) a pag. 79), 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 il canale come se fosse *a banda infinita* o perfetto, e quindi capace di riprodurre il segnale inalterato.

La figura a lato mostra lo spettro di densità di potenza $\mathcal{P}_x(f)$ *calcolato* per f che va da zero fino al doppio di f_b, per un segnale dati binario che ricade in una delle categorie illustrate di seguito, dette *codici di linea*. Il risultato è ottenuto generando i valori (0 o 1) per 400 simboli binari a_k in modo pseudo-casuale[19], campionando il segnale dati (15.1) con 16 campioni per periodo di bit, e valutando con questi una *stima spettrale*[20].



Ogni scelta per il codice di linea a cui corrisopnde una diversa definizione di impulso $g(t)$ ha particolari proprietà, e può essere usato per trasmettere informazioni di natura binaria sotto determinate condizioni. Elenchiamo quindi caratteristiche e proprietà di tali scelte, con riferimento agli esempi riportati in figura 15.6.

Codici unipolari Sono realizzati come segnali *sbilanciati*,[21] e codificano i due livelli logici 0 ed 1 rispettivamente con un valore nullo, od un valore positivo.

[18] In realtà al § 7.7.4 si mostra come il risultato possa essere un po' diverso nel caso di simboli statisticamente dipendenti e/o non a media nulla.

[19] La 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.

[20] Ottenuta applicando ai dati una finestra triangolare (§ 3.8.4) e quindi valutando il periodogramma (§ 7.3.1).

[21] Viene detto *sbilanciato* un segnale trasmesso mediante un collegamento (ad es. su rame) in cui uno dei due conduttori è connesso *a massa* ad entrambe le estremità, dando luogo ad una maggiore sensibilità a fenomeni di induzione elettromagnetica relativi ad altri segnali in transito nelle vicinanze (vedi *diafonia* a pag. 648), e dunque ad un peggiore SNR rispetto ai segnali (e collegamenti) bilanciati - .vedi ad es. `https://it.wikipedia.org/wiki/Linea_bilanciata`.

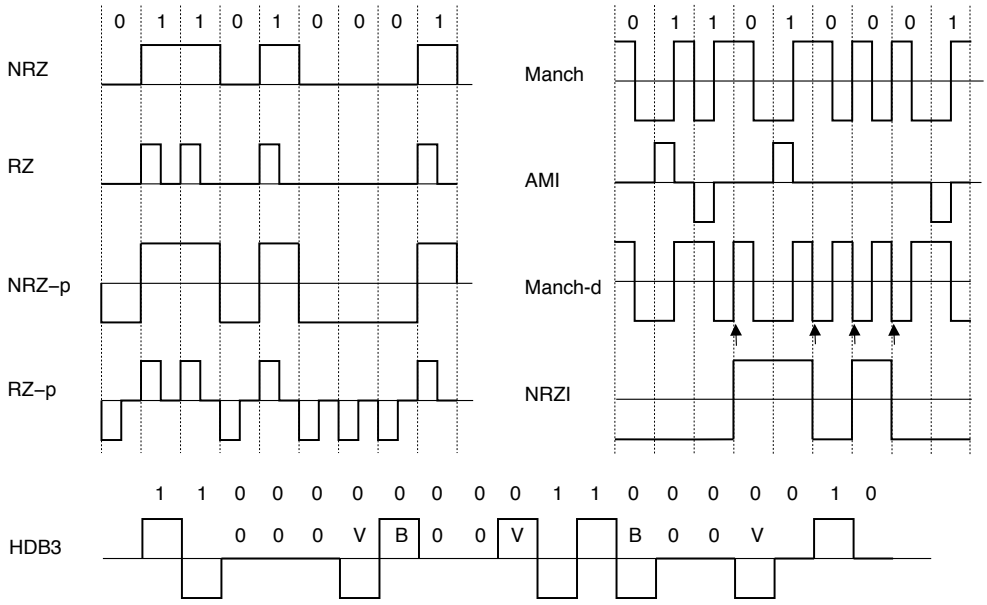


Figura 15.6: Forme d'onda di esempio associate ai codici di linea descritti nel testo

- **NRZ** o *No Return to Zero*: l'acronimo che lo descrive significa che il segnale "non torna a zero" per tutto il periodo di bit, essendo $g(t) = rect_{T_b}(t)$; pertanto lo spettro $G(f)$ è 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 (§ 15.7) del clock del ricevitore stante l'assenza in questo caso di transizioni. La mancanza di energia per $f = 1/T_b$ aggrava inoltre la situazione anche per dati qualsiasi.
- **RZ** o *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. Ma anche questo segnale si mantiene costante per lunghe sequenze di *zeri*.

Codici bipolari Usano segnali bilanciati o *antipodali*, e sono ricevuti mediante uno stadio di ingresso differenziale[22], 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.

[22]Vedi `https://it.wikipedia.org/wiki/Amplificatore_differenziale`

- **Manchester**[23]: realizza una codifica *di fase*, in quanto usa un impulso RZ a piena dinamica, in salita od in discesa, in corrispondenza dei 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 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 § 23.1.4).
- **AMI** o *Alternate Mark Inversion*: 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. La caratteristica di alternare valori positivi e negativi gli fa meritare il nome di codice *pseudo-ternario*, e denota la presenza di memoria. Da un punto di vista spettrale, l'AMI esibisce una occupazione di banda[24] 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 l'**HDB3**[25].

Codici differenziali Sono ancora di tipo bipolare, ma la forma d'onda non è più legata al valore di un solo bit, bensì dipende da quello di due bit contigui (vedi anche § 16.4). Ciò permette di risolvere l'ambiguità che si determina qualora si scambino tra loro gli estremi del collegamento[26].

- **Manchester Differenziale**: usa un impulso RZ a piena dinamica come per il Manchester, la cui polarità risulta però *invertita* rispetto all'impulso precedente se il nuovo bit è uno, mentre è mantenuta uguale nel caso arrivi uno 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.
- **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.

[23] Vedi `https://it.wikipedia.org/wiki/Codifica_Manchester`

[24] La densità spettrale mostrata in figura è relativa all'uso di una $g(t)$ di tipo RZ.

[25] La codifica HDB3 è utilizzata per trasmettere il segnale PCM a 2 Mbps (vedi § 24.3.1), 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 una *violazione* della regola dell'alternanza 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. Questi si collocano al posto del primo dei quattro zeri, e la loro polarità è scelta in modo che la sequenza delle violazioni abbia un polarità alternata; in definitiva, dopo la prima violazione, si usa sempre anche il bit di bilanciamento.

[26] In tal caso tutti gli zeri diventerebbero uni e viceversa, mentre con la codifica differenziale questo viene evitato.

15.2.2 Segnale dati limitato in banda

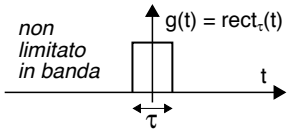
Discutiamo ora come la scelta di un impulso $g(t)$ *non* rettangolare permetta di ridurre l'occupazione di banda del relativo segnali dati, indipendentemente dalla codifica multilivello.

15.2.2.1 Requisiti per l'impulso di trasmissione

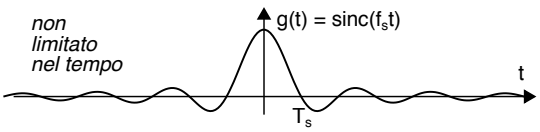
Organizziamo il ragionamento analizzando ordinatamente i requisiti che deve avere l'impulso $g(t)$ per soddisfare *tre diverse esigenze* parzialmente contrapposte, anticipando che la soluzione sarà necessariamente una forma di compromesso.

Limitazione di banda Al § 15.1.2 abbiamo osservato che (sotto opportune condizioni) il segnale dati (15.1) ha densità di potenza pari a $\mathcal{P}_x(f) = \sigma_A^2 \frac{|G(f)|^2}{T_s}$ (eq. (15.2)), e la sua ricezione inalterata è possibile solo se $x(t)$ è trasmesso per il tramite di un *canale perfetto* (pag. 8). Se al contrario la banda del segnale eccede quella del canale gli effetti di distorsione lineare non sono trascurabili e gli impulsi $g(t)$ *si deformano*[27], causando problemi di *interferenza tra simboli* (ISI).

Ad esempio adottando $g(t) = \text{rect}_\tau(t)$ con $\tau \leq T_s$ il relativo spettro di ampiezza $G(f) = \tau \text{sinc}(f\tau)$ presenta il primo passaggio per zero a frequenza $\frac{1}{\tau} \geq \frac{1}{T_s} = f_s$, e la densità di potenza $\mathcal{P}_x(f)$ può essere considerata nulla solo dopo diversi multipli di tale valore.

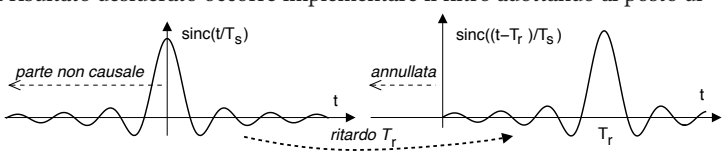


Limitazione nel tempo Il problema della limitazione di banda potrebbe essere risolto adottando un impulso elementare di tipo $g(t) = \text{sinc}(f_s t)$ che, essendo $f_s = 1/T_s$, ha trasformata $G(f) = T_s \cdot \text{rect}_{f_s}(f)$ strettamente limitata nella banda $|f| < \frac{f_s}{2}$: in tal caso se il canale presenta un comportamento ideale in tale (limitato) intervallo di frequenze il segnale dati non subisce alterazioni. Ma lo svantaggio di adottare una forma d'onda $g(t)$ limitata in frequenza è che la stessa è illimitata nel tempo, e dunque l'impulso può essere realizzato[28] solo in modo approssimato!



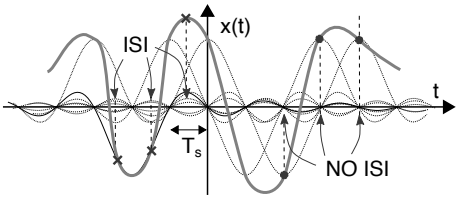
[27] Come mostrato al § 15.1.2.2, il segnale dati filtrato è basato su impulsi $\widetilde{g}(t) = g(t) * h(t)$, con una durata pari alla somma delle durate di $g(t)$ e $h(t)$. Pertanto, anche se $g(t)$ è limitato nel tempo, come nei casi descritti al § 15.2.1, l'impulso $\widetilde{g}(t)$ si può estendere a valori di $t > T_s$. Considerando ad esempio la trasmissione di soli due simboli a_0 ed a_1, si otterrebbe $x(t) = a_0\widetilde{g}(t) + a_1\widetilde{g}(t - T_s)$, e dunque $x(T_s) = a_0\widetilde{g}(T_s) + a_1\widetilde{g}(0)$ dipenderà da entrambi i simboli anziché solamente da a_1, osservando quindi un errore pari a $a_0\widetilde{g}(T_s)$, detto appunto *interferenza tra simboli*.

[28] Il requisito di causalità $h(t) = 0$ per $t < 0$ (pag. 26) a cui deve sottostare qualsiasi sistema fisico impedisce infatti di realizzare un filtro la cui risposta impulsiva $g(t)$ abbia una estensione temporale illimitata: per avvicinarsi al risultato desiderato occorre implementare il filtro adottando al posto di $g(t)$ una sua versione ritardata e limitata $g'(t) = g(t - T_R)$ con $t \geq 0$ e $g'(t) = 0$ altrimenti, come in figura.



Notiamo ora che con questa scelta $g(t)$ *passerebbe* da zero per $t = nT_s$ e quindi non provocherebbe *interferenza* tra i simboli collocati agli istanti nT_s, come verificabile notando che in tal caso l'espressione (15.1) risulta del tutto simile alla (4.1) relativa alla ricostruzione *cardinale* di un segnale campionato.

Limitazione di precisione Contrariamente al caso del campionamento, ora *non siamo interessati* al valore del segnale negli istanti *intermedi* a quelli a cui sono centrati i simboli, e desideriamo unicamente recuperare i valori originali a_k. D'altra parte anche ammettendo di poter adottare una $g(t) = \text{sinc}(f_s t)$ il recupero degli a_k può avvenire solamente campionando il segnale dati $x(t)$ *esattamente* agli istanti $t = nT_s$, dato che al di fuori di tali istanti il valore del segnale dipende dal valore delle *code* degli impulsi $g(t)$ centrati sugli altri simboli[29] come mostrato in figura, in cui la linea spessa rossa rappresenta il segnale dati $x(t)$ risultato della somma dei contributi di termini $a_k \text{sinc}(f_s(t - kT_s))$ con $a_k = \pm 1$. L'orologio (*clock*) del ricevitore deve quindi comandare il campionatore esattamente agli istanti kT_s, e non in anticipo o in ritardo, perché altrimenti si verifica ISI, tanto maggiore e da parte di tanti più simboli anche lontani, quanto più *lentamente* si attenuano le code della $g(t)$. Dato però che nessun oscillatore ha una precisione infinta e che anzi i metodi di sincronizzazione (§ 15.7) presentano piccole variazioni (*jitter*) della loro frequenza, occorre ricercare una soluzione per $g(t)$ che pur rimanendo limitata in banda presenti oscillazioni di ampiezza ridotta, in modo da tollerare meglio modesti *errori di precisione* nella determinazione degli istanti di campionamento.



Riepilogando Vorremmo soddisfare contemporaneamente le esigenze

1. occupare una banda contenuta;
2. ricorrere ad un filtro con $g(t)$ di durata ridotta e quindi poco complesso;
3. ridurre la sensibilità agli errori di campionamento.

Per i punti 2 e 3 sarebbe sufficiente adottare $g(t)$ di tipo rettangolare producendo un segnale dati del tipo $x(t) = \sum_k a_k \cdot \text{rect}_\tau(t - kT_s)$, ma questo ha lo svantaggio di occupare una banda troppo elevata, che può essere ridotta ad un valore finito (punto 1) pur di accettare una durata per $g(t)$ maggiore di T_s, come per il caso del *sinc*, che però ha eccessiva estensione temporale e code troppo ampie. Mostriamo ora come a partire dalla formalizzazione analitica dell'esigenza di non subire ISI sia possibile ottenere una soluzione di compromesso a tutti e tre i problemi.

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.

[29] Al 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.

15.2.2.2 Criterio di Nyquist per l'assenza di ISI

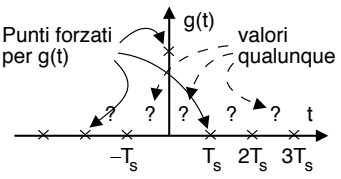
Il fatto che sia un *rect* che un *sinc* permettano di evitare interferenza intersimbolica[30] sembra suggerire che possano essere accomunate dal soddisfare un criterio generale: infatti, sia *rect* che *sinc* sono casi particolari di impulsi che rispettano le *condizioni* espresse nel seguito, e che alla fine ci permettono di *negoziare* il compromesso necessario a soddisfare le tre esigenze espresse sopra.

Condizioni di Nyquist nel tempo Torniamo a riferirci alla (15.1) per osservare che, affinché $x(t = nT_s)$ dipenda dal solo valore a_n e non dagli altri a_k con $k \neq n$, deve risultare

$$g(t) = \begin{cases} 1 & \text{se } t = 0 \\ 0 & \text{se } t = mT_s \text{ con } m \neq 0 \\ \forall & altrove \end{cases} \tag{15.3}$$

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 (15.1) si ottiene:

$$x(nT_s) = \sum_k a_k \cdot g(nT_s - kT_s) = \sum_k a_k \cdot g((n-k)T_s) = a_n$$



cond. di Nyquist nel tempo

dato che $m = n - k$ è un intero che vale zero solo quando $k = n$. Le condizioni (15.3) prendono il nome di *condizioni di Nyquist per l'assenza di interferenza intersimbolo* (ISI) nel dominio del *tempo*. Se una forma d'onda $g(t)$ soddisfa tali condizioni, allora viene detta *impulso di Nyquist*([31]).

Condizioni di Nyquist in frequenza Dalle condizioni di Nyquist *nel tempo* (15.3) 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 eq. (3.31)) si ottiene:

$$1 = G(f) * \tfrac{1}{T_s} \cdot \Pi_{\frac{1}{T_s}}(f) = G(f) * \tfrac{1}{T_s} \cdot \sum_k \delta\left(f - k\tfrac{1}{T_s}\right)$$

Indicando con $f_s = \frac{1}{T_s}$ la frequenza di simbolo, ed eseguendo la convoluzione tra $G(f)$ e gli impulsi centrati in $f = kf_s$, risulta infine

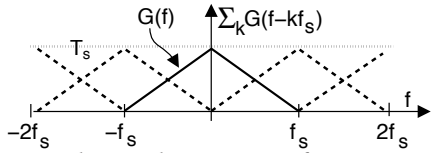
$$\sum_k G(f - kf_s) = T_s \tag{15.4}$$

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 se $G(f)$ manifesta *simmetria dispari* rispetto ad $f_s/2$. In tal caso $G(f)$ può essere descritta come la risposta in frequenza di un *filtro di Nyquist*. Notiamo che

[30] Almeno, in assenza di distorsione lineare!

[31] Ad esempio, l'impulso rettangolare è di Nyquist, in quanto $\text{rect}_{T_s}(t) = \begin{cases} 1 & se\ |t| < \frac{T_s}{2} \\ 0 & se\ t = kT_s \end{cases}$.

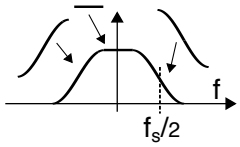
seppure $G(f)$ possa essere qualsiasi, anche non limitata in banda, il nostro interesse è appunto per le $G(f)$ limitate in banda, come quella triangolare dell'esempio a fianco.



condizioni di Nyquist in frequenza

15.2.2.3 Filtro a coseno rialzato

Descrive una famiglia parametrica di filtri di Nyquist limitati in banda, detti a *coseno rialzato* in quanto la $G(f)$ è realizzata mediante 2 semiperiodi di coseno raccordati da una retta, come mostrato a lato. La fig. 15.7-a)([32]) illustra l'andamento di $G(f)$ per diverse scelte del parametro $0 < \gamma < 1$ chiamato coefficiente *di roll-off*[33], che rappresenta l'indice *di dispersione* del ramo di coseno attorno alla frequenza $f_s/2$, detta *frequenza di Nyquist*[34].



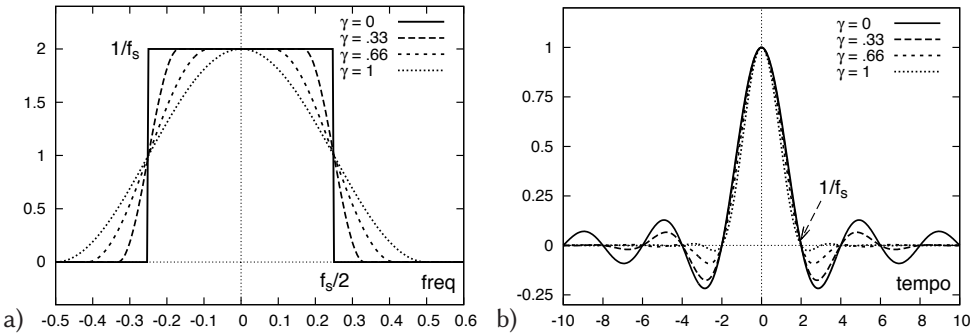
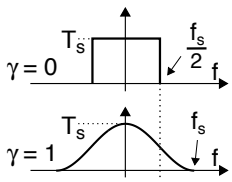


Figura 15.7: a) - filtro a coseno rialzato e b) - impulso di Nyquist per $f_s = 0.5$, variando γ

La banda occupata a frequenze positive da $G(f)$ può quindi essere espressa in funzione di γ come

$$B = \frac{f_s}{2}(1+\gamma) = \frac{f_b}{2\log_2 L}(1+\gamma) \tag{15.5}$$

e varia da *un minimo* $B|_{\gamma=0} = f_s/2$ in corrispondenza di un $G(f)|_{\gamma=0} = T_s \mathrm{rect}_{f_s}(f)$ rettangolare, ad un *massimo* pari a $B|_{\gamma=1} = f_s$ a cui corrisponde una



$$G(f)|_{\gamma=1} = \tfrac{T_s}{2}\left[1+\cos\left(2\pi\tfrac{1}{2f_s}f\right)\right]\cdot \mathrm{rect}_{2f_s}(f) \quad \text{con } |f| < f_s$$

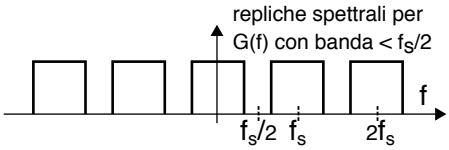
che rappresenta esattamente un periodo di coseno (in f) di periodo $2f_s$, *rialzato*.

[32] La fig. 16.2 a pag. 498 mostra la stessa funzione su di una scala quadratica e in decibel.

[33] Il termine ROLL-OFF può essere tradotto come "*rotola fuori* ".

[34] Molto intimamente legata alla *velocità* di Nyquist definita al § 4.1 come la *minima* frequenza di campionamento $f_c = 2W$ per un segnale analogico di banda W, mentre *la frequenza* di Nyquist si riferisce invece a *metà della massima* frequenza di segnalazione $f_s/2 = B$ per un segnale dati che transita su di un canale limitato in banda B, per motivi presto chiari. Come evidente due aspetti dello stesso fenomeno, ma in contesti differenti, vedi `https://en.wikipedia.org/wiki/Nyquist_rate`.

Filtro a banda minima Il caso di $\gamma = 0$ individua una $G(f) = T_s \text{rect}_{f_s}(f)$ e come già discusso a pag. 449 corrisponde ad un impulso $g(t) = sinc(f_s t)$. Tale scelta viene detta *a banda minima* poiché non è possibile occupare una banda *inferiore*, dato che in tal caso non sarebbero verificate le condizioni di Nyquist in frequenza, in quanto nella (15.4) resterebbero dei *"buchi"*.



Abbiamo già osservato alla nota (28) a pagina 449 come la realizzazione di $G(f)$ *a banda minima* sia complicata, dato che la corrispondente $g(t) = sinc(f_s t)$ va a zero con $t \to \infty$ come $\frac{1}{\pi f_s t}$, sviluppando *code* che si estendono su di un elevato numero di simboli adiacenti: oltre a complicare la realizzazione del filtro, ciò comporta la possibilità di introdurre notevole ISI in presenza di errori negli istanti di campionamento. Ma la situazione migliora decisamente usando $\gamma > 0$, con γ via via più grande, come ora illustriamo.

Roll off γ diverso da zero In questo caso per $g(t)$ si può ottenere l'espressione generale[35]

$$g(t) = \text{sinc}(f_s t) \cdot \frac{\cos \gamma \pi t f_s}{1 - (2\gamma t f_s)^2} \tag{15.6}$$

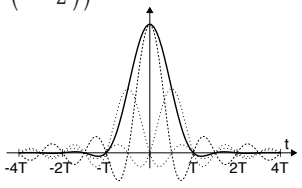
a cui corrisponde una forma d'onda *simile* a $\text{sinc}(f_s t)$, ma che va a zero molto più rapidamente, come verificabile osservando la parte destra di Fig. 15.7. Pertanto se $\gamma \to 1$ le oscillazioni di $g(t)$ sono molto più smorzate, ed anche in presenza di errori negli istanti di campionamento $t = kT_s$ ogni impulso estende la sua *influenza* ad un numero di simboli limitrofi *molto ridotto* rispetto al caso $\gamma = 0$.

Per verificare visivamente quanto affermato, aiutiamoci con la fig. 15.8 che in alto a sinistra mostra l'andamento di un segnale dati realizzato adottando la $g(t)$ fornita dalla (15.6), calcolata per $\gamma = 0.5$, e per simboli a_k a due valori, pari a 0 e 1. Notiamo che al di fuori degli istanti di simbolo $t = kT_s$ il segnale può assumere valori arbitrari, anche oltre la dinamica degli a_k. La rappresentazione fornita dal diagramma ad occhio per questo segnale dati, mostrato in alto a destra in fig 15.8, permette di valutare meglio la precisione di temporizzazione che è necessaria per evitare ISI, e che è pari a metà della *apertura orizzontale* dell'occhio.

[35] Non ho trovato questi passaggi già svolti in nessun posto, qualche lettore può aiutare? Tutto quel che sono riuscito a calcolare è relativo al caso $\gamma = 1$, per cui $g(t) = \mathcal{F}^{-1}\left\{\frac{T}{2}\left[1 + \cos(\pi T f)\right]\right\} \cdot \text{rect}_{2/T}(f)$ che fornisce

$$g(t) = \frac{T}{2}\left[\delta(t) + \frac{1}{2}\delta\left(t - \frac{T}{2}\right) + \frac{1}{2}\delta\left(t + \frac{T}{2}\right)\right] * \frac{2}{T}\text{sinc}\left(\frac{2}{T}t\right) =$$
$$= \text{sinc}\left(\frac{2}{T}t\right) + \frac{1}{2}\text{sinc}\left(\frac{2}{T}\left(t - \frac{T}{2}\right)\right) + \frac{1}{2}\text{sinc}\left(\frac{2}{T}\left(t + \frac{T}{2}\right)\right)$$

che una volta graficato, conferma l'andamento di fig. 15.7. Osserviamo che per $t \to 1/2\gamma f_s$ il denominatore di (15.6) si annulla, ma lo stesso avviene anche per il numeratore, che in tal caso tende a $\cos\frac{\pi}{2}$, dando luogo alla forma $\frac{0}{0}$, ed il cui limite sembra tendere a poco meno di uno.



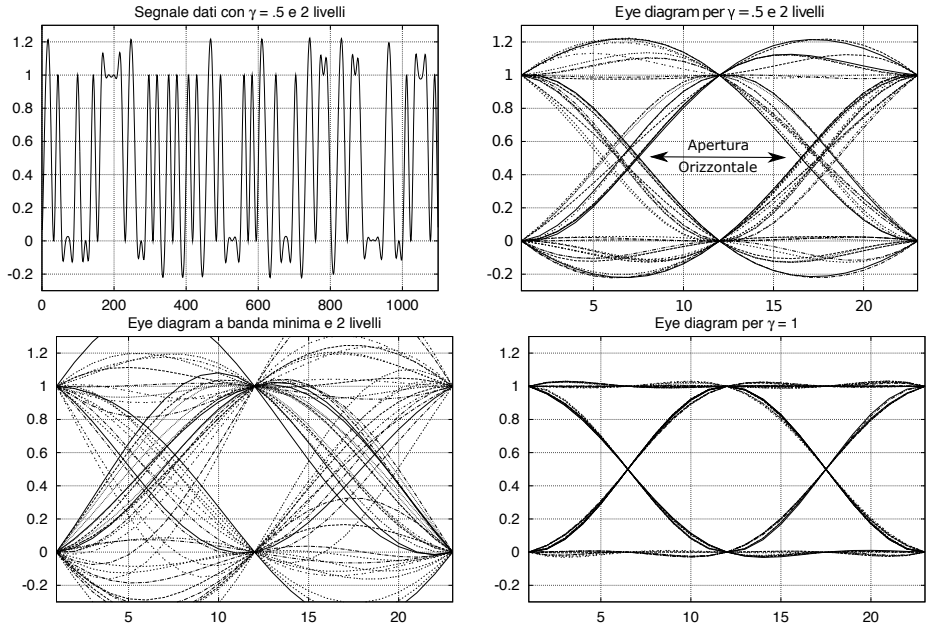


Figura 15.8: Segnale dati e *diagramma ad occhio* per diversi valori di *roll-off*

Gli ultimi due diagrammi nella parte inferiore di fig. 15.8 permettono il confronto tra le scelte relative alla banda minima e *massima*: notiamo che nel primo caso (a destra) $g(t)$ è un *sinc*, e l'apertura orizzontale dell'occhio si è ristretta, mentre nel secondo (a sinistra, con $\gamma = 1$) l'occhio è alla sua apertura massima, ed i tracciati sono quasi del tutto identici tra loro, indipendentemente dai simboli precedenti e successivi.

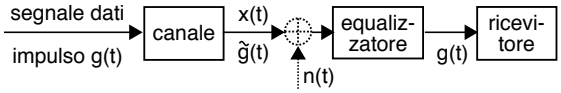
15.3 Equalizzazione

Torniamo ad occuparci del problema legato all'attraversamento da parte del segnale dati di un canale che presenta una $H(f)$ *non ideale* (vedi pag. 231), subendo quindi *distorsione lineare* (§ 8.2) e causando così per il sistema di trasmissione la comparsa di ISI, dato che (vedi § 15.1.2.2) il segnale ricevuto $x(t)$ in uscita dal canale risulta ora realizzato mediante un impulso *distorto* $\tilde{g}(t) = g(t) * h(t)$ anziché $g(t)$. Fortunatamente la distorsione lineare è (almeno in linea di principio) completamente reversibile, e nel caso in cui $H(f)$ sia nota può essere *compensata* facendo transitare $x(t)$ attraverso un *filtro di equalizzazione* $H_{eq}(f)$ tale che in cascata ad $H(f)$ ripristini le condizioni di canale perfetto, cioè tale che

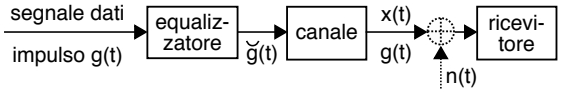
$$\begin{cases} H(f) H_{eq}(f) &= a e^{-j2\pi f\tau} \\ h(t) * h_{eq}(t) &= a\delta(t-\tau) \end{cases} \qquad \text{e dunque} \qquad H_{eq}(f) = \frac{a e^{-j2\pi f\tau}}{H(f)}$$

Come noto scrivere $H(f) H_{eq}(f)$ oppure $H_{eq}(f) H(f)$ è la stessa cosa, e dunque il filtro di equalizzazione può essere posto sia in trasmissione che in ricezione, con le seguenti conseguenze.

Al ricevitore Se il lato ricevente *conosce* $H(f)$ può calcolare la $H_{eq}(f)$ teorica e *sintetizzare* un filtro (vedi § 5.2) che la approssimi[36]. Se invece il ricevitore *non* la conosce, $H(f)$ può essere *stimata* (producendo $\hat{H}(f)$) a partire dal segnale ricevuto, facendo precedere la trasmissione vera e propria da una fase di *apprendimento*, durante la quale sono trasmessi dati che anche il ricevitore conosce, in modo da poter utilizzare il segnale di errore per stimare la distorsione lineare introdotta dal canale, ovvero la $H(f)$.



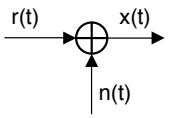
Al trasmettitore Il problema principale della metodica precedente è che anche il rumore presente in ingresso al canale passa attraverso il filtro di equalizzazione, trasformandosi da bianco a *colorato*, e questo peggiora le prestazioni. Pertanto se il trasmettitore conosce *a priori* la $H(f)$ del canale, oppure questa è stimata al lato ricevente ed esiste un canale *di ritorno*, può essere preferibile attuare l'equalizzazione *in partenza*, utilizzando al posto dell'impulso $g(t)$ *originario* un impulso definito come $\breve{g}(t) = g(t) * h_{eq}(t)$, e tale quindi da ripresentarsi come $g(t)$ in uscita dal canale, essendo $h_{eq}(t) * h(t) = a\delta(t-\tau)$.



La trattazione delle tecniche di equalizzazione prosegue al § 15.5.1 dove sono chiariti alcuni aspetti qui solamente accennati, ed al § 18.4 in cui si discutono soluzioni di tipo numerico.

15.4 Probabilità di errore per trasmissioni di banda base

Fin qui abbiamo trascurato di prendere in considerazione gli effetti del rumore additivo, a cui si è accennato al § 15.1.1, e che provoca la ricezione di un segnale $x(t) = r(t) + n(t)$. Al segnale utile $r(t)$ risulta dunque sovrapposto un diverso segnale $n(t)$ indicato come disturbo o rumore (*noise*[37]), *membro* di un processo ergodico (vedi § 6.3), con densità di probabilità del primo ordine *gaussiana* (vedi § 6.2.4) a media nulla, e spettro di densità di potenza *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 § 8.4) di ricondurle tutte ad un'unica fonte di rumore (equivalente) in ingresso al decisore. Come appare dalla figura a pag. 440, l'effetto del rumore è quello di causare degli *errori* nelle decisioni sui livelli, e quindi sui simboli e sui bit ricevuti.

Sviluppiamo dunque una analisi per valutare la *probabilità* di questi errori, in funzione delle grandezze che vi contribuiscono, in modo a poter successivamente affrontare problematiche di progetto, vedi § 19.1.

[36] Le esigenze di mantenere basso l'ordine del filtro tentando al contempo di rispondere ai requisiti sulla fase oltre che sul modulo impediscono di ottenere una sintesi perfetta di $H_{eq}(f)$.

[37] Vedi il § 8.4.2.1 per una descrizione della sua natura fisica.

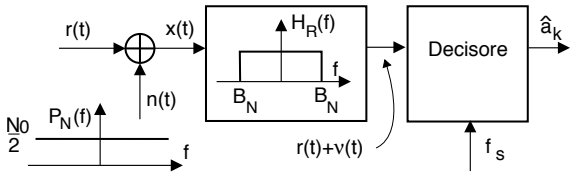
15.4.1 Banda di ricezione e dinamica del rumore

Come anticipato il disturbo $n(t)$ è la realizzazione di un processo ergodico gaussiano a valor medio nullo e con spettro di densità di potenza *bianco* o costante[38]

$$\mathcal{P}_N(f) = N_0/2$$

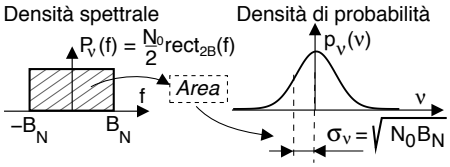
spesso indicato come *Additive White Gaussian Noise* o AWGN[39].

Allo scopo di limitare la potenza del rumore alla minima possibile il ricevitore vero e proprio è preceduto da un filtro passa-basso ideale[40] con risposta in frequenza $H_R(f)$ limitata in una banda $\pm B_N$ (detta *banda di rumore*, vedi § 14.1.2), in modo da lasciar passare *per intero* le componenti frequenziali del segnale $r(t)$ e limitare la banda del rumore $\nu(t)$ in uscita da $H_R(f)$ al minimo. Il rumore *filtrato* $\nu(t)$ è anch'esso un processo gaussiano ergodico (vedi nota 36 a pag. 208) a media nulla, la cui potenza vale[41]



$$\mathcal{P}_\nu = \int_{-\infty}^{\infty} \mathcal{P}_N(f)\,|H_R(f)|^2\,df = \int_{-B_N}^{B_N} \frac{N_0}{2}\,df = N_0 B_N$$

In virtù della ergodicità di $\nu(t)$ il valore di $\mathcal{P}_\nu$ eguaglia quello del momento di secondo ordine $m_\nu^{(2)} = E\left\{(\nu)^2\right\}$ di una v.a. ν ottenuta campionando una sua qualsiasi realizzazione; dato inoltre che $n(t)$ e dunque ν sono a media nulla, si ha[42] $m_\nu^{(2)} = \sigma_\nu^2$ e quindi $\mathcal{P}_\nu$ individua anche *la dinamica* dei valori della v.a. di rumore sovrapposta ai valori di segnale, come esemplificato in figura.



15.4.2 Dinamica del segnale e decisione a massima verosimiglianza

Proseguiamo l'analisi descrivendo il segnale ricevuto nella forma

$$r(t) = \textstyle\sum_k a[k] \cdot g(t - kT_s) \tag{15.7}$$

con $g(t)$ che è un impulso di Nyquist (15.3); si assume inoltre una perfetta sincronizzazione temporale (§ 15.7) in modo da poter considerare l'ISI assente. Gli elementi della sequenza $\{a[k]\}$ sono v.a. discrete, i.i.d. con d.d.p. uniforme, che assumono uno tra L possibili valori a_i equispaziati in un intervallo con dinamica $\Delta = a_L - a_1$, in modo da poter scrivere $a_i = a_1 + \frac{\Delta}{L-1} \cdot (i-1)$ con $i = 1, 2, .., L$.

[38] Al § 8.4.2.1 si illustra come in realtà $\mathcal{P}_N(f)$ non è costante per qualsiasi valore di f fino ad infinito, ma occupa una banda grandissima ma limitata: altrimenti, avrebbe una potenza infinita.

[39] I due aggettivi *White* e *Gaussian* non sono per nulla *inscindibili*, nel senso che un processo può essere gaussiano ma non bianco, o bianco ma non gaussiano!

[40] Al § 15.5 si descrive un diverso modo di progettare $H_R(f)$, in modo da minimizzare la probabilità di errore anziché la potenza di rumore, e che di fatto realizza un *filtro adattato*, descritto al § 7.6. Come vedremo al § 15.5, nel caso di un impulso *a banda minima* i due approcci portano al medesimo risultato.

[41] Per i dettagli relativi al filtraggio di processi, ci si può riferire al § 7.4.1.

[42] vedi eq. (6.9) a pag. 152

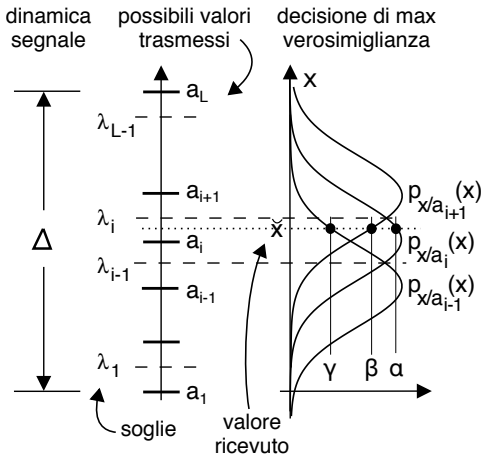
Agli istanti $t = kT_s = k/f_s$ multipli del periodo di simbolo T_s il decisore acquisisce il valore di segnale più rumore $x\,(kT_s) = (r\,(t) + \nu\,(t))|_{t=kT_s}$ ed anziché ritrovare il valore $a[k] = a_{i(k)}$ del simbolo trasmesso, osserva la realizzazione di una v.a. gaussiana $\check{x} = x\,(kT_s)$ con valor medio $a_{i(k)}$ e varianza $\sigma_\nu^2 = N_0 B_N$, essendo $\nu\,(t)$ a media nulla. Per stabilire quale valore a_i sia stato (più probabilmente) associato al simbolo $k-$esimo il ricevitore effettua quindi una decisione *di massima verosimiglianza* (o ML, vedi § 6.6.2.1) confrontando tra loro le densità di probabilità *condizionate* alle diverse ipotesi che sia stato trasmesso uno tra i simboli a_i:

$$P_{X/a_i}\,(\check{x}) = \frac{1}{\sqrt{2\pi}\sigma_\nu} \mathrm{e}^{-\frac{(\check{x}-a_i)^2}{2\sigma_\nu^2}} \tag{15.8}$$

e scegliendo per l'$\widehat{a_i}$ tale che $P_{X/\widehat{a_i}}\,(\check{x})$ è la più grande, ossia

$$\widehat{a_i} = \arg\max_{a_i} \left\{P_{X/a_i}\,(\check{x})\right\}$$

Il criterio di massima verosimiglianza equivale pertanto (vedi figura) a definire $L-1$ soglie di decisione λ_i, $i = 1, 2, .., L-1$ poste *a metà* tra i valori a_i ed a_{i+1}[43], e decidere per il valore a_i se il segnale ricevuto $\check{x}$ cade all'interno dell'intervallo compreso tra λ_{i-1} e λ_i ([44]), dato che (con riferimento alla notazione in figura) ciò corrisponde ad imporre



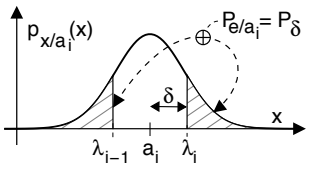
$$\alpha = P_{X/a_i}\,(\check{x}) > \beta = P_{X/a_{i+1}}\,(\check{x}) > \gamma = P_{X/a_{i-1}}\,(\check{x})$$

15.4.3 Probabilità dell'errore gaussiano

A seguito dell'applicazione del criterio di massima verosimiglianza il decisore commette errore quando, a fronte della trasmissione di un simbolo a_i, il campione di rumore filtrato $\nu\,(kT_s)$ assume un valore abbastanza elevato da far oltrepassare a $\check{x}$ una soglia di decisione, ovvero qualora $|\nu\,(kT_s)| > \delta$ in cui δ è la metà dell'intervallo tra due soglie e cioè $\delta = |\lambda_i - a_i| = \frac{\Delta}{2(L-1)}$. La probabilità di questo errore si dice *condizionata* alla trasmissione di a_i e vale

$$P_{e/a_i} = 2\int_{\lambda_i}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_\nu} \mathrm{e}^{-\frac{(x-a_i)^2}{2\sigma_\nu^2}} dx = P_{\delta} \tag{15.9}$$

che chiameremo P_δ, e che rappresenta (vedi figura) la somma delle aree tratteggiate.



[43] La proprietà di *equidistanza* delle soglie dal valore di simboli 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 *falso allarme* e di *perdita*, vedi § 6.6.1.

[44] Chiaramente, tutti i valori x minori di λ_1 provocano la decisione a favore di a_1, e quelli maggiori di λ_{L-1} indicano la probabile trasmissione di a_L.

Lo stesso valore P_{δ} 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, in quanto in tali casi esiste solamente una delle due soglie il cui superamento determina una decisione errata, e dunque scriviamo $P_{e/a_1} = P_{e/a_L} = \frac{1}{2}P_{\delta}$. Applicando ora alla (15.9) il cambiamento di variabile descritto al § 6.2.4, si ottiene $P_{\delta} = \text{erfc}\left\{\frac{\lambda_i - a_i}{\sqrt{2}\sigma_\nu}\right\}$, ed esprimendo l'intervallo $\lambda_i - a_i$ in funzione della dinamica di segnale Δ troviamo

$$P_{\delta} = \text{erfc}\left\{\frac{\Delta}{2\sqrt{2}\sigma_\nu\,(L-1)}\right\} \tag{15.10}$$

Per arrivare all'espressione della probabilità di errore *incondizionata*, ovvero indipendente dall'identità del simbolo trasmesso, occorre eseguire una operazione di *valore atteso* (§ 6.2.2) rispetto a tutti gli indici i, con $i = 1, 2, ..., L$, cioè pesare le diverse probabilità di errore condizionate P_{e/a_1} con le rispettive probabilità $P_i = Pr\{a_i\}$ degli eventi condizionanti a_i. Avendo assunto l'ipotesi di valori a_i *equiprobabili* risulta $P_i = \frac{1}{L}$ e quindi

$$P_e = E_{a_i}\left\{P_{e/a_i}\right\} = \sum_{i=1}^{L} P_i P_{e/a_i} = \frac{1}{L}\left[(L-2)\,P_{\delta} + 2\frac{1}{2}P_{\delta}\right] = \left(1 - \frac{1}{L}\right)P_{\delta} \tag{15.11}$$

in cui si è tenuto conto del diverso valore della probabilità condizionata per i livelli intermedi e per i due agli estremi.

15.4.4 Parametri di sistema e di trasmissione

Il risultato ottenuto, benché già idoneo a valutare la P_e con i dati con cui è stata impostata l'analisi, deve attraversare qualche ulteriore passaggio per poter esprimere P_e in funzione dei parametri di sistema[45] potenza *di segnale* $\mathcal{P}_R$, densità di potenza di *rumore* $\mathcal{P}_\nu(f) = {}^{N_0}/_2$ e *velocità binaria* f_b, nonché dei parametri *di trasmissione*[46] L e γ, in modo da poter affrontare gli aspetti di *bilancio di collegamento* (cap. 19). Approfondiamo a tale scopo alcune relazioni per giungere alla definizione di un nuovo parametro riassuntivo.

Legame tra potenza del segnale $\mathcal{P}_R$ e dinamica Δ Al § 15.8.1 si ottiene che, sotto le ipotesi (che manterremo valide anche nel seguito) in cui

- si adotti un impulso di Nyquist a coseno rialzato con roll-off γ;
- i valori dei simboli $a[k]$ siano statisticamente indipendenti, equiprobabili, a media nulla e distribuiti uniformemente su L livelli con dinamica $a_L - a_1 = \Delta$;

la relazione tra $\mathcal{P}_R$ e Δ risulta[47]

[45] Sono detti *di sistema* in quanto indipendenti dalla natura della trasmissione, infatti $\mathcal{P}_R$ dipende da amplificatori e mezzi trasmissivi, $\mathcal{P}_\nu(f)$ dall'entità dei disturbi additivi presenti in uscita dal canale, mentre f_b è imposta dal contratto di servizio con il produttore di contenuti, o sorgente informativa.

[46] Questi sono invece parametri *negoziati* allo scopo di ottemperare ai vincoli relativi alla banda occupata ed alla precisione del temporizzatore.

[47] Anche se il risultato sarà dimostrato al § 15.8.1, 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*.

$$\mathcal{P}_R = \frac{\Delta^2}{12}\frac{L+1}{L-1}\left(1-\frac{\gamma}{4}\right) \tag{15.12}$$

Essendo il termine $\frac{L+1}{L-1}\left(1-\frac{\gamma}{4}\right)$ decrescente per $L \geq 2$ e $\gamma \geq 0$, la potenza ricevuta assume il valore massimo $\mathcal{P}_R = \frac{\Delta^2}{4}$ nel caso di trasmissione binaria a banda minima, ossia per $L = 2$ e $\gamma = 0$. Per essere utilizzata nella (15.10), la (15.12) deve prima essere invertita in modo da esprimere $\triangle$ in funzione di $\mathcal{P}_R$

$$\triangle = \sqrt{12\frac{L-1}{L+1}\frac{\mathcal{P}_R}{(1-\gamma/4)}} \tag{15.13}$$

Facciamo ora entrare in gioco anche la conoscenza di f_b, introducendo un nuova grandezza:

Energia per bit o E_b Dato che la potenza rappresenta l'energia sviluppata per unità di tempo, e che in un secondo *entrano* f_b bit, possiamo pensare $\mathcal{P}_R$ *suddivisa* tra i bit presenti, in modo da definire una quantità detta *energia per bit*

$$E_b = \mathcal{P}_R T_b = \frac{\mathcal{P}_R}{f_b} \tag{15.14}$$

che riassume in sé i parametri di sistema *potenza di segnale* e *velocità binaria*, mentre non dipende dai *parametri di trasmissione* L e γ, e consente di sostituire $\mathcal{P}_R = E_b f_b$ nella (15.13).

Dipendenza di $\mathcal{P}_\nu$ da L e γ Ora nella (15.10) l'unico termine rimasto incognito sembra essere σ_ν, pari a $\mathcal{P}_\nu$ per via del valor medio nullo del rumore. D'altra parte la potenza di rumore $\mathcal{P}_\nu = N_0 B_N$ dipende anche da L e γ attraverso la (15.5) ovvero $B_N = \frac{f_b(1+\gamma)}{2\log_2 L}$, ma vorremmo mantenere separati i contributi dei parametri *di sistema* da quelli *di trasmissione*. Allora, anziché tentare di esprimere la (15.10) in funzione di $SNR = \frac{\mathcal{P}_R}{\mathcal{P}_\nu}$, introduciamo un diverso *rapporto di qualità.*

Definizione di Eb/No e suo contributo all'SNR Esprimendo le potenze $\mathcal{P}_\nu$ e $\mathcal{P}_R$ in funzione di $T_b = 1/f_b$, e considerando sempre un segnale dati a coseno rialzato, le eq. (15.5) e (15.14) permettono di scrivere

$$\mathcal{P}_\nu = N_0 B_N = \frac{N_0\,(1+\gamma)}{T_b 2\log_2 L} \quad \text{e} \quad \mathcal{P}_R = \frac{E_b}{T_b} \tag{15.15}$$

in modo da ottenere

$$SNR = \frac{\mathcal{P}_R}{\mathcal{P}_\nu} = \frac{E_b}{T_b}\frac{T_b 2\log_2 L}{N_0\,(1+\gamma)} = \frac{E_b}{N_0}\frac{2\log_2 L}{(1+\gamma)} \tag{15.16}$$

Quindi, mentre SNR dipende anche da L e da γ, il rapporto $\frac{E_b}{N_0}$ coinvolge solo i parametri di sistema $\mathcal{P}_R$, f_b ed N_0: sarà questa la variabile *indipendente* rispetto alla quale valutare la P_e.

15.4.5 Probabilità di errore per simbolo

Non resta ora che inserire la (15.13) nella espressione di $P_{\hat{\delta}}$ (eq. 15.10), ricordare che $\sigma_\nu^2 = \mathcal{P}_\nu$, e tenere conto della (15.16), in modo da ottenere la probabilità di decidere per un simbolo a_j diverso da quello trasmesso[48]:

$$P_e^{simb} = \left(1 - \frac{1}{L}\right) \operatorname{erfc}\left\{\sqrt{\frac{E_b}{N_0} \frac{3\log_2 L}{(L^2-1)\,(1+\gamma)\,\left(1-\frac{\gamma}{4}\right)}}\right\} \tag{15.17}$$

la cui dipendenza da $\frac{E_b}{N_0}$ (espresso in dB, vedi § 8.1) è graficata alla Fig 15.9 per tre condizioni operative.

In particolare notiamo che per $L = 2$ e $\gamma = 0$ la (15.17) diviene

$$P_e = \frac{1}{2}\operatorname{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\} \tag{15.18}$$

mentre, a parità di E_b/N_0, scelte progettuali diverse da $L = 2$ e $\gamma = 0$ determinano immancabilmente un peggioramento della P_e: tali scelte possono essere comunque adottate per soddisfare esigenze di risparmio di banda (aumentando L)[49], e per ridurre i termini di interferenza intersimbolica (aumentando γ).

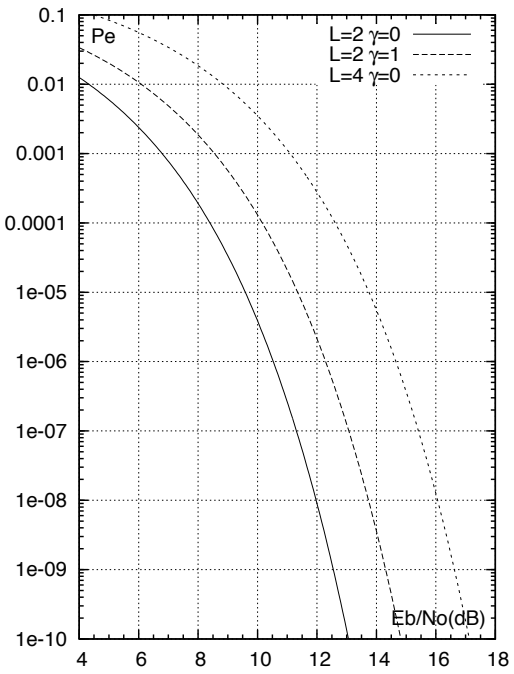


Figura 15.9: Andamento di P_e vs. E_b/N_0

Due domande riassuntive:

- perché P_e peggiora se aumento i livelli? *Risposta* ([50]).
- perché P_e peggiora se aumento γ? *Risposta* ([51]).

[48] Per completezza sviluppiamo i passaggi, piuttosto banali anche se non ovvi:

$$P_e = \left(1-\tfrac{1}{L}\right) P_{\hat{\delta}} = \left(1-\tfrac{1}{L}\right) \operatorname{erfc}\left\{\tfrac{\Delta}{2\sqrt{2}\sigma_\nu (L-1)}\right\} = \left(1-\tfrac{1}{L}\right) \operatorname{erfc}\left\{\sqrt{12\tfrac{L-1}{L+1}\tfrac{\mathcal{P}_R}{(1-\gamma/4)}}\tfrac{1}{2\sqrt{2\mathcal{P}_\nu}(L-1)}\right\} =$$
$$= \left(1-\tfrac{1}{L}\right) \operatorname{erfc}\left\{2\sqrt{3\tfrac{L-1}{L+1}\tfrac{1}{(1-\gamma/4)}}\tfrac{1}{2\sqrt{2}}\sqrt{\tfrac{\mathcal{P}_R}{\mathcal{P}_\nu}}\tfrac{1}{(L-1)}\right\} = \left(1-\tfrac{1}{L}\right) \operatorname{erfc}\left\{\sqrt{\tfrac{3}{2}\tfrac{L-1}{L+1}\tfrac{1}{(L-1)^2}\tfrac{1}{(1-\gamma/4)}SNR}\right\} =$$
$$= \left(1-\tfrac{1}{L}\right) \operatorname{erfc}\left\{\sqrt{\tfrac{3}{2}\tfrac{1}{L^2-1}\tfrac{1}{(1-\gamma/4)}\tfrac{E_b}{N_0}\tfrac{2\log_2 L}{1+\gamma}}\right\} = \left(1-\tfrac{1}{L}\right) \operatorname{erfc}\left\{\sqrt{\tfrac{E_b}{N_0}\tfrac{3\log_2 L}{(L^2-1)(1+\gamma)\left(1-\frac{\gamma}{4}\right)}}\right\}$$

[49] Aumentando L l'argomento di (15.17) diminuisce in quanto $\left(L^2-1\right)$ cresce più velocemente di $\log_2 L$.

[50] Perché a parità di $\mathcal{P}_R$ gli intervalli di decisione sono più ravvicinati, le "code" della gaussiana sottendono un'area maggiore, e questo peggioramento prevale sul miglioramento legato alla diminuzione di σ_ν conseguente alla riduzione della banda di rumore.

[51] 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.

15.4.6 Relazione con il filtro adattato

Qualche lettore può chiedersi come mai si sia utilizzato come filtro di ricezione un semplice passa basso, anziché operare come descritto al § 7.6. Tale opzione viene esplorata al § 15.5, ma possiamo notare fin da subito l'equivalenza tra i risultati (15.18) e (7.27). Infatti l'energia di un singolo impulso $\mathcal{E}_g$ equivale all'energia per bit E_b, ed un segnale dati a media nulla e simboli binari corrisponde ad una segnalazione antipodale. Quanto all'adozione di un impulso di Nyquist a coseno rialzato con $\gamma = 0$, ovvero a *banda minima* (§ 15.2.2.3), ciò corrisponde ad aver posto $G(f) = T_s \operatorname{rect}_{f_s}(f)$, ovvero proprio il passa basso ideale qui adottato in ricezione, che si rivela essere anche *adattato* nel caso appunto di trasmissione binaria a banda minima. Viceversa, il passa basso ideale non è più adattato qualora si scelga $g(t)$ con $\gamma > 0$, e questo è il motivo della dipendenza della (15.17) dal parametro γ.

15.4.7 Compromesso banda - potenza

Osservando le fig. 15.9 e 15.12 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*. Si può infatti dimostrare (vedi pag. 564) che *è possibile trasmettere senza errori* (ricorrendo a tecniche di codifica di canale ottimali) purché la velocità di trasmissione f_b non ecceda la *capacità di canale*, definita come

$$C = B \log_2 \left(1 + \frac{\mathcal{P}_R}{N_0 B}\right) \tag{15.19}$$

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 minor *banda passante* B dispone di una minore capacità, in quanto anche se in tal caso l'argomento di $\log_2(.)$ aumenta, il logaritmo cresce più lentamente di quanto non decresca B che compare a fattore nella (15.19); pertanto per mantenere la stessa capacità è necessario trasmettere con una maggiore potenza di segnale $\mathcal{P}_R$. Per questo motivo qualora sussistano 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 potenza. L'argomento viene approfondito a pag. 565.

Coerentemente con queste osservazioni, un ulteriore aumento di banda occupata si può ottenere con l'aggiunta di bit di ridondanza, come avviene applicando le tecniche di *codifica di canale* introdotte al § 15.6 ed approfondite cap. 17, dato che a questo corrisponde un *aumento* della velocità di trasmissione complessiva. Mostreremo in tale sede come ciò consenta di *ridurre* la probabilità di errore, e dunque migliorare la *fedeltà* del flusso binario, anche a parità di potenza ricevuta.

15.4.8 Diagramma ad occhio in presenza di rumore

Si tratta dello stesso tipo di grafico già descritto a pag. 454, e che ora ci aiuta a valutare in modo visivo la qualità di una trasmissione numerica. In fig. 15.10 sono riportati i grafici per un segnale dati a 4 livelli, in presenza di due diversi valori per la potenza

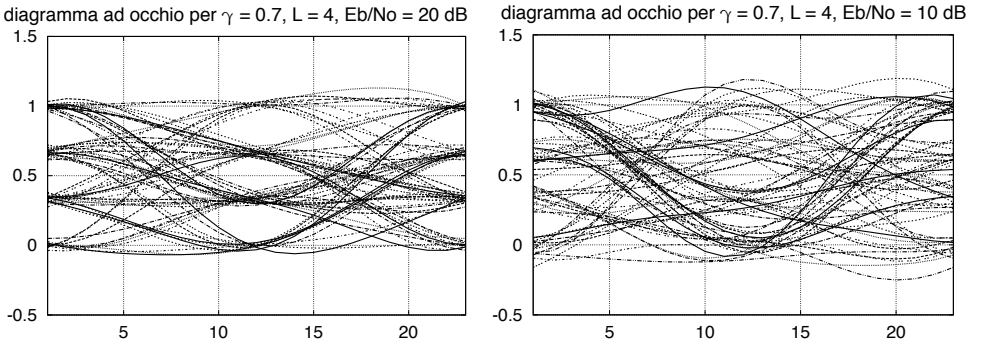
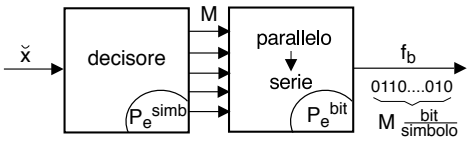


Figura 15.10: Diagramma ad occhio con E_b/N_0 pari a 20 e 10 dB, $\gamma = .7$, $L = 4$

di rumore: notiamo che al peggiorare del rapporto $\frac{E_b}{N_0}$ da 20 a 10 dB la zona priva di traiettorie (*l'occhio*) riduce la sua estensione verticale (*tende a chiudersi*). Con un tale approccio la qualità di un segnale numerico può essere valutata in modo approssimato, qualora si disponga di un oscilloscopio, esaminando il *grado di apertura dell'occhio.*

15.4.9 Valutazione della probabilità di errore per bit

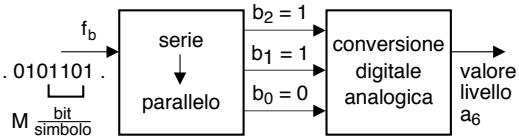
La probabilità di errore P_e^{simb} (15.17) si riferisce all'evento di decidere per la ricezione del *simbolo* a_i quando invece ne è stato trasmesso un altro, mentre ora intendiamo valutare la probabilità che sia errato *un qualunque bit* presente nel flusso a velocità f_b, ricostruito dopo la *serializzazione* (vedi figura a lato) della codifica binaria associata al simbolo a_i emesso dal decisore.



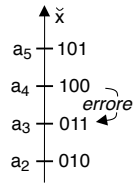
Precisiamo subito che quando la decisione per a_i è errata significa che in realtà è stato trasmesso a_{i-1} o a_{i+1} e non un altro simbolo qualsiasi, dato che la probabilità che il rumore provochi il salto di *due o più* livelli è molto inferiore a quella di *un salto singolo.* Questa circostanza ha permesso di ideare il procedimento (che ora illustriamo) di associare ad ogni simbolo (o livello) una particolare *codifica binaria*, capace di garantire la presenza di *un solo* bit errato per ogni simbolo errato.

15.4.9.1 Codice di Gray

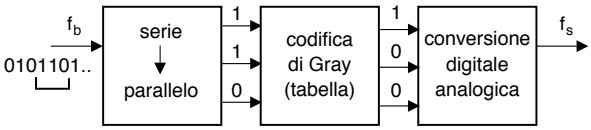
Per illustrare il problema di cui questo codice è soluzione, riprendiamo in esame la fig. 15.4-a) dove si mostra un segnale dati multilivello con $L = 4$ per il quale i valori a_i sono associati a coppie di bit b_1b_0 che sono la semplice *codifica binaria* dell'indice i corrispondente, ovvero $i = b_1 2^1 + b_0 2^0$: ciò significa che i valori a_i possono essere prodotti da un semplice convertitore D/A (pag. 97) alimentato da una parola di M bit $b_{M-1} \dots b_1 b_0$ per una trasmissione ad $L = 2^M$ livelli, come mostrato a lato per $M = 3$.



Prendiamo quindi in esame la situazione (per $M = 3$) mostrata a a lato, e consideriamo ad esempio di trasmettere il livello a_4 a cui è associata la codifica 100, e che il decisore a causa del rumore commetta l'errore di ritenere di aver ricevuto il livello a_3, associato alla sequenza 011: in tal caso avremmo sbagliato tutti e tre i bit!



Per evitare di osservare un numero di bit errati che dipende dal simbolo trasmesso e dal segno del rumore, la conversione D/A viene fatta precedere da una *riscrittura* della parola di M bit attuata consultando una tabella dove è memorizzato il *codice di Gray*.



Possiamo immaginare l'operazione come quella di un accesso a una *memoria associativa*[52], in cui la parola originaria costituisce *la chiave* con cui individuare la parola *codificata* da utilizzare al suo posto, come rappresentato nella tabella a lato per $M = 3$. Notiamo che la colonna di sinistra ha la proprietà di codificare righe adiacenti mediante configurazioni binarie che differiscono tra loro *in una sola posizione*, ovvero per un solo bit. Per analizzare la conseguenza di ciò, osserviamo che ora al posto della parola 110 di ingresso (quarta riga) si usa il codice 100 a cui il DAC fa carrispondere il livello a_4, lo stesso dell'esempio precedente.

Ingresso	Codifica
100	111
101	110
111	101
110	100
010	011
011	010
001	001
000	000

Decodifica In ricezione si attua la trasformazione inversa che utilizza la tabella *al contrario*, individuando nella seconda colonna la riga in cui compare la *codifica binaria* associata al livello ricevuto, e sostituendo ad essa la parola nella prima colonna. In assenza di errori si riottiene la parola binaria originale; se invece si verifica un errore, ovvero ad es. come prima al posto di a_4 si decide per a_3 (011), il *decodificatore* di Gray al lato ricevente individua tale *chiave* alla 5^a riga della seconda colonna, a cui fa corrispondere la sequenza 010 che trova alla prima colonna, e che infatti differisce dall'originale (110) per un solo bit (il primo).

In presenza di un errore sul simbolo il procedimento illustrato produce *sempre* un solo bit errato. Ciò comporta che con M bit a simbolo la probabilità di osservare un bit errato si riduce di M volte rispetto a quella di errore sul simbolo, ossia risulta $P_e^{bit} = P_e^{simb}/M$, dato che

$$P_e^{bit} = \frac{\text{n.bit errati}}{\text{n.bit totali}} = \frac{\text{n.simboli errati}}{\text{M·n.simboli}} = P_e^{simb}\frac{1}{M} \tag{15.20}$$

Esempio Con $L = 256$ livelli ovvero $M = 8$ bit/simbolo la P_e sul bit si riduce di $\log_2 L = 8$ volte.

[52] Mentre in un *array* gli elementi sono inviduati in base alla loro posizione od *indice*, una memoria associativa *non è ordinata* e restituisce l'elemento *associato* alla chiave, come ad esempio `colore[banana]=giallo`.

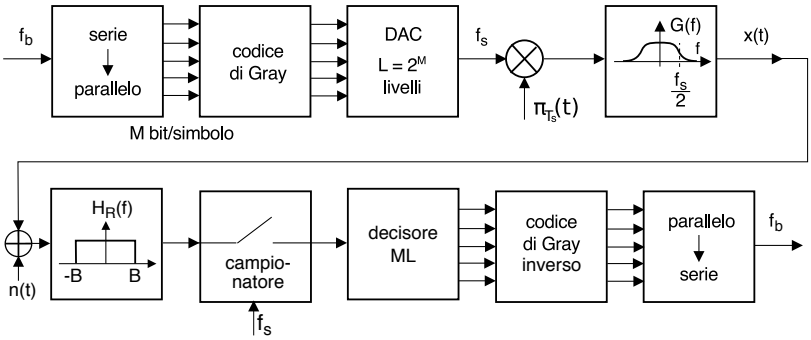


Figura 15.11: Co-decodifica di linea per un segnale dati multilivello a coseno riazato e codifica di Gray

Riassumendo La figura 15.11 mostra l'intera sequenza di operazioni necessarie a generare un segnale dati multilivello, con codifica di Gray ed impulso a coseno rialzato, e quindi riceverlo recuperando la sequenza trasmessa. Ricordiamo che mentre al flusso binario di ingresso compete una velocità di f_b bit/secondo, la sequenza multilivello 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 $B = \frac{f_s(1+\gamma)}{2} = \frac{f_b(1+\gamma)}{2\log_2 L}$, vedi eq. (15.5).

15.4.9.2 Probabilità di errore per bit

Alla luce dell'evidente vantaggio di ottenere un solo bit errato per ogni simbolo errato il codice di Gray discusso al § precedente viene adottato in modo *sistematico*, e la (15.20) può essere letta "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}·Pr{bit errato/simbolo errato}* = $P_e^{simb} \cdot \frac{1}{\log_2 L}$" . L'espressione (15.17) della P_e *per bit* nel caso si adotti una codifica di Gray diviene quindi

$$P_e^{bit} = \frac{1}{\log_2 L}\left(1 - \frac{1}{L}\right) \operatorname{erfc}\left\{\sqrt{\frac{E_b}{N_0}\frac{3\log_2 L}{(L^2-1)\,(1+\gamma)\,\left(1-\frac{\gamma}{4}\right)}}\right\} \tag{15.21}$$

Le curve in fig. 15.12 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 $\gamma \neq 0$ equivalgono ad un peggioramento[53] per $\frac{E_b}{N_0}\big|_{dB}$ pari a $10\log_{10}(1+\gamma)\left(1-\frac{\gamma}{4}\right)$, o detto in altri termini, conseguono la stessa P_e^{bit} del caso $\gamma = 0$, a patto di incrementare $\frac{E_b}{N_0}\big|_{dB}$ della stessa quantità[54].

Dimensionamento di una trasmissione numerica Una tipica metodologia operativa di progetto può basarsi sull'imporre un determinato valore di P_e^{bit}, una volta nota

[53] Infatti con $\gamma > 0$ l'argomento di erfc {.} si riduce. Ma non di molto: per $\gamma = 1$ il peggioramento risulta di 1.76 dB.

[54] Una volta scelto un valore per L è individuta la curva da usare in fig. 15.12, ed una volta imposta una P_e sulle ordinate l' $\frac{E_b}{N_0}\big|_{dB}$ necessario a conseguire tale P_e con $\gamma = 0$ si ottiene sulle ascisse seguendo la curva. Aumentando γ l'argomento di erfc {.} nella 15.21 si riduce, e ciò equivale a spostarsi verso sinistra sull'asse delle ascisse della stessa quantità di dB, a cui corrisponde (seguendo la curva) un aumento della P_e. Per ristabilire la P_e desiderata non resta quindi altro da fare che aumentare $\frac{E_b}{N_0}\big|_{dB}$ dello stesso numero di dB.

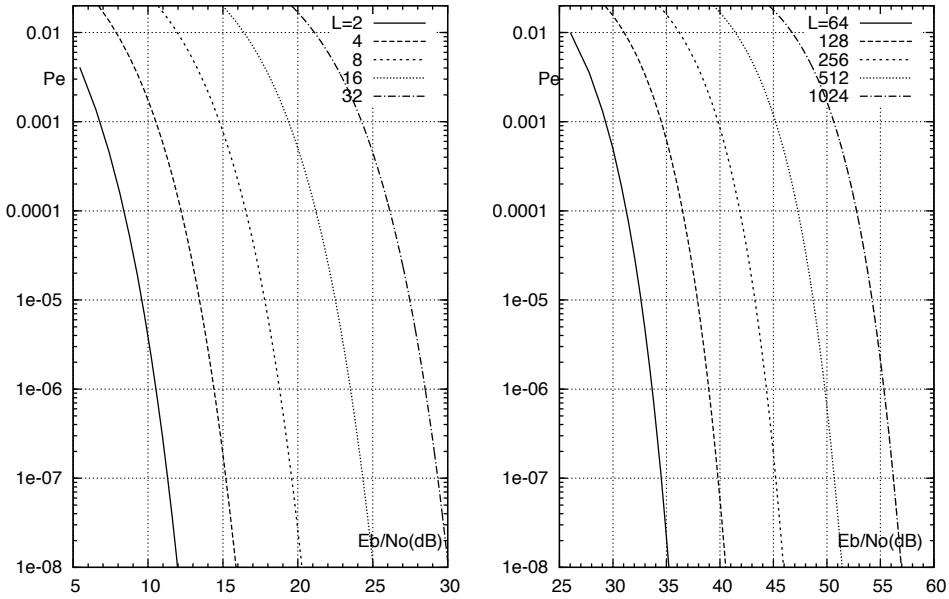


Figura 15.12: Probabilità di errore sul bit per trasmissione multilivello a banda minima e con codifica di Gray

la banda disponibile B e la velocità f_b richiesta. In tal caso

- in base a B e f_b si può determinare il valore di L mediante la (15.5), nell'ipotesi di adottare $\gamma = 0$;
- in base alle curve di fig. 15.12 ed al valore di L individuato, si determinano i valori di E_b/N_0 (in dB) necessari per ottenere la P_e^{bit};
- 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 $W_{R_{min}} = E_b \cdot f_b$.

Esempio Un canale analogico con banda a frequenze positive $B = 500$ KHz è utilizzato per realizzare la trasmissione numerica di un flusso binario a velocità $f_b = 10$ Mbps adottando una codifica di linea multilivello con codice di Gray ed impulso a banda minima. Al punto di ricezione è presente un rumore gaussiano bianco a media nulla e densità di potenza $\mathcal{P}_n(f) = 10^{-12}\ \frac{W}{Hz}$, la cui potenza è limitata dal ricevitore mediante un filtro passa basso con la medesima banda del canale. Desiderando una $P_e \leq 10^{-5}$, determinare la potenza di segnale che è necessario ricevere.

Svolgimento Per prima cosa determiniamo il numero di livelli: sapendo che $B = \frac{f_b}{2}\frac{1}{\log_2 L}$ si ottiene $\log_2 L = \frac{f_b}{2B} = 0.5 \cdot 10^7 \frac{1}{5 \cdot 10^5} = 10$, e dunque $L = 2^{10} = 1024$ livelli. Dalle curve $P_e\,(E_b/N_0)$ otteniamo quindi che per avere $P_e = 10^{-5}$ con 1024 livelli occorre un $E_b/N_0\big|_{dB} \geq 54$ dB, ossia $E_b/N_0 \geq 10^{5.4}$. Osservando infine che $N_0 = 2\mathcal{P}_n(f)$ si ottiene la potenza del segnale come $\mathcal{P}_x = E_b \cdot f_b = \frac{E_b}{N_0} \cdot N_0 \cdot f_b = 10^{5.4} \cdot 2 \cdot 10^{-12} \cdot 10 \cdot 10^6 = 2 \cdot 10^{0.4} = 5.2$ Watt.

A pag. 490 viene proposto un diverso esercizio che comprende anche alcuni concetti introdotti alla sezione 15.6.

15.5 Ricevitore ottimo

In questa sezione rimettiamo in discussione i risultati ottenuti ai §§ 15.2.2.3 e 15.4.5. Infatti come è stato illustrato al § 7.6 in relazione al *filtro adattato* in presenza di rumore bianco, il valore di SNR al 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 $H_R(f) = G^*(f)$. Al contrario, nello schema adottato per la figura a pag. 456 il filtro di ricezione ha l'unico scopo di limitare la banda del rumore, ed è sempre un passa-basso ideale, indipendentemente dalla scelta fatta per $g(t)$. In tal caso, 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 indipendenti*[55], e quindi la P_e che si ottiene *non è* la minima possibile[56].

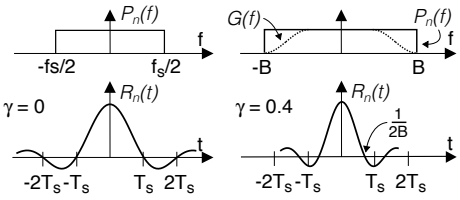
Per rendere incorrelati i campioni di rumore e ridurre la P_e al minimo, realizzando al contempo le condizioni di Nyquist in ricezione, tentiamo di verificare anche le condizioni $H_R(f) = H_T^*(f)$ di filtro adattato, decomponendo il filtro a coseno rialzato $G(f)$ in parti uguali tra trasmettitore e ricevitore e dando quindi luogo allo schema di figura 15.13, in cui

$$H_T(f) = H_R(f) = \sqrt{G(f)}$$

In tal modo al decisore giunge esattamente lo stesso segnale di prima[57], mentre la

[55] Infatti il segnale $n(t)$ uscente da $H_R(f) = rect_{2B}(f)$ ha autocorrelazione $\mathcal{R}_N(\tau) = \mathcal{F}^{-1}\left\{|H_R(f)|^2\right\} = 2B\text{sinc}(2B\tau)$ (vedi § 7.2.4), 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 $\tau = \frac{1}{2B} = \frac{1}{f_s(1+\gamma)}$, mentre invece il segnale 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 $\mathcal{R}_N(T_s) = 2B\text{sinc}(1+\gamma)$.



[56] Al § 6.5.1 si dimostra come delle v.a. gaussiane incorrelate siano anche statisticamente *indipendenti*, mentre nel nostro caso i campioni di rumore sono correlati, e statisticamente dipendenti. In accordo alla trattazione della *regressione* (§ 7.7.1) e della *predizione lineare* (§ 10.1.2.2), osserviamo che la *dipendenza statistica* tra campioni di rumore implica la possibilità di ridurre l'incertezza relativa ai nuovi valori a partire dalla conoscenza dei valori passati. Il *vero* valore di un campione *di rumore* può essere calcolato sottraendo al valore s_{k-1} del segnale ricevuto all'istante di simbolo $k-1$, il valore del simbolo *deciso* senza commettere errore; da questo risultato è possibile *predire* il successivo campione di rumore come $\hat{n}_k = n_{k-1}\frac{\mathcal{R}_N(T_s)}{\mathcal{R}_N(0)}$, che viene quindi sottratto al successivo valore s_k osservato. In tal modo, anche se la regressione non è *esatta*, l'ampiezza (e la varianza) del rumore residuo sono comunque ridotte, ed altrettanto la probabilità di errore del decisore.

[57] Infatti quando $G(f)$ è tutta al trasmettitore il segnale generato (e ricevuto) ha espressione (15.7) (vedi anche la (15.1)); indicando ora $g^{\surd}(t) = \mathcal{F}^{-1}\left\{\sqrt{G(f)}\right\}$, ed eseguendo un calcolo del tutto analogo a quello svolto in § 15.1.2.2, 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 $h_T(t) * h_R(t) = g^{\surd}(t) * g^{\surd}(t) = g(t)$ per la proprietà di prodotto in frequenza.

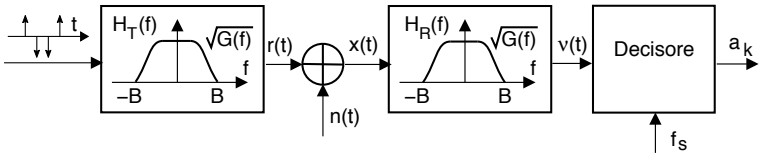


Figura 15.13: Ricevitore ottimo con impulso a radice di coseno rialzato

densità di potenza del rumore a valle di $H_R(f)$ non è più costante, ma ora vale

$$\mathcal{P}_\nu(f) = \frac{N_0}{2}|H_R(f)|^2 = \frac{N_0}{2}G(f) \tag{15.22}$$

Pertanto i campioni di rumore presi a distanza T_s sono incorrelati (e quindi statisticamente indipendenti perché gaussiani, vedi § 6.5.1) in quanto $\mathcal{R}_\nu(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_\nu(f)\}$ 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.

Prestazioni Per ottenere risultati comparabili con quelli ottenuti per $H_R(f) = \text{rect}_{f_s(1+\gamma)}(f)$ consideriamo un filtro a coseno rialzato $G(f)$ per il quale $\max|G(f)| = 1$ (anziché T_s come al § 15.2.2.3), e 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 vale[58]

$$\mathcal{P}_\nu = \int_{-\infty}^{\infty}\mathcal{P}_\nu(f)\,df = \frac{N_0}{2}\int_{-\infty}^{\infty}G(f)\,df = \frac{N_0}{2}f_s = \frac{N_0}{2T_b\log_2 L} \tag{15.23}$$

riducendosi di un fattore $(1+\gamma)$ se confrontata con (15.15), e causando un aumento equivalente per l'SNR; lo stesso fattore $(1+\gamma)$ è quindi rimosso anche nella (15.21), portando a

$$p_e^{bit} = \frac{1}{\log_2 L}\left(1-\frac{1}{L}\right)\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}\frac{3\log_2 L}{(L^2-1)\left(1-\frac{\gamma}{4}\right)}}\right\} \tag{15.24}$$

il valore della probabilità di errore sul bit ottenuta 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. 15.12. D'altra parte, il fatto che per $\gamma = 0$ la (15.24) coincida con la (15.21) *non è* un risultato inatteso: infatti, se $\gamma = 0$ si attua una trasmissione *a banda minima*, e dunque un $H_R(f)$ rettangolare passabasso realizza esattamente un filtro adattato!

Conseguenze Notiamo che 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 $H_R(f) = \sqrt{G(f)}$. La figura 15.14 mostra poi l'andamento di $g^{\surd}(t) = \mathcal{F}^{-1}\left\{\sqrt{G(f)}\right\}$ a confronto con una $g(t)$ a coseno rialzato, per valori di roll-off pari a 0.5 ed 1, ottenuta mediante IFFT della corrispondente risposta in frequenza di modulo unitario nell'origine. Si può notare

[58] Il risultato si può ottenere visivamente, a partire dalla $G(f)$ a coseno rialzato mostrata in fig. 15.7 a pag. 452 ma con altezza 1, e in base alle sue proprietà di simmetria attorno a $\pm f_s/2$: il risultato dell'integrale $\int_{-\infty}^{\infty}G(f)\,df$ è quindi pari all'area di un rettangolo di altezza 1 e base $f_s = 1/T_s$.

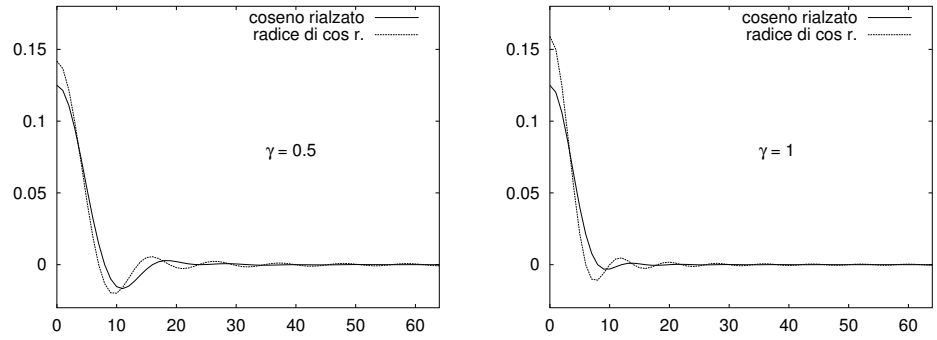


Figura 15.14: Confronto della risposta impulsiva del filtro ottimo e subottimo

un aumento sia della durata che dell'ampiezza delle oscillazioni: questa circostanza determina una maggiore complessità realizzativa del filtro di trasmissione, che deve avere una risposta impulsiva più lunga[59].

15.5.1 Equalizzazione del ricevitore ottimo

Ulteriori considerazioni possono essere svolte qualora il canale trasmissivo presenti una risposta in frequenza $H(f)$ *non ideale*, rendendo necessaria l'adozione di un filtro di equalizzazione che in accordo con quanto discusso al § 15.3 può essere semplicemente inglobato in quello di trasmissione, sintetizzando lo stesso come $H_T(f) = \sqrt{G(f)}/H(f)$. In tal caso la risposta in frequenza complessiva[60] $H_T(f) H(f) H_R(f) = G(f)$ torna ad essere quella di un filtro di Nyquist, il segnale al punto di decisione è esente da ISI, e la densità di potenza del rumore sovrapposto è ancora espressa dalla (15.22), in modo che i campioni della componente di rumore negli istanti di simbolo sono tuttora incorrelati, e la probabilità di errore *dovrebbe essere* la minima possibile.

Tuttavia la presenza del canale distorcente e dell'equalizzatore determinano, a parità di altre condizioni (potenza di segnale, velocità binaria e livello di rumore), un peggioramento della P_e causato dalla riduzione del rapporto E_b/N_0 da utilizzare nella (15.24) di una quantità α_{dB}, che al § 15.8.2 è valutata in

$$\alpha_{dB} = 10 \log_{10} \int_{-B}^{B} \frac{G(f)}{|H(f)|^2} df \quad dB \tag{15.25}$$

Equalizzazione distribuita Una soluzione alternativa al problema dell'equalizzazione è quella di suddividerne il compito in parti uguali sia al lato di trasmissione che a quello di ricezione, realizzando $H_T(f) = H_R(f) = \sqrt{G(f)}/H(f)$: anche così la risposta in frequenza complessiva $H_T(f) H(f) H_R(f)$ è pari a quella di un filtro a coseno rialzato[61], ma stavolta il rumore al punto di decisione ha densità spettrale

[59] Per una analisi degli effetti della limitazione temporale dell'impulso $g_{\surd}(t)$, vedere il contributo disponibile presso `https://engineering.purdue.edu/~ee538/SquareRootRaisedCosine.pdf`.

[60] A meno di un contributo di fase lineare $e^{j2\pi f\tau}$ necessario a garantire la causalità dell'insieme.

[61] Anche in questo caso, a meno di un termine di fase lineare, che viene omesso per non appesantire la notazione.

$\mathcal{P}_\nu(f) = \frac{N_0}{2}|H_R(f)|^2 = \frac{N_0}{2}\frac{G(f)}{|H(f)|}$ e dunque i campioni di rumore agli istanti di decisione sono correlati, vedi nota 55. Cionostante, al § 15.8.2.1 si mostra che la scelta di suddividere l'equalizzazione è migliore: in questo caso il peggioramento di E_b/N_0 ammonta a

$$\beta_{dB} = 20\log_{10}\int_{-B}^{B}\frac{G(f)}{|H(f)|}df$$

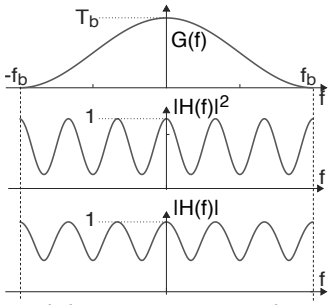
e risulta essere $\beta_{db} \leq \alpha_{dB}$, con l'uguaglianza solo se $|H(f)|$ = costante, nel qual caso $\alpha_{dB} = \beta_{db} = 0$ dB. Per una discussione del risultato, si rimanda al § 15.8.2.1.

Esercizio Una trasmissione numerica binaria antipodale con potenza $\mathcal{P}_x = 1\ Volt^2$ e velocità $f_b = 10$ Mbps adotta un impulso a coseno rialzato con roll-off $\gamma = 1$, mentre la densità di potenza del rumore bianco in ingresso al ricevitore è pari a $\mathcal{P}_n(f) = 0.5 \cdot 10^{-8}$. Determinare la probabilità di errore qualora

1. il ricevitore non sia ottimizzato ed il canale sia perfetto;
2. il filtro a coseno rialzato sia ripartito tra trasmettitore e ricevitore;
3. sia presente un canale con $|H(f)|^2 = 0.4 \cdot \cos\left(2\pi 3\frac{f}{f_b}\right) + 0.6$ e l'equalizzazione sia tutta al lato trasmittente;
4. come 3., ma con l'equalizzazione ripartita tra i due estremi del collegamento.

1. - Per il valore di $\frac{E_b}{N_0}$ otteniamo $\frac{E_b}{N_0} = \frac{\mathcal{P}_x}{f_b}\frac{1}{2P_n(f)} = \frac{1}{10^7}\frac{1}{10^{-8}} = 10$ pari a 10 dB. Per valutare la P_e utilizziamo la seconda curva di fig. 15.9 che raffigura la (15.21) per $\gamma = 1$ ed $L = 2$ (il nostro caso), ovvero $P_e^{bit} = \frac{1}{2}\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}\frac{1}{1.5}}\right\}$. Dunque per $\frac{E_b}{N_0}\big|_{dB} = 10$ dB troviamo una $P_e \simeq 10^{-4}$;

2. - siamo nelle condizionei di ricevitore ottimo, e l'espressione della P_e è data dalla (15.24) ovvero (per $\gamma = 1$ ed $L = 2$) $P_e^{bit} = \frac{1}{2}\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}\frac{1}{\left(1-\frac{\gamma}{4}\right)}}\right\} = \frac{1}{2}\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}\frac{1}{0.75}}\right\}$; essendo il denominatore sotto radice dimezzato, ciò corrisponde ad un equivalente incremento sotto radice di un fattore 2 ovvero di 3 dB, ed in corrispondenza del nuovo $\frac{E_b}{N_0}\big|_{dB} = 10 + 3 = 13$ dB troviamo una $P_e \simeq 10^{-7}$;



3. - in presenza di un canale distorcente equalizzato al trasmettitore, il ricevitore ottimo subisce un peggioramento di prestazioni equivalente ad un decremento di $\frac{E_b}{N_0}\big|_{dB}$ della quantità $\alpha_{dB} = 10\log_{10}\int_{-B}^{B}\frac{G(f)}{|H(f)|^2}df$ che, valutata per via numerica *con un programmino Octave*, risulta pari a circa 3.5 dB; pertanto al nuovo $\frac{E_b}{N_0}\big|_{dB} = 13 - 3.5 = 9.5$ dB corrisponde una $P_e \simeq 3\cdot 10^{-4}$;

4. - qualora l'equalizzazione sia ripartita tra i due estremi il decremento equivalente di $\frac{E_b}{N_0}\big|_{dB}$ è pari alla quantità $\beta_{dB} = 20\log_{10}\int_{-B}^{B}\frac{G(f)}{|H(f)|}df$ che valutata per via numerica risulta pari a circa 3.15 dB, e quindi al nuovo $\frac{E_b}{N_0}\big|_{dB} = 13 - 3.15 = 9.85$ dB corrisponde una $P_e \simeq 1.5 \cdot 10^{-4}$;

Equalizzazione al ricevitore Spesso *non è possibile* equalizzare il canale al trasmettitore od in modo ripartito[62] e l'equalizzazione deve essere svolta tutta al lato

[62] Può essere che $H(f)$ non sia nota a priori e che dunque debba essere *stimata* al ricevitore (§ 18.4), ma non sia disponibile un canale di ritorno per comunicarla al trasmettitore; oppure si tratti di una

ricevente imponendo $H_T(f) = \sqrt{G(f)}$ e $H_R(f) = \sqrt{G(f)}/H(f)$; si può mostrare che il peggioramento in dB del rapporto E_b/N_0 rispetto al caso di assenza di distorsione è dato anche ora dalla (15.25), mentre il rumore che perviene al decisore ha densità spettrale

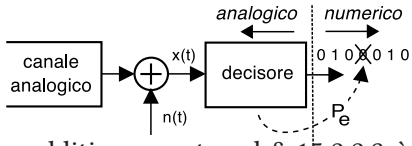
$$\mathcal{P}_\nu(f) = \frac{N_0}{2}|H_R(f)|^2 = \frac{N_0}{2}\frac{G(f)}{|H(f)|^2}$$

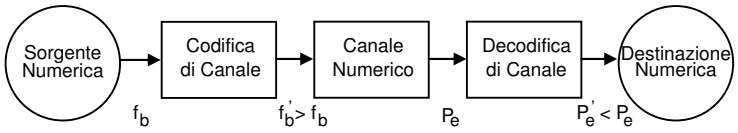
L'argomento viene ripreso al § 18.4, dove si approfondiscono le tecniche di filtraggio *adattivo* che sintetizzano $H_R(f)$ in modo da minimizzare alcune diverse funzioni obbiettivo.

Ora che abbiamo esaurito la discussione su natura e entità degli errori di trasmission, occupiamoci di trattare come questi possano essere *mitigati*, ovvero come poterci *convivere*.

15.6 Gestione degli errori di trasmissione

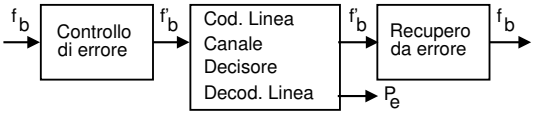
La figura a lato, già proposta a pag. 6, ricorda ancora una volta come le conseguenze prodotte dagli errori del decisore siano alcuni bit *diversi* da quelli trasmessi. Al § 15.4.9.2 siamo giunti al calcolo probabilità di errore P_e^{bit} dovuta a rumore additivo, mentre al § 15.2.2.3 è stato illustrato 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 anche mostrata la figura riproposta sotto e che illustra come la P_e, qualunque ne sia la causa, possa essere *ridotta* adottando tecniche di *codifica di canale* e *controllo di errore,* discusse nel seguito.





15.6.1 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. Vedremo che la quantità di bit aggiunti, indicata anche come *ridondanza* (§ 15.6.2.1), può essere appena sufficiente a permettere di *accorgersi* della presenza di errori di trasmissione, o (se più elevata), può mettere il lato ricevente anche in grado di *correggere* fino ad una certa percentuale dei bit errati. Sussistono dunque due diverse modalità di gestione degli errori:



trasmissione broadcast (§ 11.1.1.1), e la $H(f)$ è differente per ognuno dei ricevitori, oppure ancora $H(f)$ varia nel tempo, e la sua equalizzazione deve essere modificata di continuo.

Forward error correction o FEC Qualora il sistema di trasmissione sia da considerare *unidirezionale*[63], e dunque sprovvisto di un canale di ritorno idoneo a chiedere la *ritrasmissione* dei dati errati, l'unica soluzione consiste nell'aggiunta di una ridondanza sufficiente a correggere direttamente la maggior parte degli eventuali errori. La soluzione prende il nome di correzione di errore *in avanti* o FEC, e si è sviluppata nel contesto della *codifica di canale* (vedi § 15.6.2), appannaggio del mondo delle telecomunicazioni.

Automatic repeat request o ARQ 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, allora ci si può accontentare di un minor grado di ridondanza, sufficiente ad *accorgersi* degli errori, ovvero a *rivelarli*, ma non a correggerli. Infatti è ora possibile invocare la *ritrasmissione* del dato errato, dando luogo ad una *strategia di richiesta* di ripetizione[64] o ARQ (vedi § 22.6): tale approccio si è sviluppato nel contesto delle reti di computer e della trasmissione dati, ed ha dato origine ai *protocolli a finestra* (§ 23.1.2.3).

Suddivisione in parole Le unità informative su cui operano FEC e ARQ in generale non sono singoli bit, ma gruppi di bit denominati *parole* o WORD[65], e per questo siamo interessati a valutare come la P_e^{bit} calcolata al § 15.4.9.2 influenzi il *numero* di errori in una parola, dato che da ciò dipende la possibilità di rilevarli e/o correggerli.

15.6.1.1 Errori su parole

Occupiamoci quindi di determinare la probabilità $P(i,n)$ di trovare $0 \leq i \leq n$ bit errati *in una parola* di n bit, qualora ciascun bit possa essere errato con probabilità $p = P_e^{bit}$ e nell'ipotesi che gli eventi di errore siano statisticamente indipendenti (§ 6.1.5), ovvero che il verificarsi o meno di un errore su di un bit non condizioni gli altri.

Per i casi $i = 0$ (tutti giusti) ed $i = n$ (tutti sbagliati) il risultato è immediato, in quanto risulta $P(0,n) = (1-p)^n$ e $P(n,n) = p^n$. Per $0 < i < n$ ci troviamo in un

[63] Oltre al caso banale in cui la comunicazione sia effettivamente *half-duplex* (pag. 6), il canale deve essere considerato unidirezionale anche qualora la trasmissione a distanza riguardi informazioni generate in *tempo reale* e *consumate* immediatamente in ricezione, come nel caso televisivo o telefonico, in cui l'attesa di una ritrasmissione introdurrebbe, oltre ad una temporanea interruzione, anche un ritardo aggiuntivo a tutto ciò che viene dopo, impossibile da sostenere in una applicazione interattiva.

Un altro caso di applicazione della tecnica FEC riguarda ad es. il caso di informazioni memorizzate in forma numerica su *data storage*, come as esempio *CD/DVD, chip di memoria, hard disk....* in cui pur se possibile ri-leggere le informazioni, ciò non cambierebbe nulla, in quanto l'errore è attribuibile al supporto rovinato, e non al rumore. Per questo i dispositivi di memoria aggiungono una ridondanza ai propri dati, usata per rimediare al possibile deterioramento della loro conservazione, o per segnalare la cella di memoria come inaffidabile.

[64] L'aggettivo *automatic* si riferisce al fatto che spesso la gestione della ritrasmissione avviene a carico di uno strato protocollare (§ 22.5.2.3) di livello *inferiore* a quello che effettivamente consuma il messaggio, che in definitiva neanche si avvede della presenza del meccanismo di ritrasmissione.

[65] In generale questo raggruppamento è indipendente da quello in simboli operato dal codificatore di linea multilivello, così come non riflette altre suddivisioni come i bit di un campione quantizzato (§ 4.3) o gli intervalli temporali di una multiplazione (§ 24.2) mediante trame (§ 24.3.1) o pacchetti § 22.5.1.

classico caso di *prove ripetute*[66], in quanto si tratta di osservare l'evento di errore (con probabilità p) su n ripetizioni. La probabilità che ci siano esattamente (ad es. i primi) i bit errati ha valore $p^i\,(1-p)^{n-i}$, ma dato che i bit errati possono essere comunque distribuiti su n, e che vi sono $\binom{n}{i} = \frac{n(n-1)\cdots(n-i+1)}{i!}$ modi di scegliere i oggetti su n, il risultato cercato è espresso dalla d.d.p. di Bernoulli (vedi § 22.1)

$$P\,(i,n) = \frac{n\,(n-1)\cdots(n-i+1)}{i!}\,p^i\,(1-p)^{n-i} \tag{15.26}$$

che, se $np < 0.1$, può essere approssimata come[67]

$$P\,(i,n) \approx \frac{n\,(n-1)\cdots(n-i+1)}{i!}\,p^i \tag{15.27}$$

La (15.27) applicata al caso di un singolo errore ($i = 1$) su n fornisce $P\,(1,n) \simeq np$[68], ovvero la probabilità di *un solo* bit errato *su* n è circa pari ad n volte la P_e^{bi} nel flusso binario. D'altro canto, per la probabilità di *due* bit errati su n la (15.27) fornisce $P\,(2,n) \simeq \frac{1}{2}n\,(n-1)\,p^2$ e, sempre se $np < 0.1$, osserviamo che $P\,(2,n) \ll np \simeq P\,(1,n)$, ovvero inferiore a quella di un solo bit errato. Più in generale, risulta che

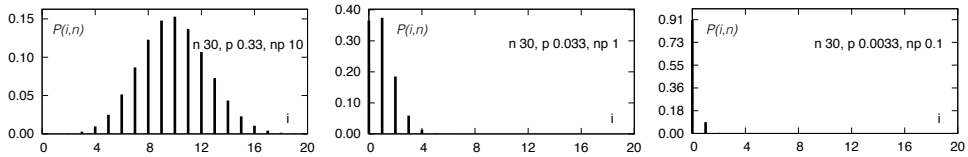
$$P\,(i+1,n) \ll P\,(i,n) \tag{15.28}$$

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/o di n, l'approssimazione perde validità, e la probabilità $P\,(i,n)$ può invece *aumentare* con i, e in tal caso il sistema di trasmissione è praticamente inusabile. E' questo il motivo per cui non è opportuno che la lunghezza n delle parole[69] ecceda il limite $np < 0.1$ imposto dalla probabilità di errore sul bit $p = P_e^{bit}$ offerta dal collegamento.

L'esposizione prosegue introducendo per prime definizioni e soluzioni che rea-

[66] Vedi ad es. `https://it.wikiversity.org/wiki/Prove_ripetute`

[67] Dal confronto tra (15.26) e (15.27) osserviamo che l'approssimazione consiste nel considerare $(1-p)^{n-i} \simeq 1$. Per verificare che ciò sia lecito qualora $np \ll 1$, cerchiamo il valore di p tale che $(1-p)^n > 0.9$, da cui discende che anche $(1-p)^{n-i} > 0.9$ per qualunque i. Dato che si può scrivere (vedi nota 2 a pag. 768) $(1-p)^n = \sum_{k=0}^{n}\binom{n}{k}p^k > 1-np$, otteniamo la condizione $(1-p)^n > 1-np > 0.9$, da cui si ottiene $np < 1-0.9 = 0.1$: ad esempio, se $n = 1000$ occorre che sia almeno $p = 10^{-4}$. Altrimenti l'approssimazione non è valida, e la (15.28) deve essere verificata; qui sotto mostriamo la d.d.p. di Bernoulli (15.26) per diversi valori di np, evidenziando come qualora $np < 0.1$ essa risulti monotona decrescente con i.



[68] In linea di principio, dato che la probabilità che solo *il primo* bit su n sia sbagliato è pari a $p\,(1-p)^{n-1}$ e che lo stesso risultato si ottiene anche per gli altri $n-1$ casi possibili, la probabilità $P\,(1,n)$ di *un solo* (generico) bit sbagliato su n è pari a $P\,(1,n) = np\,(1-p)^{n-1}$, che si approssima come np qualora si consideri $(1-p)^{n-1} \simeq 1$ in virtù della condizione $np \ll 1$.

[69] Così come non è opportuno aumentare di troppo la dimensione di un *pacchetto dati*, anche se in tal modo si riduce l'*overhead*, vedi § 22.5.1.

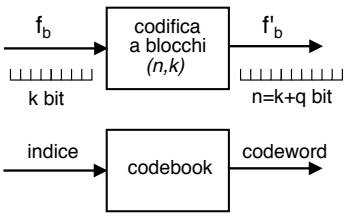
lizzano la *correzione* degli errori, il cui approfondimento è svolto al § 17.4, mentre al § 15.6.3 sono illustrate tre tecniche comunemente utilizzate per la *detezione* di errore; la descrizione dei protocolli ARQ è infine rimandata al § 22.6.

15.6.2 Correzione di errore e codifica di canale

Affrontiamo i criteri con cui scegliere i bit di ridondanza allo scopo di realizzare un sistema di tipo FEC basato sulle tecniche di *codifica di canale*, peraltro idonee ad essere utilizzate anche per il caso dei sistemi ARQ. Descriviamo quindi due semplici metodi di correzione di errore, ossia il *codice a ripetizione* e *l'interleaving*.

15.6.2.1 Codice a blocco

La *codifica a blocco* opera raggruppando k bit consecutivi del messaggio originale, distinti dai precedenti e dai successivi blocchi di k. Per ogni k bit da trasmettere il codificatore produce $n = k + q$ bit in uscita che sono trasmessi al posto dei k di ingresso, da cui dipendono in modo univoco.



Codeword e codebook L'operazione di *mappatura* svolta dal codificatore può essere pensata o realizzata come l'accesso ad una tabella (o memoria) denominata CODEBOOK (o *cifrario*), dove i k bit da codificare rappresentano un *indice* che individua 2^k differenti righe, in cui si trovano scritte le *parole di codice* (CODEWORD), costituite ognuna da $n = k + q$ bit. Un codice siffatto è detto *codice* (n, k).

Ridondanza Esprime la proporzione percentuale tra il numero dei q bit di protezione aggiunti rispetto ai k (di informazione) in ingresso al codificatore, ovvero

$$\rho = \frac{q}{k} \cdot 100\ \%$$

ed è una misura del grado di protezione offerto dal processo di codifica.

Esempio In una trasmissione con ridondanza del 50% per ogni coppia di bit di informazione ne viene inserito uno di protezione.

La presenza di ridondanza fa sì che il numero 2^k di codeword esistenti sia *inferiore* a quello delle 2^n possibili configurazioni di n bit, in modo che se un bit di una codeword viene modificato a causa di un errore del decisore, la nuova parola di n bit individua una configurazione che nel codebook *non esiste*, permettendo così di rivelare e/o correggere l'errore. Ma cosa succede se avviene più di un errore per codeword? Per rispondere, introduciamo un nuovo concetto:

Distanza di Hamming E' indicata come $d_H\left(c_i, c_j\right)$ e misura la *dissimilarità* tra due codeword c_i, c_j, espressa come il numero di posizioni di bit in cui esse sono diverse. Viene calcolata eseguendo *l'OR esclusivo* delle rispettive rappresentazioni binarie, e contando il numero di *uni* del risultato.

Esempio Con $c_i = 011010$ e $c_j = 010110$ si ottiene il risultato mostrato a lato, che evidenzia come le codeword differiscano in due posizioni di bit, e dunque $d_H(c_i, c_j) = 2$.

$$\begin{array}{r} 011010 \oplus \\ 010110 = \\ \hline 001100 \end{array}$$

Distanza del codice Individua la *minima* distanza di Hamming d_m tra tutte le possibili coppie di codeword di uno stesso codebook

$$d_m = \min_{i \neq j} d_H(c_i, c_j)$$

e permette di valutare la capacità di rivelazione e correzione del codebook, in quanto rappresenta il minimo numero di errori necessario a trasformare una codeword in un'altra, almeno nel caso peggiore di due parole a distanza d_m.

Rivelazione e correzione di errore Un codice con distanza d_m può

- *rivelare* al massimo $d_m - 1$ errori, in quanto se ne avvengono di più *si finisce* in un'altra codeword, e
- *correggere* fino a $\frac{d_m - 1}{2}$ errori, oltre i quali si finisce *più vicini* ad una altra codeword.

Mentre la prima azione è possibile ogniqualvolta si osservi una codeword c_α non presente nel codebook, per intraprendere la seconda occorre operare una decisione relativa a quale sia la codeword $\hat{c}$ *realmente* trasmessa in base ad un criterio di *minima distanza*, ovvero scegliendo la codeword più vicina (nel senso della distanza di Hamming) a quella ricevuta:

$$\hat{c} = \arg\min_{c_i} d_H(c_i, c_\alpha)$$

Esempio Nel caso in cui $d_m = 3$, la presenza di *un solo* bit errato su n fa sì che la codeword ricevuta differisca da quella trasmessa per *un* bit, mentre mantiene almeno *due bit* di differenza rispetto a tutte le altre possibili, permettendo così al ricevitore di correggere l'errore. Se sono invece presenti *due* errori, la parola ricevuta diviene più vicina ad una codeword *diversa* da quella trasmessa, ed in tal caso la procedura di correzione sceglierebbe una codeword errata.

E' quindi un po' come se attorno ad ogni codeword fosse costruita *una sfera* contenente tutte le parole di n bit che *non sono codeword* e che sono *distanti* dalla codeword al massimo $\frac{d_m - 1}{2}$ bit: la ricezione di ciascuna di esse comporta la decisione per la codeword al centro della sfera.

Massima distanza minima Per un generico codice a blocco (n, k) la distanza *del codice* d_m rispetta la diseguaglianza

$$d_m \leq q + 1$$

ovvero $d_m \leq n - k + 1$ dato che $n = k + q$. Infatti i 2^k bit da proteggere sono qualunque, e dunque il contributo di questi k bit alla distanza tra codeword è uno; tale distanza può aumentare al massimo di una ulteriore unità per ognuno dei q bit aggiunti.

Esempio Aggiungendo $q = 3$ bit di ridondanza si ha $d_m \leq q + 1 = 4$, e se per una particolare scelta del codebook (ad es. di Hamming, § 17.4.1.1) si ottiene $d_m = 3$, saremo in grado di correggere un errore e rivelarne 2.

Probabilità di errore residua per codeword Da quanto illustrato fin qui occorre stabilire *a priori* se utilizzare il codice a fini di correzione oppure di detezione. Indicando con e_M il massimo numero di errori che è possibile correggere o rivelare, può essere definita una probabilità *residua* di errore *su parola* P_e^r per descrivere il caso in cui, anche dopo l'esecuzione delle procedure di controllo di errore[70], siano ancora presenti errori, perché in numero superiore ad e_M e dunque eccedenti la capacità correttiva del codice. Risulta quindi che $P_e^r = P(i > e_M, n)$ che, nelle condizioni di validità della (15.27), è pari a $P(e_M + 1, n)$. In linea generale, la *valutazione* dell'errore residuo *sul bit* dipende dal tipo di controllo di errore attuato.

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

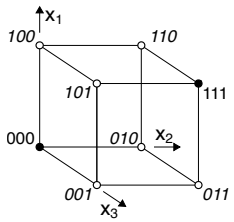
$$f_b' = \frac{f_b}{R_c} > f_b \tag{15.29}$$

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 blocco è molto vasto[71], e fornisce molteplici soluzioni, la cui trattazione esauriente eccede il livello di approfondimento del presente capitolo (vedi però il § 17.4.1); gli stessi tre casi di controllo di errore trattati al § 15.6.3 (parità, somma di controllo e CRC) possono essere inquadrati nel contesto dei *codici a blocco*. Qui ci limitiamo ad un esempio di codice a correzione molto elementare, il codice *a ripetizione*, mentre al § 17.4.1.1 è illustrata una tecnica nota come codice *di Hamming*, in grado di conseguire una efficienza di gran lunga migliore.

15.6.2.2 Codice a ripetizione n:1

Realizza un codice a blocco molto semplice e con proprietà correttive, per il quale $k = 1$ e le uniche due codeword sono di n bit tutti uguali al bit in ingresso; ponendo ad esempio $n = 3$ si ottiene il codice *a ripetizione* 3 : 1, con codeword 000 ed 111, di cui in figura è riportata la disposizione (rappresentate come pallini) in uno spazio vettoriale descritto dai tre bit di codice $x_1x_2x_3$, ed in cui i cerchi vuoti indicano gli 8-2=6 vettori binari che *non sono* codeword, scritti in *corsivo*. Se gli errori sono sufficientemente distanti nel tempo la correzione può basarsi su di una "votazione a maggioranza" (*majority voting*): sempre nel caso di $n = 3$ si



[70] Che nel caso di rivelazione richiede la ritrasmissione della parola errata.

[71] Senza pretesa di esaustività, possiamo annoverare l'esistenza dei codici *di Hamming, di Hadamard, BCH, Reed-Solomon, Reed-Muller, di Golay, di Gallager, turbo, a cancellazione, a fontana, punturati...*

ottiene $d_m = 3$[72], ed il codice è in grado di correggere un errore e rivelarne due[73].

Notiamo come questo codice sia particolarmente poco efficiente, dato che per esso si ottiene un tasso di codifica $R_c = \frac{k}{n} = \frac{1}{n}$; d'altra parte, il codice a ripetizione è uno dei pochi per cui $d_m = q + 1$, e non meno.

Esercizio Calcolare la probabilità di errore residua P^r per un codice a ripetizione 3:1, in presenza di una $P_e^{bit} = p$. **Risp.** Come discusso il codice può correggere un errore singolo, mentre in presenza di un errore doppio la decisione a maggioranza modifica anche il terzo bit, ed un errore triplo passa inosservato. La (15.27) approssima la probabilità che due o più bit su tre siano errati come[74] $P(2,3) \simeq \frac{1}{2}3 \cdot 2p^2 = 3p^2$ che è la P^r cercata *per codeword*. Dato che per il codice a ripetizione ad ogni codeword corrisponde un solo bit del messaggio originario, lo stesso valore di $P^r = 3p^2$ è anche la P_e^{bit} residua. Ad esempio, in corrispondenza di una $p = 10^{-4}$ iniziale, dopo decodifica si ottiene una $P_e^{bit} = 3 \cdot 10^{-8}$.

Compromesso banda - potenza Torniamo su questo argomento (§ 15.4.7) in quanto l'adozione di un codice di canale aumenta di fatto la f_b eq. (15.29) e dunque l'occupazione di banda eq. (15.5); non solo, ma a parità di potenza $\mathcal{P}_x$ ricevuta l'aumento di f_b comporta la riduzione di $E_b = {}^{P_x}/_{f_b}$ e dunque di ${}^{E_b}/_{N_0}$ della stessa frazione, e quindi un *peggioramento* della P_e di base su cui opera il decodificatore di linea. Evidentemente il miglioramento apportato dalla codifica FEC sopperisce anche al peggioramento dovuto all'aumento di banda, che la teoria dell'informazione prevede possa essere ripagato da un minor bisogno di potenza per ottenere le stesse prestazioni.

Esercizio Proseguendo l'esercizio precedente, osserviamo che il codice a ripetizione determina una f_b tripla, e dunque un ${}^{E_b}/_{N_0}$ ridotto di un terzo, ovvero di circa -4.7 dB inferiore rispetto al caso non codificato. Dalle curve di fig. 15.12 riscontriamo che il miglioramento sulla P_e (da 10^4 a $3 \cdot 10^{-8}$) dovuto al codice corrisponde più o meno allo stesso numero di dB. Per codici più efficienti (con minore aumento di banda come ad esempio il codice di Hamming, vedi § 17.4.1.1) il bilanciamento tra i due effetti conferma il risultato anticipato dal compromesso banda-potenza, qui applicato alla codifica di canale.

15.6.2.3 Interleaving

Le capacità di correzione fino ad ora discusse sono valide purché gli eventi di errore siano *statisticamente indipendenti*. In realtà gli errori possono presentarsi in maniera *non* indipendente, e concentrati in un breve intervallo di tempo: questa circostanza è indicata come *caso degli errori a pacchetto* [75]. In tal caso, si usa ricorrere alla tecnica

[72] Con riferimento alla figura, 3 è il numero di vertici da attraversare, ovvero di errori da subire, per passare da una codeword all'altra.

[73] Poniamo di dover trasmettere 0110. La sequenza diventa 000 111 111 000 e quindi, a causa di errori, ricevo 000 1*0*1 11*0* *1*00. Votando a maggioranza, ricostruisco la sequenza corretta 0 1 1 0.

[74] Volendo essere *esatti* la probabilità di 2 bit errati su 3 è data dalla d.d.p. binomiale (§ 22.1) ed è pari a $\binom{3}{2}p^2(1-p) = 3p^2(1-p)$, a cui va sommata la probabilità di 3 bit errati, pari a p^3. Pertanto $P^r = 3p^2(1-p) + p^3 = 3p^2 - 3p^3 + p^3 = 3p^2 - 2p^3 \simeq 3p_e^2$, approssimazione legittima se $np = 3p \ll 1$.

[75] In inglese, errori a *burst* (*scoppio*). Dovuti a rumori e disturbi di tipo *impulsivo*, ad esempio a causa di scintille come per motori elettrici o candele di accensione, o fenomeni di fast fading (§ 20.4.4).

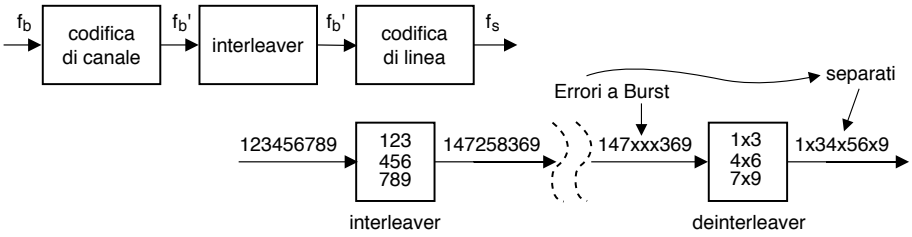


Figura 15.15: Schema di principio della trasmissione a dati intercalati

nota come *scrambling*[76] o *interleaving*[77] 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 si manifestano su bit *vicini* si riflettano in errori su bit... *lontani,* e quindi appartenenti a codeword differenti, come illustrato in fig. 15.15 in cui i bit sono scritti per righe e letti per colonne. Una analisi più approfondita viene svolta a pag. 577. 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, e dunque *non è* una forma di codifica FEC.

La trattazione della codifica di canale di tipo FEC prosegue in modo più approfondito al § 17.4, dove sono illustrate le tecniche più avanzate.

15.6.3 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 può esserne richiesta la ritrasmissione.

15.6.3.1 Controllo di parità

Viene comunemente usato nell'ambito della trasmissione asincrona (§ 15.7.1) e sincrona orientata al carattere (pag. 485) per rivelare errori sul bit, e consiste nell'aggiungere alla parola da trasmettere un unico ulteriore bit, in modo che in totale ci sia un numero *pari* di uni[78], applicando così 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 casi[79] dopo aver raggruppato i bit pervenuti il ricevitore esegue un *controllo* detto appunto *di parità*, semplicemente contando[80] il numero di uni, ed

[76] Letteralmente: arrampicamento, ma anche "arruffamento", vedi *scrambled eggs,* le uova strapazzate dell'*english breakfast.*

[77] LEAVE = *foglia, sfogliare, rastrellare,* ed il termine potrebbe essere tradotto come *intercalamento.*

[78] Ad 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.

[79] Il ricevitore deve comunque essere al corrente del fatto se la parità sia ODD o EVEN !

[80] La *conta* può essere realizzata in forma algoritmica o circuitale, eseguendo la somma modulo due di tutti i bit che compongono la parola (ovvero complementando il risultato, nel caso di parità *dispari*). La somma modulo due è equivalente all'operazione di OR esclusivo, viene a volte indicata con il simbolo $\oplus$, e corrisponde alla regola $0 \oplus 0 = 0$, $0 \oplus 1 = 1$, $1 \oplus 0 = 1$, $1 \oplus 1 = 0$.

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 errato. Se invece si è 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. 473) relativa ad un codebook ottenuto aggiungendo ad ogni possibile parola di k bit il corrispondente di parità, è pari a due[81].

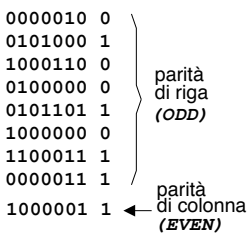
Esempio: indicando la probabilità di errore sul bit con p (es 10^{-3}) ed applicando la (15.27) si ottiene che la probabilità di $i = 2$ errori su $n = 10$ bit vale $P_e^{word} = \frac{1}{2} n\,(n-1)\,p^2 = 4.5 \cdot 10^{-5}$, che rappresenta il *tasso residuo* di errore *su parola* legato all'uso di un bit di parità: essendo due i bit errati su 10, la $P_e^{bit} = P\{err/word_e\} \cdot P_e^{word}$ risulta pari a $\frac{2}{10} \cdot 4.5 \cdot 10^{-5} = 0.9 \cdot 10^{-5}$, un bel risultato rispetto al 10^{-3} di partenza.

Il concetto di parità può essere esteso calcolando q bit di parità, ognuno a partire da un *diverso sottoinsieme* dei k bit di ingresso, con sottoinsiemi eventualmente sovrapposti. Un codice del genere prende il nome di codice di *Hamming*, descritto al § 17.4.1.1.

15.6.3.2 Somma di controllo o *checksum*

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 68.

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* o *checksum*), 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.



A volte, si preferisce calcolare la somma di controllo mediante una operazione di somma *modulo uno*[82], 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.

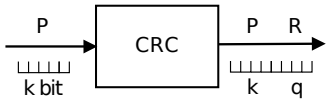
15.6.3.3 Codice polinomiale e CRC

L'utilizzo di una somma di controllo può produrre risultati scadenti nel caso di una distribuzione temporale dei bit errati particolarmente sfavorevole, mentre la tecnica nota come *Cyclic Redundancy Check* (o CRC)[83] garantisce prestazioni *più uniformi*.

[81] Considerando 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 almeno due bit.

[82] La 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.

[83] Tale denominazione indica un'azione di controllo (*check*) realizzata mediante l'aggiunta di una *ridondanza* ottenuta applicando un codice *ciclico* - vedi § 17.4.1.2.



Questo metodo consiste nell'aggiungere ad una parola P di k bit che si desidera trasmettere, un gruppo R di $q < k$ ulteriori bit *di protezione*, calcolati a partire dai primi k, in modo da permettere la detezione di eventuali errori; sotto questo aspetto, il CRC rientra nella categoria dei *codici a blocco* (§ 15.6.2.1).

L'aggettivo *polinomiale* trae origine dalla associazione tra un numero binario B di $n+1$ bit, indicati con b_i, $i = 0, 1, \ldots n$, ed un polinomio[84] $B(x)$ a coefficienti binari nella variabile x, di grado n, con espressione

$$B(x) = b_n x^n + b_{n-1} x^{n-1} + \ldots + b_1 x^1 + b_0$$

Un *codice polinomiale* è definito a partire da un *polinomio generatore* $G(x)$ di grado q, i cui coefficienti binari identificano una parola $G = g_q g_{q-1} \ldots g_1 g_0$ di $q+1$ bit.

Indicando ora con P la sequenza dei k bit p_i da proteggere, *aggiungiamo* a destra di questi un gruppo di q bit pari a zero, ottenendo una nuova parola $P \cdot 2^q$ lunga $k+q$ bit, che quindi dividiamo per G (mediante aritmetica modulo due[85]), ottenendo un quoziente Q, ed un resto R con al massimo q bit. Pertanto, possiamo scrivere

$$\frac{P \cdot 2^q}{G} = Q \oplus \frac{R}{G}$$

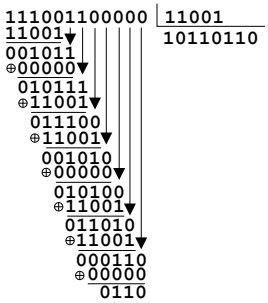
Le sequenze Q ed R costituiscono rispettivamente i coefficienti dei polinomi quoziente $Q(x)$ e resto $R(x)$, ottenibili dalla divisione di $P(x) \cdot 2^q$ per $G(x)$. I q bit R del resto sono quindi utilizzati come *parola di protezione*, in modo da esprimere la sequenza T da trasmettere come $T = P \cdot 2^q \oplus R$ di $k+q$ bit, ovvero con i k bit più significativi pari a P ed i q 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*

$$\frac{T}{G} = \frac{P \cdot 2^q \oplus R}{G} = \frac{P \cdot 2^q}{G} \oplus \frac{R}{G} = Q \oplus \frac{R}{G} \oplus \frac{R}{G} = Q$$

[84] L'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.

[85] Per fissare le idee, consideriamo $k = 8$ bit a da proteggere, pari a $P = 11100110$, $q = 4$ bit di CRC, ed un generatore $G = 11001$. La sequenza $P \cdot 2^q$ 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 $T = P \cdot 2^q \oplus R = 11100110\,0110$ con $k + q = 12$ bit.

La divisione modulo 2 si realizza come mostrato nella figura a lato: considerando i bit più significativi di $P \cdot 2^q$ e G, l'uno nell'uno ci sta una volta, e scriviamo uno come primo bit di Q. Riportiamo ora G sotto $P \cdot 2^q$, ed anziché sottrarre i bit, ne calcoliamo l'OR-ESCLUSIVO $\oplus$ bit-a-bit, ottenendo 00101, a cui aggiungiamo un uno *abbassando* il successivo bit (1) di $P \cdot 2^q$. Stavolta l'uno nello zero ci sta zero volte, e dunque aggiungiamo uno zero a Q, riportiamo cinque zeri (come la lunghezza di G) allineati sotto al resto parziale, eseguiamo l'EXOR, ed abbassiamo un'altra cifra (1) di P. Il confronto ora è tra il quinto bit da destra del resto parziale (1) ed il bit più significativo (il quinto, 1) di G, ottendo la terza cifra di Q (1). Ripetiamo il procedimento, e quando tutti i bit del divisore sono stati usati, l'ultima operazione $\oplus$ fornisce il resto R cercato.



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 q posizioni.

Nel caso invece in cui si siano verificati errori, indichiamo con E la sequenza binaria di errore, di lunghezza $k + q$ 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

$$\frac{R}{G} = \frac{T \oplus E}{G} = \frac{T}{G} \oplus \frac{E}{G} = Q \oplus \frac{E}{G} \tag{15.30}$$

e quindi si verifica la presenza di un resto diverso da zero[86], 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. Nel caso in cui $q = 1$ si ricade nel caso del controllo di parità (§ 15.6.3.1), a cui corrisponde $G(x) = x + 1$.

Per applicare il metodo, sia il trasmettitore che il ricevitore devono utilizzare lo stesso generatore $G(x)$, per il quale esistono diverse scelte standardizzate[87]. Si può dimostrare che scegliendo $G(x)$ in modo opportuno, il metodo discusso permette di rivelare

- tutti gli errori singoli;
- se $G(x)$ contiene il temine noto +1, tutti gli errori a burst che si estendono per q o meno bit;
- se $x + 1$ è un fattore di $G(x)$, tutti gli errori in numero dispari;
- se $G(x)$ è un polinomio primitivo[88], tutti gli errori doppi;
- se $G(x)$ è un polinomio primitivo di grado $q - 1$ moltiplicato per il fattore $x + 1$, tutti gli errori doppi entro un intervallo di $2^{q-1} - 1$ bit, e tutti gli errori in numero dispari.

Calcolo del CRC L'aspetto che ha reso popolare questo metodo è la maniera in cui è possibile calcolare i q bit di controllo, che vengono essi stessi indicati come CRC. Infatti la divisione binaria illustrata alla nota 85 è realizzabile a livello circuitale in modo relativamente semplice[89].

Si tratta di utilizzare un registro a scorrimento *controreazionato* (vedi figura seguente), in cui i k bit da proteggere sono immessi ad uno ad uno da destra verso sinistra,

[86] Dalla (15.30) sembrerebbe che il resto sia E, ma dato che $E(x)$ può avere grado $> q$, esso è divisibile per $G(x)$, e dunque il resto *non è* E - altrimenti, *sarebbe possibile correggerlo!!*

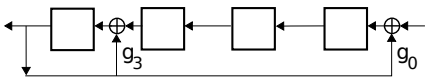
[87] Ecco quattro scelte utilizzate nei sistemi di trasmissione:

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 q genera un CRC di q 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.

[88] https://it.wikipedia.org/wiki/Polinomio_primitivo

[89] Vedi ad es. `http://en.wikipedia.org/wiki/Computation_of_cyclic_redundancy_checks`



seguiti da q zeri consecutivi. Per ogni valore immesso, quelli già presenti *scorrono* a sinistra nel registro, ed il bit che *trabocca*, oltre a rappresentare una cifra del quoziente, alimenta gli OR esclusivi presenti nel registro; questi ultimi sono posti in corrispondenza dei coefficienti di $G(x)$ pari ad uno, tranne che per quello corrispondente ad x^q.

Al termine dell'inserimento dei $k+q$ bit di $P \cdot 2^q$, lo stato del registro a scorrimento costituisce proprio il resto R, da usare come CRC. L'esempio in figura rappresenta il caso mostrato alla nota 85, in cui di $G(x) = x^4 + x^3 + 1$: con un po' di pazienza, è possibile verificare che il circuito effettivamente svolge i calcoli prescritti dalla procedura di divisione, e che si ottiene lo stesso risultato.

15.7 Sincronizzazione dati

Come mostrato al § 15.1.1 il segnale dati deve essere campionato al ricevitore con cadenza pari alla frequenza di simbolo f_s il più possibile in prossimità del centro dell'intervallo di simbolo, in modo da contrastare gli effetti della limitazione di banda (vedi fig. 15.3) e dell'ISI (§ 15.2.2.2); per questo motivo occorre che il temporizzatore mostrato in fig. 15.1 determini gli istanti di campionamento più idonei, effettuando la *sincronizzazione di simbolo*[90]. Le diverse scelte per l'onda elementare $g(t)$ discusse al § 15.2.1, determinano differenti gradi di "difficoltà" nel conseguimento di tale obiettivo.

Un diverso aspetto della sincronizzazione riguarda il problema di ricostruire la struttura *sintattica*[91] del segnale binario, in primo luogo per permettere il corretto svolgimento delle operazioni di decodifica FEC e di controllo di errore. Inoltre, la sequenza di bit 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 segnale[92], oltre alle quali possono essere state introdotte ulteriori strutture sintattiche legate ai protocolli di multiplazione sia a circuito (§ 24.3.4 e § 24.4) che a pacchetto (§ 22.5.1). Lo *strato di collegamento*, il secondo della pila ISO-OSI, si occupa appunto di segmentare il flusso di bit ricevuto in accordo a tali strutture, coinvolgendo essenzialmente il sotto-strato MAC (§ 23.1.4), che deve quindi essere in grado di conseguire la sincronizzazione di *simbolo*, di *parola* e di *trama*.

Nel seguito analizziamo le esigenze e le soluzioni di sincronizzazione che emergono

[90] In 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 una stessa *motherboard*.

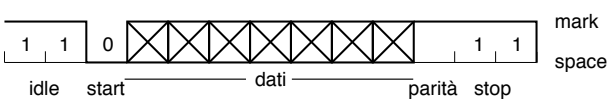
[91] Una sintassi definisce un linguaggio, prescrivendo le regole con cui possono essere costruite sequenze di simboli noti (l'alfabeto), e l'analisi delle sequenze eseguita nei termini degli elementi definiti dalla sintassi, ne permette una interpretazione semantica. Il parallelismo linguistico porta 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*).

[92] In appendice 15.8.4 è riportata la codifica in termini di sequenze binarie dei caratteri stampabili, definita dallo standard ASCII; al § 24.3.1 si mostra la struttura della *trama PCM*, che trasporta i campioni di più sorgenti analogiche campionate.

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.

15.7.1 Trasmissione asincrona

Viene adottata, ad esempio, nel caso di un terminale *stupido*[93] 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*[94].



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, indicati con una croce a significare la loro variabilità) basandosi su di un suo orologio indipendente. Segue poi un bit di *parità* (se prevista, vedi § 15.6.3.1) ed uno o due di *stop* (realizzati come MARK), presenti per assicurare una durata minima dello stato di IDLE, prima della trasmissione del carattere successivo.

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 elevata[95], che viene *risvegliato* in corrispondenza del bit di start. Tale semplicità operativa causa 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. 15.16 è mostrato uno schema di funzionamento per trasmettitore e ricevitore, in cui le parole *entrano* in modo parallelo nel trasmettitore, che le *incapsula* tra un bit di start e due di stop, disponendo la parola dal *less* al *most significant bit*. Il risultato viene quindi prelevato in modalità seriale e da questo generato il segnale dati a velocità $f_s = T_{xC}$, in cui $T_{xC} =$ `TxClock` è la frequenza (CLOCK) di *trasmissione* (T_x).

Sincronizzazione di bit e di parola Il ricevitore dispone di un orologio interno operante ad un ritmo R_{xC} (`RxClock`) multiplo di quello di trasmissione ovvero $R_{xC} = N \cdot T_{xC}$, che *decrementa* un contatore, il quale viene inizializzato con il valore $N/2$

[93] Un 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. Per lungo tempo ogni computer disponeva di interfacce seriali RS-232 che possono funzionare sia in modalità asincrona che sincrona, oggi soppiantate dalle prese USB.

[94] In tal caso la linea ".. IS MARKING TIME" (sta *marcando* il tempo).

[95] Ovviamente, occorre stabilire un accordo a priori a riguardo la velocità, ossia della frequenza, sia pure approssimata, della trasmissione.

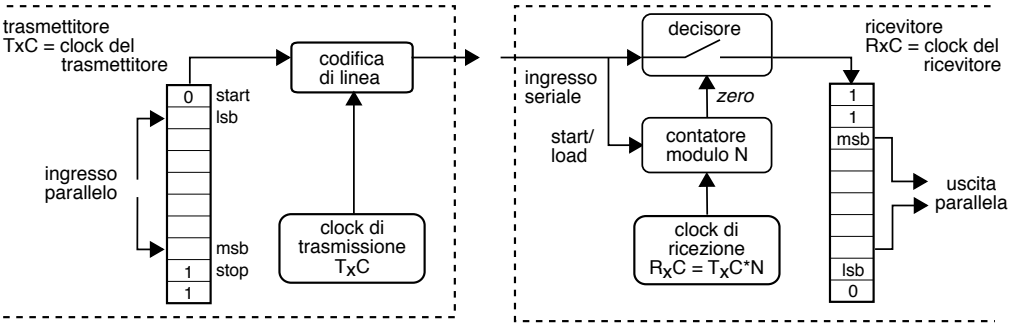
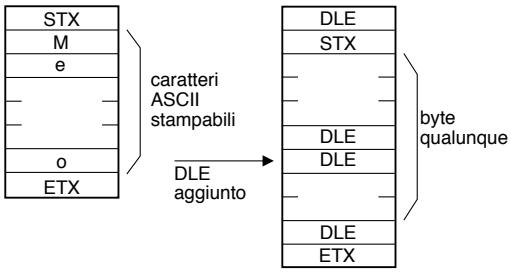


Figura 15.16: Trasmissione asincrona

in corrispondenza del fronte di *discesa* del bit di start.[96]. Il contatore quindi si azzera dopo $^{N R_{xC}}/_{2}$ periodi di R_{xC}, ovvero a metà del bit di start, determinando l'istante in cui campionare il segnale da parte del decisore, l'esito della cui decisione viene accodato in un registro di ricezione.

Da quel punto in poi il contatore viene inizializzato con il valore N, in modo che le successive *letture* del segnale di ingresso cadano 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 trasmettere non sono tutte quelle possibili in base alla lunghezza di parola adottata[97], 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. 494), 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 trasmessi[98], causando un *troncamento* prematuro del messaggio. In tal caso, sia

[96] Per accorgersi di questa e delle altre transizioni, il ricevitore può ad esempio sfruttare un circuito che approssimi l'operazione di derivata, di cui constatare il superamento di una soglia.

[97] Una 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.

[98] Cioè, i dati trasmessi, che ora riempiono tutto lo spazio delle configurazioni possibili, contengono al loro interno la configurazione che è propria del carattere ETX.

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.

15.7.2 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 successivo 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.

15.7.2.1 Sincronizzazione di simbolo

La figura 15.17-a mostra uno schema idoneo ad estrarre la temporizzazione RxC dal segnale ricevuto, basata sull'uso di un circuito DPLL (*Digital Phase Locked 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. 15.17-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, e la ricarica del contatore 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. 15.17-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 simbolo[99].

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

[99] In 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 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 controreazione, che mantiene il clock locale *al passo* con quello dei dati in arrivo. Un diverso caso particolare di questo stesso principio è analizzato al § 16.11.1, ed anche ai § 12.2.2.2 e 12.3.2.1 a proposito del PLL.

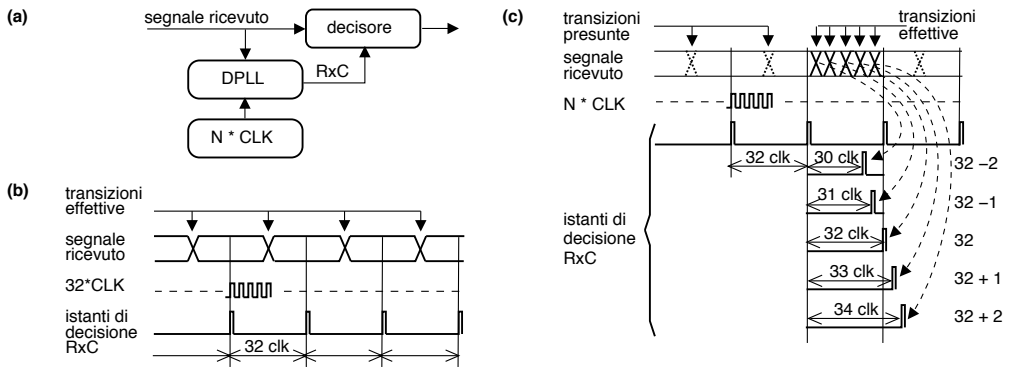


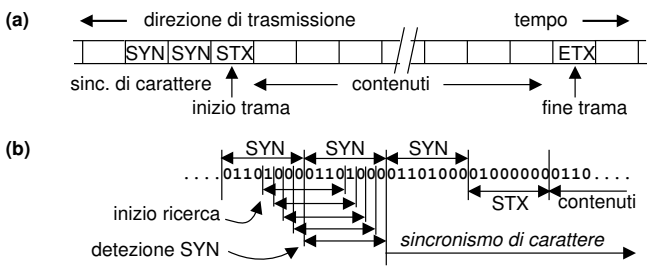
Figura 15.17: Funzionamento del *DPLL*: (a) schema circuitale; (b) condizione di sincronismo; (c) correzione di fase

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 di dati fittizi o *trailer*, allo scopo di permettere l'acquisizione del sincronismo di simbolo. La durata del trailer dipende dalla velocità di convergenza della procedura, in cui sono imposte correzioni di maggiori entità per errori di fase più elevati.

15.7.2.2 Sincronizzazione di parola e di trama

Distinguiamo due casi.

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 permette sia di conseguire (o mantenere) il sincronismo di bit, che di consentire l'individuazione dei confini di carattere, e quindi il sincronismo di carattere.

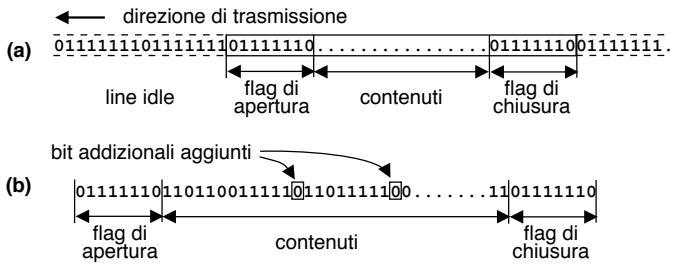


La figura a lato mostra **(a)** che la sincronizzazione di trama è ottenuta come per il caso asincrono, racchiudendo il blocco da trasmettere entro una coppia di caratteri STX-ETX. Una volta che il ricevitore ha conseguito il sincronismo di bit, passa in una modalità di ricerca, verificando **(b)** se l'allineamento di 8 bit consecutivi corrisponde al carattere SYN, ed in caso negativo, ripete il tentativo bit a bit.

Dopo aver individuato il SYN, il ricevitore consegue l'allineamento sul carattere, ed inizia ad attendere 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 asincrono,

e cioè sia l'STX che l'ETX vengono fatti precedere da un DLE, ed all'interno dei dati si esegue il *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, così da trasportare l'informazione in modo più universale e *trasparente*. 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 a lato mostra in **(a)** un esempio di trama, ed in **(b)** la soluzione del *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. Viceversa, 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.

La procedura ora descritta può fallire qualora applicata ad un flusso binario affetto da errori, ed in tal caso occorre attendere la successiva opportunità di sincronizzazione. Tale problematica può essere risolta evitando la necessità del *flag byte* finale grazie alla presenza di un campo *lunghezza* nella intestazione di pacchetto, come avviene ad es. per la trama *ethernet*, vedi § 23.1.4.2.

15.8 Appendici

15.8.1 Potenza di un segnale dati

Al § 15.4.4 si è affermato che ad un segnale dati

$$s(t) = \sum_n a[n]\, g(t - nT)$$

(in cui $g(t)$ è la risposta impulsiva di un filtro a coseno rialzato con roll-off γ, e $a[n]$ è una sequenza di v.a. discrete, statisticamente indipendenti, a media nulla ed uniformemente distribuite su L valori in una dinamica $-\frac{\Delta}{2} \le a_i \le \frac{\Delta}{2}$) corrisponde una potenza

$$\mathcal{P}_s = \frac{\Delta^2}{12}\frac{L+1}{L-1}\left(1 - \frac{\gamma}{4}\right) \tag{15.31}$$

Svolgiamo qui i passi necessari per arrivare al risultato (15.31). Al § 7.7.4 si è mostrato che per lo stesso segnale risulta $\mathcal{P}_s(f) = \sigma_A^2 \frac{|G(f)|^2}{T}$, e dunque

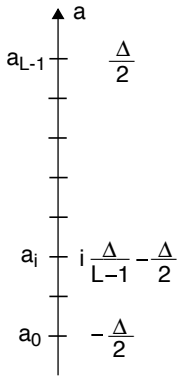
$$\mathcal{P}_s = \int \mathcal{P}_s(f)\, df = \int \sigma_A^2 \frac{|G(f)|^2}{T} df$$

Si può mostrare[100] che

$$\int |G(f)|^2 \, df = T\left(1 - \frac{\gamma}{4}\right) \tag{15.32}$$

e quindi $\mathcal{P}_s = \sigma_A^2\left(1 - \frac{\gamma}{4}\right)$; resta pertanto da calcolare

$$\sigma_A^2 = E_A\left\{(a - m_A)^2\right\} = E_A\left\{a^2\right\}$$



in virtù della media nulla. I valori che la v.a. a può assumere corrispondono a quelli dei diversi livelli del segnale dati, esprimibili come $a_i = i \cdot \frac{\Delta}{L-1} - \frac{\Delta}{2}$ (vedi figura), in modo che per $i = 0, 1, \cdots, L-1$ corrispondano ad L valori uniformemente spaziati entro l'intervallo $\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)$; inoltre, ogni possibile valore a_i ricorre con probabilità $p(a_i) = p_i = 1/L$. Possiamo dunque sviluppare i conti:

$$\begin{aligned}
\sigma_A^2 = E_A\left\{a_i^2\right\} &= \textstyle\sum_{i=0}^{L-1} p_i \cdot a_i^2 = \\
&= \tfrac{1}{L}\textstyle\sum_{i=0}^{L-1}\left(i\frac{\Delta}{L-1} - \frac{\Delta}{2}\right)^2 = \frac{\Delta^2}{L}\sum_{i=0}^{L-1}\left(\frac{i^2}{(L-1)^2} + \frac{1}{4} - \frac{i}{L-1}\right) = \\
&= \frac{\Delta^2}{L}\left(\frac{1}{(L-1)^2}\sum_{i=0}^{L-1}(i)^2 + \frac{L}{4} - \frac{1}{L-1}\sum_{i=0}^{L-1} i\right) = (^{101}) = \\
&= \frac{\Delta^2}{L}\left(\frac{L}{4} - \frac{1}{L-1}\frac{L(L-1)}{2} + \frac{1}{(L-1)^2}\frac{(L-1)L(2(L-1)+1)}{6}\right) = \Delta^2\left(\frac{1}{4} - \frac{1}{2} + \frac{2L-2+1}{6(L-1)}\right) = \\
&= \Delta^2\frac{6L-6-12L+12+8L-8+4}{24(L-1)} = \Delta^2\frac{2L+2}{24(L-1)} = \frac{\Delta^2}{12}\frac{L+1}{L-1}.
\end{aligned}$$

15.8.2 Prestazioni del ricevitore ottimo equalizzato

Mostriamo quanto affermato al § 15.5.1, ovvero che in presenza di un canale non perfetto con risposta in frequenza $H(f) \neq \mathrm{a}e^{-j2\pi f\tau}$ le prestazioni del ricevitore ottimo *equalizzato al trasmettitore*, in cui cioè

$$H_T(f) = \sqrt{G(f)}/H(f) \quad \text{e} \quad H_R(f) = \sqrt{G(f)}$$

(vedi fig. 15.18-a), subiscono una degradazione del rapporto E_b/N_0 (a parità di parametri di sistema, § 15.4.4) valutabile come una perdita di $10\log_{10}\int_{-B}^{B}\frac{G(f)}{|H(f)|^2}df$ dB.

Per semplificare l'analisi consideriamo una segnalazione binaria antipodale in cui il segnale trasmesso ha espressione

$$x(t) = \textstyle\sum_k a_k \cdot h_T(t - kT_b) \quad \text{con} \quad a_k = \{+d, -d\} \quad \textit{equiprobabili}$$

[100] La 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`

[101] Facciamo uso delle relazioni $\sum_{n=1}^{N} n = \frac{N(N+1)}{2}$ e $\sum_{n=1}^{N} n^2 = \frac{N(N+1)(2N+1)}{6}$, che sono ovviamente ancora valide anche qualora la somma parta da $n = 0$.

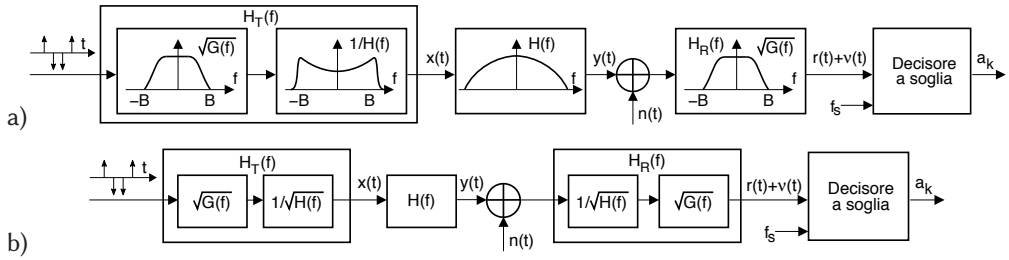


Figura 15.18: Ricevitore ottimo equalizzato: a) - al trasmettitore; b) - ripartito

ed $h_T(t) = \mathcal{F}^{-1}\left\{\sqrt{G(f)}/H(f)\right\}$ è l'impulso a radice di coseno rialzato equalizzato, mentre il segnale in ingresso al decisore è

$$r(t) + \nu(t) = \sum_k a_k \cdot g(t - kT_b) + \nu(t)$$

in cui $g(t)$ è l'impulso a coseno rialzato e $\nu(t)$ è un rumore gaussiano bianco a media nulla, filtrato attraverso $H_R(t)$, dunque con potenza

$$\mathcal{P}_\nu = \int \mathcal{P}_\nu(f)\,df = \frac{N_0}{2}\int |H_R(f)|^2 df = \frac{N_0}{2}\int G(f)\,df = \frac{N_0}{2} = \sigma_\nu^2$$

in quanto $\int G(f)\,df = 1$, vedi nota 58 ma considerando qui $G(f)\big|_{f=0} = 1/f_s = T_b$.

In tali condizioni agli istanti di campionamento il decisore osserva una v.a. gaussiana con media $\pm d$ (a seconda dell'a_k per quel simbolo) e varianza σ_ν^2, e dunque commette errore quando $\nu > d$ o $\nu < -d$ (a seconda di a_k, vedi fig. al § 7.6.1) ovvero con probabilità (vedi eq. (6.11) e § 7.6.1) $P_e = \frac{1}{2}\mathrm{erfc}\left\{\frac{d}{\sqrt{2}\sigma_\nu}\right\}$, tanto minore quanto maggiore è il rapporto $R = d^2/2\sigma_\nu^2$: esprimiamo dunque quest'ultimo in funzione della potenza trasmessa $\mathcal{P}_x = \int \mathcal{P}_x(f)\,df$. Sapendo che $\mathcal{P}_x(f) = \sigma_a^2 \cdot \frac{\mathcal{E}_{h_T}(f)}{T_b}$ (eq. (15.2)) dove[102] $\sigma_a^2 = d^2$ e $\mathcal{E}_{h_T}(f) = G(f)/|H(f)|^2$ troviamo che la potenza trasmessa vale

$$\mathcal{P}_x = \int \mathcal{P}_x(f)\,df = \frac{d^2}{T_b}\int_{-B}^{B} \frac{G(f)}{|H(f)|^2} df$$

da cui otteniamo $d^2 = \frac{\mathcal{P}_x T_b}{\int_{-B}^{B} \frac{G(f)}{|H(f)|^2} df}$, potendo così scrivere

$$R = \frac{d^2}{2\sigma_\nu^2} = \frac{\mathcal{P}_x T_b}{2\int_{-B}^{B} \frac{G(f)}{|H(f)|^2} df \;\cdot\; N_0/2} = \frac{E_b}{N_0}\frac{1}{\int_{-B}^{B} \frac{G(f)}{|H(f)|^2} df}$$

In assenza di distorsione lineare si ha[103] $|H(f)| = 1$ e dunque il denominatore vale 1[104], riottenendo il risultato noto (7.27) per un filtro adattato con segnalazione antipodale. Se invece $|H(f)|$ non è costante (ovvero $|H(f)| \leq 1$) il denominatore è più grande; pertanto la presenza del canale $H(f)$ determina la riduzione del rapporto E_b/N_0 di un fattore $\alpha = \int_{-B}^{B} \frac{G(f)}{|H(f)|^2} df$, ovvero della quantità α_{dB} (15.25) qualora il rapporto sia espresso in dB.

[102] $\sigma_a^2 = E\left\{(a_k)^2\right\} = \frac{1}{2}d^2 + \frac{1}{2}(-d)^2 = d^2$

[103] A meno di un valore costante, ininfluente ai fini della valutazione che stiamo svolgendo.

[104] Sempre in quanto $\int G(f)\,df = 1$

15.8.2.1 Equalizzazione distribuita

Qualora si realizzi invece

$$H_T(f) = H_R(f) = \sqrt{G(f)/H(f)}$$

in modo da ripartire l'equalizzazione in parti uguali ad entrambi i lati del collegamento come mostrato in fig. 15.18-b), la densità di energia dell'impulso usato in trasmissione vale $\mathcal{E}_{h_T}(f) = G(f)/|H(f)|$ e dunque il segnale trasmesso $x(t)$ ha una potenza

$$\mathcal{P}_x = \int \sigma_a^2 \cdot \frac{\mathcal{E}_{h_T}(f)}{T_b} df = \frac{d^2}{T_b} \int_{-B}^{B} \frac{G(f)}{|H(f)|} df$$

fornendo così $d^2 = \frac{\mathcal{P}_x T_b}{\int_{-B}^{B} \frac{G(f)}{|H(f)|} df}$, mentre ora la potenza del rumore filtrato attraverso $H_R(f)$ risulta

$$\mathcal{P}_\nu = \int \mathcal{P}_\nu(f)\, df = \frac{N_0}{2} \int |H_R(f)|^2 df = \frac{N_0}{2} \int_{-B}^{B} \frac{G(f)}{|H(f)|} df = \sigma_\nu^2$$

da cui

$$R = \frac{d^2}{2\sigma_\nu^2} = \frac{\mathcal{P}_x T_b}{2 \int_{-B}^{B} \frac{G(f)}{|H(f)|} df \cdot \frac{N_0}{2} \int \frac{G(f)}{|H(f)|} df} = \frac{E_b}{N_0} \frac{1}{\left| \int_{-B}^{B} \frac{G(f)}{|H(f)|} df \right|^2}$$

che qualora $|H(f)| \leq 1$, determina la riduzione del rapporto $\frac{E_b}{N_0}$ per un fattore β pari a $\left| \int_{-B}^{B} \frac{G(f)}{|H(f)|} df \right|^2$ ovvero un suo decremento in dB di $\beta_{dB} = 20 \log_{10} \int_{-B}^{B} \frac{G(f)}{|H(f)|} df$.

Mostriamo infine che $\beta \leq \alpha$: ricordando infatti la disuguaglianza di Schwartz (pag. 55)

$$\left| \int_{-\infty}^{\infty} X(f)\, Y(f)\, df \right|^2 \leq \int_{-\infty}^{\infty} |X(f)|^2 df \cdot \int_{-\infty}^{\infty} |Y(f)|^2 df$$

ed identificando $X(f)$ con $\sqrt{G(f)}$ e $Y(f)$ con $\sqrt{G(f)}/|H(f)|$ troviamo che

$$\left| \int_{-\infty}^{\infty} \sqrt{G(f)} \frac{\sqrt{G(f)}}{|H(f)|} df \right|^2 = \left| \int_{-\infty}^{\infty} \frac{G(f)}{|H(f)|} \right|^2 \leq \int_{-\infty}^{\infty} G(f)\, df \cdot \int_{-\infty}^{\infty} \frac{G(f)}{|H(f)|^2} df = \int_{-\infty}^{\infty} \frac{G(f)}{|H(f)|^2} df$$

di nuovo in quanto $\int_{-\infty}^{\infty} G(f)\, df = 1$, e quindi

$$\beta = \left| \int_{-\infty}^{\infty} \frac{G(f)}{|H(f)|} \right|^2 \leq \int_{-\infty}^{\infty} \frac{G(f)}{|H(f)|^2} df = \alpha$$

con l'uguaglianza valida qualora $\sqrt{G(f)} = k \frac{\sqrt{G(f)}}{|H(f)|}$, ovvero quando $|H(f)| = 1/k =$ costante.

15.8.2.2 Discussione

A prima vista le migliori prestazioni dell'equalizzazione distribuita rispetto a quella localizzata al trasmettitore appaiono contraddittorie, vista la correlazione dei campioni del rumore e la mancata realizzazione delle condizioni di filtro adattato nel primo caso, in quanto l'impulso che giunge al ricevitore ha densità spettrale

$$H_T(f)\, H(f) = \sqrt{G(f)\, H(f)} \neq H_R(f) = \sqrt{G(f)/H(f)}$$

In realtà quest'ultima affermazione non è vera, perché nulla vieta di considerare (vedi fig. 15.18-b) $H_R(f) = \sqrt{G(f)}$ ed inglobare il termine $1/\sqrt{H(f)}$ assieme al canale ed all'altra mezza equalizzazione, per riottenere un filtro adattato. A mio avviso le migliori prestazioni rispetto all'equalizzazione al trasmettitore sembrano essere dovute ad una minore "distorsione totale" subìta dal segnale in transito, in quanto anziché invertire *per intero* $H(f)$ per poi subirne la distorsione sempre per intero, nel caso distribuito la massima distorsione in ogni singolo passo è limitata a $\sqrt{H(f)}$ od al suo inverso. Mentre per quanto riguarda la correlazione dei campioni di rumore al decisore, questo significa solamente che è possibile fare *ancora di meglio*, vedi la nota 56 a pag. 466, così come il § 18.4.4.

15.8.3 Esercizio

Un sistema di trasmissione basato sul campionamento e sulla trasmissione numerica è rappresentato in figura 15.19. 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 $\mathcal{P}_y^{Max} = 1$ Volt^2; 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 $\mathcal{P}_N(f) = \frac{N_0}{2} = 4.61 \cdot 10^{-14}$ Volt^2/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.

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 $P_e^a \simeq 3\left(P_e^b\right)^2$.

6) Indicare la frequenza binaria f_a per le sequenze $\{a\}$ ed $\{a'\}$.

Supponiamo ora che $P_e^a = 0$, e si desideri un $SNR_Q = \mathcal{P}_x/\mathcal{P}_{z-x} = 10000$. 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_Q 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*).

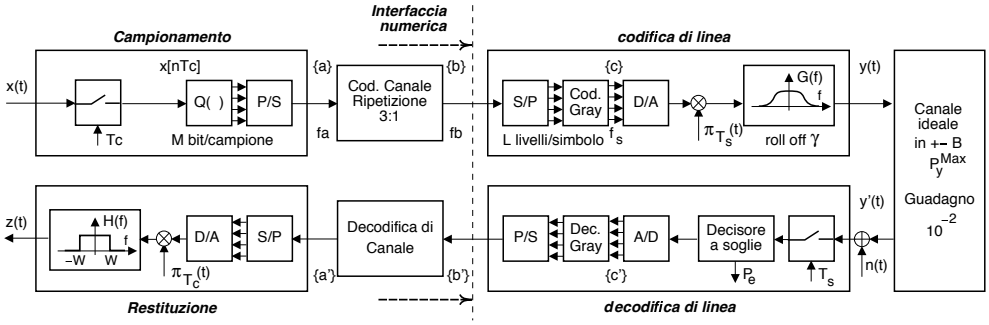


Figura 15.19: Sistema di trasmissione a cui si riferisce l'esercizio

2) Osserviamo che in questo caso la (15.17) 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'$erfc\{.\}$, questo può essere riscritto come

$$\begin{aligned} y &= \sqrt{\frac{E_b}{N_0} \frac{3\log_2 L}{(L^2-1)(1+\gamma)\left(1-\frac{\gamma}{4}\right)}} = \sqrt{\frac{\mathcal{P}_{y'}}{f_s \cdot \log_2 L} \frac{1}{N_0} \frac{3 \cdot \log_2 L}{(L^2-1)\,1.31}} \\ &= \sqrt{\frac{\mathcal{P}_{y'}}{f_s \cdot N_0} \cdot \frac{2.29}{L^2-1}} \end{aligned}$$

avendo tenuto conto che se $\gamma = 0.5$, allora $(1+\gamma)\left(1-\frac{\gamma}{4}\right) \simeq 1.31$. Inoltre, se $L \gg 1$ (come verificheremo), la (15.17) può essere approssimata come $P_e \simeq erfc\{y\}$, e dunque per $P_e = 10^{-4}$ la figura di pag. 154 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:
  - $P_e^b = \frac{P_e^c}{N_b} = \frac{10^{-4}}{10} = 10^{-5}$;
  - $P_e^b \cdot f_b = P_e^c \cdot f_s = 10^{-5} \cdot 420 \cdot 10^3 = 10^{-4} \cdot 42 \cdot 10^3 = 4.2 \ \frac{errori}{secondo}$

5) Ogni bit di $\{a'\}$ è sbagliato solo se sono sbagliati 2 o più bit in un gruppo di 3; come mostrato al § 15.6.1.1, 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^o e 2^o, 1^o e 3^o, 2^oe 3^o). In definitiva, risulta $P_e^a \simeq 3 \left(P_e^b\right)^2 = 3 \cdot 10^{-10}$.

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 $P_e^a \cdot f_a = 3 \cdot 10^{-10} \cdot 140 \cdot 10^3 = 4.2 \cdot 10^{-5} \ \frac{errori}{secondo}$.

7) Sappiamo che per un processo uniforme l'SNR 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 $SNR_q' = (L' - 1)^2 = \left(2^{2M} - 1\right)^2 = \left(2^{14} - 1\right)^2 \simeq 2.68 \cdot 10^8$, ovvero $SNR_q'\,(dB) = 84.3$ dB.

15.8.4 Codifica di carattere

Il codice ASCII (*American Standard Code for Information Interchange*) 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 15.1 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.

15.8.4.1 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.

dec	*hex*	*char*	*dec*	*hex*	*char*	*dec*	*hex*	*char*	*dec*	*hex*	*char*
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

Tabella 15.1: Tabella di codici e caratteri ASCII

Capitolo 16

Modulazione numerica

È giunto il momento di mettere assieme gli aspetti delle trasmissioni modulate (capp. 11 e 12) con quelli della trasmissione dati (cap. 15), e discutere delle tecniche necessarie a trasmettere in forma modulata un segnale di natura simbolica. Il contesto applicativo può variare su un ampio ventaglio di casi, come le forme di broadcast digitale (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 multiplazione 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 codifica di linea di *banda base* (§ 15.2), e dunque è necessario produrre un segnale modulato per trasporre la banda del segnale in accordo ai vincoli imposti dal canale. Ora non vengono però semplicemente applicate le tecniche esposte al cap. 12, ma queste sono rese specifiche alla caratteristica del segnale dati di essere costituito da sequenze di simboli appartenenti ad un alfabeto finito, da *mappare* (agli istanti di simbolo) su un insieme finito di punti nello spazio, che per segnali modulati è lo spazio dell'inviluppo complesso.

Il capitolo si sviluppa affrontando per prime le tecniche *a portante singola* basate su modulazione di ampiezza, di fase ed in quadratura, di cui si individuano l'occupazione spettrale e le prestazioni conseguibili. Dopo aver discusso delle particolarità e delle possibilità offerte dalla codifica differenziale, vengono trattate la modulazione di frequenza, a simboli ortogonali, e sviluppata la teoria della demodulazione incoerente. Si passa quindi alla analisi della tecnica OFDM, la relativa architettura di mo-demodulazione, la valutazione delle prestazioni, più una serie di aspetti particolari come temporizzazione, equalizzazione, codifica differenziale, criterio di ottimizzazione del *bit loading*, trasmissione codificata ed adozione di portanti pilota. A questo fa quindi seguito la trattazione dei sistemi a *spettro espanso*, con le sue sequenze pseudo-casuali, le problematiche di sincronizzazione, l'analisi della tecnica DSSS in presenza di rumore e di tono interfe-

rente, il suo utilizzo ai fini dell'accesso multiplo, accennando inoltre alle tecniche di *frequency* e *time hopping* o UWB. Il capitolo si chiude con una panoramica su ulteriori possibilità operative, dall'*offset keying* a MSK e CPK, risposta parziale e *trellis coding*.

16.1 Modulazione di ampiezza

In questo caso l'informazione numerica è impressa sulla portante alterando le ampiezze di una (o entrambe, come mostrato al § 16.3) delle componenti analogiche di bassa frequenza.

16.1.1 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:

$$x_{BPSK}(t) = a \sin(\omega_0 t + \varphi(t)) \quad \text{dove} \quad \varphi(t) = \sum_{k=-\infty}^{\infty} \varphi_k \text{rect}_{T_b}(t - kT_b) \tag{16.1}$$

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 (§ 12.3), è facile mostrare come possa essere realizzata mediante una comune modulazione di ampiezza BLD-PS (§ 12.1.1) con segnalazione antipodale (§ 7.6.1). Se definiamo infatti un segnale $m(t)$ come un codice di linea NRZ bipolare (§ 15.2.1), 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 a quelle espresso dalla (16.1), e la sua mo-demodulazione coerente avviene mediante l'architettura mostrata alla fig. 16.1. Il segnale uscente dal moltiplicatore di demodulazione[2] 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. La parte centrale di fig. 16.1 mostra la *forma d'onda* che corrisponde alle elaborazioni previste, mentre nella parte inferiore sono raffigurate le densità spettrali corrispondenti (espresse in dB, vedi § 8.1), tenendo conto[3] di eq. (15.2), di fig. 3.6, e della mo-demodulazione BLD-PS (§ 12.1.1.1).

Una buona caratteristica di questa tecnica è il valore *costante* dell'ampiezza della portante modulata, che permette di utilizzare la massima potenza al trasmettitore, appena inferiore al valore che inizia a produrre fenomeni di distorsione (§ 13.3). L'aspetto

[1] Letteralmente, *slittamento di tasto a due fasi.*

[2] Qui e nel seguito assumiamo di disporre di una portante di demodulazione omodina o coerente (§ 12.2.1), 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 16.11.

[3] Il segnale di banda base $m(t) = \sum_k a_k \cdot g(t - kT_b)$ in cui $g(t) = \text{rect}_{T_b}(t)$ ed i simboli a_k sono a media nulla ed indipendenti, ha una densità di potenza $\mathcal{P}_m(f) = \frac{\sigma_A^2}{T_b} \text{sinc}^2(fT_b)$, il cui andamento è mostrato in fig. 3.6 di pag. 79.

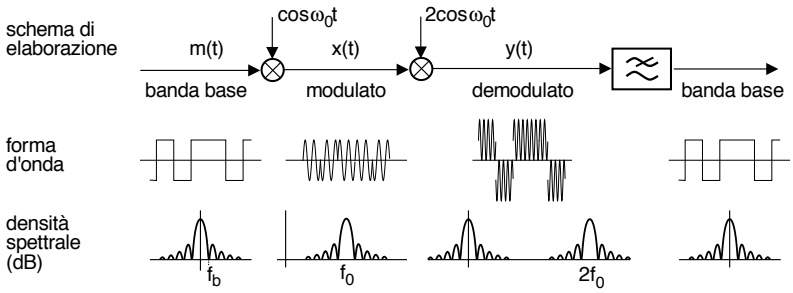
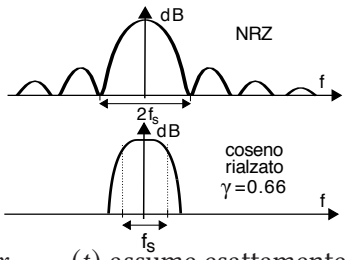


Figura 16.1: Architettura di mo-demodulazione BPSK, forma d'onda, e densità spettrale

negativo è l'elevata occupazione di banda, legata all'uso di forme d'onda rettangolari per $m(t)$ che, nel caso di trasmissione su canali con limitazioni di banda, causa vincoli sulla massima frequenza binaria. Pertanto il metodo è particolarmente indicato nel caso di collegamenti in cui è limitata la potenza di trasmissione, ma non la banda[4].

Alternative per l'impulso di banda base $g(t)$ Riprendendo i concetti discussi al § 15.2, in questo capitolo il segnale dati di *banda base* (ovvero pre-modulazione) è realizzato mediante una delle seguenti possibilità di scelta per l'impulso $g(t)$:

- NRZ o rettangolare (pag. 447), che determina una occupazione di banda multipla di $f_b = 1/T_b$;
- con trasformata a coseno rialzato (§ 15.2.2.3), con una banda pari a $\frac{f_s}{2}(1+\gamma)$;
- a banda minima (pag. 453), che riduce l'occupazione di frequenza ad $\frac{f_s}{2}$, ma presenta difficoltà realizzative.



Gli aspetti prima evidenziati per il BPSK sono *sovvertiti* qualora il segnale $m(t)$ sia basato su forme d'onda $g(t)$ con una limitata occupazione di banda, come per il coseno rialzato, con una banda *a frequenze positive* pari a $B_{BPSK} = f_b(1+\gamma)$, doppia rispetto al caso di banda base, a causa della modulazione BLD, mentre l'ampiezza del segnale modulato *non è più costante*. Infatti in corrispondenza degli istanti kT_b l'ampiezza di $x_{BPSK}(t)$ assume esattamente uno dei valori (± 1) del segnale dati $m(t)$, ma nell'intervallo tra due istanti $kT_b < t < (k+1)T_b$ l'ampiezza dipende della somma di tutte *le code* delle funzioni $g(t)$ relative ai simboli trasmessi (vedi fig. 15.8 a pag. 454).

16.1.2 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 (§ 15.1.2.4), producendo un segnale modulato di espressione

$$x_{L-ASK}(t) = m(t)\cos(2\pi f_0 t) \quad \text{dove} \quad m(t) = \sum_{k=-\infty}^{\infty} a_k \cdot \text{rect}_{T_s}(t - kT_s)$$

[4]Come ad esempio i collegamenti satellitari, vedi § 25.3.

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 zero[5]. L'ampiezza di L-ASK subisce dunque variazioni, come mostrato nella figura soprastante per un caso con $L = 8$, in cui è rappresentato anche un diagramma detto *costellazione*, che rappresenta i valori assunti dall'inviluppo 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 in questo caso, se $m(t)$ è generato mediante un impulso a coseno rialzato anziché con uno NRZ[7], la densità spettrale assume il noto andamento (vedi pag. 452) riportato ora in figura 16.2, assieme ai corrispettivi valori in dB. Pertanto la banda a frequenze positive occupata da $x_{L-ASK}(t)$ con $g(t)$ a coseno rialzato vale

$$B_{L-ASK} = f_s(1+\gamma) = \frac{f_b}{\log_2 L}(1+\gamma) \qquad (16.2)$$

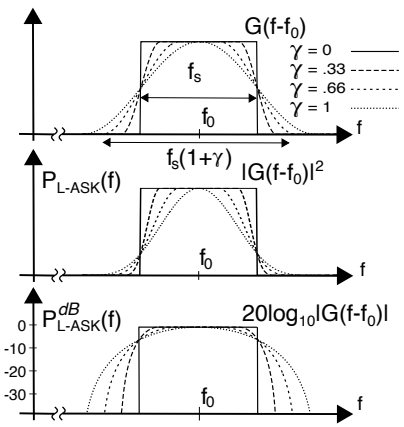


Figura 16.2: Densità spettrale di L-ASK in diverse unità di misura

Efficienza spettrale o densità di informazione E' definita come il rapporto ρ tra la frequenza binaria e la banda occupata

$$\rho = \frac{f_b}{B} \qquad (16.3)$$

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

$$\rho_{L-ASK} = \frac{f_b}{B} = \log_2 L \qquad (16.4)$$

mentre per altre forme di modulazione e/o di impulsi si ottengono altri valori[8], confrontando i quali si valuta la bontà di un metodo rispetto all'altro nei termini dell'utilizzo di banda.

[5] Per chi si sta chiedendo quanto valgono questi livelli, diciamo che il livello i-esimo (con $i = 0, 1, \cdots, L-1$) corrisponde al valore $a^i = i \cdot \frac{\Delta}{L-1} - 1$. Verificare per esercizio con $\Delta = 2$ ed $L = 4$.

[6] Ad 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.

[7] Notiamo che mentre per il BPSK scegliere il primo al posto del secondo comporta perdere i benefici di una ampiezza costante, nel caso multilivello l'ampiezza è intrinsecamente variabile.

[8] Vedi tabella 16.1 a pag. 519.

Esempio Se compariamo il risultato per ρ_{L-ASK} con quello relativo ad una trasmissione numerica di banda base (vedi eq. (15.5)), notiamo un *peggioramento* di un fattore 2, dovuto all'uso di una AM-BLD.

Come per il caso analogico, la banda potrebbe essere *dimezzata* adottando una AM-BLU, ma troveremo invece tra breve che si preferisce seguire approcci diversi, come ad esempio PSK e QAM.

16.1.3 Valutazione delle prestazioni

Dopo alcune considerazioni relative al legame tra SNR ed occupazione di banda, la valutazione della P_e fa tesoro di quanto ottenuto al § 15.4.

SNR, E_b/N_0 ed efficienza spettrale Nell'analisi delle prestazioni che affronteremo la probabilità di errore per simbolo P_e (*simbolo*) o per bit P_e (*bit*) è espressa in funzione della grandezza $\frac{E_b}{N_0}$ introdotta a pag. 459, e che rappresenta l'equivalente del rapporto segnale rumore *di sistema* $SNR_0 = \frac{\mathcal{P}_x}{N_0 W}$ definito al § 14.2.1.1[9], nel senso che come questo consente il confronto tra tecniche diverse[10]. D'altra parte, una trasmissione AM-BLD numerica che occupi una banda a frequenze positive B si presenta in ingresso al decisore con un

$$SNR = \frac{\mathcal{P}_x}{\mathcal{P}_n} = \frac{E_b f_b}{2B \cdot N_0/2} = \frac{E_b}{N_0}\frac{f_b}{B} = \rho\frac{E_b}{N_0} \tag{16.5}$$

in cui $\mathcal{P}_n$ è limitata da un filtro di ricezione, e ρ è l'efficienza spettrale definita alla (16.3): pertanto E_b/N_0 è anche indicato come SNR *normalizzato* o SNR *per bit*. Nel caso di $g(t)$ a banda minima (§ 15.2.2.3) la (16.4) fornisce $\rho_{L-ASK} = \log_2 L$ e dunque $SNR = \log_2 L \frac{E_b}{N_0}$, mentre a pag. 459, eq. (15.16), si deriva la relazione tra E_b/N_0 e SNR per il caso particolare di un segnale dati *a coseno rialzato*.

Invertendo la (16.5) si ottiene $E_b/N_0 = SNR/\rho$ evidenziando come, a parità di SNR, al miglioramento dell'efficienza spettrale ρ corrisponda una diminuzione di $\frac{E_b}{N_0}$, che a sua volta è causa di un peggioramento della probabilità di errore, in accordo con il *compromesso banda-potenza*, vedi § 15.4.7.

Probabilità di errore BPSK e L-ASK La P_e viene calcolata per un segnale L-ASK in funzione di E_b/N_0, al variare del numero di livelli, ottenendo il caso BPSK per $L = 2$.

Al § 16.1.2 abbiamo osservato come l'L-ASK sia ottenibile mediante una modulazione AM-BLD di un segnale dati di banda base (vedi anche fig. 16.3), e come discusso al § 14.2.1.1, l'SNR in uscita dal filtro di ricezione (e dunque l'E_b/N_0, vedi eq. (16.5)) per una modulazione AM-BLD è pari al rapporto SNR_0 tra potenza ricevuta e potenza di rumore nella banda del segnale *modulante*. Pertanto l'SNR (e l'E_b/N_0) dopo demodulazione di L-ASK è pari a quello che si avrebbe per il segnale dati di banda base da cui ha origine. Le prestazioni per un segnale dati *di banda base* a coseno rialzato sono ricavate al

[9] Ricordiamo che $\mathcal{P}_x$ 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.

[10] Infatti come discusso a pag. 459 $E_b = \frac{\mathcal{P}_x}{f_b}$, come N_0, dipende solamente da *parametri di sistema* ($\mathcal{P}_x$ e f_b), mentre invece non dipende dai *parametri di trasmissione* L e γ e dal tipo di modulazione.

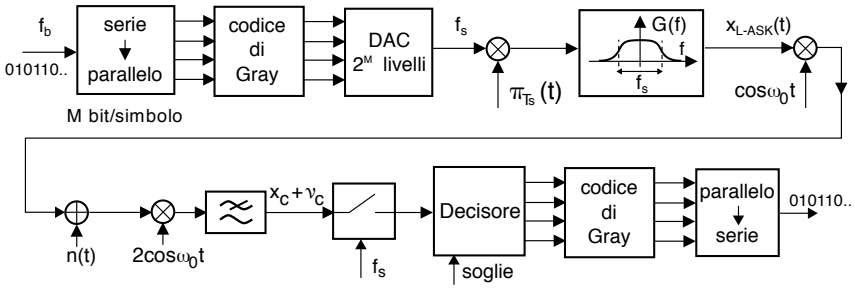


Figura 16.3: Schema di mo-demodulazione per un segnale L-ASK

§ 15.4.9, che riportiamo sotto come probabilità di errore *per simbolo* dell'L-ASK[11]

$$P_e^{L-ASK}(simbolo) = \left(1 - \frac{1}{L}\right) \text{erfc}\left\{\sqrt{3\frac{E_b}{N_0}\frac{log_2 L}{(L^2-1)}}\right\} \tag{16.6}$$

valida per un segnale con $\gamma = 0$, ossia a *banda minima*[12]. Le curve di $P_e^{L-ASK}(bit)$ in funzione di ${}^{E_b}/_{N_0}|_{dB}$ sono quelle di fig. 15.12 a pag. 465, dove si tiene anche conto dell'uso di un codice di Gray (§ 15.4.9.1) per associare i livelli a configurazioni binarie.

Come anticipato, per $L = 2$ la (16.6) esprime le prestazioni del BPSK, ovvero

$$P_e^{BPSK}(bit) = \frac{1}{2}\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\} \tag{16.7}$$

che come prima si riferisce al caso di banda minima, ed i cui valori sono graficati in fig. 16.4, identica alla (15.18) ottenuta per il caso di banda base, ed alla (7.27) relativa al filtro adattato. Per completare i confronti osserviamo che ora all'aumentare di L la banda (16.2) (per $\gamma = 0$)

$$B_{L-ASK} = f_s = \frac{f_b}{\log_2 L}$$

si riduce, mentre la P_e (16.6) aumenta: ciò può tor-

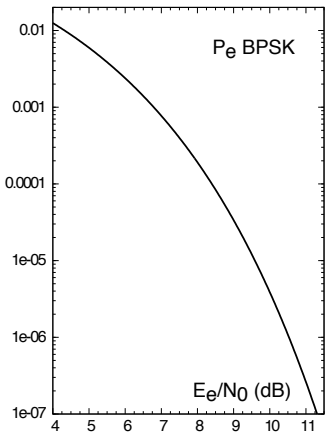
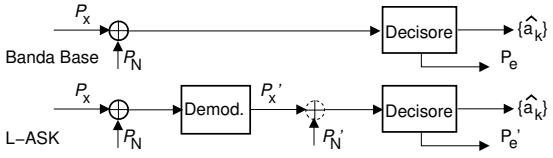


Figura 16.4: Prestazioni BPSK

[11] Forniamo qui una contro-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 § 14.2.1) $P'_x = P_{x_c} = k_a^2 P_M = 2P_x$ 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 $P'_N = P_{n_c} = \sigma^2_{n_c} = \sigma^2_n = \frac{N_0}{2}4W$ 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.

[12] Se $\gamma \neq 0$, valgono le considerazioni svolte al § 15.4.9.

nare 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 § 16.5.1 vedremo come nella tecnica di FSK ortogonale lo stesso compromesso operi in direzione opposta, ovvero riuscendo a migliorare P_e al prezzo di aumentare l'occupazione di banda.

16.2 Modulazione di fase

Nel caso in cui l'informazione viene rappresentata dalla fase della portante ci si può riferire a simboli quaternari, o multilivello.

16.2.1 Modulazione QPSK ed L-PSK

Questi due acronimi si riferiscono alla possibilità di usare rispettivamente *quattro* oppure $L > 4$ scelte diverse[13] 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

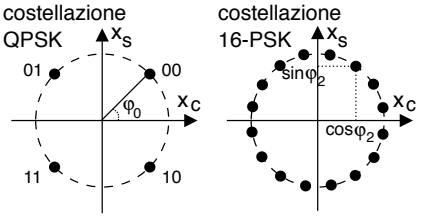
$$\underline{x}_{L-PSK}(t) = ae^{j\varphi(t)} = a\cos\varphi(t) + ja\sin\varphi(t) \tag{16.8}$$

in cui

$$\varphi(t) = \sum_{k=-\infty}^{\infty} \varphi_k \mathrm{rect}_{T_s}(t - kT_s) \quad \text{e} \quad \varphi_k \in \{\varphi_0, \varphi_1, \ldots, \varphi_{L-1}\} \tag{16.9}$$

La generica fase

$$\varphi_i = \frac{\pi}{L} + i \cdot \frac{2\pi}{L} \quad \text{con} \quad i = 0, 1, \cdots L-1$$



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'inviluppo complesso ricevuto in assenza di rumore negli istanti di simbolo $t = kT_s$ ed a cui si fanno corrispondere gruppi di bit in accordo alla codifica di Gray (§ 15.4.9.1). Lo stesso valore di fase è quindi mantenuto per tutto il periodo di simbolo se $\varphi(t)$ è realizzata mediante rettangoli come indicato nella (16.9). Ma l'espressione (16.8) 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 quadratura[14], suggerendo l'implementazione del modulatore secondo lo schema di fig. 16.5, 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 impiegati 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 $x_{L-PSK}(t)$, la cui densità di potenza

[13] Il caso in cui $L = 2$ ricade nel BPSK già discusso

[14] che *non è* una modulazione AM-BLU dato che $x_s \neq \widehat{x}_c$

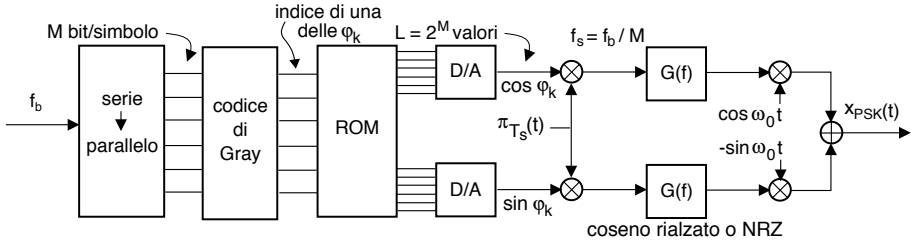
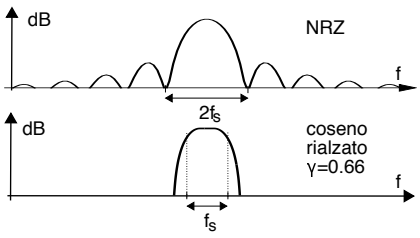


Figura 16.5: Modulatore L-PSK

in tal caso acquisisce un andamento $\left(\frac{\sin x}{x}\right)^2$ centrato in f_0 e con lobo principale di estensione[15] pari ad $2f_s = {}^{2f_b}/_{M}$, come rappresentato in figura per una densità di potenza in dB. L'occupazione di banda può essere limitata a

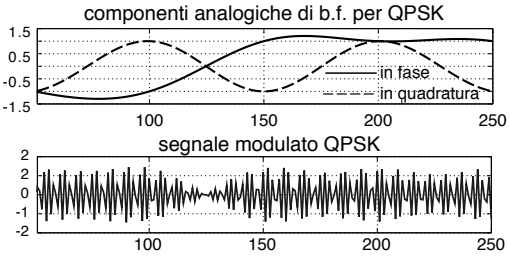


$$B_{L-PSK} = f_s\,(1+\gamma)$$

se si realizza $\varphi\,(t)$ mediante impulsi $g\,(t)$ a coseno rialzato, potendo così scrivere

$$\underline{x}_{L-PSK}\,(t) = a \sum_{k=-\infty}^{\infty} e^{j\varphi_k} \cdot g\,(t - kT_s) \tag{16.10}$$

Dinamica delle ampiezze Adottando un impulso $g\,(t)$ a coseno rialzato anziché rettangolare, $\underline{x}\,(t)$ passa dai punti della costellazione *solo* negli istanti di simbolo, mentre nell'intervallo tra due di essi segue traiettorie di ampiezza variabile[16], come illustrato nella figura a fianco, non permettendo al segnale modulato di mantenere un'ampiezza costante come dovrebbe avere la modulazione angolare[17]. Pertanto la scelta tra NRZ o coseno rialzato dipende dalla necessità di limitare la dinamica delle ampiezze, piuttosto che l'estensione della banda.



Traiettoria dell'inviluppo complesso Altrettanto interessante può essere riflettere sulla fig. 16.6 in cui si mostrano x_c ed x_s ancora 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 e polari; al centro è mostrato la corrispondente evoluzione per $\underline{x}\,(t)$, mentre a destra si mostra la traiettoria di $\underline{x}\,(t)$ visualizzata per 100 simboli.

[15] Infatti un rettangolo di durata T_s ha trasformata $sinc\,(T_s f)$, con il primo zero in $f = {}^1/_{T_s} = f_s$, e la modulazione AM-BLD produce un raddoppio della banda occupata.

[16] Se viceversa $g\,(t) = \mathrm{rect}_{T_s}\,(t)$, $|\underline{x}|$ giace su di un cerchio, spostandosi *istantaneamente* da un punto all'altro della costellazione

[17] Nella parte superiore della figura è mostrato l'andamento delle c.a. di b.f. per 5 simboli di un QPSK realizzato adottando $g\,(t)$ a coseno rialzato con $\gamma = 0.5$, e si può notare che ognuna di esse assume il valore ± 1 in corrispondenza di ogni periodo di simbolo. Nella parte inferiore è riportato il corrispondente segnale modulato, che come si vede non è affatto ad ampiezza costante.

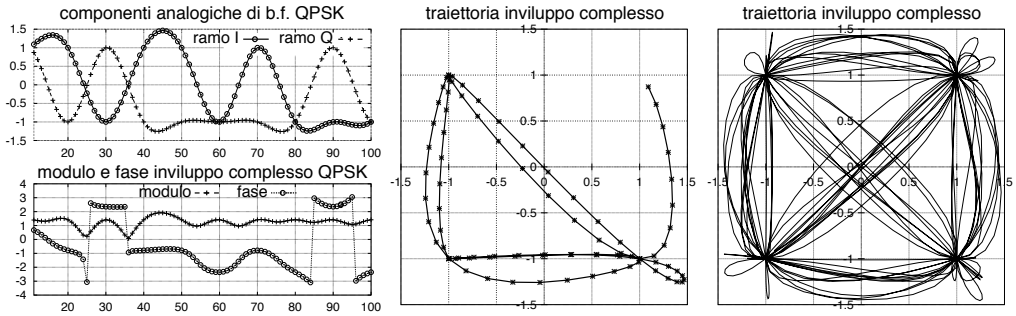


Figura 16.6: Andamento delle c.a. di b.f. x_c ed x_s per dieci simboli di segnale QPSK, in coordinate cartesiane e polari (a sn.), traiettoria dell'inviluppo complesso $\underline{x}(t)$ per 10 simboli (al centro) e per 100 simboli (a ds.)

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.

16.2.2 Prestazioni QPSK

In questo caso (PSK con 4 livelli) il demodulatore coerente è costituito da due rami indipendenti in fase e quadratura (vedi fig. 16.8), che operano ad una frequenza di simbolo f_s pari a metà di quella binaria. In virtù di come la codifica di Gray dispone la costellazione (vedi fig. 16.7), ogni ramo del decisore opera su di un solo bit della coppia associata al simbolo trasmesso; le due decisioni vengono poi riserializzate. Entrambi i rami si comportano quindi come un demodulatore L-ASK (§ 16.1.2) con L=2, ovvero un BPSK, e per un segnale dati a coseno rialzato *a banda minima* ($\gamma = 0$) la probabilità di errore relativa ad ogni

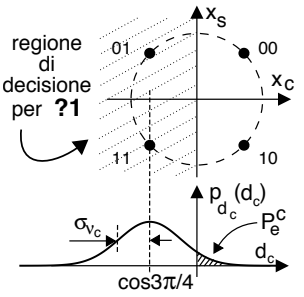


Figura 16.7: Costellazione QPSK e calcolo della P_e^c

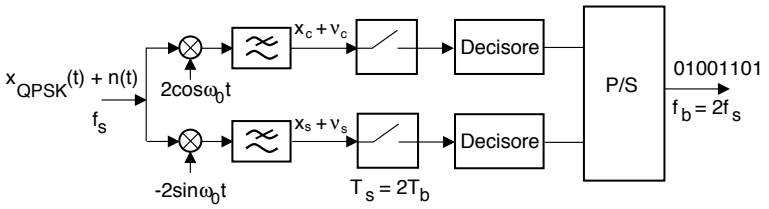


Figura 16.8: Demodulatore QPSK

singolo ramo è espressa[18] dalla (16.7), fornendo

$$P_e^c = P_e^s = \frac{1}{2}\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$$

che rappresenta la probabilità che una componente analogica di b.f. ottenuta demodulando il segnale con sovrapposto rumore, valutata all'istante di decisione kT_s, giaccia nell'area mostrata in basso[19] della figura 16.7. La probabilità di errore (per impulso a banda minima) *in un bit* della sequenza re-serializzata risulta quindi[20]

$$P_e^{QPSK}(bit) = P_e^c \cdot Pr\{c\} + P_e^s \cdot Pr\{s\} = \frac{1}{2}\left(P_e^c + P_e^s\right) = \frac{1}{2}\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\} \qquad (16.11)$$

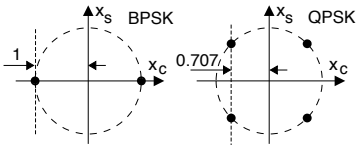
in cui $Pr\{c\} = Pr\{s\} = 1/2$ sono le probabilità che il bit ricevuto provenga dal ramo in fase o da quello in quadratura; d'altra parte, la probabilità di errore *per simbolo* risulta $P_e(simbolo) \simeq P_e^c + P_e^s = \text{erfc}\left\{\sqrt{E_b/N_0}\right\}$ (trascurando nuovamente la probabilità di un errore contemporaneo su entrambi i rami, vedi nota 20).

Osserviamo quindi come il QPSK consenta di ottenere *le stesse prestazioni* del BPSK, graficate in fig. 16.4, utilizzando solo *metà banda*, in virtù del T_s doppio*:*

$$B_{QPSK} = f_s = \frac{f_b}{2}(1+\gamma)$$

Quest'ultima osservazione permette di scrivere $\rho_{QPSK} = \frac{f_b}{B_{QPSK}} = 2 = 2 \cdot \rho_{BPSK}$, e suggerisce un ulteriore[21] punto di vista rispetto all'invarianza delle prestazioni rispetto al BPSK: 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.

[18] In effetti, dovremmo verificare che l'attuale valore di E_b/N_0 sia lo stesso del caso BPSK: mentre per N_0 al § 14.1.3 si mostra che è vero, per quanto riguarda E_b (a prima vista) sembra che il suo valore 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 le fasi formano un angolo di 45^o rispetto agli assi, moltiplicando per $\cos\frac{\pi}{4} = \sqrt{2}/2$ le coordinate cartesiane, e riducendo dunque la potenza delle c.a. di b.f. di un fattore 2, e così pure il valore di E_b. In realtà, la durata doppia del periodo di simbolo $T_s = 2T_b$ compensa questa riduzione, e dunque anche l'E_b' di ogni ramo $E_b' = P_x T_s$ si mantiene invariato.



[19] All'istante di decisione k su ciascun ramo si osserva una v.a. gaussiana $d_{c,s}$ con varianza $\sigma^2_{\nu_{c,s}}$ (vedi fig. 14.2 a pag. 415) e valor medio $x_c(kT_s) = \cos\varphi_k$ e $x_s(kT_s) = \sin\varphi_k$, dove φ_k è la fase del punto di costellazione trasmesso all'istante $t = kT_s$, vedi eq. (16.10). Nell'esempio di figura la decisione per il secondo dei due bit del simbolo cambia in funzione del segno di d_c, e dunque si commette errore sul ramo in fase se ad es. si trasmette `x1,` ma il rumore su quel ramo ha un valore sufficientemente positivo da far superare lo zero in corrispondenza dell'istante di decisione.

[20] In 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 tanto più trascurabile rispetto a P_e^c quanto più quest'ultimo è piccolo.

[21] Ulteriore rispetto al commento alla nota 18, dove il ragionamento è svolto sull'E_b/N_0, mentre ora sull'SNR.

Ottimalità di BPSK e QPSK Se queste tecniche sono attuate mediante un $G(f) = rect_{f_s/2}(f)$ ovvero a banda minima, in presenza di demodulazione *coerente* il filtro passabasso che precede il campionatore del decisore costituisce *un filtro adattato* (vedi § 7.6.1) con segnalazione *antipodale*, e le prestazioni sono quelle ottime, come si verifica confrontando l'eq. (7.27) con le (16.7) e (16.11). Se viceversa $G(f)$ *non è* a banda minima, ma di Nyquist, la coerenza di demodulazione permette comunque di adottare in ricezione un filtro adattato (§ 15.5) al posto del passabasso mostrato in fig. 16.3, e dunque ottenere le medesime prestazioni.

16.2.3 Prestazioni L-PSK

In questo caso il demodulatore ha una differente architettura (vedi fig. 16.9), 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 $d_{c,s}^k$ le c.a. di b.f. demodulate all'istante $t = kT_s$, la decisione avviene calcolando $\varphi_k^d = \arctan \frac{d_s^k}{d_c^k}$ e stabilendo all'interno di quale regione di decisione $\widehat{\varphi}_k$ cada la fase ricevuta φ_k^d, a cui il decisore fa corrispondere una codeword di Gray (fig. 16.10). 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 $\log_2 L$, mentre la spaziatura tra le regioni di decisione diminuisce con legge lineare rispetto ad L. Pertanto, 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

$$P_e^{L-PSK}(simbolo) = \operatorname{erfc}\left\{\sin\left(\frac{\pi}{L}\right)\sqrt{\frac{E_b}{N_0}\log_2 L}\right\} \tag{16.12}$$

che rappresenta la probabilità di decidere di aver ricevuto un $\widehat{\varphi}_k \neq \varphi_k$ (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. 16.9).

Confrontando il risultato con quello (eq. 16.6) per l'ASK, osserviamo che l'assenza del termine $\left(1 - \frac{1}{L}\right)$ è dovuto alla *circolarità* della costellazione, che il termine $\sin\left(\frac{\pi}{L}\right)$ 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 (16.12) è una approssimazione valida se $P_e \ll 1$, e via via più accurata con L crescente.

Nella tabella che segue si riporta il risultato del confronto, per uno stesso valore

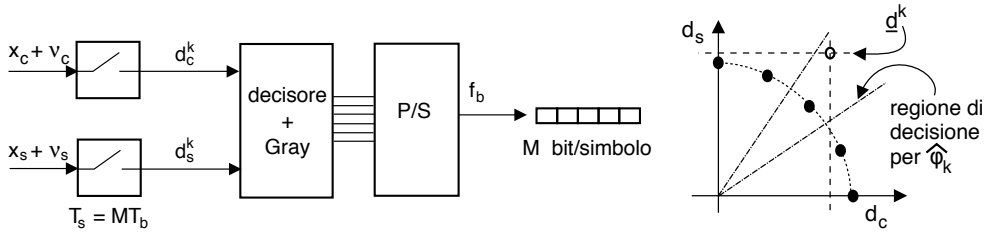


Figura 16.9: Demodulatore L-PSK

di P_e, dei valori $\frac{E_b}{N_0}$ necessari per L-PSK (16.12), contro quelli necessari (16.6) per L-ASK: si è eseguito il rapporto tra gli argomenti degli erfc {}, si è elevato al quadrato, indicato come Δ, ed il risultato è espresso in dB: quest'ultimo rappresenta (a parte il termine $\left(1 - \frac{1}{L}\right)$ 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.

L	$\Delta_{E_b/N_0} = \frac{1}{3}\left(L^2 - 1\right)\sin^2\frac{\pi}{L}$	$\Delta_{E_b/N_0}(dB)$
4 (QPSK)	2.5	4
8	3.07	4.88
16	3.23	5.1
32	3.28	5.2
64	3.29	5.2

E' opportuno osservare che, qualora si desideri ottenere un valore di probabilità di errore *per bit*, questo è pari a

$$P_e\,(bit) = \frac{P_e\,(simbolo)}{\log_2 L}$$

a patto di associare, a livelli contigui, gruppi di bit differenti in una sola posizione, come previsto dal codice di Gray[22] (mostrato nella figura 16.10), in modo che l'errore tra due fasi φ_k contigue 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. 16.10, sono calcolate in questo modo.

16.3 Modulazione QAM

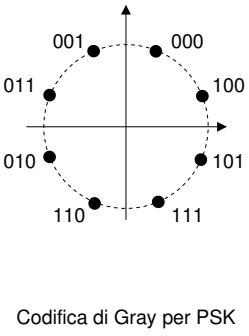
Questo acronimo sta per *Quadrature Amplitude Modulation*, ed individua la tecnica di modulazione che utilizza due portanti in quadratura come il PSK

$$x_{QAM}\,(t) = x_c\,(t)\cos\omega_0 t - x_s\,(t)\sin\omega_0 t$$

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. 16.11 , 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

[22] vedi il § 15.4.9.1



Codifica di Gray per PSK

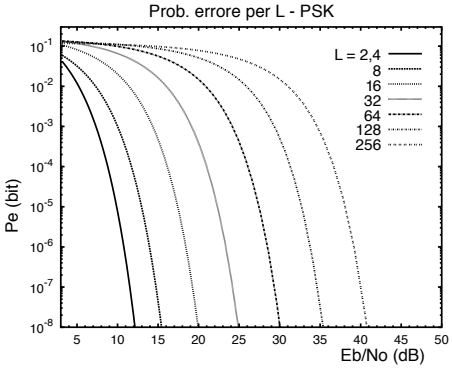


Figura 16.10: Prestazioni L-PSK con codice di Gray

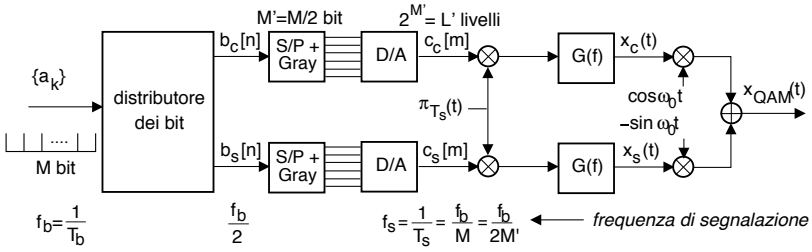
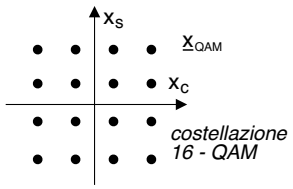


Figura 16.11: Architettura di un modulatore QAM

alternativamente sui due rami (sequenze $b_c[n]$ e $b_s[n]$ in figura) a velocità dimezzata[23], 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 poi si ottengono (mediante codifica di Gray) i valori c_c e c_s con $L' = 2^{M'} = 2^{M/2} = \sqrt{2^M} = \sqrt{L}$ livelli, che attraversando in forma impulsiva 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[24], che rappresentano le coordinate (nel piano dell'inviluppo 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

$$T_s = T_b \cdot M = \frac{1}{f_b} \log_2 L$$

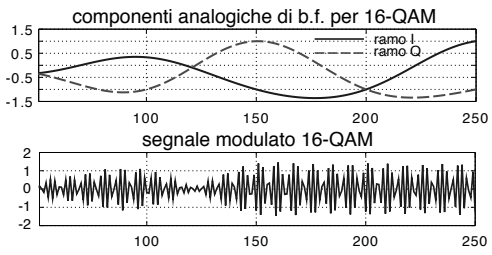


Esempio il 16-QAM si ottiene con $m = 4$ bit/simbolo ($L = 16 = 2^M = 2^4$) e sui due rami sono presenti $L' = \sqrt{16} = 4$ livelli, ottenendo il risultato di una costellazione quadrata di $L = 4\text{x}4 = 16$ punti.

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

$$B_{QAM} = \frac{f_b}{\log_2 L}(1+\gamma) = f_s(1+\gamma)$$

e quindi uguale a quella di ASK e PSK con uguale numero di livelli (di cui condivide quindi anche l'efficienza spettrale).



Dinamica delle ampiezze Nella parte superiore della figura a lato è mostrato l'andamento di x_c ed x_s per 5 simboli di

[23] In pratica, l'indice n si incrementa ogni due incrementi dell'indice k.

[24] Per come si è impostata la distribuzione dei bit tra i rami L deve risultare un quadrato perfetto. Sebbene sia possibile realizzare anche costellazioni di forma *non quadrata*, vedi ad es. AA.VV., *A Survey on Design and Performance of Higher-Order QAM Constellations* presso `https://arxiv.org/pdf/2004.14708.pdf`, la soluzione quadrata è preferita per semplicità realizzativa.

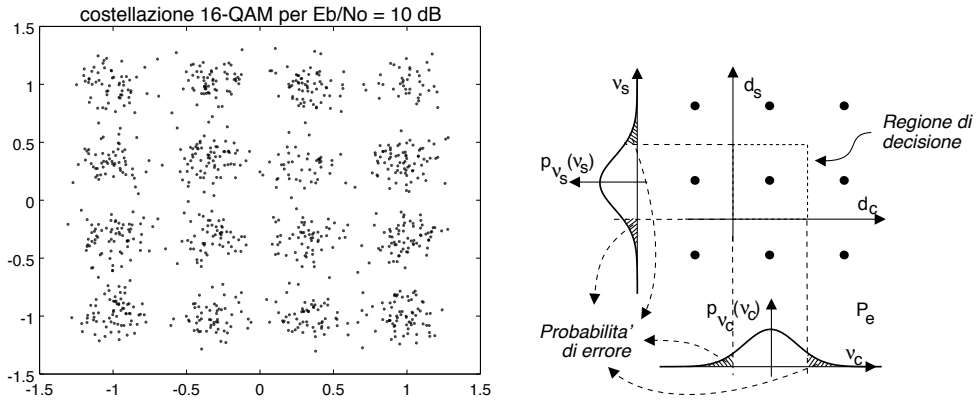


Figura 16.12: Costellazione 16-QAM in presenza di rumore (a sn), calcolo della P_e sui rami (a ds)

un 16-QAM realizzato a partire da un segnale dati con $\gamma = 0.5$, e si può notare che in corrispondenza di ogni periodo di simbolo entrambe assumono uno tra i valori $-1, -0.33, 0.33, 1$. Nella parte inferiore è riportato il corrispondente segnale modulato, che come si vede non è affatto ad ampiezza costante.

16.3.1 Prestazioni di QAM

Nella parte sinistra di fig. 16.12 è mostrata la costellazione per un 16-QAM in presenza di rumore ($E_b/N_0 = 10$ dB). La *distanza* tra due punti di costellazione è maggiore (a parità di L) del caso PSK, e 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 segnale QAM viene demodulato secondo lo schema di fig. 16.13, che ci permette di constatare come su ciascuno dei due rami in fase e quadratura transita un segnale ASK multilivello con $L' = \sqrt{L}$, e dunque possiamo adottare[25] l'espressione (16.6) (relativa ad un impulso $g(t)$ a banda minima) per definire la probabilità di errore su ciascuno dei due rami, pari a

$$P_\alpha \quad = \quad P_e^c\,(simbolo) = P_e^s\,(simbolo) = \left(1 - \frac{1}{L'}\right) \operatorname{erfc}\left\{\sqrt{3\frac{E_b}{N_0}\frac{\log_2 L'}{(L')^2 - 1}}\right\}$$

Ricordando ora che $L' = \sqrt{L} = (L)^{1/2}$, e dunque $\log_2 L' = \frac{1}{2}\log_2 L$, si ottiene

[25] Per applicare la (16.6) dobbiamo verificare se il valore di E_b/N_0 è lo stesso nei due casi (vedi nota 18). Mentre per N_0 non vi sono dubbi, notiamo (vedi § 12.4.5 per il caso di c.a. di b.f. incorrelate) che la potenza ricevuta $\mathcal{P}_x$ si dimezza su entrambi i rami, così come la f_b, e pertanto si ottiene $E_b' = \frac{\mathcal{P}_x/2}{f_b/2} = \frac{\mathcal{P}_x}{f_b} = E_b$.

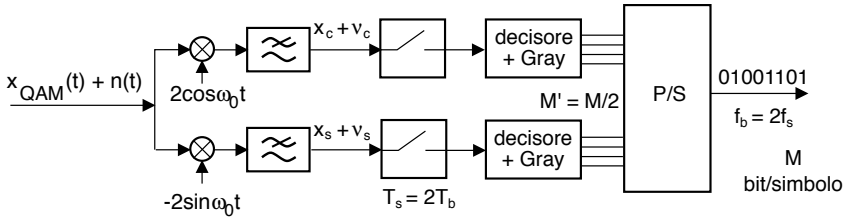


Figura 16.13: Demodulatore QAM

$$P_\alpha = \left(1 - \frac{1}{\sqrt{L}}\right) \operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{N_0}\frac{\log_2 L}{L-1}}\right\} \tag{16.13}$$

La probabilità di errore (a simbolo) complessiva, cioè la probabilità che il segnale ricevuto $\underline{d} = \underline{x} + \underline{\nu}$ cada fuori della regione di decisione relativa all'$\underline{x}$ trasmesso (vedi parte destra di fig. 16.12), risulta $P_e\,(simbolo) \simeq P_\alpha + P_\alpha = 2P_\alpha$, 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 ai livelli dei due rami, consente di esprimere la probabilità di errore per bit come

$$P_e^{QAM}\,(bit) \quad = \quad \frac{P_e\,(simbolo)}{\log_2 L} = \frac{2}{\log_2 L}\left(1 - \frac{1}{\sqrt{L}}\right) \operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{N_0}\frac{\log_2 L}{L-1}}\right\} \tag{16.14}$$

In figura 16.14 troviamo le curve dei valori di $P_e\,(bit)$, 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 riceva 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 desideri 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?

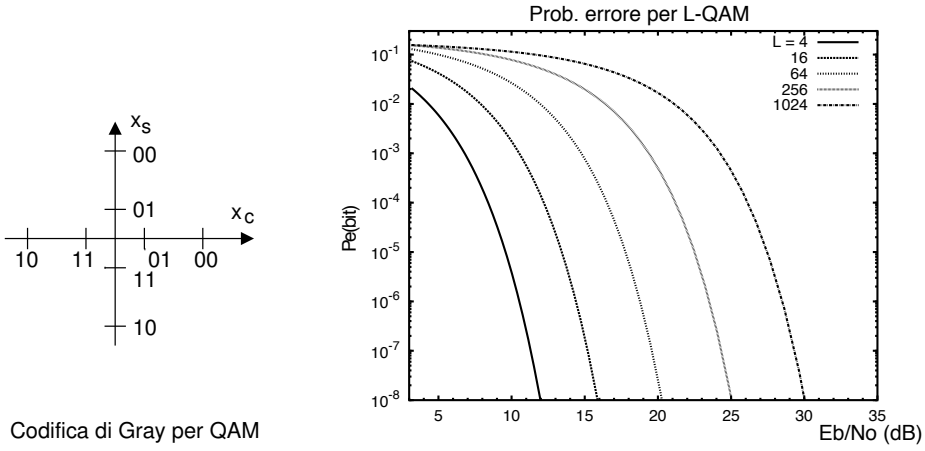


Figura 16.14: Prestazioni L-QAM con codice di Gray

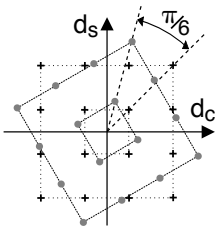
Soluzione

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 utilizzare $M = 8$ bit/simbolo, ovvero raddoppiare $\log_2 L$, 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 $f'_b = \frac{1}{T'_b} = \frac{10^{0.35}}{T_b} = 10^{0.35} \cdot f_b = 10^{6.35} \simeq 2.24$ Mb/sec.

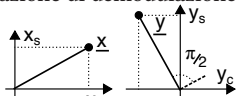
   E se $\gamma \neq 0$? La trattazione del caso di banda base (pag. 464), mostra che l'argomento sotto radice della erfc {} subisce un peggioramento di $(1+\gamma)\left(1-\frac{\gamma}{4}\right)$, che (per esempio) con $\gamma = 0.5$ fornisce 1.31, che deve essere compensato da una uguale diminuzione di 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.

16.4 Codifica differenziale

Se la portante con cui si effettua la demodulazione presenta un errore di fase[26], il piano dell'inviluppo complesso subisce una rotazione, producendo decisioni sistematicamente errate a causa dello spostamento dei punti di costellazione ottenuti campionando le c.a. di b.f. (pallini rossi) rispetto a quelli che si otterrebbero nel caso di demodulazione *coerente* (crocette nere). Per rimediare al problema, si può estendere il principio della codifica *differenziale* (espresso a pag. 448 per segnali di banda base) al caso delle modulazioni numeriche, rendendo la decisione su quale punto di costellazione sia stato ricevuto indipendente dalla



[26] Il cui effetto su $x_c(t)$ è stato discusso al § 12.2.3.1. Facciamo ricadere in questa casistica l'ambiguità di fase dei sistemi di recupero portante come descritto al § 12.2.2.1, ma anche la distorsione di fase introdotta dal canale non selettivo in frequenza, vedi § 13.1.2.4. Che un errore di fase si traduca in una rotazione dell'inviluppo complesso può essere mostrato considerando che l'operazione di demodulazione omodina corrisponde a valutare $\mathcal{R}\left\{\underline{x}e^{j\omega_0 t}\right\}$, mentre una portante $\cos(\omega_0 t + \varphi)$ corrisponde a $\mathcal{R}\left\{\underline{x}e^{j\omega_0 t}e^{j\varphi}\right\}$, equivalente alla demodulazione coerente di $\underline{y} = \underline{x}e^{j\varphi}$ ossia un inviluppo complesso *ruotato*. In figura, un caso per cui $\varphi = \frac{\pi}{2}$.



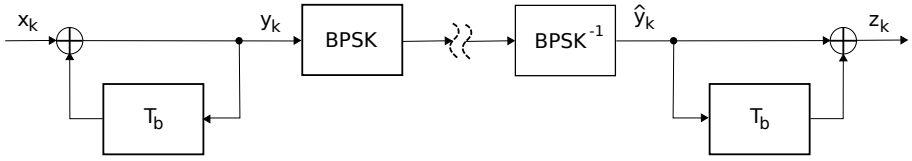


Figura 16.15: Codifica e detezione differenziali

fase della portante di demodulazione, ma dipendente invece dalla fase dell'inviluppo complesso osservata *per il simbolo precedente*. Ciò si realizza modificando il criterio con cui sono determinati i punti di costellazione da trasmettere, scegliendoli ora in funzione *di una coppia* di simboli consecutivi, anziché di uno solo.

16.4.1 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. 16.15, simile[27] a quella a pag. 264, mostra la sequenza delle operazioni necessarie, e che consistono nel trasformare il messaggio binario x_k in quello y_k in base alla relazione

$$y_k = x_k \oplus y_{k-1} \tag{16.15}$$

(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

$$\begin{aligned} x_k &= 001011011010010 \\ y_k &= 001101101100011 \\ z_k &= /01011011010010 \end{aligned}$$

$$z_k = \hat{y}_k \oplus \hat{y}_{k-1} \tag{16.16}$$

che, in assenza di errori (ossia se $\hat{y}_k = y_k$ per tutti i k), permette di ottenere nuovamente i valori del messaggio originario[28] x_k a partire dalla sequenza z_k, come la figura del riquadro precedente consente di verificare tramite un esempio.

Se assumiamo ora di rappresentare lo zero con una fase nulla, e l'uno con una fase di π, possiamo riscrivere la (16.1) come

$$x_{BPSK}(t) = a\cos(\omega_0 t + \varphi(t)) \quad \text{con} \quad \varphi(t) = \pi \cdot \sum_{k=-\infty}^{\infty} x_k \text{rect}_{T_b}(t - kT_b) \tag{16.17}$$

e ponendo per semplicità il periodo di bit pari ad un ciclo di portante, possiamo confrontare in fig. 16.16 la forma d'onda BPSK associata alla sequenza originaria x_k, con quella ottenibile utilizzando nella (16.17) y_k anziché x_k ed indicata come DBPSK, in cui la D sta appunto per *differenziale*. Pertanto quando il demodulatore BPSK di fig. 16.15 riceve il segnale DBPSK, in assenza di errori si produce in uscita la sequenza y_k, e quindi il circuito mostrato realizza l'operazione $z_k = \hat{y}_k \oplus \hat{y}_{k-1}$, permettendo

[27] La similitudine non è per nulla casuale, dato che qualora il predittore mostrato a pag. 264 sia realizzato mediante un elemento di ritardo, i due schemi di elaborazione coincidono.

[28] A 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, sostituendo la (16.15) nella (16.16) e in assenza di errori (ossia $\hat{y}_k = y_k$) si ottiene $\hat{z}_k = \hat{y}_k \oplus \hat{y}_{k-1} = x_k \oplus y_{k-1} \oplus y_{k-1} = x_k$.

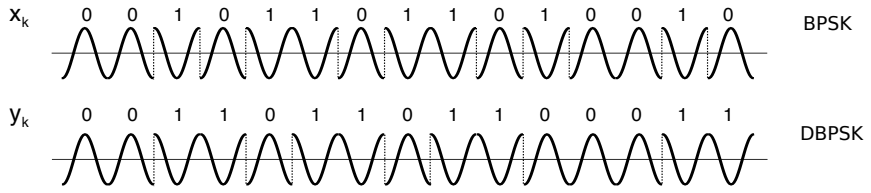


Figura 16.16: Segnale BPSK e BPSK *differenziale*

di ottenere la sequenza originaria. Verifichiamo ora che il segnale di partenza viene recuperato anche se la portante di demodulazione presenta un errore di fase di π, tale da causare *l'inversione* della forma d'onda e quindi di tutti i bit decodificati, producendo un messaggio $\hat{y}_k = \bar{y}_k = 110010010011100$ (la sopralineatura rappresenta l'inversione logica). Infatti, applicando la (16.16) alla sequenza $\hat{y}_k = \bar{y}_k$ si ottiene di nuovo la sequenza originaria, dato che $a \oplus b = \bar{a} \oplus \bar{b}$.

D'altra parte lo schema di fig. 16.15 non è l'unico possibile, osservando che la sequenza x_k può anche essere derivata *direttamente* dall'esame visivo della forma d'onda DBPSK, in quanto i bit della sequenza y_k *cambiano* nel caso in cui il corrispondente bit di x_k è un uno, mentre *non cambiano* se è uno zero, e dunque lo stesso accade per il segno della forma d'onda DBPSK. Il demodulatore può essere dunque sostituito da uno in fase e quadratura, in modo da calcolare la fase a partire dalle c.a. di b.f. come $\varphi = \arctan \frac{d_s}{d_c}$, e quindi determinare la sequenza x_k in base alle *variazioni* di fase, senza dover applicare la (16.16), consentendo di applicare il principio anche in presenza di errori di fase qualsiasi, e non solo pari a π.

Prestazioni DBPSK Esaminiamo ora cosa accade in presenza di errori: supponiamo di ricevere un messaggio $y_k = 00\underline{0}101101100011$, in cui il terzo bit (sottolineato) è errato. Calcolando $z_k = \hat{y}_k \oplus \hat{y}_{k-1}$ questa volta si ottiene $z_k = /0\underline{0}1111011010010$ 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 prima vista possiamo dire che a parità di E_b/N_0, il DBPSK è affetto da un tasso di errore *circa doppio* di quello del BPSK. Una analisi più approfondita (che omettiamo) fornisce l'espressione

$$P_e^{DBPSK}(bit) = \frac{1}{2} e^{-\frac{E_b}{N_0}} \tag{16.18}$$

che equivale ad un peggioramento di prestazioni di circa 1 dB rispetto al BPSK, ed il cui andamento è riportato a pag. 518.

16.4.2 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 che segue[29] per $L = 4$([30]), ovvero nel caso della modulazione DQPSK. L'inviluppo complesso di tale segnale assumerà quindi, 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	$-\pi/2$

Anche nel caso del QAM si può applicare una forma di codifica differenziale, ma lo schema di corrispondenza tra gruppi di bit e punti della costellazione è più complesso[31], e non viene qui trattato.

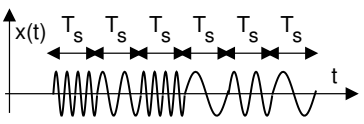
Infine, la modulazione differenziale può essere proficuamente sfruttata nella trasmissione OFDM (§ 16.8.8), al fine di evitare la necessità di equalizzazione.

16.5 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 a_k ottenuto raggruppando M bit, uno tra $2^M = L$ possibili valori (o *livelli*) di frequenza f_k, da sommare a quello della portante in accordo all'espressione

$$x_{FSK}(t) = \cos\left[2\pi\left(f_0 + m(t)\right)t\right] \quad \text{dove} \quad m(t) = \Delta \cdot \sum_{k=-\infty}^{\infty} f_k \cdot \text{rect}_{T_s}(t - kT_s)$$

in cui ogni elemento della sequenza f_k assume uno tra gli L valori $\{0, 1, 2, \ldots, L-1\}$. Si tratta in altri termini 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



$$x_{FSK}(t) = \sum_{k=-\infty}^{\infty} \cos\left[2\pi\left(f_0 + \Delta f_k\right)t\right] \cdot \text{rect}_{T_s}(t - kT_s) \tag{16.19}$$

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 VCO ed un PLL (vedi § 12.2.2.2 e 12.3.2.1) riportato (per $L = 2$) in figura 16.17, 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

[29] Tratta da *Andrea Goldsmith*, Wireless Communications, pag. 151.

[30] Nel caso di $L > 4$ la tabella si modifica molto semplicemente scrivendo accanto al codice di Gray al L livelli, la sequenza crescente delle fasi differenziali $\Delta\theta_k = k\frac{2\pi}{L-1}$.

[31] Vedi ad es. K. WESOLOWSKI, *Introduction to Digital Communication Systems*, Wiley, pag. 328.

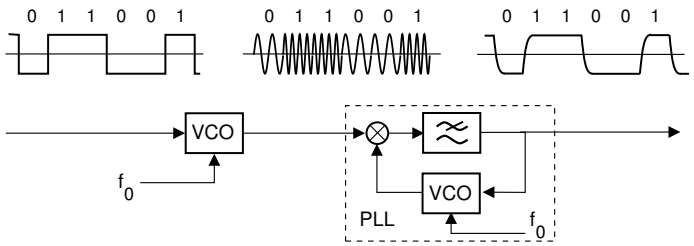


Figura 16.17: Modem FSK a bassa velocità

di L, occorre ridurre T_s, in modo da aumentare $f_s = \frac{1}{T_s}$ e quindi $f_b = f_s M = f_s \log_2 L$. In tal caso può essere necessario ricorrere ad un demodulatore più complesso, come descritto appresso.

16.5.1 Modulazione FSK ortogonale

Nel caso in cui si realizzi la condizione $\Delta = {}^{l}/_{2T_s}$ con l intero, le diverse frequenze $f_0 + \Delta f_k$ sono *ortogonali*[32], e può essere adottato un *demodulatore a correlazione* (vedi § 7.6.2), realizzato mediante un banco di correlatori ed una decisione di massimo (fig. 16.18), in cui l'$n-$esimo correlatore esegue

$$\int_0^{T_s} \cos\left[2\pi\left(f_0 + m\Delta\right)t\right] \cos\left[2\pi\left(f_0 + n\Delta\right)t\right] dt \tag{16.20}$$

dove $f_0 + m\Delta$ rappresenta la frequenza (incognita) in arrivo, mentre $f_0 + n\Delta$ è una delle frequenze possibili, con $n \in \{0, 1, 2, \ldots, L-1\}$. Essendo tali frequenze tra loro ortogonali[33] entro l'intervallo di integrazione, al termine del calcolo una sola delle uscite sarà diversa da zero. Come discusso al § 7.6 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* (§ 6.6.2.1). Indicando infatti con $r_n,\ n = 0, 1 \cdots, L-1$ la grandezza prodotta

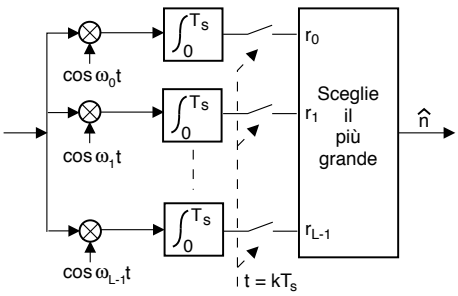


Figura 16.18: Demodulatore FSK a correlazione

[32] La discussione al riguardo è sviluppata al § 16.12.1, definendo anche le condizioni di demodulazione *incoerente* e *coerente*, ovvero se le portanti generate in ricezione $\cos\left[2\pi\left(f_0 + \Delta f_k\right)t + \phi_k\right]$ presentino 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 $\frac{1}{T_s}$.

[33] La condizione di ortogonalità tra le forme d'onda associate ai diversi simboli corrisponde alla intercorrelazione nulla tra le forme d'onda in un periodo (§ 7.1.4), ed infatti scegliendo opportunamente Δ ed f_0 (vedi § 16.12.1) l'integrale (16.20) vale $\mathcal{R}_{nm} = \begin{cases} .5 \cdot T_s & \text{se } n = m \\ 0 & \textit{altrimenti} \end{cases}$. Ciò si dimostra (ricordando che $\cos^2\alpha = \frac{1}{2} + \frac{1}{2}\cos 2\alpha$), notando che per $m = n$ l'uscita del correlatore vale $\frac{1}{2}\int_0^{T_s}\left(1 + \cos\left(4\pi\left(f_0 + m\Delta\right)t\right)\right)dt$, e scegliendo opportunamente f_0 e Δ (vedi § 16.12.1), 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 di nuovo in virtù di § 16.12.1, contiene solo termini a media nulla.

dal campionatore *n−esimo*, e con $\mathbf{r} = [r_0, r_1, \cdots r_{L-1}]$ il vettore aleatorio $L-$dimensionale corrispondente, la scelta del maggiore tra gli r_n corrisponde a scegliere l'ipotesi $f_{\hat{n}}$ che rende massima la $p\left(\mathbf{r}/f_{\hat{n}}\right)$[34].

Occupazione di banda In generale, se ogni diversa f_k è equiprobabile l'FSK ha una densità spettrale del tipo[35] mostrato alla figura 16.19. Se $L \gg 1$, l'occupazione di banda complessiva risulta quindi (circa) uguale a $L \cdot \Delta$. Nel caso di *modulazione coerente* (vedi nota 32) 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

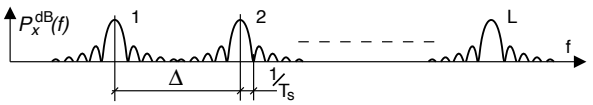


Figura 16.19: Densità spettrale per FSK *lento*

$$B_{FSK}^{coerente} \simeq L \cdot \frac{f_s}{2} = \frac{f_b}{2} \cdot \frac{L}{\log_2 L} \tag{16.21}$$

mentre per quanto riguarda l'efficienza spettrale si ottiene

$$\rho_{FSK} = \frac{f_b}{B} = f_b \cdot \frac{2}{f_b} \cdot \frac{\log_2 L}{L} = \frac{2}{L} \log_2 L$$

ossia $\frac{L}{2}$ volte *peggiore* dell'L-ASK (pag. 498). 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 rappresentato in fig. 16.17, di facile realizzazione ed economico: ad esempio, veniva usato per salvare su *compact cassette* audio i dati degli *home computer* degli anni '70[36]

Nel caso a due livelli l'efficienza spettrale è *quasi* $\rho_{BFSK} = \frac{2}{L} \log_2 L\Big|_{L=2} = 1$, come per il caso del BPSK[37]. Al contrario, al crescere di L l'efficienza spettrale diviene sempre peggiore.

---

[34]Infatti il vettore $\mathbf{r}$ ha una d.d.p. condizionata $p\left(\mathbf{r}/f_n\right)$ gaussiana multidimensionale a componenti indipendenti, e dunque (vedi § 6.5.1) si fattorizza nel prodotto di L gaussiane monodimensionali con uguale varianza e media nulla, tranne per la componente $n = m$ relativa all'ipotesi realmente occorsa, che presenta una media non nulla. Pertanto per ogni possibile ipotesi f_n la $p\left(\mathbf{r}/f_n\right)$ è *concentrata* sulla $n-esima$ componente, e dunque decidere per $\hat{n} = \arg\max_n \{r_n\}$ equivale a scegliere $\hat{n} = \arg\max_n \{p\left(\mathbf{r}/f_n\right)\}$.

[35]Difatti la (16.19) può essere riscritta come la somma di L segnali $x_k(t)$, uno per ogni possibile valore di f_k, costituiti da un codice di linea RZ che modula la corrispondente $f_0 + \Delta f_k$, a cui corrisponde dunque un tono intermittente. Essendo i simboli indipendenti e (in virtù della portante) a media nulla, la (7.46) di pag. 227 si riduce alla nota forma $\mathcal{P}_{X_k}(f) = \frac{\sigma_A^2}{T_s} |G_k(f)|^2$ in cui $G_k(f) = \mathcal{F}\{g_k(t)\}$ e $g_k(t) = \cos\left[2\pi\left(f_0 + \Delta f_k\right)t\right] rect_{T_s}(t)$; applicando ora il risultato di fig. 3.5 a pag. 77 si ottiene la densità di potenza mostrata in fig. 16.19.

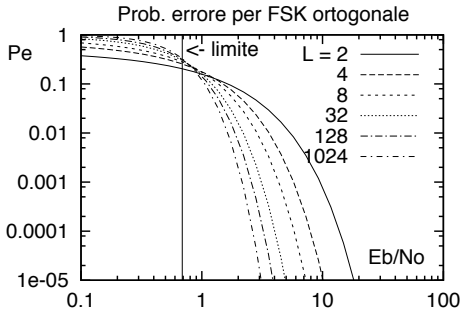
[36]tipo: Sinclair Spectrum, Commodore Vic20 e 64 ... Come noto, le cassette audio soffrono di variazioni di velocità di trascinamento del nastro (*wow & flutter*), ma il PLL non ne risente.

[37]Tranne che, essendo ora presenti solo 2 frequenze, l'approssimazione (16.21) non è più corretta. In particolare, con riferimento alla fig. 16.19, è tanto meno corretta quanto più f_s è elevata, che corrisponde ad oscillazioni del $sinc^2$ più estese in frequenza.

La probabilità di errore Si può dimostrare che l'uso dello schema di fig. 16.18 e di portanti di demodulazione ortogonali e coerenti[38] permette di ottenere una

$$P_e^{FSK}\,(simbolo) = 1 - \frac{1}{\sqrt{\pi^L}} \int_{-\infty}^{\infty} \mathrm{e}^{-z^2} \left(\int_{-\infty}^{z+\sqrt{\log_2 L \cdot E_b/N_0}} \mathrm{e}^{-y^2} dy \right)^{L-1} dz \qquad (16.22)$$

che deve essere valutata per via numerica, e che può essere resa *piccola a piacere*, nei limiti previsti dalla teoria dell'informazione[39], semplicemente aumentando L[40] (e dunque T_s). La figura a lato mostra i valori della (16.22) in funzione di E_b/N_0 per diversi valori di L, e illustra come all'aumentare di quest'ultimo sia necessaria sempre minor potenza per ottenere la stessa P_e, a patto che risulti



$$E_b/N_0 > \ln 2 = 0,69$$

che rappresenta il valore noto come *limite di Shannon-Hartley*, ricavato a pag. 565.

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, a cui si aggiunge il ritardo temporale necessario ad accumulare gli $M = \log_2 L$ bit che realizzano un simbolo con un enorme numero L di livelli.

Discussione sull'ottimalità per $L \to \infty$ Osserviamo innanzitutto che il ricevitore a correlazione commette errore quando il rumore sovrapposto al segnale di ingresso è 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 della frequenze di segnalazione, e quindi si riduce la probabilità di errore. La nota 14 a pag. 564 propone una interpretazione analitica di questo fenomeno.

Chiaramente, all'aumentare di L aumenta proporzionalmente la complessità del ricevitore, che deve disporre di un numero di correlatori crescente. Pertanto, le

[38] Ovvero qualora siano soddisfatte le condizioni per f_0 e Δ valutate al § 16.12.1 per il caso di demodulazione coerente, e si verifichi la *sincronizzazione* tra le forme d'onda in ingresso ai correlatori del banco.

[39] Ovvero tenendo conto che (vedi § 17.2) f_b non può superare la capacità di canale (eq. (17.18)), che a sua volta non può superare il limite C_∞ espresso dalla (17.20).

[40] Vedi la discussione seguente per una motivazione informale di questo comportamento.

prestazioni ideali per L che tende ad infinito rivestono solamente un interesse teorico.

16.6 Demodulazione incoerente

Nel caso in cui la portante di demodulazione non abbia la stessa fase di quella ricevuta, ci si trova nelle condizioni esposte al § 12.2.3.1, ossia il piano dell'inviluppo complesso risulta ruotato, rendendo impraticabili le tecniche di modulazione di fase, a meno di non adottare tecniche differenziali (§ 16.4).

In realtà è ancora praticabile la tecnica OOK (*on off keying*), ovvero una modulazione PAM della portante con un impulso NRZ polare, oltre a quella dell'FSK *incoerente*. Per entrambe si tratta di rivelare la presenza/assenza di una sinusoide nel rumore, per la durata di un bit T_b o di un simbolo T_s, e si ricorre allo schema di demodulazione discusso al § 12.2.4 e riportato a fianco, in cui la portante di demodulazione è una di quelle dell'FSK, oppure l'unica nel caso di OOK, ed il generico passa basso è realizzato come un integratore, ovvero con risposta impulsiva[41] $h(t) = \frac{1}{T_s} rect_{T_s}(t)$, ovvero ancora come un *filtro adattato* all'impulso di trasmissione $g(t) = rect_{T_s}(t)$[42], in modo da scrivere il segnale ricevuto come

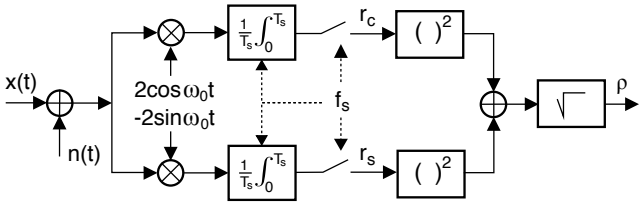


Figura 16.20: Demodulatore incoerente di inviluppo

$$x(t) = \sum_{k=-\infty}^{\infty} a_k \text{rect}_{T_s}(t - kT_s) \cos(\omega_0 t + \theta)$$

in cui $a_k = A$ se la frequenza f_0 è attiva durante il simbolo k, oppure $a_k = 0$ nel caso opposto. Il rumore $n(t)$ in ingresso, con densità di potenza $\frac{N_0}{2}$, rende le grandezze di osservazione r_c e r_s due v.a., che nel caso di segnale presente hanno valor medio[43] $m_c = A\cos\theta$ e $m_s = A\sin\theta$, oppure zero per segnale assente, mentre in entrambi i casi e per entrambi i rami la varianza risulta pari a[44] $\sigma^2 = {}^{N_0}/_{T_s}$.

La decisione se sia presente o meno la frequenza è basata sul *modulo* dell'inviluppo complesso $\underline{r} = r_c + jr_s$, ovvero $\rho = \sqrt{r_c^2 + r_s^2}$, ed attuata mediante l'approccio di massima verosimiglianza esposto al § 14.4.2. Nelle ipotesi poste, il caso in cui $a_k = 0$ corrisponde

[41] Ad esempio realizzato mediante un *integrate and dump* (pag. 215), che deve essere *resettato* a fine T_s.

[42] Il fattore ${}^1/_{T_s}$ che compare nell'espressione di $h(t)$ rende l'energia dell'impulso complessivo $g(t) * h(t) = tri_{2T_s}(t)$ (vedi eq. (3.26)) *normalizzata* rispetto a T_s (vedi § 3.8.8).

[43] Infatti il segnale demodulato (as es.) sul ramo in fase ha ampiezza costante $A\frac{1}{T_s}\cos\theta$, che risulta moltiplicata per T_s quando integrato su tale periodo.

[44] Infatti il filtro adattato ha una $|G(f)|^2 = sinc^2(T_s f)$, e dunque il rumore alla sua uscita (vedi § 14.1.3 e pag. 206) presenta una densità di potenza $\mathcal{P}_n(f) = N_0 sinc^2(T_s f)$. La potenza di rumore perciò risulta pari a $P_n = \sigma^2 = \frac{N_0}{T_s}$, in quanto $n(t)$ è a media nulla, e $\int_{-\infty}^{\infty} sinc^2(T_s f) = \frac{1}{T_s}$. Quest'ultimo risultato può essere verificato considerando che $sinc^2(T_s f)$ ha antitrasformata $\frac{1}{T_s} tri_{2T_s}(t)$, e che per la proprietà del valore iniziale (pag. 66) $\int_{-\infty}^{\infty} sinc^2(T_s f) = \frac{1}{T_s} tri_{2T_s}(t=0) = \frac{1}{T_s}$.

ad osservare una v.a. di Rayleigh (pag. 427) con d.d.p.

$$p_P(\rho) = \frac{\rho}{\sigma^2} e^{-\frac{\rho^2}{2\sigma^2}}$$

mentre se $a_k = A$ si osserva[45] una v.a. di Rice, con d.d.p.

$$p_P(\rho) = \frac{\rho}{\sigma^2} e^{-\frac{\rho^2+A^2}{2\sigma^2}} \cdot I_0\left(\frac{\rho A}{\sigma^2}\right)$$

ed in questa circostanza si è trovato (eq. 14.20) che se le due ipotesi di segnale presente (H_1) ed assente (H_0) sono equiprobabili e la soglia di decisione è posta pari a $A/2$, la probabilità di errore può essere approssimata come $P_e = \frac{1}{2} e^{-\frac{A^2}{8\sigma^2}}$.

Nel caso dell'OOK osserviamo che $A^2/2$ è la potenza di una sinusoide di ampiezza A, ma se questa per metà del tempo (gli $a_k = 0$) *è spenta* la potenza si dimezza, e così risulta $E_b = P_s T_b = \frac{A^2}{4} T_b$; sostituendo dunque $\frac{A^2}{4} = E_b/T_b$ e $\sigma^2 = N_0/T_b$ nell'espressione della P_e otteniamo

$$P_e^{OOK}(bit) = P_e^{2-FSK}(bit) = \frac{1}{2} e^{-\frac{E_b}{2N_0}} \tag{16.23}$$

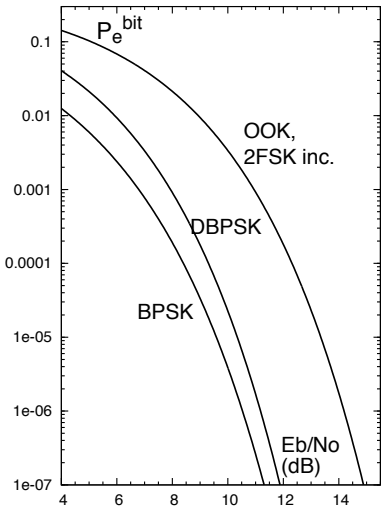


Figura 16.21: Confronto di prestazioni per demodulazione binaria coerente, differenziale, e incoerente

La stessa espressione descrive anche le prestazioni per una modulazione FSK a due livelli: in tal caso infatti la decisione avviene confrontando due v.a. ρ_1, ρ_2 con distribuzione una di Rice e l'altra di Rayleigh, ottenute duplicando lo schema di fig. 16.20 per le due frequenze (ortogonali) utilizzate, associate ad uno stesso bit uno o zero, e dunque il problema statistico è identico al precedente. La figura 16.21 permette il confronto delle prestazioni tra le tecniche di modulazione binarie presentate.

Nel caso poi di una modulazione L-FSK *incoerente*, la decisione avviene scegliendo tra L v.a. ρ_i di cui una distribuita Rice e tutte le altre Rayleigh; lo sviluppo analitico è un po' più complesso, e fornisce un risultato che seppur peggiore del caso coerente, lo approssima abbastanza bene per L ed E_b/N_0 elevati.

16.7 Schema riassuntivo delle prestazioni

La tabella 16.1 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

[45] La discussione a pag. 428 fa riferimento ad una sola v.a. (quella in fase) a media A, mentre nel caso attuale sia ha $\rho = A$ ma con una fase qualsiasi. Per le proprietà di simmetria radiale del problema, la conclusione è valida anche nel nostro caso.

Metodo	*bit/simbolo*	$P_e(bit)$ *con codifica di Gray ed impulso a banda minima*	*Banda RF*	ρ *[bit/sec/Hz]*
BPSK	1	$\frac{1}{2}\operatorname{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$	$f_b\,(1+\gamma)$	$\frac{1}{(1+\gamma)}$
QPSK	2	$\frac{1}{2}\operatorname{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$	$\frac{f_b}{2}\,(1+\gamma)$	$\frac{2}{(1+\gamma)}$
OOK, 2-FSK	1	$\frac{1}{2}\exp\left(-\frac{E_b}{2N_0}\right)$ *(incoerente, impulso RZ)*	$\sim 2f_b$	$\sim \frac{1}{2}$
L-ASK	M	$\frac{1}{M}\left(1-\frac{1}{L}\right)\operatorname{erfc}\left\{\sqrt{3\frac{E_b}{N_0}\frac{M}{(L^2-1)}}\right\}$	$f_b\frac{(1+\gamma)}{M}$	$\frac{M}{(1+\gamma)}$
L-PSK	M	$\frac{1}{M}\operatorname{erfc}\left\{\sin\left(\frac{\pi}{L}\right)\sqrt{\frac{E_b}{N_0}M}\right\}$	$f_b\frac{(1+\gamma)}{M}$	$\frac{M}{(1+\gamma)}$
L-QAM	M	$\frac{2}{M}\left(1-\frac{1}{\sqrt{L}}\right)\operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{N_0}\frac{M}{L-1}}\right\}$	$f_b\frac{(1+\gamma)}{M}$	$\frac{M}{(1+\gamma)}$
L-FSK incoerente	M	*vedi pag.* 518	$f_b\frac{L}{M}$	$\frac{M}{L}$
L-FSK coerente	M	*eq* (16.22) *a pag. 516*	$f_b\frac{L}{2M}$	$\frac{M}{2L}$

Tabella 16.1: Confronto tra metodi di modulazione numerica a portante singola

$M = \log_2 L$ bit. Se $\gamma \neq 0$, si deve aggiungere un termine $(1+\gamma)\left(1-\frac{\gamma}{4}\right)$ al denominatore sotto radice, procedendo come indicato al § 15.4.9.

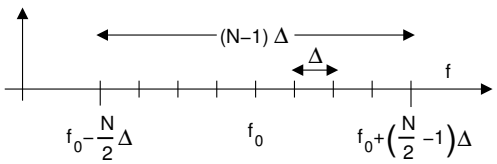
16.8 Modulazione OFDM

La sigla sta per ORTHOGONAL FREQUENCY DIVISION MULTIPLEX, ossia *multiplazione a divisione di frequenza ortogonale*. Si tratta della tecnica di modulazione numerica *multiportante* adottata per le trasmissioni ADSL[46], DVB-T, WIFI e per telefonia mobile; si contraddistingue per la particolarità di utilizzare in modo *ottimo* la banda del canale, e di ricondurre l'operazione di equalizzazione ad un prodotto tra vettori.

16.8.1 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 evoluzione[47] della modulazione FSK, in cui le diverse frequenze sono spaziate tra loro di Δ Hz come descritto dall'espressione

$$f_n = f_0 + \Delta \cdot \left(n - \frac{N}{2}\right) \qquad (16.24)$$



[46] ADSL = *Asymmetric Digital Subscriber Line*, vedi § 24.9.4.

[47] La 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.

con $n = 0, 1, ..., N-1$ e sono utilizzate contemporaneamente, mentre su ognuna di esse si realizza una modulazione numerica a due o più livelli (es. QPSK o QAM) con impulso NRZ rettangolare di durata T.

Indicando con $\underline{a}_n^k = a_{n_c}^k + j a_{n_s}^k$ le coordinate nel piano dell'inviluppo complesso di un generico punto della costellazione realizzata per la portante f_n all'istante $t = kT$, il segnale OFDM può essere scritto come

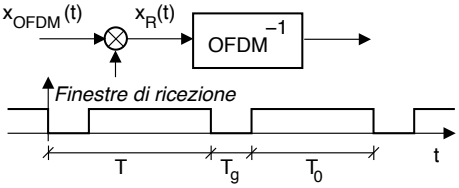
$$x_{OFDM}(t) = \sum_k \mathrm{rect}_T(t - kT) \sum_{n=0}^{N-1} \left(a_{n_c}^k \cos \omega_n (t - kT) - a_{n_s}^k \sin \omega_n (t - kT) \right) \quad (16.25)$$

$$= \sum_{k=-\infty}^{\infty} \delta(t - kT) * \left(\mathrm{rect}_T(t) \sum_{n=0}^{N-1} \left(a_{n_c}^k \cos \omega_n t - a_{n_s}^k \sin \omega_n t \right) \right) \quad (16.26)$$

in cui la prima sommatoria (su k) identifica gli istanti di simbolo, e la seconda (su n) le diverse portanti. Tale segnale presenta[48] un inviluppo complesso rispetto a f_0 pari a

$$\underline{x}_{OFDM}(t) = \sum_{k=-\infty}^{\infty} \delta(t - kT) * \left(\mathrm{rect}_T(t) \sum_{n=0}^{N-1} \underline{a}_n^k e^{j2\pi \left[\Delta \left(n - \frac{N}{2} \right) \right] t} \right) \quad (16.27)$$

L'espressione (16.26) non vincola la durata T di un simbolo ad un valore particolare; deve però risultare $T \geq T_0 = 1/\Delta$, in quanto il ricevitore opera sul segnale $x_R(t)$ ottenuto per moltiplicazione con una finestra temporale di estensione $T_0 = 1/\Delta$, allo scopo di rendere ortogonali tra loro[49] le frequenze $f_n = f_0 + \Delta \cdot (n - N/2)$, e mettere il ricevitore in grado 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. 514.



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"[50] risiede nel fatto che, in presenza di un canale che introduce distorsione lineare, la parte iniziale di ogni

[48] Osserviamo innanzitutto che per un segnale

$$x(t) = \cos \omega_1 t = \frac{1}{2} \left(e^{j\omega_1 t} + e^{-j\omega_1 t} \right)$$

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

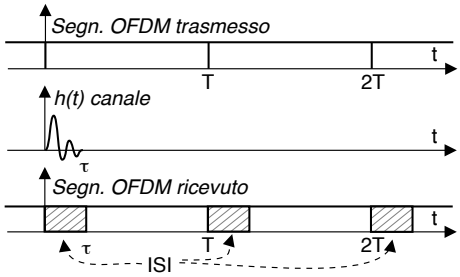
$$\underline{z}(t) = a_{n_c}^k e^{j(\omega_n - \omega_0)t} + j a_{n_s}^k e^{j(\omega_n - \omega_0)t} = \underline{a}_n^k e^{j2\pi(f_n - f_0)t}$$

Applicando ora la (16.24) si ottiene $f_n - f_0 = \Delta \cdot \left(n - \frac{N}{2} \right)$ e quindi la (16.27).

[49] Come mostrato per il caso *incoerente* discusso al § 16.12.1

[50] Infatti 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.

simbolo risulta corrotta (vedi figura a lato) 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



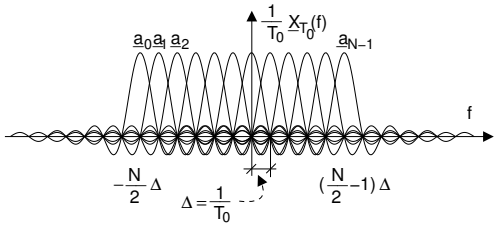
$$\underline{x}_{T_0}(t) = \mathrm{rect}_{T_0}(t) \cdot \sum_{n=0}^{N-1} \underline{a}_n e^{j2\pi\left[\Delta\left(n-\frac{N}{2}\right)\right]t} \tag{16.28}$$

e calcoliamone la trasformata per determinare l'occupazione di banda:

$$\underline{X}_{T_0}(f) = T_0 \,\mathrm{sinc}(fT_0) * \sum_{n=0}^{N-1} \underline{a}_n \delta\left(f - \Delta\left(n - \frac{N}{2}\right)\right) = \tag{16.29}$$

$$= T_0 \sum_{n=0}^{N-1} \underline{a}_n \,\mathrm{sinc}\left(\left(f - \Delta\left(n - \frac{N}{2}\right)\right) T_0\right) \tag{16.30}$$

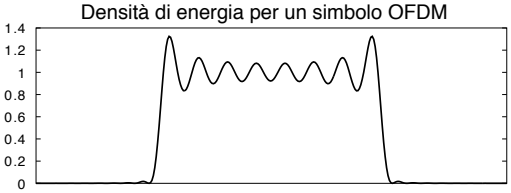
Otteniamo pertanto la costruzione grafica mostrata alla figura a lato, che evidenzia come ogni funzione *sinc* sia moltiplicata per uno dei coefficienti $\underline{a}_n$, che potrebbero quindi essere ri-ottenuti in ricezione campionando (in modo complesso) $\underline{X}(f)$ su frequenze spaziate di Δ.

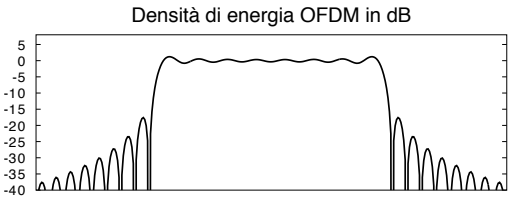


Per quanto riguarda la densità di potenza $\mathcal{P}_{\underline{x}_R}(f)$ dell'inviluppo complesso $\underline{x}_R(t)$ ricevuto e finestrato, consideriamo l'espressione (vedi § 7.3.1)

$$\mathcal{P}_{\underline{x}_R}(f) = \frac{1}{T} E\left\{\left|\underline{X}_{T_0}(f)\right|^2\right\} \tag{16.31}$$

in cui $\left|\underline{X}_{T_0}(f)\right|^2$ è la densità di energia di un simbolo OFDM (eq. (16.30)): 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 inviluppo complesso $\mathcal{P}_{\underline{x}_R}$

$$\mathcal{P}_{\underline{x}_R} = \int \mathcal{P}_{\underline{x}_R}(f)\, df \tag{16.32}$$

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 (16.26) attribuisce una potenza $\mathcal{P}_n$.

Per calcolare la (16.32) mediante la (16.31) utilizzando l'espressione di $\underline{X}_{T_0}(f)$ fornita dalla (16.30), 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 § 4.1.2) sussiste la condizione

$$\int_{-\infty}^{\infty} \text{sinc}((f - n\Delta)T_0)\,\text{sinc}((f - m\Delta)T_0)\,df = \begin{cases} 0 & \text{se} \quad n \neq m \\ \frac{1}{T_0} & \text{se} \quad n = m \end{cases}$$

Pertanto, introducendo una insignificante[51] traslazione di f pari a $\frac{N}{2}\Delta$, si ha

$$\begin{aligned} \mathcal{P}_{\underline{x}_R} &= \frac{1}{T}\int E\left\{\left|\underline{X}_{T_0}(f)\right|^2\right\} df = \\ &= \frac{1}{T}\int T_0^2 \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} E\left\{\underline{a}_n \underline{a}_m^*\right\} \text{sinc}((f - \Delta n)T_0)\,\text{sinc}((f - \Delta m)T_0)\,df = \\ &= \frac{T_0^2}{T}\sum_{n=0}^{N-1} E\left\{\underline{a}_n^2\right\}\int \text{sinc}^2((f - \Delta n)T_0)\,df = \frac{T_0^2}{T}\sum_{n=0}^{N-1} E\left\{\underline{a}_n^2\right\}\frac{1}{T_0} = \frac{T_0}{T}\sum_{n=0}^{N-1}\sigma_{\underline{a}_n}^2 \end{aligned}$$

in quanto i termini *incrociati* prodotti dalla doppia sommatoria si annullano[52].

Scegliendo il lato della costellazione QAM in modo opportuno[53], si può ottenere $\sigma_{\underline{a}_n}^2 = E\left\{\underline{a}_n^2\right\} = 2\mathcal{P}_n$, in cui $\mathcal{P}_n$ è la potenza per la n-esima portante QAM; considerando infine (vedi eq. (11.22) a pag. 353) 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.

16.8.2 Architettura di modulazione

Una caratteristica fondamentale della modulazione OFDM è quella di essere realizzata senza *oscillatori e integratori*, ma attraverso l'uso della elaborazione numerica. Con riferimento alla figura 16.22 , il flusso binario a frequenza f_b viene *parallelizzato* per formare simboli ad $L = 2^M$ livelli a frequenza $f_s = f_b/M = f_b/\log_2 L$. Questi M bit/simbolo

[51] Equivalente a definire l'inviluppo complesso con riferimento ad una portante a frequenza pari alla prima delle f_n.

[52] Vedi nota 22 a pag. 47

[53] Al § 15.8.1 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 $\sigma_{\underline{a}_n}^2 = E\left\{\left(a_{n_c} + ja_{n_s}\right)^2\right\} = 2\sigma_a^2 = \frac{2A^2}{3}\frac{\sqrt{L}+1}{\sqrt{L}-1}$; volendo eguagliare tale valore a $2\mathcal{P}_n$, occorre quindi scegliere $A = \sqrt{3\mathcal{P}_n\frac{\sqrt{L}-1}{\sqrt{L}+1}}$.

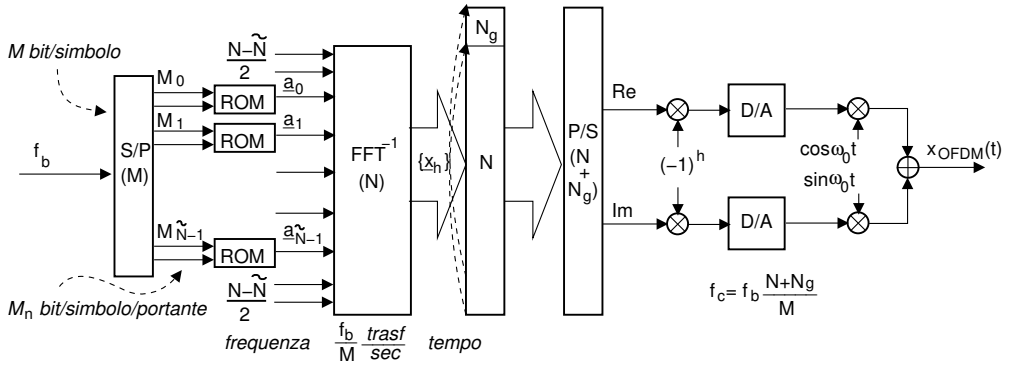
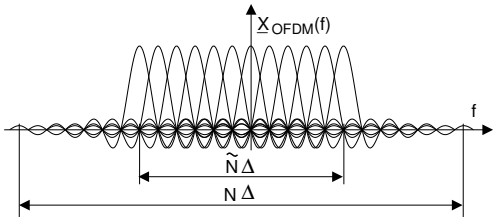


Figura 16.22: Architettura di un modulatore OFDM numerico

sono suddivisi in $\tilde{N}$ gruppi di M_n ($n = 0, 1, ..., \tilde{N}-1$) 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 (16.30) dia luogo ad un inviluppo complesso praticamente limitato in banda (vedi figura) tra (circa) $\pm\frac{N}{2}\cdot\Delta$ Hz, che può essere pertanto rappresentato dai suoi campioni $\underline{x}_{T_0}(hT_c)$ presi a frequenza $f_c = N\cdot\Delta\ \frac{campioni}{secondo}$. Il blocco indicato come FFT^{-1} ha esattamente il ruolo di valutare i campioni temporali di $\underline{x}_{T_0}$, calcolando[54]

[54]La (16.33) è in qualche modo simile alla formula di ricostruzione (2.7) (vedi pag. 39) per il segnale (complesso) periodico limitato in banda $\pm\frac{N}{2}F$

$$\underline{x}(t) = \sum_{m=-N/2}^{N/2} X_m e^{j2\pi mFt}$$

che calcolata per $t = hT_c = \frac{h}{NF}$ fornisce $\underline{x}(hT_c) = \sum_{m=-N/2}^{N/2} X_m e^{j2\pi\frac{m}{N}h}$. Ponendo ora $n = m + \frac{N}{2}$ e $Y_n = X_{n-\frac{N}{2}}$ otteniamo

$$\underline{x}(hT_c) = \sum_{n=0}^{N-1} Y_n e^{j2\pi\frac{n-N/2}{N}h} = e^{-j\pi h}\sum_{n=0}^{N-1} Y_n e^{j2\pi\frac{n}{N}h}$$

dato che $\left(n-\frac{N}{2}\right)\frac{1}{N} = \frac{n}{N} - \frac{1}{2}$. Osservando ora che dalla (16.28) con $T_c = \frac{1}{N\Delta}$ si ha

$$\underline{x}(hT_c) = \sum_{n=0}^{N-1} \underline{a}_n e^{j2\pi\Delta\left(n-\frac{N}{2}\right)\frac{h}{N\Delta}} = \sum_{n=0}^{N-1} \underline{a}_n e^{j2\pi\Delta\frac{n-N/2}{N}h} = e^{j\pi h}\sum_{n=0}^{N-1} \underline{a}_n e^{j2\pi\frac{n}{N}h}$$

e che $e^{-j\pi h} = (-1)^h$, si ottiene la (16.33). 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 § 4.5), e consentono il calcolo di una serie di campioni in frequenza a partire da campioni nel tempo e viceversa.

$$\sum_{n=0}^{N-1} \underline{a}_n e^{j2\pi \frac{n}{N} h} = \frac{1}{(-1)^h} \underline{x}_{T_0}(hT_c) \tag{16.33}$$

Il risultato della FFT^{-1} è quindi una sequenza di valori complessi $\left\{\underline{x}_h\right\}$, che a meno di un segno alterno rappresentano i campioni dell'inviluppo complesso $\underline{x}_{T_0}(t)$ espresso dalla (16.28) relativamente ad un simbolo. Dopodiché, il preambolo da trasmettere durante il tempo di guardia T_g è ottenuto *aggiungendo* in testa a $\left\{\underline{x}_h\right\}$ un gruppo di campioni prelevati dalla coda[55].

Infine, le parti reale ed immaginaria di $\left\{\underline{x}_h\right\}$ 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.

16.8.3 Efficienza dell'OFDM

Come vedremo al § 16.8.9, 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 vuote ed il preambolo) ha gli stessi scopi di quella introdotta dal roll-off γ di un impulso a coseno rialzato, in quanto evita che si verifichino fenomeni di interferenza tra simboli, e realizza un segnale limitato in banda. Osserviamo che l'efficienza migliora all'aumentare di T e di N, dato che T_g ed $N - \tilde{N}$ sono fissi.

Esercizio Un flusso binario a velocità $f_b = 1$ Mbps è trasmesso mediante modulazione OFDM con portante 1 GHz, caratterizzata da: $\tilde{N} = 464$ portanti attive su $N = 512$ totali, $M_n = 2$ bit a portante, con modulazione QPSK, e $T_g = 28$ μsec di tempo di guardia.
Calcolare: 1) il numero di bit/simbolo M ed il corrispondente periodo di simbolo T e 2) la spaziatura tra portanti $\Delta = 1/T_0$ e la corrispondente occupazione di banda.

1. $M = M_n \cdot \tilde{N} = 2 \cdot 464 = 928$ bit/simbolo, e
   $T = 1/f_b \cdot M = 10^{-6} \cdot 928 = 928$ μsec;
2. $T_0 = T - T_g = 928 - 28 = 0.9$ msec, dunque $\Delta = 1/T_0 \simeq 1.11$ KHz, e
   $B = N \cdot \Delta = 512 \cdot 1.11 \cdot 10^3 \simeq 568$ KHz.

16.8.4 Architettura di demodulazione

Per ottenere gli elementi della sequenza $\left\{\underline{a}_n\right\}$ e quindi il gruppo di M bit che hanno originato il simbolo si adotta l'architettura mostrata in figura 16.23, che svolge una

[55] In effetti la (16.33) fornisce un risultato periodico rispetto ad h, con periodo N, ossia con periodo $N \cdot T_c = N\frac{1}{f_c} = N\frac{1}{\Delta N} = \frac{1}{\Delta} = T_0$ per la variabile temporale. Per questo motivo il *preambolo* dell'OFDM è detto anche *estensione ciclica*.

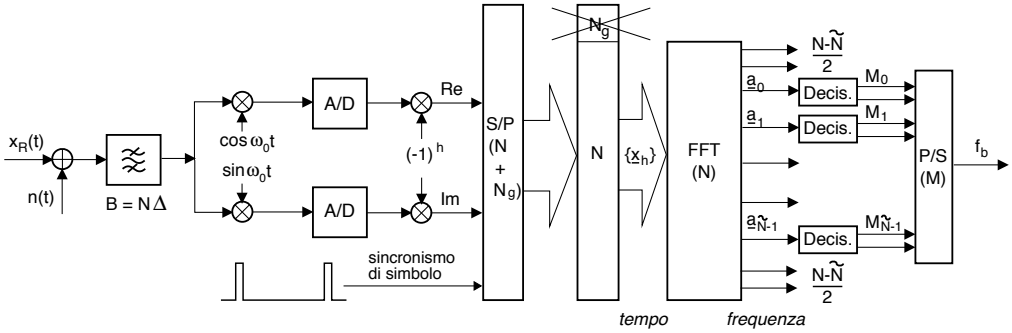


Figura 16.23: Architettura di un demodulatore OFDM numerico

azione del tutto inversa a quella del modulatore. Innanzitutto il ricevitore deve acquisire il sincronismo di frequenza (vedi § 16.8.11), 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, e dopo 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 (vedi nota 54), ottenendo i valori

$$\frac{1}{N}\sum_{h=0}^{N-1} \underline{x}_h \mathrm{e}^{-j2\pi\frac{h}{N}n} = \underline{X}_{T_0}\left(\left(n-\frac{N}{2}\right)\Delta\right) = \underline{a}_n \tag{16.34}$$

Solo gli $\tilde{N}$ valori centrali sono avviati verso altrettanti decisori, che determinano il punto di costellazione più vicino all'$\underline{a}_n$ ricevuto per ogni portante, associandovi il relativo codice di M_n bit, ed il risultato finale è quindi serializzato per produrre gli M bit che hanno dato origine al simbolo.

16.8.5 Prestazioni

Al § 16.12.2 viene svolta una laboriosa analisi per arrivare a valutare l'espressione della P_e^{bit} in caso di tempo di guardia $T_g = 0$ ed in presenza di rumore additivo gaussiano limitato alla stessa banda del segnale; il risultato è confrontato con quello ottenibile per una trasmissione QAM che occupi la medesima banda dell'OFDM, trasporti lo stesso flusso f_b, utilizzi ovviamente una sola portante con un adeguato numero di livelli, e adotti un impulso a coseno rialzato che determini la stessa (in)efficienza spettrale legata nell'ODFM alla presenza delle $\tilde{N}$ portanti spente. Il risultato è che le prestazioni *sono identiche.*

E allora dov'è la convenienza ? E' il tema delle prossime sottosezioni!

16.8.6 Sensibilità alla temporizzazione

Con l'OFDM *non siamo* nelle condizioni di demodulazione coerente come per l'FSK (§ 16.5.1), e le portanti del simbolo OFDM ricevuto mantengono ortogonalità (§ 16.12.1) purché finestrate su di un periodo $T_0 = \frac{1}{\Delta}$. Pertanto nel caso in cui il ricevitore non acquisisca una perfetta sincronizzazione di simbolo, se l'ISI introdotta dal canale ha

una durata minore di T_g, la FFT di demodulazione può operare su di un gruppo di campioni presi a partire dalla coda del preambolo, riducendo così la sensibilità rispetto agli errori di sincronizzazione.

16.8.7 Equalizzazione

Consideriamo il caso in cui la trasmissione attraversi un canale la cui $h(t)$ è descritta da un inviluppo 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 (16.30) si altera a causa del filtraggio, ed i suoi campioni $\underline{a}_n$ restituiti dalla (16.34) si modificano in

$$\tilde{\underline{a}}_n = \underline{a}_n \cdot \underline{H}_n$$

dove $\underline{H}_n = H_n e^{j\varphi_n} = \underline{H}\left(f - \Delta\left(n - \frac{N}{2}\right)\right)$ sono i campioni (complessi) di $\underline{H}(f)$. Come anticipato, l'equalizzazione si riduce a svolgere un semplice prodotto tra la sequenza dei valori ricevuti $\tilde{\underline{a}}_n$ e quella di equalizzazione $\frac{1}{H_n} e^{-j\varphi_n}$ che inverte l'effetto del canale, ovviamente purché si conosca $\underline{H}(f)$, od una sua stima.

16.8.8 Codifica differenziale

Nel caso in cui l'entità della distorsione lineare introdotta dal canale non sia *eccessiva* si può evitare del tutto lo stadio di equalizzazione e ricorrere ad una *codifica differenziale* (§ 16.4), che risulta particolarmente semplice qualora le sottoportanti siano modulate PSK o QPSK. In tal caso infatti il processo di demodulazione per ogni sottoportante non risente di variazioni di guadagno, ovvero variazioni di $H_n = \left|\underline{H}_n\right|$, e dunque devono essere compensate le sole variazioni di fase φ_n tra una portante e l'altra, ognuna delle quali determina la corrispondente rotazione (vedi nota a pag. 17) del piano dell'inviluppo complesso su cui sono riferiti gli $\underline{a}_n$, rispetto alla disposizione degli assi per la portante $n-1$.

Acquisendo dunque un primo riferimento di fase da una *portante pilota* (§ 16.8.11) sempre accesa senza trasportare informazione, si può prendere quello per demodulare la portante successiva, acquisire da questa un nuovo riferimento di fase, e iterare il procedimento per tutte le portanti. Questo procedimento si attua applicando la teoria del § 16.4 alla sequenza simbolica di valori complessi $\underline{a}_n$ da trasmettere, sostituendo nelle (16.15) e (16.16) l'OR esclusivo con una operazione di prodotto, ed aggiungendo una operazione di coniugato, come mostrato in fig. 16.24, in cui R rappresenta un ritardo unitario.

Dal lato della trasmissione le portanti sono quindi modulate a partire dalla sequenza

$$\underline{d}_n = \underline{a}_n \cdot \underline{d}_{n-1}$$

con $n = 0, 1, \cdots, \tilde{N}-1$, avendo posto $\underline{a}_0 = 1$. In assenza di rumore e di distorsione lineare la sequenza $\underline{d}_n$ è ricevuta inalterata, ed è così disponibile in uscita dal demodulatore

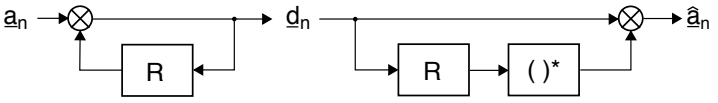


Figura 16.24: Codifica differenziale per simboli complessi, R rappresenta un *ritardo*

OFDM; da essa si ottiene quindi

$$\underline{\hat{a}}_n = \underline{d}_n \cdot \underline{d}^*_{n-1} = \underline{a}_n \cdot \underline{d}_{n-1} \cdot \underline{d}^*_{n-1} = \underline{a}_n \cdot \left|\underline{d}_{n-1}\right|^2 \tag{16.35}$$

che presenta la stessa fase di $\underline{a}_n$.

In presenza di distorsione lineare al posto di $\underline{d}_n$ si riceve invece $\underline{\tilde{d}}_n = \underline{d}_n \cdot \underline{H}_n$ in cui $\underline{H}_n = H_n e^{j\varphi_n}$ sono i campioni della risposta in frequenza del canale, e dunque la (16.35) fornisce

$$\underline{\hat{a}}_n = \underline{\tilde{d}}_n \cdot \underline{\tilde{d}}^*_{n-1} = \underline{a}_n \cdot \underline{d}_{n-1} \cdot \underline{H}_n \cdot \underline{d}^*_{n-1} \cdot \underline{H}^*_{n-1} = \underline{a}_n \cdot \left|\underline{d}_{n-1}\right|^2 \cdot \Delta H_n \cdot e^{j\Delta\varphi_n}$$

in cui $\Delta H_n = \left|\underline{H}_n \underline{H}^*_{n-1}\right|$ (ma il modulo non ci interessa), e $\Delta\varphi_n = \varphi_n - \varphi_{n-1}$ è la differenza tra i valori della risposta di fase del canale valutata per due portanti contigue, e rappresenta l'entità di cui è ruotato il piano dell'inviluppo per i simboli trasportati dalle due portanti. Pertanto, se questa è di lieve entità (essendo le portanti vicine), produce un errore trascurabile.

Accenniamo brevemente all'ulteriore possibilità di applicazione del principio differenziale, oltre che portante per portante, anche ad interi simboli OFDM consecutivi: in questo caso il vettore di simboli $\underline{a}^k_n$ da trasmettere all'istante k viene combinato con i valori del vettore trasmesso al simbolo precedente $k-1$, ovvero $\underline{d}^k_n = \underline{a}^k_n \cdot \underline{d}^{k-1}_n$. In questo modo possono essere contrastati i fenomeni tempo-varianti che modificano il canale, per una stessa portante n, simbolo dopo simbolo.

16.8.9 Distribuzione ottima di potenza

Si riferisce alla possibilità dell'OFDM di assegnare valori di potenza differenti alle diverse portanti, che consente di sfruttare *al massimo* la capacità trasmissiva dal canale, anche in presenza di attenuazione selettiva e/o rumore colorato.

La trasmissione numerica con una f_b elevata, eseguita utilizzando un tecnica ad *una sola* portante, deve necessariamente occupare una banda molto ampia, rendendo scarsamente applicabile la semplificazione di cui al § 13.1.2.4; in tal caso $H(f)$ presenta distorsione di ampiezza, la cui equalizzazione (§ 18.4) 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 per tenere conto dell'andamento incostante di $\mathcal{P}_N(f)$ il calcolo della *capacità di canale*[56] $C = W \log_2(1 + \mathcal{P}_R/WN_0)$, valido in presenza di un rumore bianco $\mathcal{P}_N(f) = N_0/2$ e con una potenza ricevuta $\mathcal{P}_R$ in una banda (positiva) W, si modifica come segue. Se consideriamo il canale scomposto in *infinite sottobande* entro le quali le densità di potenza di segnale e di rumore possono ritenersi costanti, l'espressione della capacità diviene

$$C = \sup_{\mathcal{P}_R(f)} \int_{f \in I_f} \log_2 \left(1 + \frac{\mathcal{P}_R(f)}{\mathcal{P}_N(f)}\right) df \tag{16.36}$$

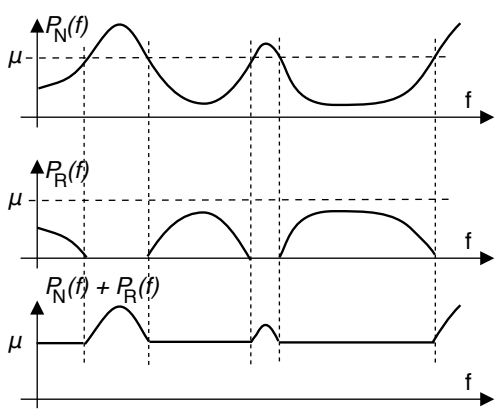
[56] Come discusso ai § 17.2 e 17.3 la teoria di Shannon asserisce che $f_b = C$ è la massima velocità di trasmissione per cui si può (teoricamente) conseguire una probabilità di errore nulla, e che il canale consegue capacità C massima a seguito di una scelta appropriata su come trasmettere il messaggio.

in cui I_f rappresenta l'insieme delle frequenze in cui è presente il segnale, ovvero $I_f = \{f : \mathcal{P}_R(f) > 0\}$. La (16.36) significa che se $\mathcal{P}_N(f)$ in ingresso al canale non è pari ad una costante, la *massima* capacità trasmissiva C (e dunque velocità f_b) può essere raggiunta *sagomando in modo opportuno* la densità di potenza $\mathcal{P}_R(f)$ del segnale ricevuto. Nell'OFDM ciò equivale a distribuire la potenza totale $\mathcal{P}_R$ in modo non uniforme tra le portanti, a patto che la $\mathcal{P}_R(f)$ che rende massima (16.36) rispetti i vincoli

$$\int_{f \in I_f} \mathcal{P}_R(f)\, df = \mathcal{P}_R \quad \text{e} \quad \mathcal{P}_R(f) \geq 0 \tag{16.37}$$

Questo problema di *ottimo vincolato* ammette la soluzione (vedi § 16.12.3)

$$\mathcal{P}_R(f) + \mathcal{P}_N(f) = \begin{cases} \mu & \text{se} \quad \mathcal{P}_N(f) < \mu \\ \mathcal{P}_N(f) & \text{se} \quad \mathcal{P}_N(f) \geq \mu \end{cases} \tag{16.38}$$



detta anche *riempimento d'acqua* (WATER-FILLING) perché (vedi figura) determina una maggiore potenza di segnale nelle regioni di frequenza dove il rumore è minore, un po' come se $\mathcal{P}_R$ fosse un volume d'acqua *versata sopra* $\mathcal{P}_N(f)$. La costante μ viene determinata in modo 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 dal particolare formatore di impulsi $G(f)$ scelto 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 (16.38) viene resa nota al modulatore, è possibile avvicinarsi alla velocità massima permessa dalla (16.36).

Bit loading 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 $n = 0, 1, ..., \tilde{N} - 1$ tali da rendere le $P_{e/n} = P_e$. Questo risultato può essere ottenuto scegliendo le potenze $\mathcal{P}_n$ in accordo alla (16.38), e quindi trasmettere (o *caricare*) più bit M_n sulle portanti n per le quali $\mathcal{P}_n$ è maggiore.

16.8.10 Modulazione codificata

Abbiamo appena mostrato come, conoscendo la $\mathcal{P}_N(f)$ e la $H(f)$ del canale, sia possibile equalizzare $\mathcal{P}_x(f) = \frac{\mathcal{P}_R(f)}{|H(f)|^2}$ e al contempo soddisfare (16.38) e rendere massima la f_b. Ma nel caso di collegamenti tempo-varianti la $H(f)$ non è nota, ed anche se lo fosse non esiste garanzia che rimanga costante. In tal caso 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 la presenza di una $H(f)$ tempo-variante rende il processo di equalizzazione particolarmente complesso, dato che deve *inseguire* le variazioni di $H(f)$. Se l'equalizzazione non è perfetta insorge ISI, e la trasmissione può divenire rapidamente così piena di errori da renderne impossibile la correzione anche adottando codici di canale.

Nel caso dell'OFDM, al contrario, l'andamento di $H(f)$ determina un peggioramento di prestazioni solamente per quelle portanti per le quali $\left|H(f)\right|$ si è ridotto[57]. Pertanto l'applicazione di un codice di canale (§ 17.4) al blocco di M bit che costituisce un simbolo, seguito da una operazione di scrambling (§ 15.6.2.3), 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).

16.8.11 Portanti pilota

Fin qui abbiamo assunto che il ricevitore OFDM mostrato in fig. 16.23 operi in condizioni di sincronismo sia per quanto riguarda la portante di demodulazione, sia per 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 alle altre, allo scopo di facilitare la sincronizzazione in frequenza. In figura 16.25 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 permettere una migliore stima della $H(f)$ del canale attraversato.

[57] Si consideri ad esempio il caso in cui $H(f)$ ha origine da un fenomeno di cammini multipli, che determina un andamento di $H(f)$ selettivo in frequenza (§ 20.4.5).

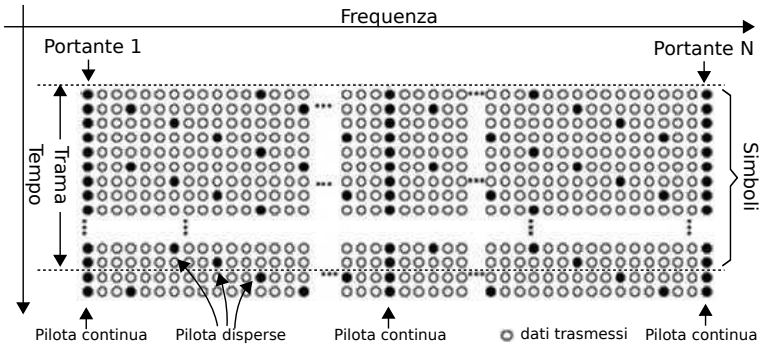
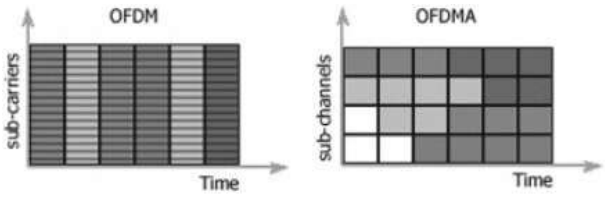


Figura 16.25: Allocazione delle portanti OFDM in un sistema di trasmissione DVB-T

16.8.12 Accesso multiplo OFDMA

Lo sviluppo di protocolli di gestione e coordinamento delle risorse impiegate da più utenti mobili per comunicare con una stessa stazione radio-base (§§ 11.1.1.3, 16.9.2.5) rende possibile assegnare *differenti sottoinsiemi* di portanti ai diversi utenti, permettendo di ripartire la banda a disposizione in percentuali variabili tra gli stessi. Ogni utente semplicemente *spenge* le portanti a lui non intestate prima di eseguire l'IFFT, così come ne scarta il risultato della decodifica. Si determinano così i vantaggi



- rispetto ad un sistema a divisione di tempo (§ 24.2.1, 24.3.1, 22.5.2.1) si ha un minor ritardo in quanto non occorre attendere il proprio time-slot o *contendere* la risorsa in comune, bensì i vari utenti trasmettono *in contemporanea*;
- l'assegnazione delle portanti può variare simbolo per simbolo, permettendo di allocare dinamicamente più banda agli utenti con maggiori esigenze;
- gli utenti con esigenze minori oltre alla banda risparmiano anche in potenza trasmessa;
- la riduzione del numero di portanti attive diminuisce la dinamica del segnale OFDM nel tempo, mitigando le problematiche di non-linearità.

Per far fronte a fenomeni di fading selettivo (§ 20.4.5) la qualità del canale tra stazione base ed ogni utente è monitorata di continuo affinché l'allocazione delle portanti possa essere modificata dinamicamente, e mantenuta una adeguata qualità di servizio per tutti: in questo senso si realizzano gli scopi della radio cognitiva (pag. 543).

La modalità di trasmissione OFDMA è stata adottata nei sistemi WiFi 802.11ax,[58] WiMAX 802.16e, e di telefonia mobile LTE e 5G.

16.9 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 opportuna *trasformazione* del messaggio da trasmettere, in modo che questo occupi una banda molto maggiore di quella originaria, e sulla manipolazione inversa in ricezione. Il peculiare aumento della banda occupata è indicato con il termine di *spread spectrum*[59], e la tecnica di trasmissione risultante prende anche il nome di multiplazione a divisione di codice o CDM (*code division multiplex*).

Sebbene la doppia operazione di *spreading/despreading* non produca nessun vantaggio effettivo nei riguardi delle prestazioni ottenibili qualora la ricezione sia disturbata

[58] `https://en.wikipedia.org/wiki/IEEE_802.11ax-2021`

[59] *To spread* = spalmare, vedi ad es. lo *spread butter*.

dalla sola presenza di rumore additivo gaussiano, si ottengono invece i seguenti altri benefici:

- altre trasmissioni e/o disturbi *a banda stretta* che occupano la stessa regione di frequenza occupata dal segnale espanso causano 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;
- per conoscere il contenuto della trasmissione occorre poter riprodurre in ricezione una esatta replica della trasformazione attuata.

16.9.1 Sequenze pseudo-casuali

La trasformazione che produce l'espansione spettrale si basa sull'utilizzo di una sequenza cosiddetta *pseudo-noise* o PN (§ 16.9.3), ovvero le cui caratteristiche statistiche *si avvicinano* a quelle di un rumore stazionario bianco e cioè a valori incorrelati, tranne che questi non sono casuali ma *deterministici*, in modo che la loro ripetizione ciclica rende la sequenza PN riproducibile dal lato ricevente. La fig. 16.26-a) mostra una parte di un possibile segnale dati $p(t)$ pseudo casuale, bipolare, di durata $L \cdot T_p$, la cui espressione può essere posta nella forma

$$p(t) = \sum_{k=0}^{L-1} a_k g\left(t - kT_p - \theta\right) \tag{16.39}$$

con θ v.a. uniforme tra $\pm T_p/2$, basata sulla ripetizione di impulsi NRZ bipolari $g(t) = rect_{T_p}(t)$ di durata T_p con polarità stabilita dagli L valori a_k, scelti pari a ± 1 in modo da avvicinarsi alle condizioni[60]

- media nulla e varianza unitaria, cioè $m_A = 0$, $\sigma_A^2 = 1$;
- una autocorrelazione $\mathcal{R}_a(n)$ la più piccola possibile con $n \neq 0$, *mimando* così la proprietà di indipendenza statistica.

[60] Data la sequenza deterministica $a_k = \{a_0, a_1, \cdots a_{L-1}\}$ di lunghezza L, media e varianza sono definiti come

$$m_A = \frac{1}{L}\sum_{k=0}^{L-1} a_k, \qquad \sigma_A^2 = \frac{1}{L}\sum_{k=0}^{L-1}(a_k - m_A)^2$$

mentre l'autocorrelazione tra coppie di elementi a distanza n è definita da

$$\mathcal{R}_a(n) = \frac{1}{L-n}\sum_{k=0}^{L-n-1} a_k a_{k+n}$$

Considerando invece la sequenza periodica ottenuta ripetendo gli a_k, possiamo definire la stessa grandezza come

$$\mathcal{R}_a(n) = \frac{1}{L}\sum_{k=0}^{L-1} a_k a_{(k+n) \bmod L}$$

detta anche *autocorrelazione ciclica.*

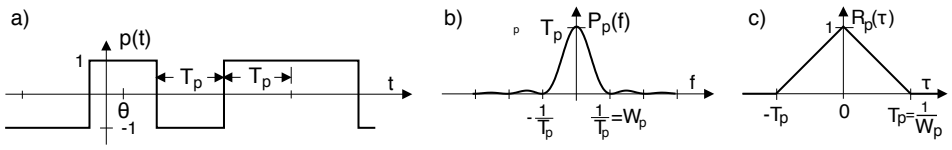


Figura 16.26: a) - sequenza pseudonoise; b) - densità di potenza; c) - autocorrelazione

Al § 7.7.4 abbiamo mostrato che un segnale simile a $p(t)$ ed espresso dalla (16.39), nel caso in cui gli a_k siano v.a. indipendenti a media nulla, presenta uno spettro di densità di potenza[61]

$$\mathcal{P}_p(f) = \sigma_A^2 \frac{\mathcal{E}_G(f)}{T_p} = T_p \text{sinc}^2\left(fT_p\right) \tag{16.40}$$

rappresentato in fig. 16.26-b), e per il quale la frequenza $W_p = \frac{1}{T_p}$ ne approssima l'occupazione di banda: prendiamo dunque questo risultato come una accettabile approssimazione per $p(t)$. Dalla (16.40) consegue che l'autocorrelazione di $p(t)$ si esprime come[62]

$$\mathcal{R}_p(\tau) = \mathcal{F}^{-1}\left\{\mathcal{P}_p(f)\right\} = \text{tri}_{2T_p}(\tau) \tag{16.41}$$

mostrata in fig. 16.26-c), e che appunto si azzera per $\tau \geq T_p$. Sebbene le sequenze pseudo-noise utilizzate realmente (§ 16.9.3) non aderiscano esattamente a queste caratteristiche, vi si avvicinano in modo soddisfacente per gli scopi delle telecomunicazioni.

Chip rate L'estensione temporale T_p di un simbolo di $p(t)$ è indicata come *periodo di chip*[63], e ci si riferisce ai suoi simboli come *chip*, per distinguerli dai bit; pertanto, la frequenza $f_p = W_p = 1/T_p$ è detta *chip rate*.

16.9.2 Modulazione per sequenza diretta

Ottiene l'espansione spettrale eseguendo il prodotto $\tilde{x}(t) = x(t)\, pn(t)$ tra un segnale di banda base $x(t)$ e la ripetizione ciclica del segnale PN $pn(t) = \sum_{i=-\infty}^{\infty} p\left(t - iLT_p\right)$, con il risultato di effettuare una modulazione AM-BLD-PS; l'operazione nel suo insieme prende il nome di *Direct Sequence Spread Spectrum* (o DSSS).

Sebbene l'effetto di espansione spettrale sia valido per $x(t)$ qualsiasi, affrontiamo l'analisi con riferimento ad un segnale $x(t)$ numerico binario NRZ antipodale ossia polare[64], il cui periodo di bit $T_b \gg T_p$ ne determina una densità di potenza $\mathcal{P}_x(f)$ del tipo di (16.40) ma con banda $W_x \ll W_p$. La fig. 16.27 illustra la situazione, facendo anche notare come scegliendo $T_b = LT_p$ e moltiplicando i 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, e con una banda che è quella di un segnale dati a frequenza $f_p = W_p \gg f_b$. Osserviamo che il

[61] Avendo scelto $g(t) = \text{rect}_{T_p}(t)$, risulta $G(f) = T_p \text{sinc}\left(fT_p\right)$ e quindi

$$\mathcal{E}_G = |G(f)|^2 = T_p^2 \text{sinc}^2\left(fT_p\right)$$

che diventa la (16.40) dato che $\sigma_A^2 = 1$.

[62] Applicando il teorema di Wiener si ottiene (vedi tabella a pag. 88)

$$\mathcal{R}_p(\tau) = \mathcal{F}^{-1}\left\{\mathcal{P}_p(f)\right\} = \mathcal{F}^{-1}\left\{T_p \text{sinc}^2\left(fT_p\right)\right\} = tri_{2T_p}(\tau)$$

[63] Oltre che indicare un circuito integrato, la parola *chip* è la stessa usata per le patine fritte olandesi, e prima ancora per *scheggia*, *frammento* o *truciolo*.

[64] Il *prodotto* tra due segnali dati di tipo *polare* a frequenza f_b e $f_p = Lf_b$, è equivalente a creare il segnale dati partendo dall'*or esclusivo* $\oplus$ delle corrispondenti rappresentazioni binarie fatte da zeri ed uni, come mostrato dalle tabelle poste a lato.

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

a	b	$a \cdot b$
-1	-1	1
-1	1	-1
1	-1	-1
1	1	1

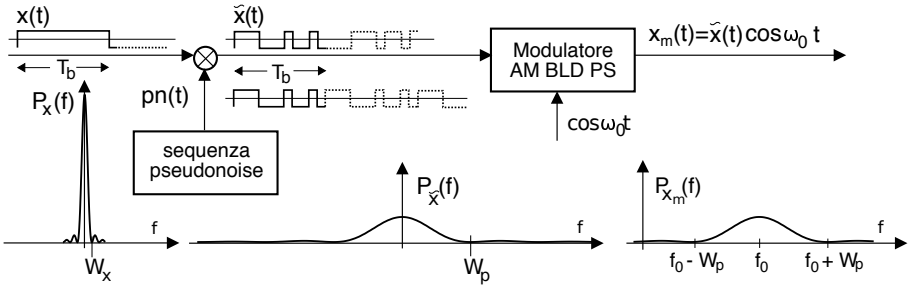


Figura 16.27: Generazione di un segnale modulato DSSS

segnale *allargato* $\tilde{x}(t)$ è così chiamato anche perché la potenza $\mathcal{P}_{\tilde{x}}$ è la stessa[65] $\mathcal{P}_x$ di $x(t)$, che ora risulta però *spalmata* sulla banda W_p di $pn(t)$.

L'effetto di espansione spettrale può essere verificato anche osservando che la densità di potenza $\mathcal{P}_{\tilde{x}}(f)$ è il risultato della convoluzione in frequenza[66]

$$\mathcal{P}_{\tilde{x}}(f) = \mathcal{P}_x(f) * \mathcal{P}_{pn}(f) \simeq \int_{-W_x}^{W_x} \mathcal{P}_x(\lambda)\,\mathcal{P}_{pn}(f-\lambda)\,d\lambda$$

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 $\mathcal{P}_{pn}(f-\lambda) \simeq \mathcal{P}_{pn}(f)$, e quindi

$$\mathcal{P}_{\tilde{x}}(f) \approx \left[\int_{-W_x}^{W_x} \mathcal{P}_x(\lambda)\,d\lambda\right] \mathcal{P}_{pn}(f) = \mathcal{P}_x \mathcal{P}_{pn}(f)$$

Infine, $\tilde{x}(t)$ è usato per modulare AM-BLD-PS una portante a frequenza f_0, producendo il segnale $x_m(t) = \tilde{x}(t)\cos 2\pi f_0 t$.

16.9.2.1 Guadagno di processo

E' il termine adottato per indicare il rapporto

$$G_p = \frac{W_p}{W_x} = \frac{T_b}{T_p} = \frac{f_p}{f_b} \tag{16.42}$$

tra la banda del segnale *allargato* e quella del segnale di partenza. Il *processing gain* varia tipicamente tra 10 e 10000 volte, ossia tra 10 e 40 dB, e come vedremo rappresenta una misura del miglioramento dell'SNR nel caso di presenza di segnali interferenti.

16.9.2.2 Despreading

Proseguiamo l'analisi considerando lo schema di ricevitore schematizzato in fig. 16.28, nella cui parte sinistra è mostrato il segnale modulato ricevuto $x_m(t) = \tilde{x}(t)\cos\omega_0 t$

[65] Considerando $x(t)$ realizzazione di un processo ergodico indipendente da $pn(t)$, la potenza di $\tilde{x}(t)$ risulta (§ 7.5.3) $\overline{\tilde{x}^2} = E\left\{x^2(t)\,pn^2(t)\right\} = \overline{x^2} = \mathcal{P}_x$, dato che dalla (16.41) si ha $E\left\{pn^2(t)\right\} = 1$.

[66] L'autocorrelazione del prodotto di processi indipendenti è pari al prodotto delle autocorrelazioni (§ 7.5.3), ed a questo si applica la proprietà di equivalenza tra prodotto nel tempo e convoluzione in frequenza, applicato alle trasformate delle autocorrelazioni, in base al teorema di Wiener (§ 7.2.1).

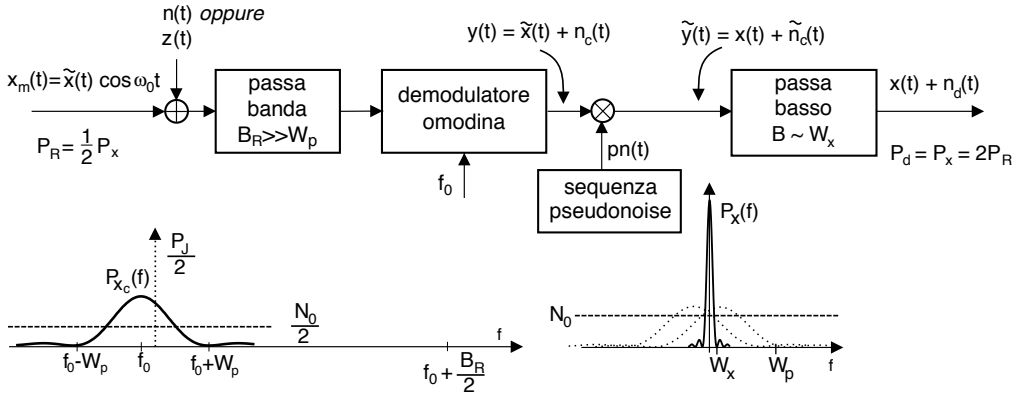


Figura 16.28: Ricevitore DSSS con rumore additivo $n(t)$ o interferenza $z(t)$

con potenza[67] $\mathcal{P}_R = \frac{1}{2}\mathcal{P}_x$, a cui si sovrappone un disturbo gaussiano $n(t)$ (od un interferente a banda stretta $z(t)$). Entrambi (segnale e disturbo) attraversano quindi il filtro passabanda di ricezione, caratterizzato da una banda di rumore $B_R \gg W_p \gg W_x$ in quanto deve lasciar passare l'intero spettro *allargato*, compresi i suoi lobi laterali. 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 disturbo. 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

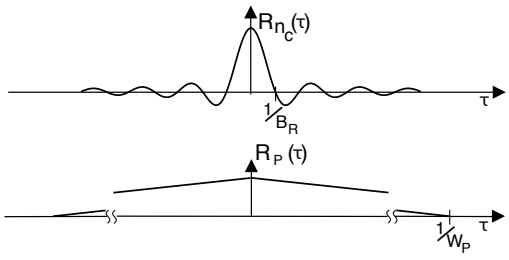
$$\tilde{y}(t) = [\tilde{x}(t) + n_c(t)]\, pn(t) = x(t)\, pn^2(t) + n_c(t)\, pn(t) = x(t) + \tilde{n}_c(t)$$

in virtù dei valori ± 1 assunti da $pn(t)$. Pertanto, mentre il messaggio $x(t)$ è tornato quello precedente all'allargamento, $n(t)$ e/o il disturbo $z(t)$ subiscono le *spreading* descritto al § 16.9.2. 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 $\mathcal{P}_d = \mathcal{P}_x = 2\mathcal{P}_R$, mentre per il termine di disturbo additivo $n_d(t)$ è stata rimossa la potenza che cade al di fuori della banda di segnale.

16.9.2.3 Prestazioni in presenza di rumore

La componente in fase (dopo demodulazione omodina) del rumore bianco $n(t)$ con densità di potenza $\mathcal{P}_n(f) = {}^{N_0}/_2$ ha densità $\mathcal{P}_{n_c}(f) = N_0 rect_{B_R}(f)$ (vedi § 14.1.3) e dunque autocorrelazione

$$\mathcal{R}_{n_c}(\tau) = N_0 B_R \mathrm{sinc}(B_R \tau)$$



Allo scopo di valutare la densità di potenza $\mathcal{P}_{\tilde{n}_c}(f)$ del rumore $\tilde{n}_c(t)$ 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

[67] $\tilde{x}(t)\cos(\omega_0 t + \varphi)$ 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, vedi § 7.5.3

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

$$\mathcal{P}_{\tilde{n}_c}(f) \simeq \mathcal{P}_{n_c}(f) = N_0 rect_{B_R}(f)$$

La componente di rumore $n_d(t)$ in uscita dall'ultimo passa basso con banda $\simeq W_x$ ha pertanto una potenza $N_d \simeq 2N_0W_x$, permettendo di valutare il rapporto segnale-rumore dopo demodulazione come

$$\left(\frac{\mathcal{P}_x}{\mathcal{P}_{n_c}}\right)_d = \frac{2\mathcal{P}_R}{2N_0W_x} = \frac{\mathcal{P}_R}{N_0W_x}$$

ossia proprio pari all'*SNR di sistema* (pag. 416), mostrando come la concatenazione delle operazioni di spreading e despreading *non alteri* le prestazioni del processo di modulazione nei confronti del rumore bianco.

16.9.2.4 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. Come caso limite consideriamo un *tono* interferente sinusoidale $z(t)$ (o *jammer*) centrato a frequenza $f_0 + f_z$ ossia

$$z(t) = \sqrt{2\mathcal{P}_j}\cos(\omega_0 + \omega_z)t$$

con potenza $\mathcal{P}_z = \mathcal{P}_j$. Dopo demodulazione si ottiene $z_c(t) = \sqrt{2\mathcal{P}_j}\cos\omega_z t$ e

$$\mathcal{P}_{z_c}(f) = \frac{\mathcal{P}_j}{2}\left[\delta(f - f_z) + \delta(f + f_z)\right] \tag{16.43}$$

Moltiplicando quindi il tono interferente demodulato $z_c(t)$ per $pn(t)$ come necessario per il despreading, si ottiene un disturbo $\tilde{z}_c(t)$ con densità di potenza $\mathcal{P}_{\tilde{z}_c}(f) = \mathcal{P}_{z_c}(f) * \mathcal{P}_p(f)$, mostrata alla riga centrale di fig. 16.29[68], permettendo di apprezzare l'effetto di *allargamento* subito dal tono interferente. 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 caso limite di $f_z = 0$. Pertanto il limite superiore della potenza interferente uscente dal passa basso con banda W_x è

$$\mathcal{P}_{z_d} = \int_{-W_x}^{W_x} \mathcal{P}_{\tilde{z}_c}(f)\,df \le 2W_x\frac{\mathcal{P}_j}{W_p}$$

e dunque il rapporto segnale-interferente diviene

$$\left(\frac{\mathcal{P}_x}{\mathcal{P}_{z_d}}\right)_d \ge 2\mathcal{P}_R\frac{W_p}{2W_x\mathcal{P}_j} = \frac{\mathcal{P}_R}{\mathcal{P}_j}\frac{W_p}{W_x}$$

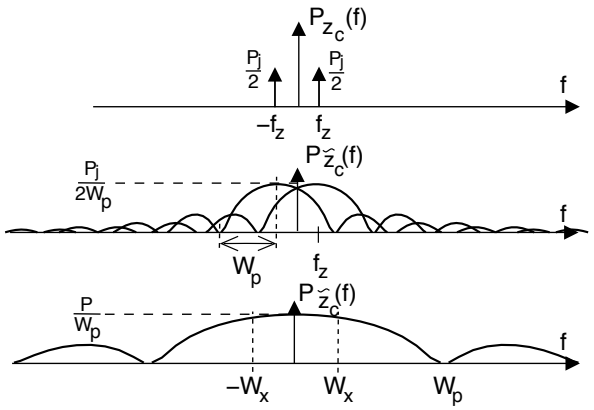


Figura 16.29: Despreading di un tono interferente

[68] Il risultato si ottiene tenendo conto delle eq. (16.40) e (16.43), effettuando la convoluzione, e ricordando che $T_p = 1/W_p$.

che rappresenta un miglioramento esattamente pari al guadagno di processo G_p, eq. (16.42).

16.9.2.5 Accesso multiplo CDMA

Una frequente applicazione della tecnica spread spectrum è quella di permettere la comunicazione *contemporanea* di una pluralità di soggetti, possibile qualora ognuno di essi adotti una diversa sequenza PN: tale approccio prende il nome di CDMA (*Code Division Multiple Access*). Mostriamo ora che con questo approccio ogni comunicazione subisce (a causa delle altre) solo un modesto innalzamento del rumore di fondo, tanto più piccolo quanto minore è il valore della *intercorrelazione* tra i codici PN utilizzati. Indichiamo con $z(t)$ il termine interferente (dopo demodulazione) dovuto alla presenza di N diversi utenti, ognuno con un diverso codice $pn_i(t)$ e segnale dati $x_i(t)$, che può essere scritto come

$$z(t) = \sum_{i=1}^{N} A_i x_i(t - \tau_i)\, pn_i(t - \tau_i) \cos\theta_i$$

in cui A_i, τ_i e $\cos\theta_i$ sono rispettivamente ampiezza, ritardo di simbolo e fase della portante relativi all'i-esimo utente. Assumendo ora eguali tra loro le ampiezze del segnale utile $x(t)$ e degli interferenti, dopo il despreading otteniamo

$$\tilde{y}(t) = x(t) + \left[\sum_{i=1}^{N} x_i(t - \tau_i)\, pn_i(t - \tau_i) \cos\theta_i\right] pn(t)$$

Se realizziamo ora il filtro passa basso di fig. 16.28 come un integratore esteso ad un periodo di bit, ovvero un filtro adattato al segnale NRZ[69], il valore della sua uscita campionata al termine della durata del k-esimo periodo di bit risulta

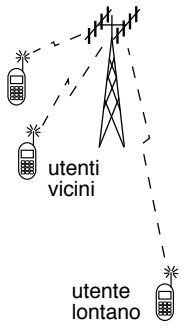
$$\begin{aligned} d(kT_b) &= T_b x(kT_b) + \sum_{i=1}^{N} \left[\cos\theta_i \int_{(k-1)T_b}^{kT_b} x_i(t - \tau_i)\, pn_i(t - \tau_i)\, pn(t)\, dt\right] \\ &= T_b x(kT_b) + z_d(kT_b) \end{aligned}$$

in cui $z_d(kT_b)$ rappresenta il termine di interferenza complessiva da parte di tutti gli altri N utenti, indicata come *interferenza multiutente* o MUI (multi-user interference). Dato che i valori di x_i possono essere ± 1, l'integrale calcola in effetti *l'intercorrelazione* $\mathcal{R}_{p_0 p_i}(\tau_i)$ (§ 7.1.4) tra la sequenza PN usata per la propria trasmissione e le sequenze PN usate dagli altri, calcolata per un ritardo τ_i. Pertanto, scegliendo la famiglia di sequenze pseudo-noise in modo che esibiscano una intercorrelazione molto ridotta (in teoria nulla, se le PN fossero esattamente *ortogonali*), l'effetto degli interferenti si riduce in egual misura.

Controllo di potenza Qualora un utente di un sistema CDMA sia sensibilmente più lontano dal ricevitore rispetto agli altri, se tutti trasmettono con la stessa potenza l'attenuazione subìta dal segnale dell'utente lontano fa si che il termine MUI aumenti

[69] Eventualmente realizzato come descritto a pag. 215, supponendo inoltre che sia verificata la condizione di sincronizzazione temporale.

di importanza, anche in presenza di intercorrelazione bassa, causando un importante degrado della qualità della trasmissione. Questo fenomeno è indicato come *effetto near-far*. Per ovviare al problema un sistema CDMA viene usualmente corredato di un meccanismo di *controllo di potenza*, espletato dalla stazione radio base[70], che misurando la potenza ricevuta da ciascun utente, ne richiede la diminuzione ai vicini e/o l'aumento ai lontani, in modo da ricevere la medesima potenza da ciascuno di essi.



Prestazioni multi-utente con PN incorrelate Consideriamo il caso in cui le trasmissioni CDMA di K diversi utenti siano tutte ricevute con la medesima potenza $\mathcal{P}_x$, e le sequenze PN utilizzate da ciascuno di essi abbiano una intercorrelazione nulla. Allora, per una generica trasmissione la potenza interferente $\mathcal{P}_{n_d}$ risulta ridotta rispetto a quella effettivamente ricevuta di una quantità pari al guadagno di processo, e quindi il *rapporto segnale-interferenza* (indicato come SIR) risulta circa pari a[71]

$$SIR = \frac{\mathcal{P}_x}{\mathcal{P}_{n_d}} = \frac{\mathcal{P}_x}{(K-1)\,\mathcal{P}_x/G_p} = \frac{G_p}{K-1}$$

Dato che le PN effettivamente utilizzate *non presentano* intercorrelazione nulla, il risultato mostrato costituisce una approssimazione limite rispetto alla quale valutare la qualità delle prestazioni effettivamente ottenute. Nel caso di trasmissione a due livelli la prob. di errore (16.7) minima (a causa degli interferenti) diviene quindi

$$P_e^{BPSK}(bit) = \frac{1}{2}\mathrm{erfc}\left\{\sqrt{SIR}\right\} = \frac{1}{2}\mathrm{erfc}\left\{\sqrt{\frac{G_p}{K-1}}\right\}$$

Infine, per tener conto allo stesso tempo sia dell'effetto degli interferenti che del rumore gaussiano comunque presente, può essere usato il *rapporto segnale-rumore più interferente* (o SINR) definito come

$$SINR = \frac{\mathcal{P}_x}{\mathcal{P}_{n_d} + \mathcal{P}_n} = \left(\frac{\mathcal{P}_{n_d} + \mathcal{P}_n}{\mathcal{P}_x}\right)^{-1} = \left(\frac{K-1}{G} + \frac{N_0}{E_b}\right)^{-1}$$

ossia pari *al parallelo* degli SNR, come discusso a pag. 244.

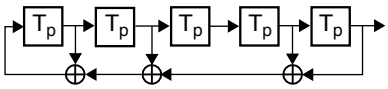
Esempio In un sistema CDMA-DSSS si desidera una $P_e = 10^{-6}$, a cui la tecnica di modulazione adottata fa corrispondere un $E_b/N_0 = 13$ dB. Trascurando il rumore termico, determinare il massimo numero K di utenti contemporaneamente attivi se $G_p = 30$ dB.
Imponendo $E_b/N_{0I} = G_p/(K-1) = 10^{1.3} = 20$ si ottiene $K = (G_p{+}20)/20 = 1020/20 = 51$.

16.9.3 Sequenze pseudo casuali

Accenniamo brevemente ad alcune tipologie di sequenze pseudo noise.

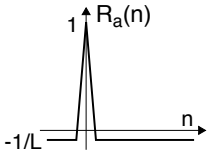
[70] Ossia l'antenna con cui tutti telefonini nella medesima cella sono in comunicazione.

[71] In tal caso infatti i $K-1$ interferenti sono assimilabili ad un rumore gaussiano (in virtù del teorema centrale del limite) con potenza complessiva $(K-1)\,\mathcal{P}_x$ e limitato in banda alla stessa banda W_p del segnale utile. Dopo il despreading la densità spettrale interferente $N_{0I}/2$ *si allarga* su di una banda $G_p W_p$, e si riduce di ampiezza dello stesso fattore G_p. Pertanto il filtro passa basso a valle del despreading lascia passare una potenza interferente pari a $(K-1)\,\mathcal{P}_x/G_p$.



Sequenze di massima lunghezza Una prima possibilità è quella delle sequenze-m, o *di massima lunghezza*, ottenute mediante dei registri a scorrimento controreazionati[72] con m ritardi, simili a quelli discussi a proposito del CRC (pag. 480) ma con la struttura mostrata in figura, in cui non è presente nessun ingresso esterno ed il bit che *rientra* è calcolato in base all'OR esclusivo di una combinazione di bit di stato. Dato che con m bit si ottengono 2^m configurazioni dello stato, ma che quella *tutti zeri* arresterebbe il processo di generazione, le sequenze di *massima lunghezza*[73] sono composte da $L = 2^m - 1$ bit (ognuno dei quali corrisponde ad una diversa configurazione dello stato) che si ripetono ciclicamente, e sono ottenute per particolari scelte[74] di quali bit far partecipare alla controreazione.

Tra le proprietà positive di questa famiglia annotiamo la *quasi equiprobabilità* dei bit uno e zero, la equa distribuzione delle sequenze di bit uguali[75], ed una *autocorrelazione ciclica* $\mathcal{R}_a(n) = \frac{1}{L}\sum_{k=0}^{L-1} a_k a_{(k+n) \bmod L}$ che vale 1 per $n = 0$ ed $-1/L$ altrimenti[76] (vedi figura). D'altra parte, l'*intercorrelazione* ciclica tra due diverse sequenze-m (di uguale lunghezza L) presenta valori massimi che sono una percentuale apprezzabile di $\mathcal{R}_a(0)$, rendendo necessario individuare altre soluzioni per i casi di accesso multiplo.



Sequenze di Gold e Kasami Le sequenze *di Gold* si ottengono eseguendo l'OR esclusivo bit a bit di due diverse[77] sequenze-m **a** e **b** di uguale lunghezza L; ripetendo il procedimento per tutti i $2^m - 1$ possibili scorrimenti temporali di **b** rispetto ad **a**, ed includendo **a**, si ottengono 2^m diverse sequenze, con una intercorrelazione massima pari a $\sqrt{2/L}$. Una soluzione lievemente diversa è quella *di Kasami*, in cui una delle due sequenze-m di partenza viene *decimata* ciclicamente, e che produce $2^{m/2}$ sequenze, con intercorrelazione massima pari a $1/\sqrt{L}$.

Sequenze di Walsh-Hadamard[78] Si tratta di sequenze ortogonali, ovvero per le quali risulta $\sum_{k=0}^{L-1} a_k b_k = 0$, ossia **a** e **b** sono sequenze *incorrelate* qualora allineate, e che sono generate mediante l'algoritmo iterativo schematizzato alla figura che segue, che individua un numero di $L = 2^m$ sequenze, di lunghezza L.

[72] Vedi `https://en.wikipedia.org/wiki/Linear-feedback_shift_register`

[73] In quanto in linea di principio il periodo della sequenza può essere inferiore al massimo.

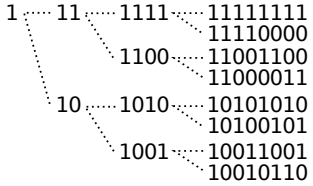
[74] Anche in questo caso come al § 15.6.3.3 la posizione degli XOR può essere associata ad un polinomio generatore, e per produrre una sequenza di massima lunghezza occorre scegliere un *polinomio primitivo*, vedi `https://en.wikipedia.org/wiki/Primitive_polynomial_(field_theory)`. A parità di m, cambiando polinomio si ottengono sequenze differenti ma di uguale lunghezza, ed il loro numero massimo N aumenta all'aumentare di m con legge $N = {}^{(2^m-2)}/_m$.

[75] Indicando con *run* una sequenza di bit uguali, su $2^m - 1$ bit si trova un run di uni lungo m, un run di zeri lungo $m - 1$, e quindi 2^{m-i-2} run sia di zeri che di uni di lunghezza i, per $1 < i \leq m - 2$.

[76] L'autocorrelazione si intende calcolata a partire da valori *bipolari*, ossia ottenuti a partire dalla sequenza binaria facendo corrispondere ± 1 ai valori 0, 1, vedi nota 64.

[77] La coppia di sequenze-m non è qualsiasi, ma scelta tra quelle con una intercorrelazione massima ridotta, chiamate *sequenze preferite*.

[78] `https://en.wikipedia.org/wiki/Hadamard_code`



Possono dunque essere usate nel contesto di un sistema di accesso multiplo qualora gli apparati possano essere sincronizzati tra loro, come per il collegamento *in discesa* tra una stazione radio base ed i terminali radiomobili associati ad essa[79]. Il lato meno positivo di queste sequenze è una autocorrelazione che presenta diversi picchi secondari, e dunque non sono idonee ad assolvere la funzione di sincronizzazione (§ 16.11.1). D'altra parte, la proprietà di ortogonalità può altresì essere sfruttata per realizzare una *segnalazione* ortogonale (§ 7.6.2) nel contesto di una comunicazione punto-punto.

Sequenze di Barker Presentano valori di autocorrelazione (*non ciclica*)

$$\mathcal{R}_a(n) = \frac{1}{L} \sum_{m=0}^{L-1-|n|} a_m b_{m+|n|}$$

con valori $\mathcal{R}_a(0) = 1$ e $|\mathcal{R}_a(n)| \leq 1/L$ per $1 \leq n < L$, e come le sequenze-m esibiscono buone proprietà rispetto al bilanciamento ed alle corse. L'aspetto negativo è che la massima lunghezza di sequenza conosciuta è $L = 13$, e con questa lunghezza, ne esiste solo una! Nonostante ciò, sono utilizzate ad esempio nei sistemi di accesso WiFi.

16.9.4 Frequency Hopping

Si tratta di una diversa tecnica *spread spectrum*, in cui la sequenza PN è di tipo *multilivello*, ed è utilizzata in uno schema L-FSK incoerente (§ 16.5) per cambiare in continuazione la frequenza portante a cui avviene la trasmissione, tipicamente FSK anch'essa (vedi figura 16.30), da cui il nome di *saltando di frequenza* (traduzione letterale di Frequency Hopping). Per una corretta ricezione, è necessaria una accurata sincronizzazione temporale tra la PN usata in trasmissione e quella in ricezione.

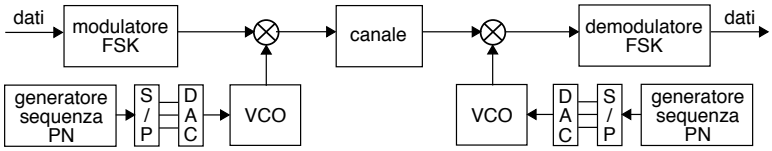


Figura 16.30: Schema di una trasmissione Frequency Hopping

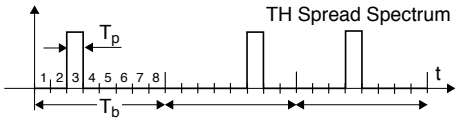
Anche in questo caso si verifica un fenomeno di espansione spettrale, ma stavolta non tutta la banda è occupata in modo *permanente* come nel DSSS, ma anzi durante ogni *salto* si occupa solo la banda necessaria alla modulazione *non allargata*. In questo caso un disturbo a banda stretta provoca interferenza solo durante il salto che occupa la sua stessa frequenza, e dunque può essere facilmente contrastata adottando una codifica di canale (§ 17.4). 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.

[79] Ma in tal caso, anziché *accesso multiplo*, potremmo definire la modalità di trasmissione come un *broadcast ortogonale*.

16.9.5 Time Hopping o UWB

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 UWB[80]); 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 PN[81]. La figura a lato mostra un segnale THSS di banda base, in cui per ogni bit viene trasmesso un chip con $T_p \ll T_b$, posizionato su (ad es.) una di otto possibili posizioni, in maniera pseudo casuale.

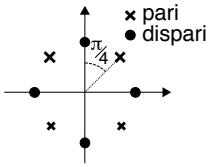


16.10 Altre possibilità

Diamo un accenno ad altre tecniche di modulazione numerica, che non sono state sviluppate in questo capitolo.

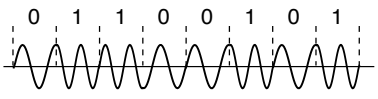
Offset keying[82] Una variante del QPSK detta OQPSK, in cui la temporizzazione dei rami I e Q viene *sfasata* di metà del periodo di simbolo, in modo che la fase dell'inviluppo complesso non vari per più di $\pi/2$ ogni T_s, e dunque il modulo dell'inviluppo complesso non può più annullarsi: ciò si traduce in una dinamica delle ampiezze ridotta, riducendo così i problemi legati alla distorsione non lineare (§ 8.3).

Modulazione $\pi/4$ Un'altra variante del QPSK, in cui ogni simbolo viene mappato alternativamente su due costellazioni QPSK *ruotate* di $\pi/4$, in modo che i possibili salti di fase tra simboli contigui possono essere di ±45ř e ±135ř anziché 90 e 180 come nel QPSK. L'alternanza tra le due costellazioni avviene in base ad una modulazione differenziale, realizzata associando ad ogni simbolo la rotazione riportata in tabella[83]. La trasmissione può quindi essere ricevuta adottando un ricevitore in fase e quadratura, dato che per decidere il simbolo ricevuto non è necessario un riferimento di fase assoluto, ma ci si basa su quello del simbolo precedente. Inoltre, la sincronizzazione è semplificata in quanto c'è un cambio di fase ad ogni simbolo, anche qualora siano ...tutti uguali.



$a_k a_{k-1}$	$\Delta\theta$
11	45ř
01	135ř
00	-135ř
10	-45ř

Minimum shift keying - MSK[84] Individua una modulazione FSK binaria *coerente* in cui l'intervallo di frequenza Δ assume il *minimo* valore $\frac{1}{2T_s}$ (vedi § 16.12.1), ottenendo un segnale modulato che mantiene una *continuità di fase* tra bit contigui, come mostrato in figura in cui $f_0 = 1.25 \cdot f_b$ e $f_1 = 1.75 \cdot f_b$, e quindi



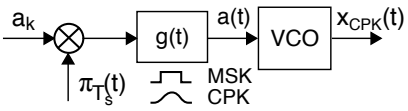
[80] http://en.wikipedia.org/wiki/Ultra-wideband
[81] http://en.wikipedia.org/wiki/Time-hopping
[82] https://en.wikipedia.org/wiki/Phase-shift_keying#Variants
[83] Dunque ad esempio la sequenza 001011 produce una sequenza di fasi −135ř, −180, −135ř
[84] https://en.wikipedia.org/wiki/Minimum-shift_keying

$\Delta = 0.5 f_b = {}^1/_{2T_b}$: questa caratteristica consente una riduzione della banda occupata, in virtù dell'assenza di brusche variazioni di ampiezza.

Modulazione a fase continua - CPK[85] Come per il caso precedente si realizza un segnale privo di discontinuità, facendo evolvere la fase dell'inviluppo complesso *con continuità* tra il valore iniziale e quello finale, nell'arco di un periodo di simbolo. Per fissare le idee, consideriamo l'uscita di un VCO

$$x_{CPK}(t) = A \cdot \sin\left(2\pi f_0 t + 2\pi\Delta \int_0^t a(\tau)\, d\tau\right)$$

alimentato da un segnale dati $a(t) = \sum_{k=1}^{L} a_k g(t - kT_s)$ come in figura.



Scegliendo $g(t) = \text{rect}_{T_b}(t)$ ed $a_k \in \{0, 1\}$ si ottiene l'MSK, e la variazione della fase avviene in modo lineare (integrale di un rettangolo) realizzando una FM; nel CPK si utilizza invece un sagomatore privo di discontinuità (ad es. un coseno rialzato nel tempo, ovvero $g(t) = \frac{1}{2}(1 + \cos 2\pi {}^t/_{T_s}) \text{rect}_{T_b}(t)$), realizzando un segnale modulato (per così dire) sia in fase che in frequenza. Questa ulteriore *dolcezza* dell'inviluppo complesso[86] determina un ulteriore risparmio di banda; d'altra parte la *fase di partenza* per ciascun simbolo dipende da quelli precedentemente trasmessi, e dunque il segnale deve essere decodificato nella sua interezza e non simbolo per simbolo.

Risposta parziale[87] Si tratta ancora di una modulazione angolare e dunque adotta lo schema mostrato sopra, ma stavolta $g(t)$ ha una durata maggiore di T_s: ciò riduce ancor di più la banda, ma introduce interferenza intersimbolica (ISI) in modo *controllato*, nel senso che *è noto* come i simboli precedenti incidono sul valore dell'attuale. Per questo, la decodifica può avvenire mediante una MLSD (§ 18.4.5), con un lieve peggioramento di prestazioni (in presenza di rumore) rispetto alla assenza di ISI. In questa categoria rientra il GMSK[88], una forma di MSK a risposta parziale in cui $g(t)$ ha un andamento *gaussiano* (ma ovviamente troncato nel tempo), e che è utilizzato diffusamente (GSM, 802.11 FHSS, BLUETOOTH) in virtù della ridottissima occupazione spettrale[89].

Modulazione codificata a traliccio - TCM[90] E' una tecnica che combina la codifica di canale (§ 17.4) con il processo di modulazione, e che anziché aumentare il numero di bit da trasmettere e quindi la banda, aumenta il *numero di punti* di costellazione per simbolo. A prima vista ciò comporterebbe un peggioramento di prestazioni, ma queste sono compensate dal guadagno di codifica associato alla ridondanza introdotta, che si manifesta in un vincolo sui possibili valori della sequenza di simboli, che viene a dipendere anche dal valore di un certa quantità di bit precedenti.

[85] `http://en.wikipedia.org/wiki/Continuous_phase_modulation`

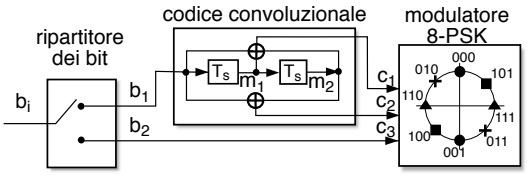
[86] Dato che a differenza di MSK, tra due simboli anche la derivata di $\underline{x}_{CPK}(t)$ è continua.

[87] Vedi ad es. `http://complextoreal.com/wp-content/uploads/2013/01/qpr.pdf`

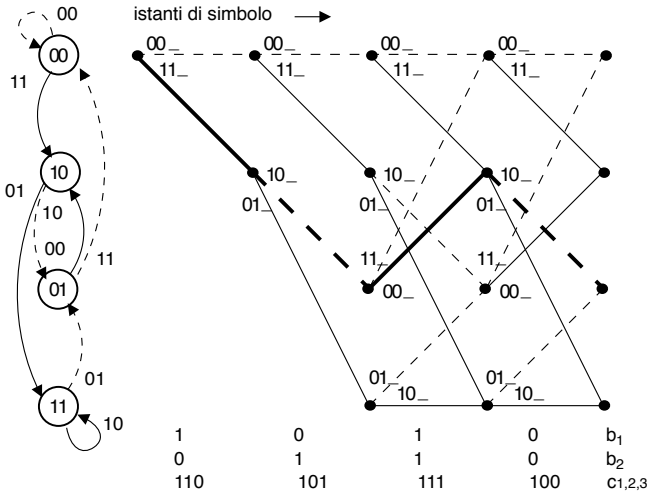
[88] Vedi ad es. T. Turletti, *GMSK in a nutshell*, `Citeseerx` 1996

[89] Un po' come realizzare un segnale dati *a banda minima*, senza ricorrere ad un passa basso ideale.

[90] Vedi ad es. `http://complextoreal.com/wp-content/uploads/2013/01/tcm.pdf`



Per fissare le idee riferiamoci alla figura a lato: i bit in arrivo b_i sono inviati alternativamente (b_1) ad un codificatore convoluzionale (2,1,2) (§ 17.4.2) con tasso $R_c = \frac{1}{2}$, e (b_2) ad un modulatore 8-PSK. In questo esempio il codificatore è realizzato come in fig. 17.4 a pag. 580, e dunque produce due bit c_1c_2 in uscita per ognuno che ne entra, in funzione di due bit precedenti m_1m_2 (o *di stato*); anche questo risultato viene inviato al modulatore 8-PSK, le cui 8 possibili fasi sono state *suddivise* in quattro partizioni come in figura, ed assegnate alle configurazioni binarie (non di Gray) indicate, in modo che per ogni partizione il bit di ingresso non codificato $c_3 = b_2$ sceglie uno tra due punti *più distanti* possibile.



Le possibili sequenze $c_{1,2}$ in uscita dal codificatore sono schematizzate mediante l'automa mostrato alla figura a lato, in cui gli stati sono etichettati con il valore dei bit $m_{1,2}$, e le transizioni con il codice di uscita $c_{1,2}$ corrispondente all'ingresso b_1 pari a zero (linea a tratti) o uno (linea continua), così come calcolate in base agli EX-OR. Le possibili sequenze $c_{1,2,3}$ in ingresso al modulatore sono descritte da tutti i percorsi di attraversamento del *traliccio*[91] disegnato in figura a destra dell'automa, in cui le righe corrispondono allo stato $m_{1,2}$ e le colonne agli istanti di simbolo, mentre le transizioni (continue o tratteggiate a seconda se b_1 è 1 o 0) sono etichettate con i bit associati alla costellazione 8-PSK, a meno del bit b_2 che è indicato dal sottolineato_. In definitiva, la sequenza di simboli PSK corrispondente ad un ingresso (ad es.) 10 01 11 00 è 110 101 111 100, come mostrato in basso in figura, e corrispondente alla linea spessa.

La ricezione di questo segnale si svolge in due fasi: nella prima si individua il punto di costellazione, per ogni istante di simbolo k ed ogni partizione p, più vicino al segnale ricevuto $\underline{r}_k$, e si valuta la relativa *verosimiglianza logaritmica* $-\log\left[p\left(\underline{r}_k/p\right)\right]$, riportandone il valore sull'arco del traliccio a cui si riferisce (vedi anche §§ 17.4.2.3 e 18.4.5). Nella seconda si individua nel traliccio il percorso di minimo costo mediante l'applicazione dell'algoritmo di Viterbi, illustrato a pag. 581.

Notiamo come le scelte fatte abbiano messo in corrispondenza le transizioni uscenti (od entranti) in uno stesso stato con costellazioni disposte *a croce*, ovvero con la massima

[91] TRELLIS in inglese, da cui il nome del metodo *trellis coded modulation*.

distanza tra i punti. Per ogni periodo di simbolo solo 4 delle 8 fasi sono possibili. Si può mostrare che il semplice schema dell'esempio permette un miglioramento di E_b/N_0 di 3 dB, e quasi altri 3 possono essere aggiunti per codificatori di maggior complessità.

Sistemi MIMO[92] 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* (§ 20.3.3.1) in grado di aumentare la capacità di un collegamento radio grazie allo sfruttamento del fenomeno dei *cammini multipli*. Approfonditi al cap. 21.

Radio Cognitiva[93] Attiene a come trasmettere *intelligentemente* in modo da usare i migliori canali radio a disposizione, in presenza di altre comunicazioni contemporanee: 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.

16.11 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 simbolo[94], ed evitare l'insorgenza di ISI. Le due problematiche possono essere affrontate l'una di seguito all'altra, adottando le soluzioni già esposte[95]. D'altra parte sono ora possibili varianti, come ad esempio il *Costas loop*[96] 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 demodulazione[97].

Qualora la portante di demodulazione presenti una ambiguità di fase residua si può applicare la tecnica della codifica differenziale esposta al § 16.4, oppure inserire una sequenza di simboli noti (o *flag*) all'inizio della trama trasmissiva, in modo che il confronto tra i valori previsti 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 § 15.7.2.1.

[92] Vedi ad es. `http://complextoreal.com/wp-content/uploads/2013/01/mimo.pdf`

[93] `http://en.wikipedia.org/wiki/Cognitive_radio`

[94] Od alla fine, come nel caso di un ricevitore a correlazione, o basato su di un filtro adattato.

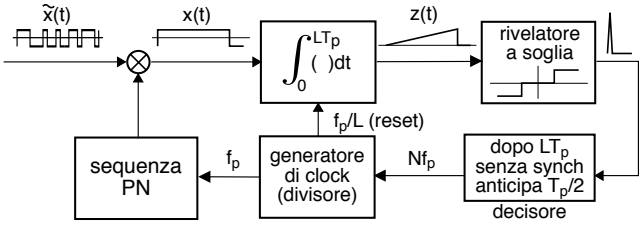
[95] Per il recupero della portante si possono usare circuiti del tipo di § 12.2.2.1, mentre l'uso del PLL (§ 12.2.2.2) non è possibile a causa della assenza di residui di portante. Una volta acquisto il sincronismo di frequenza, quello di simbolo può essere ottenuto mediante schemi operanti in banda base, come quelli al § 15.7.2.1.

[96] `http://en.wikipedia.org/wiki/Costas_loop`

[97] Vedi as es. `http://en.wikipedia.org/wiki/Carrier_recovery#Decision-directed`

16.11.1 Sincronizzazione per sistemi a spettro espanso

In questo caso occorre considerare anche la modalità di acquisizione della fase della sequenza PN necessaria al despreading, che avviene in due passi: il primo effettua una ricerca sequenziale ed ha una precisione di metà del periodo di chip T_p, mentre il secondo riduce l'errore mediante una tecnica a controreazione, e lo mantiene sotto controllo.



Schema di un *sliding correlator*

Considerando già avvenuta la sincronizzazione di portante e la demodulazione, il primo passo può essere attuato mediante lo schema in figura detto *sliding correlator*, in cui l'integratore (ad es. un integrate and dump) calcola la correlazione tra la PN in arrivo e quella generata localmente. In caso di coincidenza temporale il risultato $z(t)$ è una rampa (negativa se il bit fosse stato -1), mentre per un ritardo θ tra la seq. locale e quella ricevuta il risultato dell'integrazione all'istante LT_p è l'autocorrelazione $R_{\tilde{x}}(\theta)$, più termini di rumore a valor atteso nullo. Alla conclusione del periodo di bit $T_b = LT_p$ il generatore di clock resetta l'integratore, non prima però che il rivelatore a soglia abbia prodotto un impulso di sincronismo, inibendo il componente di decisione e terminando la ricerca. In assenza di sincronizzazione, il decisore fa invece avanzare la fase della PN di $T_p/2$, tentando di nuovo l'aggancio. Il metodo richiede dunque $2L^2T_p$ secondi (nel caso peggiore), tempo che può essere ridotto eseguendo più ricerche in parallelo.

Dopo aver acquisto la sincronizzazione *lasca* il controllo della generazione della PN passa ad un circuito indicato come *delay locked loop* o DLL, che può essere realizzato in diversi modi, e di cui discutiamo il funzionamento dello schema *canonico* mostrato in fig. 16.31.Un generatore pseudonoise produce la sequenza PN_0 affetta dall'errore di temporizzazione residuo $-T_p/2 \leq |\theta| \leq T_p/2$, che ai fini della discussione assumiamo come un ritardo, e dunque scriviamo $\text{PN}_0 = \tilde{x}(t-\theta)$; lo scopo del DLL è di rendere $\theta = 0$. Da PN_0 sono derivate due nuove sequenze, $\text{PN}_{early} = \tilde{x}(t-\theta+\delta)$ e $\text{PN}_{late} = \tilde{x}(t-\theta-\delta)$, che differiscono da PN_0 rispettivamente per un anticipo od un ritardo $\delta < T_p/2$, e che sono moltiplicate per il segnale allargato di banda base $\tilde{x}(t)$. L'integratore del ramo superiore (*early*) calcola pertanto l'autocorrelazione[98]

$$R_{\tilde{x}}(\delta-\theta) = \int_0^{LT_p} \tilde{x}(t)\,\tilde{x}(t-\theta+\delta)\,dt$$

tra $\tilde{x}(t)$ e la sua copia *in anticipo* di $\delta - \theta$, mentre quello inferiore (*late*) calcola $R_{\tilde{x}}(-\delta-\theta)$ tra $\tilde{x}(t)$ e la sua copia *in ritardo* di $\delta + \theta$ (vedi lato destro di fig. 16.31): questi due risultati sono campionati[99] a fine sequenza, ed indicati con z_{early} e z_{late};

[98] L'esempio si riferisce ad una sequenza PN di massima lunghezza, della cui autocorrelazione si è discusso a pag. 537.

[99] Il blocco che valuta il valore assoluto dell'uscita dell'integratore è necessario in quanto $\tilde{x}(t)$ trasporta anche l'informazione $x(t)$, che determina l'eventuale inversione di segno della sequenza PN, e dunque un

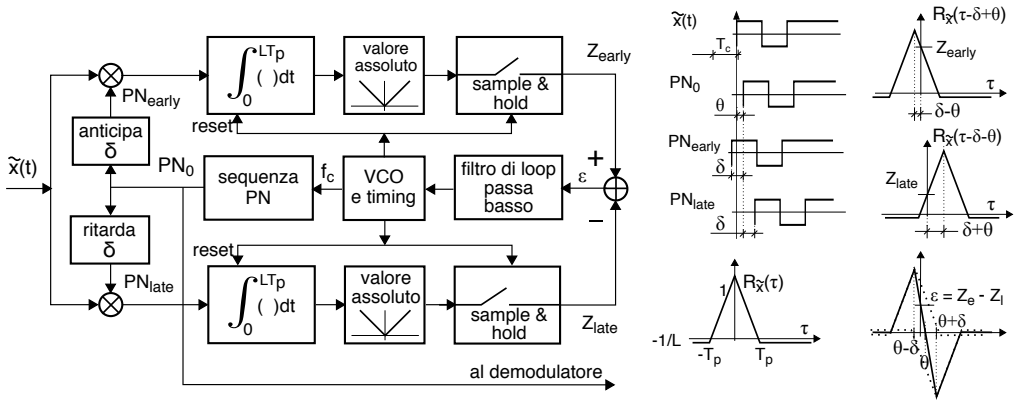
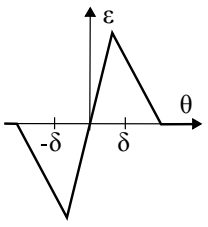


Figura 16.31: Delay locked loop e forme d'onda relative

notiamo che se $\theta = 0$, si ha $z_{early} = z_{late} = R_{\tilde{x}}(\delta)$.

Proseguendo con l'esempio, osserviamo come l'evenienza $z_{early} > z_{late}$ indichi che la fase di PN_{early} è più prossima a quella del segnale ricevuto di quanto non lo sia quella di PN_{late}, e dunque PN_0 è in ritardo, ovvero stiamo andando... troppo piano[100]. In questo caso la differenza $\varepsilon = z_{early} - z_{late}$ risulta positiva (vedi costruzione grafica in basso a ds in fig. 16.31), e ciò determina (attraverso il filtro di loop) una *accelerazione* del VCO, che tende quindi a ridurre il ritardo iniziale θ. Se si fosse verificato l'opposto ($\theta < 0$ e dunque $z_{early} < z_{late}$) sarebbe risultato $\varepsilon < 0$, a cui corrisponde un *rallentamento* del VCO.

Man mano che θ si riduce, PN_{early} e PN_{late} tendono a disporsi simmetricamente in anticipo ed in ritardo di δ rispetto a PN_0, in modo che z_{early} e z_{late} prendono il medesimo valore e la loro differenza ε si annulla, la f_c prodotta dal VCO è costante, e PN_0 è esattamente in fase con $\tilde{x}(t)$. Mentre θ si azzera, la costruzione grafica in basso a destra di fig. 16.31 *transla* verso sinistra, in modo che corrisponda $\varepsilon = 0$ per $\theta = 0$. Ciò consente di riformulare l'approccio grafico come nella figura a lato, che rappresenta l'errore ε in funzione del disallineamento θ, detta anche *curva di discriminazione* o *curva-s*, e che evidenzia l'intervallo $|\theta| < \delta$ entro il quale il sistema converge. Questa figura è simile a quella discussa al § 12.2.2.2 a proposito del PLL, ed infatti l'analisi delle prestazioni ha molto in comune. In particolare una diminuzione della banda del filtro di loop, se da un lato riduce le capacità di inseguire rapide variazioni di ritardo (come potrebbe essere per effetto doppler), d'altro canto attenua l'influenza del rumore sulla varianza della stima di θ.



Una alternativa possibile opera direttamente sul segnale modulato, e per questo indicata come DLL *incoerente*, non necessitando della sincronizzazione di portante. Al posto dell'integratore utilizza un filtro passabanda centrato sulla portante, seguito da un demodulatore di inviluppo. Una ulteriore variante, indicata come *tau-dither loop*,

cambiamento di segno per l'uscita dell'integratore.

[100] In altri termini, le tre copie della PN ($_0$, $_{early}$ e $_{late}$) dovrebbero *slittare* a sinistra, e quindi il periodo della PN deve essere *ridotto*.

affronta e risolve il problema legato ad un possibile diverso guadagno tra i due rami del DLL.

16.12 Appendici

16.12.1 Ortogonalità tra simboli sinusoidali

Al § 16.5.1 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

$$s_k(t) = \cos\left[2\pi\left(f_0 + k\cdot\Delta\right)t + \phi_k\right] \quad \text{con } k \text{ intero}$$

separati da un intervallo di frequenza Δ, in cui è inclusa una differenza (o errore) di fase aleatorio ϕ_k tra le forme d'onda, in modo da esaminare le differenze tra il caso di modulazione coerente ed incoerente, ovvero con $\phi_k = 0\ \forall k$, oppure ϕ_k v.a. incorrelate uniformi tra 0 e 2π.

Iniziamo dunque con lo sviluppare l'espressione dell'integrale di intercorrelazione tra due di questi segnali

$$\rho = \int_0^{T_s} s_n(t)\, s_m(t)\, dt = \int_0^{T_s} \cos\left[2\pi\left(f_0 + m\Delta\right)t\right] \cos\left[2\pi\left(f_0 + n\Delta\right)t + \phi\right] dt$$

che quando si annulla indica la condizione di ortogonalità. Facendo uso della relazione $\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta) + \cos(\alpha-\beta)\right]$ e riferendoci per semplicità al caso di due frequenze contigue (ponendo $m = 0$ ed $n = 1$) si ottiene

$$\rho(\Delta,\phi) = \frac{1}{2}\int_0^{T_s} \cos\left[2\pi\left(2f_0 + \Delta\right)t + \phi\right] dt + \frac{1}{2}\int_0^{T_s} \cos\left[2\pi\Delta t - \phi\right] dt \quad (16.44)$$

Verifichiamo che il primo integrale della somma assume un valore nullo indipendentemente da ϕ quando $2f_0 + \Delta = \frac{k}{T_s}$, poiché in tal caso in un intervallo T_s entrano un numero intero k 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 $\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$:

$$\begin{aligned}
&\int_0^{T_s} \cos(2\pi\Delta t - \phi)\, dt = \\
&= \int_0^{T_s} \left[\cos(2\pi\Delta t)\cos\phi + \sin(2\pi\Delta t)\sin\phi\right] dt = \\
&= \left.\frac{\sin(2\pi\Delta t)}{2\pi\Delta}\right|_0^{T_s} \cdot \cos\phi - \left.\frac{\cos(2\pi\Delta t)}{2\pi\Delta}\right|_0^{T_s} \cdot \sin\phi = \\
&= T_s\left[\frac{\sin(2\pi\Delta T_s)}{2\pi\Delta T_s}\cdot\cos\phi + \frac{1-\cos(2\pi\Delta T_s)}{2\pi\Delta T_s}\cdot\sin\phi\right] \qquad (16.45)
\end{aligned}$$

Osserviamo ora che qualora $\phi = 0$ (coerenza di fase) il secondo termine della (16.45) si annulla per qualunque Δ. Se a questo punto esaminiamo solamente il primo termine, individuiamo le condizione di ortogonalità sul valore di Δ per il caso di

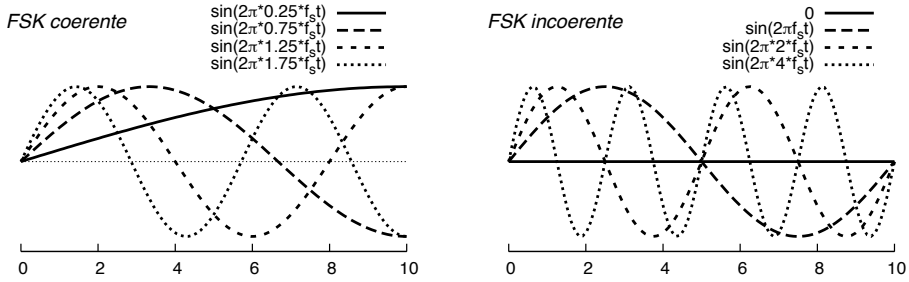


Figura 16.32: Forme d'onda ortogonali nei casi di modulazione *coerente* ed *incoerente*

Modulazione coerente Quando $\phi = 0$ il termine $\frac{\sin(2\pi\Delta T_s)}{2\pi\Delta T_s} = \text{sinc}\,(2\Delta T_s)$ della (16.45) 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 $f_0 + k\frac{f_s}{2}$.

Per fare in modo che *anche* il primo termine della (16.44) si annulli deve sussistere la relazione già osservata $2f_0 + \Delta = \frac{k}{T_s} = kf_s$, sostituendo nella quale il vincolo $\Delta = \frac{f_s}{2}$ appena trovato fornisce

$$2f_0 + \Delta = 2f_0 + \frac{f_s}{2} = kf_s$$

a cui corrisponde la la condizione

$$f_0 = f_s \frac{2k-1}{4}$$

ossia f_0 deve essere scelta come uno tra i valori $\frac{1}{4}f_s$, $\frac{3}{4}f_s$, $\frac{5}{4}f_s$, $\frac{7}{4}f_s, \ldots$, il che significa che la portante di riferimento f_0 da cui partire deve essere essa stessa ortogonale alle frequenze che codificano i simboli: infatti in tal modo ogni termine della serie dista dall'altro[101] per una frequenza pari a $\frac{f_s}{2}$, coincidente con il valore Δ necessario a che le frequenze di segnalazione siano ortogonali.

La parte sinistra della figura 16.32 quindi rappresenta, disegnate in un intervallo pari a T_s, sia le portanti f_0 che possono essere usate, sia le prime frequenze $f_k = k \cdot \Delta$ che è possibile adottare per un modulazione FSK *coerente* basata sul valore minimo di f_0 pari a $\frac{1}{4}f_s$[102].

Nel caso in cui f_0 non assuma uno dei valori individuati il primo termine di (16.44) non si annulla, ma se $f_0 \gg \frac{1}{T_s}$, risulta trascurabile rispetto al secondo. Pertanto, se $f_0 \gg f_s$ la scelta di f_0 non è più determinante.

Modulazione incoerente In questo caso si ha $\phi \neq 0$. In generale la (16.45) 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

[101] Infatti $\frac{1}{4} + \frac{1}{2} = \frac{3}{4}$, $\frac{3}{4} + \frac{1}{2} = \frac{5}{4}$ ....

[102] Possiamo notare come la spaziatura tra le frequenze di segnalazione di $\frac{f_s}{2}$ fa si 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.

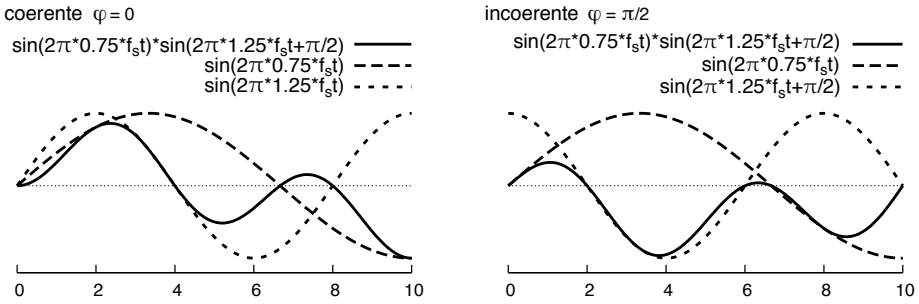


Figura 16.33: Prodotto di due frequenze ortogonali distanti $\frac{f_s}{2}$ in assenza di errore di fase (a sn) e con errore pari a $\phi = \frac{\pi}{2}$ (a ds)

quando $\phi \neq 0$ i segnali $s_k(t)$ sono ortogonali purché la spaziatura tra portati sia *doppia* della precedente, e pari cioè a $\Delta = f_s$.

Torniamo ad esaminare la (16.44): 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 $f_0 = 0,\ \frac{1}{2}f_s,\ f_s,\ \frac{3}{2}f_s, \ldots$. Notiamo come la spaziatura $\frac{f_s}{2}$ tra i possibili valori per la portante di riferimento f_0 sia identica al caso precedente, mentre la spaziatura necessaria alle frequenze di segnalazione $f_k = k \cdot \Delta$ è raddoppiata. La circostanza che ora sia ammessa anche una "portante a frequenza nulla" consente quindi di tracciare la parte destra della figura 16.32, 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 In figura 16.33 è mostrato il risultato del prodotto di due frequenze 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.

16.12.2 Prestazioni della modulazione OFDM

Il calcolo della P_e per bit accennato al § 16.8.5 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

$$P_e = \sum_{n=0}^{\tilde{N}-1} Pr(n)\, P_{e/n} = \frac{1}{M} \sum_{n=0}^{\tilde{N}-1} M_n P_{e_n} \tag{16.46}$$

16.12.2.1 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 § 16.3.1 per la modulazione QAM, che esprime P_{e_n} in funzione del numero di livelli $L_n =$

2^{M_n} e del rapporto $\left(\frac{E_b}{N_0}\right)_n$ per tale frequenza. Ma l'eq. (16.14) è ricavata considerando la densità di potenza del rumore in ingresso al ricevitore limitata da un filtro con banda pari a quella del segnale QAM, mentre ora tale filtro lascia passare l'intera banda $N\Delta$ occupata dal segnale OFDM, e quindi occorre valutare l'effetto prodotto da questo rumore sui valori $\underline{a}_n$ ottenuti mediante FFT. 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 (16.14) utilizziamo ora il rapporto SNR_n tra la quota di potenza di segnale che raggiunge l'*n-esimo* decisore, e la varianza (dovuta al rumore) della v.a. $\underline{a}_n$ su cui si basa tale decisione, ottenendo così[103]

$$P_{e/n} = \frac{2}{\log_2 L_n} P_{\alpha_n} \quad \text{in cui} \quad P_{\alpha_n} = \left(1 - \frac{1}{\sqrt{L_n}}\right) \operatorname{erfc}\left\{\sqrt{\frac{3}{2} SNR_n \frac{1}{L_n - 1}}\right\} \tag{16.47}$$

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

$$SNR_n = \frac{\mathcal{P}^c_{R_n}}{\mathcal{P}^c_{N_n}} = \frac{\mathcal{P}^s_{R_n}}{\mathcal{P}^s_{N_n}} = \frac{\frac{1}{2}\mathcal{P}_{\underline{R}_n}}{\frac{1}{2}\mathcal{P}_{\underline{N}_n}} = \frac{\mathcal{P}_{\underline{R}_n}}{\mathcal{P}_{\underline{N}_n}}$$

osserviamo che la potenza $\mathcal{P}_{\underline{R}_n}$ dell'inviluppo complesso del segnale ricevuto sulla portante n-esima è pari a

$$\mathcal{P}_{\underline{R}_n} = 2\mathcal{P}_{R_n} = 2\frac{T_0}{T}\alpha_n \mathcal{P}$$

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 $\mathcal{P}_{\underline{N}_n}$.

---

[103] La (16.47) può essere derivata dalle (16.13) e (16.14) considerando $\frac{E_b^n}{N_0^n} = \frac{SNR_n}{\log_2 L_n}$, ovvero invertendo l'eq. (15.16) $SNR = \frac{E_b}{N_0}\frac{2\log_2 L}{(1+\gamma)}$ con $\gamma = 0$ e notando che a differenza del caso di banda base, per segnali AM la banda (e la potenza di rumore) raddoppia. Ma se questa è una spiegazione troppo sintetica, ripercorriamo tutti i passaggi.

Partiamo dalla probabilità di errore condizionata (15.10) $P_{\hat{\delta}} = \operatorname{erfc}\left\{\frac{\Delta}{2\sqrt{2}\sigma_N(L-1)}\right\}$ del caso di multilivello di banda base, ed osserviamo che per un impulso rettangolare $g(t) = rect_{T_0}(t)$ la (15.12) si modifica in $\mathcal{P}_R = \frac{\Delta^2}{12}\frac{L+1}{L-1}$ in quanto $\mathcal{P}_R = \int \mathcal{P}_R(f)\,df = \int \sigma_A^2 \frac{|G(f)|^2}{T_0} df$ dove $\sigma_A^2 = \frac{\Delta^2}{12}\frac{L+1}{L-1}$ come ottenuto al § 15.8.1, mentre $\int |G(f)|^2\,df = \int T_0^2 sinc^2(fT_0)\,df = T_0$ (vedi nota 44).

In tal modo, eseguendo i passaggi di cui alla nota 48 a pag. 460 otteniamo $P_{\hat{\delta}} = \operatorname{erfc}\left\{\sqrt{\frac{12\mathcal{P}_R{}^{(L-1)}/(L+1)}{2\sqrt{2}\sigma_n(L-1)}}\right\} = \operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{1}{L^2-1}SNR}\right\}$ che conduce alla (16.47) ricordando che per il QAM ogni ramo ha $\sqrt{L}$ livelli, e che eseguendo il valore atteso rispetto alle probabilità dei simboli si ottiene $P_e(bit) = \left(1 - \frac{1}{\sqrt{L}}\right) P_{\hat{\delta}}$ (vedi eq. (15.11)).

16.12.2.2 Potenza di rumore per portante

Per quanto riguarda $\mathcal{P}_{\underline{N}_n}$, si tratta di applicare la (16.34) alla sequenza $\left\{(-1)^h \underline{n}(hT_c)\right\}$ dei campioni dell'inviluppo complesso del rumore, e determinare il valore

$$\mathcal{P}_{\underline{N}_n} = E\left\{\left(\underline{N}_n\right)^2\right\} = \sigma^2_{\underline{N}_n} \quad \text{in cui} \quad \underline{N}_n = \frac{1}{N}\sum_{h=0}^{N-1}(-1)^h \underline{n}(hT_c)\, \mathrm{e}^{-j2\pi\frac{h}{N}n}$$

tenendo conto 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 $\mathrm{e}^{-j2\pi\frac{h}{N}n}$, e che essendo $\underline{n}(t)$ ergodico è possibile scambiare medie temporali e di insieme. Sviluppando

$$\left(\underline{N}_n\right)^2 = \underline{N}_n\underline{N}_n^* = \frac{1}{N^2}\sum_{h=0}^{N-1}\sum_{k=0}^{N-1}(-1)^{h-k}\,\underline{n}(hT_c)\,\underline{n}^*(kT_c)\,\mathrm{e}^{-j2\pi\frac{h-k}{N}n}$$

e tenendo conto che

$$E\left\{(-1)^{h-k}\,\underline{n}(hT_c)\,\underline{n}^*(kT_c)\right\} = \mathrm{e}^{j\pi(h-k)}\mathcal{R}_{\underline{N}}((h-k)T_c)$$

otteniamo[104]

$$\begin{aligned}
\mathcal{P}_{\underline{N}_n} &= \frac{1}{N^2}\sum_{h=0}^{N-1}\sum_{k=0}^{N-1}\mathcal{R}_{\underline{N}}((h-k)T_c)\,\mathrm{e}^{j\pi(h-k)}\mathrm{e}^{-j2\pi\frac{h-k}{N}n} = \\
&= \frac{1}{N}\sum_{m=-(N-1)}^{N-1}\frac{N-|m|}{N}\mathcal{R}_{\underline{N}}(mT_c)\,\mathrm{e}^{j2\pi\frac{mT_c}{2T_c}}\mathrm{e}^{-j2\pi\frac{m}{N}n} = \\
&= \frac{1}{N}\sum_{m=-(N-1)}^{N-1} z(m)\,\mathrm{e}^{-j2\pi\frac{m}{N}n} \qquad (16.48)
\end{aligned}$$

in cui l'ultima riga semplifica l'espressione introducendo la sequenza $\{z(m)\}$ di lunghezza N, che si ottiene campionando

$$z(t) = \left(1 - \frac{|t|}{NT_c}\right)\mathcal{R}_{\underline{N}}(t)\,\mathrm{e}^{j2\pi\frac{t}{2T_c}} \qquad (16.49)$$

agli istanti $t = mT_c$ con $T_c = \frac{1}{N\Delta}$. Mostriamo ora come, per N sufficientemente elevato, la (16.48) possa essere calcolata in funzione dei campioni di $Z(f) = \mathcal{F}\{z(t)\}$, ed in particolare di come risulti

$$\mathcal{P}_{\underline{N}_n} \simeq \Delta \cdot Z(f)|_{f=n\Delta} \simeq 4\Delta \cdot \mathcal{P}_N(f_n)$$

[104] La 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

$$\mathcal{R}_{\underline{N}}(T_c)\,\mathrm{e}^{j\pi}\mathrm{e}^{-j2\pi\frac{1}{N}n} \qquad \text{e} \qquad \mathcal{R}_{\underline{N}}(-T_c)\,\mathrm{e}^{-j\pi}\mathrm{e}^{j2\pi\frac{1}{N}n}$$

$N-2$ volte quelli $\mathcal{R}_{\underline{N}}(2T_c)\,\mathrm{e}^{j2\pi}\mathrm{e}^{-j2\pi\frac{2}{N}n}$ e $\mathcal{R}_{\underline{N}}(-2T_c)\,\mathrm{e}^{-j2\pi}\mathrm{e}^{j2\pi\frac{2}{N}n}$, e così via.

Analizzando i termini che compaiono in (16.49), osserviamo che il prodotto $\mathcal{R}_{\underline{N}}(t)\,\mathrm{e}^{j2\pi\frac{t}{2T_c}}$ ha trasformata pari a $\mathcal{P}_{\underline{N}}(f)$, traslata in frequenza di $-\frac{1}{2T_c}=-\frac{N\Delta}{2}$, ovvero

$$\mathcal{F}\left\{\mathcal{R}_{\underline{N}}(t)\,\mathrm{e}^{j2\pi\frac{t}{2T_c}}\right\}=\mathcal{P}_{\underline{N}}\left(f-\frac{N\Delta}{2}\right)$$

mentre il termine $\left(1-\frac{|t|}{NT_c}\right)=\mathrm{tri}_{2NT_c}(t)=\mathrm{tri}_{\frac{2}{\Delta}}(t)$ possiede come noto trasformata $\mathcal{F}\left\{\mathrm{tri}_{\frac{2}{\Delta}}(t)\right\}=\frac{1}{\Delta}\mathrm{sinc}^2\left(\frac{f}{\Delta}\right)$; pertanto per N elevato il prodotto $z(t)=\mathcal{R}_{\underline{N}}(t)\,\mathrm{e}^{j2\pi\frac{t}{2T_c}}\cdot tri_{\frac{2}{\Delta}}(t)$ ha trasformata

$$Z(f)=\mathcal{P}_{\underline{N}}\left(f-\frac{N\Delta}{2}\right)*\frac{1}{\Delta}\mathrm{sinc}^2\left(\frac{f}{\Delta}\right)\simeq\mathcal{P}_{\underline{N}}\left(f-\frac{N\Delta}{2}\right)$$

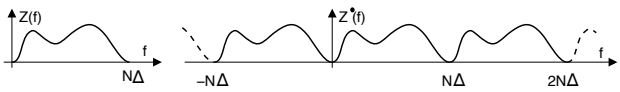
avendo approssimato $\frac{1}{\Delta}\mathrm{sinc}^2\left(\frac{f}{\Delta}\right)$ come un impulso di area unitaria, per $N\Delta$ grande rispetto a Δ.

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 $z(m)=z(mT_c)$ che compaiono nella (16.48); in particolare, per N sufficientemente elevato si ottiene che

$$\begin{aligned}\mathcal{P}_{\underline{N}_n} &= \frac{1}{N}\sum_{m=-(N-1)}^{N-1} z(m)\,\mathrm{e}^{-j2\pi\frac{m}{N}n}\simeq\Delta\cdot Z(f)|_{f=n\Delta}=\\ &= \Delta\cdot\mathcal{P}_{\underline{N}}\left(n\Delta-\frac{N\Delta}{2}\right)=\Delta\cdot\mathcal{P}_{\underline{N}}\left(\Delta\left(n-\frac{N}{2}\right)\right)=\\ &= 4\Delta\cdot\mathcal{P}_N^{+}\left(f_0+\Delta\left(n-\frac{N}{2}\right)\right)=4\Delta\cdot\mathcal{P}_N(f_n)=2\Delta\cdot\mathcal{N}_0(f_n)\end{aligned}$$

(passaggi alla nota[105]) in cui si è tenuto conto che $\mathcal{P}_{\underline{N}}(f)=4\mathcal{P}_N^{+}(f+f_0)$ e si è indicata

[105] Se campioniamo $z(t)$ con periodo $T_c=\frac{1}{N\Delta}$, il segnale $Z^{\bullet}(f)=\sum_{m=-\infty}^{\infty}Z(f-m\cdot N\Delta)$ *non presenta* aliasing (vedi figura), ed il passaggio di $z^{\bullet}(t)=\sum_{m=-\infty}^{\infty}z(mT_c)\,\delta(t-mT_c)$ attraverso un filtro di ricostruzione $H(f)=\frac{1}{N\Delta}\mathrm{rect}_{N\Delta}\left(f-\frac{N\Delta}{2}\right)$ restituisce il segnale originario. Scriviamo pertanto



$$z(t)=z^{\bullet}(t)*h(t)=\sum_{m=-\infty}^{\infty}z(mT_c)\,\delta(t-mT_c)*\mathrm{sinc}(N\Delta t)\,\mathrm{e}^{j\pi N\Delta t}$$

ed effettuiamone la trasformata:

$$\begin{aligned}Z(f) &= \mathcal{F}\left\{\sum_{m=-\infty}^{\infty}z(mT_c)\,\delta(t-mT_c)\right\}\cdot\frac{1}{N\Delta}\mathrm{rect}_{N\Delta}\left(f-\frac{N\Delta}{2}\right)\\ &= \left[\sum_{m=-\infty}^{\infty}z(mT_c)\,\mathrm{e}^{-j2\pi\frac{m}{N\Delta}f}\right]\cdot\frac{1}{N\Delta}\mathrm{rect}_{N\Delta}\left(f-\frac{N\Delta}{2}\right)\end{aligned}$$

che, calcolata alle frequenze $f=n\Delta$ con $n=0,1,...,N-1$ fornisce

$$Z(f)|_{f=n\Delta}=\frac{1}{N\Delta}\sum_{m=-\infty}^{\infty}z(mT_c)\,\mathrm{e}^{-j2\pi\frac{m}{N}n}$$

Se ora non disponiamo di tutti i campioni $z(mT_c)$, ma solo dei $2N-1$ valori con $m=-(N-1),...,0,1,...,N-1$, la relazione precedente si applica ad un nuovo segnale $z'(t)=z(t)\cdot\mathrm{rect}_{2NT_c}(t)$,

la densità di potenza in ingresso come $\mathcal{P}_N(f) = \frac{\mathcal{N}_0(f)}{2}$.

16.12.2.3 Prestazioni per portante

Siamo finalmente in grado di scrivere

$$\begin{aligned} SNR_n &= \frac{\mathcal{P}_{\underline{R}_n}}{\mathcal{P}_{\underline{N}_n}} = \frac{2\frac{T_0}{T}\alpha_n\mathcal{P}}{2\Delta\mathcal{N}_0(f_n)} = \frac{T_0}{T}\alpha_n\frac{T_0\mathcal{P}}{\mathcal{N}_0(f_n)} = \frac{T_0}{T}\alpha_n\frac{E_s}{\mathcal{N}_0(f_n)} = \\ &= \frac{T_0}{T}\alpha_n\frac{E_bM}{\mathcal{N}_0(f_n)} = \frac{T_0}{T}\frac{E_{b_n}}{E_b}\frac{E_bM}{\mathcal{N}_0(f_n)} = \frac{T_0}{T}\frac{E_{b_n}M}{\mathcal{N}_0(f_n)} \end{aligned}$$

avendo posto $T_0\mathcal{P} = E_s = E_bM$ 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

$$P_{e/n} = \frac{2}{M_n}\left(1 - \frac{1}{\sqrt{L_n}}\right)\operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{T_0}{T}\frac{E_{b_n}}{\mathcal{N}_0(f_n)}\frac{M}{L_n - 1}}\right\} \tag{16.50}$$

16.12.2.4 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 (16.50) sostituendo ad $\mathcal{N}_0(f_n)$ la costante $\mathcal{N}_0$. In questo caso il risultato $\mathcal{P}_{\underline{N}_n} = 2\Delta \cdot \mathcal{N}_0$ può essere ottenuto direttamente dalla (16.48): risulta infatti

$$\mathcal{R}_{\underline{N}}(t) = \mathcal{F}^{-1}\left\{\mathcal{P}_{\underline{N}}(f)\right\} = \mathcal{F}^{-1}\left\{4\mathcal{P}_N^+(f + f_0)\right\} = 2\mathcal{N}_0N\Delta sinc(N\Delta t)$$

e dunque $\mathcal{R}_{\underline{N}}(t) = 0$ con $t = mT_c = \frac{m}{N\Delta}$ per $m \neq 0$. Ciò in definitiva permette di scrivere

$$\mathcal{P}_{\underline{N}_n} = \frac{1}{N}\mathcal{R}_{\underline{N}}(0) = \frac{1}{N}2\mathcal{N}_0N\Delta = 2\Delta \cdot \mathcal{N}_0$$

16.12.2.5 Confronto con la portante singola

Proviamo a verificare se la modulazione OFDM è vantaggiosa in termini di prestazioni rispetto ad una QAM monoportante che trasporti il medesimo flusso binario f_b, occupi la stessa banda, ed a parità di potenza impiegata.

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 $E_{b_n} = \alpha_nE_b = \frac{E_b}{N}$ uguali per le diverse portanti; pertanto la 16.50 diviene

---

fornendo

$$Z'(f)\Big|_{f=n\Delta} = \frac{1}{N\Delta}\sum_{m=-(N-1)}^{N-1} z(mT_c)\,\mathrm{e}^{-j2\pi\frac{m}{N}n}$$

In virtù delle proprietà delle trasformate, risulta

$$Z'(f) = Z(f) * \mathcal{F}\left\{\operatorname{rect}_{2NT_c}(t)\right\} \simeq Z(f) * \delta(f) = Z(f)$$

in cui l'approssimazione è lecita per N elevato.

$$P_e^{OFDM} = P_{e/n} = \frac{2\tilde{N}}{M^{OFDM}}\left(1 - \frac{1}{\sqrt{L_n}}\right)\operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{\mathcal{N}_0}\frac{1}{\tilde{N}}\frac{M^{OFDM}}{L_n - 1}}\right\}$$

Nel caso QAM a portante singola, considerando un impulso a coseno rialzato e roll-off $\gamma = \frac{N}{\tilde{N}} - 1$ 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 § 16.3.1 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 $L_n = 2^{M_n} = 2^{\frac{M^{OFDM}}{\tilde{N}}}$ e $L = 2^{M^{QAM}} = 2^{\frac{M^{OFDM}}{\tilde{N}}} = L_n$.

16.12.3 Allocazione ottima della potenza OFDM

Affrontiamo la derivazione della (16.38) fornita a pag, 528 come soluzione al problema di trovare la $\mathcal{P}_R(f)$ che rende massima la quantità

$$C = \int_{f\in I_f} \log_2\left(1 + \frac{\mathcal{P}_R(f)}{\mathcal{P}_N(f)}\right) df \tag{16.51}$$

con $I_f = \{f : \mathcal{P}_R(f) > 0\}$, nel rispetto dei vincoli

$$\int_{f\in I_f} \mathcal{P}_R(f)\,df - \mathcal{P}_R = 0 \quad \text{e} \quad \mathcal{P}_R(f) \geq 0$$

che esprimono rispettivamente la limitazione sulla potenza totale $\mathcal{P}_R$ a disposizione e la necessità che la densità di potenza $\mathcal{P}_R(f)$ del segnale OFDM non sia negativa. Un problema di massimo vincolato siffatto viene tipicamente affrontato con la tecnica dei *moltiplicatori di Lagrange* (vedi 9.7.1) tranne che ora la presenza del vincolo di tipo disuguaglianza $\mathcal{P}_R(f) \geq 0$ comporta alcune considerazioni aggiuntive, note come *condizioni di* KARUSH–KUHN–TUCKER[106], i cui aspetti però non approfondiamo.

Scriviamo dunque la funzione *lagrangiana* (9.54) come

$$\begin{aligned} \mathcal{L}(\mathcal{P}_R(f), \lambda) &= \int \ln\left(1 + \frac{\mathcal{P}_R(f)}{\mathcal{P}_N(f)}\right) df + \lambda\left(\int \mathcal{P}_R(f)\,df - \mathcal{P}_R\right) \\ &= \int \left[\ln\left(1 + \frac{\mathcal{P}_R(f)}{\mathcal{P}_N(f)}\right) + \lambda\mathcal{P}_R(f)\right] df - \lambda\mathcal{P}_R \end{aligned}$$

in cui si sono usati i logaritmi naturali anziché in base 2 dato che ciò comporta solamente una variazione di ampiezza e non inficia il procedimento di massimizzazione. Valutiamo

[106] Vedi ad es. `https://it.wikipedia.org/wiki/Condizioni_di_Karush-Kuhn-Tucker`

quindi la derivata parziale

$$\frac{\partial \mathcal{L}}{\partial \mathcal{P}_R(f)} = \int \left[\frac{1}{1 + \frac{\mathcal{P}_R(f)}{\mathcal{P}_N(f)}} \frac{1}{\mathcal{P}_N(f)} + \lambda \right] df = \int \left[\frac{1}{\mathcal{P}_N(f) + \mathcal{P}_R(f)} + \lambda \right] df \tag{16.52}$$

in cui come al § 9.7.2 ci si è avvalsi della proprietà di *derivata sotto il segno di integrale*. Il massimo della capacità C (eq. (16.51)) si ottiene eguagliando (16.52) a zero, ovvero azzerando il termine tra parentesi quadre; per questa via otteniamo

$$\mathcal{P}_R(f) = -\frac{1}{\lambda} - \mathcal{P}_N(f)$$

ma, per rispettare i vincoli, scegliamo un valore $\mu = -\frac{1}{\lambda} > 0$ tale che scrivendo

$$\mathcal{P}_R(f) = \max\{0, \mu - \mathcal{P}_N(f)\} \tag{16.53}$$

si abbia sempre $\mathcal{P}_R(f) \geq 0$ e risulti $\int \mathcal{P}_R(f)\, df = \mathcal{P}_R$. A questo punto è immediato riconoscere la (16.53) come una notazione alternativa delle (16.38) di pag. 528.

Capitolo 17

Capacità e codifica di canale

Prima di affrontare lo studio dei mezzi trasmissivi, e dopo aver approfondito le tecniche di modulazione analogica e numerica, applichiamo alla *trasmissione* dei segnali i principi di *teoria dell'informazione* esposti al cap. 9. Lo scopo è innanzitutto quello di stabilire *i limiti* entro cui è possibile operare, ovvero quale sia *il massimo* teorico del tasso di informazione R trasmissibile su un determinato canale *rumoroso*, ovvero a cui è associata una probabilità di errore P_e. Tale massimo è noto come *capacità* C del canale, espressa in bit/secondo anche se in definitiva dipende da grandezze di natura continua come *potenza*, *banda*, e *livello di rumore* in ricezione; le conclusioni a cui si giunge sono quindi applicate ai casi studiati di trasmissione analogica e numerica, fornendo la motivazione al compromesso banda-potenza più volte citato. Il capitolo prosegue illustrando le tecniche necessarie per approcciare il più possibile da vicino il limite individuato, ovvero gli algoritmi denominati *codici di canale* che, attraverso l'introduzione (in forma appropriata) di ridondanza, consentono di rivelare e correggere gli errori di trasmissione, a patto di aumentare l'occupazione di banda, od il tempo di trasmissione. In tale ambito sono dettagliati i metodi che ricadono nelle categorie dei codici *a blocco*, *convoluzionali* ed a *decodifica iterativa*.

17.1 Dove arrivare, e come partire

E' più che lecito chiedersi ora di quanto si possa ridurre la P_e, e quanta ridondanza sia necessario aggiungere. La teoria che affronteremo risponde che finché l'intensità informativa $R = f_s \cdot H_s$ in uscita dal codificatore di sorgente (eq. (9.8)) si mantiene inferiore al valore della *capacità di canale* C (§§ 17.2 e 17.3), l'informazione può essere trasportata (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 (pur senza spiegare come fare) la teoria assicura che la ridondanza che occorre aggiungere può essere resa *trascurabile!*

Ma prima di approfondire questi risultati a dir poco *fenomenali*, svolgiamo alcune riflessioni su come

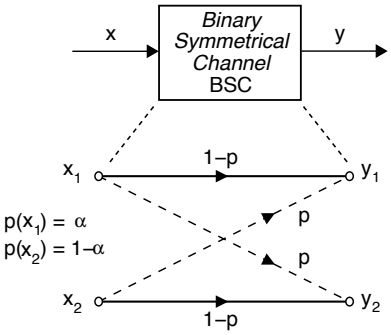
- la probabilità ed il tipo di errori introdotti da un canale numerico possono essere descritti, noto l'ingresso, nei termini di una matrice di *probabilità di transizione*;

- la decisione relativa al simbolo trasmesso si può basare, oltre che sulla conoscenza di tale matrice, anche sulle probabilità di *come sono emessi* i simboli della sorgente;
- al verificarsi di errori corrisponde una *perdita di informazione*.

17.1.1 Canale binario simmetrico

Mentre al § 15.4 si è sviluppato un lungo ragionamento per arrivare ad un valore di probabilità di errore (eq. (15.21)), in questa sede ci riferiamo al solo risultato finale, il valore $P_e^{bit} = p$ che caratterizza il *modello* raffigurato a lato e descritto dal termine BSC o *binary symmetric channel* che rappresenta appunto un canale numerico *binario* con probabilità p di introdurre errore, *indipendentemente* dal simbolo di ingresso, e per questo *simmetrico*.

In termini più formali indichiamo con x_1 e x_2 i due possibili ingressi e, qualora (con prob. $1-p$) non si verifichi errore, con y_1 e y_2 le rispettive uscite, mentre in presenza di errore (con probabilità p), in uscita si presenta il simbolo opposto.



Probabilità a priori Qualora i simboli di ingresso x_1 e x_2 non siano equiprobabili[1], indichiamo con α e $1-\alpha$ le relative prob. a priori (§ 6.1.4).

Probabilità in avanti Individuano le probabilità condizionate $p_{ji} = p\left(y_j/x_i\right)$ di osservare y_j in uscita quando in ingresso è presente x_i, e per questo dette *in avanti*.

Matrice di transizione Π I suoi elementi sono le prob. p_{ji}, e nel caso BSC la matrice è *simmetrica* in quanto

$$\begin{cases} p\left(y_2/x_1\right) = p\left(y_1/x_2\right) = p \\ p\left(y_1/x_1\right) = p\left(y_2/x_2\right) = 1-p \end{cases} \qquad \text{ovvero} \quad \Pi = \left[p_{ji}\right] = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Osserviamo due cose: la prima è che le prob. $\boldsymbol{p}_y = \left(p\left(y_1\right), p\left(y_2\right)\right)^\top$ dei simboli di uscita si calcolano come $\boldsymbol{p}_y = \Pi \cdot \boldsymbol{p}_x$; la seconda è che le definizioni date si estendono immediatamente al caso di canale $L-$ario, come nel caso multilivello.

17.1.2 Decisione a verosimiglianza ed a posteriori

Il simbolo y_i in uscita dal canale numerico NON È una variabile aleatoria, bensì una osservazione effettiva, e la decisione su quale x_j l'abbia prodotto avviene secondo un procedimento di *verifica di ipotesi* (§ 6.6.1), basata sul valore assunto da un rapporto tra valori di probabilità.

Decisione di massima verosimiglianza Qualora siano note solamente le probabilità in avanti p_{ij} ma non quelle a priori, la decisione avviene sulla base del *rapporto di verosimiglianza* (§ 6.6.2). Supponiamo che l'uscita del BSC sia ad es. il valore y_1: la

[1] Notiamo che in presenza di una codifica di sorgente efficace (PAG. 255) i simboli di ingresso dovrebbero essere pressoché equiprobabili.

decisione su quale delle ipotesi x_1 od x_2 sia più probabile in questo caso avviene in base al rapporto R_{ML} tra le probabilità *in avanti*, e prende il nome di decisione di *massima verosimiglianza* (vedi § 6.6.2.1) o *Maximum Likelihood*, ovvero

$$R_{ML}(y_1) = \frac{p(y_1/x_1)}{p(y_1/x_2)} = \frac{1-p}{p} \underset{x_2}{\overset{x_1}{\gtrless}} 1 \tag{17.1}$$

decidendo quindi per l'ipotesi *più verosimile* in funzione del valore maggiore o minore di uno per R_{ML}. Nel caso risulti $p < \frac{1}{2}$ la regola (17.1) equivale a scegliere l'ingresso concorde con l'uscita, oppure l'opposto se $p > \frac{1}{2}$ (!). Qualora invece si riceva y_2, il rapporto e la relativa regola di decisione sono definiti come $R_{ML}(y_2) = \frac{p(y_2/x_2)}{p(y_2/x_1)} \underset{x_1}{\overset{x_2}{\gtrless}} 1$.

Nel caso di trasmissione $L-$aria, infine, la ricezione di y_j porta alla decisione per $x_{\bar{i}} : \bar{i} = \arg\max_{i=1,2,\cdots L} \left\{ p\left(y_j/x_i\right) \right\}$

Decisione di massima probabilità a posteriori (MAP) Conoscendo anche le probabilità *a priori* $p(x_1)$ e $p(x_2)$, se i due simboli x_1 ed x_2 non sono equiprobabili[2], la decisione può avvenire confrontando le probabilità *a posteriori*[3] $p\left(x_j/y_i\right)$, calcolabili applicando il teorema di Bayes (vedi § 6.1.4). Facendo di nuovo il caso di aver ricevuto il simbolo y_1, scriviamo dunque

$$\begin{aligned} R_{MAP}(y_1) &= \frac{p(x_1/y_1)}{p(x_2/y_1)} = \frac{p(y_1/x_1)\,p(x_1)}{p(y_1)} \cdot \frac{p(y_1)}{p(y_1/x_2)\,p(x_2)} = \\ &= \frac{p(y_1/x_1)}{p(y_1/x_2)} \frac{p(x_1)}{p(x_2)} \underset{x_2}{\overset{x_1}{\gtrless}} 1 \end{aligned} \tag{17.2}$$

o più in generale, comprendendo anche il caso di canale $L-$ario, il criterio di decisione MAP qualora si riceva y_i è espresso come

$$x_{\bar{i}} : \bar{i} = \underset{i=1,2,\cdots L}{\arg\max} \left\{ p\left(y_j/x_i\right) p\left(x_i\right) \right\}$$

Il modo con cui le probabilità *a priori* $p(x_1)$ e $p(x_2)$ correggono la decisione ML (17.1) in MAP (17.2) per un BSC si presta a due osservazioni

- x_1 potrebbe essere *così raro* che, in presenza di una moderata probabilità di errore, si preferisce decidere sempre x_2, attribuendo l'eventuale ricezione di y_1 ad un errore del canale, piuttosto che all'effettiva trasmissione di x_1.
- in assenza di canale (ossia senza ricevere nulla) l'unica decisione possibile si basa sul confronto tra le p. a priori $p(x_1)$ e $p(x_2)$. La ricezione di un simbolo y_i apporta nuova informazione, alterando il rapporto di decisione R in misura tanto maggiore quanto minore è la probabilità di errore.

[2] In caso contrario (ovvero $p(x_1) = p(x_2) = 0.5$) la (17.2) è equivalente alla (17.1). Nei casi in cui *non si conoscano* le prob. a priori, non si può quindi fare altro che attuare una *decisione di massima verosimiglianza*.

[3] Sono indicate come *a posteriori* perché misurano la probabilità del simbolo trasmesso x *dopo* la conoscenza di quello ricevuto y.

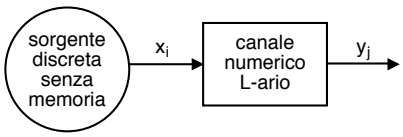
Esempio Verifichiamo le ultime osservazioni esplicitando una probabilità a posteriori in funzione di p:

$$\begin{aligned} p(x_1/y_1) &= \frac{p(x_1,y_1)}{p(y_1)} = \frac{p(y_1/x_1)\,p(x_1)}{p(y_1/x_1)\,p(x_1)+p(y_1/x_2)\,p(x_2)} = \\ &= \frac{(1-p)\cdot p(x_1)}{(1-p)\cdot p(x_1)+p\cdot p(x_2)} = \frac{p(x_1)}{p(x_1)+\frac{p}{1-p}p(x_2)} \end{aligned}$$

Se $p = 1-p = \frac{1}{2}$, il canale *è inservibile* e non trasferisce informazione: infatti si ottiene $p(x_1/y_1) = p(x_1)$ pari a quella a priori, in quanto $p(x_1)+p(x_2) = 1$. D'altra parte se $p < \frac{1}{2}$ si ottiene $p(x_1/y_1) > p(x_1)$ dato che ora $\frac{p}{1-p} < 0.5$: si assiste pertanto ad un *aumento* della probabilità di x_1 rispetto a quella a priori; se poi la probabilità di errore tende a zero ($p \to 0$) si ottiene $p(x_1/y_1) \to 1$.

17.1.3 Informazione mutua media per canale numerico $L-$ario

Approfondiamo questa nozione introdotta al § 9.4.3 e li utilizzata per definire la funzione velocità distorsione (§ 9.6.2), mostrando 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 $i = 1, 2, \cdots, L$, ed indichiamo con $y \in \{y_j\}$ (sempre per $j = 1, 2, \cdots, L$) il corrispondente simbolo ricevuto mediante un canale discreto, 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*, come[4]

$$I(x_i,y_j) = \log_2 \frac{p(x_i,y_j)}{p(x_i)\,p(y_j)} = \log_2 \frac{p(x_i/y_j)}{p(x_i)} = \log_2 \frac{p(y_j/x_i)}{p(y_j)} \quad \text{bit} \tag{17.3}$$

da cui deriva che

1. se ingresso ed uscita del canale sono *statisticamente indipendenti* si ha $p(x_i,y_j) = p(x_i)\,p(y_j)$, e di conseguenza l'informazione mutua *è nulla;*
2. se $p(y_j/x_i) > p(y_j)$ significa che l'essere a conoscenza della trasmissione di x_i rende la ricezione di y_j *più probabile* di quanto non lo fosse a priori, e corrisponde ad una informazione mutua *positiva*;
3. la definizione di informazione mutua è *simmetrica*, ovvero $I(x_i,y_j) = I(y_j,x_i)$;
4. rifrasando la 2. in virtù della 3., se $p(x_i/y_j) > p(x_j)$ allora ricevere y_j rende la trasmissione di x_i *più probabile* di quanto non lo fosse a priori, manifestando lo stesso valore di informazione mutua *positiva* del punto 2.

Per giungere ad una grandezza $I(X,Y)$ che tenga conto del comportamento *medio* del canale, ovvero per coppie ingresso-uscita qualsiasi, occorre pesare i valori di $I(x_i,y_j)$

[4]Per ottenere le diverse forme della (17.3) si ricordi che $p(x_i,y_j) = p(x_i/y_j)\,p(y_j) = p(y_j/x_i)\,p(x_i)$

con le relative probabilità congiunte, ossia calcolarne il valore atteso rispetto a tutte le possibili coppie (x_i, y_j):

$$I(X,Y) = E_{X,Y}\left\{I\left(x_i, y_j\right)\right\} = \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{p\left(x_i/y_j\right)}{p\left(x_i\right)} \tag{17.4}$$

$$= \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{p\left(y_j/x_i\right)}{p\left(y_j\right)} \tag{17.5}$$

ri-ottenendo così l'*informazione mutua media* (§ 9.4.3), misurata in bit/simbolo, e che rappresenta (in media) quanta informazione ogni simbolo ricevuto trasporta a riguardo di quello trasmesso. In virtù della simmetria di questa definizione, ci accorgiamo che il valore di $I(X,Y)$ può essere espresso[5] nelle due forme alternative

$$I(X,Y) = H(X) - H(X/Y) \tag{17.6}$$

$$= H(Y) - H(Y/X) \tag{17.7}$$

in cui l'entropia *condizionale* (§ 9.4.2)

$$H(X/Y) = \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p\left(x_i/y_j\right)} \tag{17.8}$$

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 $p\left(x_i/y_j\right)$ sia pari a 1 se $j = i$ e zero altrimenti, è facile vedere[6] 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

$$H(Y/X) = \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p\left(y_j/x_i\right)} \tag{17.9}$$

prende il nome di *noise entropy* dato che considera il processo di rumore come se fosse un segnale informativo: infatti, sebbene si possa essere tentati di 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.

Calcolo dell'informazione mutua media per il BSC Torniamo al caso binario descritto al § 17.1.1 ed usiamo la (17.7) per calcolare l'informazione mutua media in funzione della probabilità a priori $p(x_1) = \alpha$ e di quella in avanti p_e, valutando innanzitutto $H(Y)$ e $H(Y/X)$. Dal punto di vista dell'uscita del canale, i simboli y_1, y_2 costituiscono l'alfabeto di una sorgente binaria senza memoria, la cui entropia si

[5] Infatti

$$\sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{p(x_i/y_j)}{p(x_i)} = \sum_i \sum_j p\left(x_i, y_j\right) \left[\log_2 \frac{1}{p(x_i)} - \log_2 \frac{1}{p(x_i/y_j)}\right] =$$

$$= \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p(x_i)} - \sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p(x_i/y_j)}$$

L'ultimo termine è indicato come entropia condizionale $H(X/Y)$ (eq. (17.8)), mentre il penultimo è pari all'entropia di sorgente $H(X)$ dato che *saturando* la prob. congiunta $p\left(x_i, y_j\right)$ rispetto ad j, ovvero $\sum_j p\left(x_i, y_j\right) = p\left(x_i\right)$, si perviene alla (17.6) in base al risultato $\sum_i \log_2 \frac{1}{p(x_i)} \sum_j p\left(x_i, y_j\right) = \sum_i p\left(x_i\right) \log_2 \frac{1}{p(x_i)}$. Per la (17.7) il passaggio è del tutto simile.

[6] Infatti in tal caso la (17.8) diviene $\sum_i \sum_j p\left(x_i, y_j\right) \log_2 \frac{1}{p(x_i/y_j)} = \sum_i p\left(x_i, y_i\right) \log_2 1 = 0$

esprime in termini di $p(y_1)$ mediante la (9.6), ovvero $H(Y) = H_b(p(y_1))$, in cui

$$\begin{aligned} p(y_1) &= p(y_1/x_1)\, p(x_1) + p(y_1/x_2)\, p(x_2) = \\ &= (1-p_e)\,\alpha + p_e\,(1-\alpha) = p_e + \alpha - 2\alpha p_e \end{aligned}$$

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 (17.9) 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

$$I(X,Y) = H(Y) - H(Y/X) = H_b(p_e + \alpha - 2\alpha p_e) - H_b(p_e) \tag{17.10}$$

che dipende sia dalla probabilità di errore p_e, sia dalla prob. a priori 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 $p_e \to \frac{1}{2}$ allora $I(X,Y) \to 0$.

17.2 Capacità di canale discreto

Le relazioni fin qui discusse permettono di valutare la perdita di informazione causata dai disturbi, ma dipendono sia dalle probabilità *in avanti* $p(y_j/x_i)$ che descrivono il comportamento del canale, sia da quelle *a priori* $p(x_i)$, che invece attengono alle caratteristiche della sorgente. Vogliamo invece definire una grandezza che esprima esclusivamente l'attitudine (o *capacità*) del canale a trasportare informazione, indipendentemente dalle caratteristiche della sorgente. Questo risultato può essere ottenuto variando le prob. a priori in tutti i modi possibili, fino a trovare il valore

$$C_s = \max_{p(x)} I(X,Y) \quad \text{bit/simbolo} \tag{17.11}$$

che definisce la *capacità di canale per simbolo* come il massimo valore dell'informazione mutua media, ottenuto in corrispondenza della migliore sorgente possibile. Il pedice $_s$ sta per *simbolo*, e serve a distinguere il valore ora definito da quello che esprime la massima *intensità* di trasferimento dell'informazione espressa 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 valore[7]

$$C = f_s \cdot C_s \quad \text{bit/secondo} \tag{17.12}$$

L'importanza di questa quantità risiede nel *teorema fondamentale per canali rumorosi*[8] già anticipato più volte, che asserisce che 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

[7] Notiamo l'invarianza di (17.12) rispetto al numero di livelli con cui è effettuata la trasmissione: 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.

[8] `http://it.wikipedia.org/wiki/Secondo_teorema_di_Shannon`

risulti $R < C$;

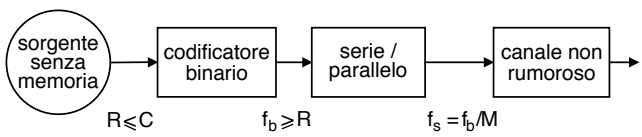
- se è accettabile una probabilità di errore p_e, si può raggiungere (con la miglior codifica possibile) una velocità $R(p_e) = \frac{C}{1-H_b(p_e)} > C$ in cui $H_b(p_e)$ è l'entropia di una sorgente binaria (9.6);
- per qualsiasi valore di p_e, non è possibile trasmettere informazione a velocità maggiore di $R(p_e)$.

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'effettiva informazione mutua media risulterà ridotta rispetto al valore della capacità, così come la massima velocità R.

Illustriamo l'applicazione di questi risultati con un paio di esempi.

17.2.1 Capacità di un canale L−ario non rumoroso

Consideriamo lo schema mostrato in figura, ovvero un canale che trasporta *senza errori* simboli con $L = 2^M$ livelli: in tal caso l'equivocazione $H(Y/X)$ è nulla, e la (17.6) permette di scrivere $I(X,Y) = H(X)$, che è massima se $P(x_i) = 1/L$ per tutti gli i, risultando così $C_s = H_{max}(X) = \log_2 L = M$ bit/simbolo, e $C = f_s \cdot C_s = f_s \cdot M$ bit/secondo.



I simboli ad L livelli sono ottenuti raggruppando M dei bit prodotti da una codifica binaria a velocità f_b, risultando $f_b \geq R = H_x$ (vedi eq. (9.9)) in funzione della ottimalità o meno del codificatore; pertanto, risulta $R \leq f_b = f_s \cdot M = C$ con l'uguaglianza valida nel caso in cui il codificatore riesca a rimuovere tutta la ridondanza dei messaggi della sorgente[9], 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 (9.13)[10], e quindi la regola del prefisso non sarebbe rispettata, causando in definitiva errori di decodifica anche in assenza di rumore!

17.2.2 Capacità del canale binario simmetrico

Esaminiamo l'effetto della presenza di rumore per questo caso particolare, per il quale a pag. 560 abbiamo valutato l'espressione dell'informazione mutua media, data dalla

[9] Ad esempio se L non è una potenza di due, un codificatore di sorgente che operi simbolo per simbolo produce necessariamente $f_b > R$, mentre se concatena più simboli (§ 9.1.4), può avvicinarsi a $f_b = R$.

[10] Infatti, potrebbe risultare $R > C$ solo se $f_b < R$, ovvero il codificatore dovrebbe produrre *meno* binit/secondo di quanti bit/secondo produca la sorgente

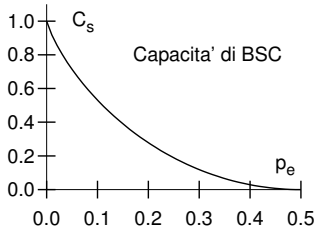
(17.10), e pari a

$$I(X,Y) = H_b(p_e + \alpha - 2\alpha p_e) - H_b(p_e)$$

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 avviene per qualunque p_e se $\alpha = \frac{1}{2}$, ossia per simboli equiprobabili. Pertanto la capacità del BSC risulta pari a

$$C_s = H_b(1/2) - H_b(p_e) = 1 - H_b(p_e)$$

il cui grafico è rappresentato alla figura a lato[11], evidenziando che $C_s \simeq 1$ bit/simbolo se $p_e \simeq 0$, ma che poi decade rapidamente a zero se $p_e \to 0.5$.

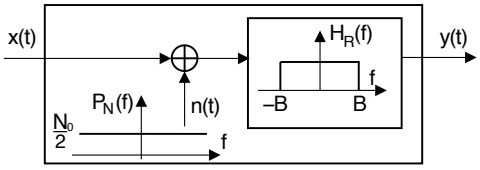


Quest'ultimo esempio in particolare ci conferma l'esigenza, in presenza di un canale rumoroso, di attuare tecniche di codifica di canale in grado di ridurre la probabilità di errore, in modo da poter sfruttare appieno la capacità che il canale presenta nel caso di p_e ridotta, e di preferire tra queste le tecniche che vi riescono mantenendo al minimo la quantità dei bit aggiuntivi, dato che altrimenti come noto aumenta la banda occupata dal segnale dati.

17.3 Capacità di canale continuo

Come anticipato fin dal § 1.2.2 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.

Canale gaussiano additivo bianco Una situazione tipica è quella raffigurata a lato, 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 $P_n = \sigma_n^2 = N_0 B$. Tale situazione viene indicata come *canale* AWGN (*additive white gaussian noise*) *limitato in banda.*



Calcolo della capacità Indicando con $p(x)$, $p(y)$, $p(x/y)$, $p(y/x)$ le d.d.p. 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 (17.4) al caso continuo porta a scrivere l'espressione dell'informazione mutua media come

$$I(X,Y) = \int\int_{-\infty}^{\infty} p_{XY}(x,y) \log_2 \frac{p_Y(y/x)}{p_Y(y)} dx dy \quad \text{bit/campione} \tag{17.13}$$

[11] Sono mostrati solo i valori per $0 \le p_e \le 0.5$ dato che successivamente l'andamento di C_s si riflette in modo speculare.

che è una misura *assoluta*[12] del trasferimento di informazione per campione di uscita. Il massimo valore di (17.13) 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 § 7.2.4), cosicché la capacità di canale risulta definita come

$$C = 2B \cdot \max_{p(x)} \{I(X,Y)\} \quad \text{bit/secondo} \tag{17.14}$$

Riscrivendo la (17.13) nella forma

$$I(X,Y) = h(Y) - h(Y/X) \tag{17.15}$$

si ottiene una espressione analoga alla (17.7) ma i cui termini sono ora da intendersi come entropia differenziale, definita al § 9.3.1. 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 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; quindi $h(Y/X)$ si riduce all'entropia differenziale di un processo gaussiano (9.20), che non dipende dal valor medio, ma solo dall'andamento di $p_N(n)$; pertanto

$$h(Y/X) = \int_{-\infty}^{\infty} p_N(n) \log_2 \frac{1}{p_N(n)} dn = \frac{1}{2} \log_2 \left(2\pi e \sigma_n^2\right) \tag{17.16}$$

come risulta per l'entropia differenziale di sorgenti gaussiane (9.20). Quindi ora il termine della (17.15) 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

$$h(Y) = \tfrac{1}{2} \log_2 \left[2\pi e \left(\sigma_x^2 + \sigma_n^2\right)\right] \tag{17.17}$$

cosicché mettendo assieme (17.15), (17.16) e (17.17), la (17.14) si riscrive come

$$\begin{aligned} C &= 2B \cdot \left\{ \tfrac{1}{2} \log_2 \left[2\pi e \left(\sigma_x^2 + \sigma_n^2\right)\right] - \tfrac{1}{2} \log_2 \left(2\pi e \sigma_n^2\right) \right\} = \\ &= B \cdot \log_2 \frac{\sigma_x^2 + \sigma_n^2}{\sigma_n^2} = B \cdot \log_2 \left(1 + \frac{P_x}{P_n}\right) \quad \text{bit/secondo} \end{aligned}$$

che è proprio il risultato tanto spesso citato, che prende il nome di *legge di Shannon-Hartley*[13] 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:

$$C = B \cdot \log_2 \left(1 + \frac{P_s}{N_0 B}\right) \quad \text{bit/secondo} \tag{17.18}$$

[12] Per il fatto di avere una ddp di y sia a numeratore che a denominatore del logaritmo, la (17.13) non soffre dei problemi discussi alla nota 26 a pag. 267.

[13] `http://en.wikipedia.org/wiki/Shannon-Hartley_theorem`

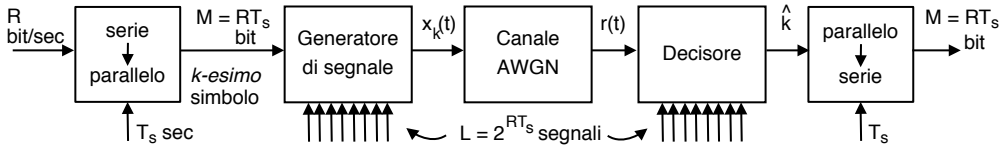


Figura 17.1: Schema ideale di codifica di canale ad errore asintoticamente nullo

che, associata al teorema fondamentale della codifica espresso al § 17.2, stabilisce il massimo tasso informativo trasmissibile senza errori su di un canale AWGN limitato in banda come $R \leq B \cdot \log_2 (1 + {}^{P_s}/_{N_0 B})$. Discutiamo ora delle conseguenze di questo risultato.

17.3.1 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 § 15.1.2.4) come mostrato alla figura 17.1 dove l'informazione generata ad una velocità R bit/secondo viene trasmessa mediante simboli emessi con periodo T_s secondi, ognuno dei quali rappresenta un gruppo di $M = RT_s$ bit, e dunque occorrono $L = 2^M$ simboli diversi.

Nella dimostrazione di Shannon ogni simbolo, anziché essere rappresentato da un valore costante come nella trasmissione multilivello, è costituito da un segnale $x_k(t)\,,\ k = 1, 2, \ldots, L$ 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 dimostrato[14] che, per $T_s \to \infty$, lo schema indicato riesce effettivamente a conseguire una $P_e \to 0$, tranne per il piccolo particolare che... occorre attendere un tempo che tende a infinito!

[14] Senza 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 $\frac{1}{T_s} \int_0^{T_s} [n(t)]^2\, dt \to \sigma_n^2$, 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 $E\left\{\varepsilon_k^{(h)}\right\} \to \sigma_n^2 + 2\sigma_x^2$ 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 $T_s \to \infty$ risulta sempre $\varepsilon_k < \varepsilon_k^{(h)}$, azzerando la probabilità di errore.

17.3.2 Minima energia per bit

In realtà uno schema di trasmissione numerica che approssima piuttosto bene quello ideale discusso al § precedente esiste veramente, ed è quello esposto al § 16.5.1 ed denominato FSK ortogonale, in cui le forme d'onda di fig. 17.1 sono sinusoidali: il grafico delle sue prestazioni a pag. 516 mostra infatti come, aumentando L, lo stesso valore di E_b/N_0 permetta di conseguire valori di P_e via via più piccoli. Lo stesso grafico mostra però l'esistenza di un valore limite sotto cui E_b/N_0 non può scendere, dovendo comunque risultare

$$\frac{E_b}{N_0} \geq \ln 2 = 0,693 \quad \text{ovvero} \quad \left.\frac{E_b}{N_0}\right|_{dB} \geq -1.6 \text{ dB} \tag{17.19}$$

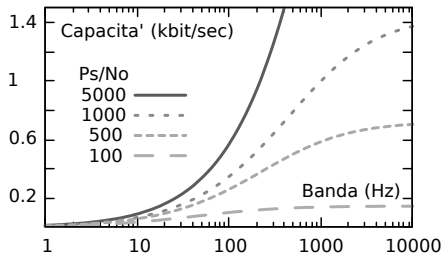
Ciò deriva dall'occupazione di banda via via crescente necessaria all'FSK qualora L aumenti: considerando che la capacità di canale per $B \to \infty$ fornita dalla (17.20) vale $C_\infty = \frac{P_s}{N_0 \ln 2}$, e che deve risultare $R \leq C$, risulta allora $\ln 2 = \frac{P_s}{N_0 C_\infty} \leq \frac{P_s}{N_0 R} = \frac{E_b}{N_0}$, ovvero la (17.19).

Ma per arrivare all'espressione di C_∞ ora citata, affrontiamo il prossimo §.

17.3.3 Compromesso banda-potenza e capacità massima

Il valore limite (17.19) trae origine da una conseguenza della (17.18) già fatta notare al § 15.4.7, ovvero la possibilità di risparmiare potenza aumentando l'occupazione di banda (o viceversa), dato che in entrambi i casi a ciò corrisponde un aumento di C. Ma ciò non avviene all'infinto, ovvero *non si può oltrepassare* un valore massimo di capacità! Infatti se nella (17.18) si aumenta B il filtro di ricezione si *allarga*, e dunque aumenta la potenza di rumore, e l'effetto finale è che per un canale con *banda infinita* non si ottiene una capacità infinita, bensì il valore

$$\begin{aligned} C_\infty &= \lim_{B\to\infty} B \cdot \log_2\left(1 + \frac{P_s}{N_0 B}\right) = \\ &= \frac{P_s}{N_0 \ln 2} \simeq 1.44 \frac{P_s}{N_0} \end{aligned} \tag{17.20}$$



che individua anche il limite *assoluto* al massimo tasso informativo R trasmissibile. In figura è mostrato l'andamento effettivo della (17.18) in funzione di B, per alcuni valori di $\frac{P_s}{N_0}$ di esempio, mentre la dimostrazione della (17.20) è riportata alla nota[15].

[15] La (17.20) si ottiene riscrivendo la (17.18) nella forma

$$C = \frac{P_s}{N_0 \frac{P_s}{N_0 B}} \cdot \frac{\ln\left(1 + \frac{P_s}{N_0 B}\right)}{\ln 2} = \frac{P_s}{N_0 \ln 2} \cdot \frac{\ln(1+\lambda)}{\lambda}$$

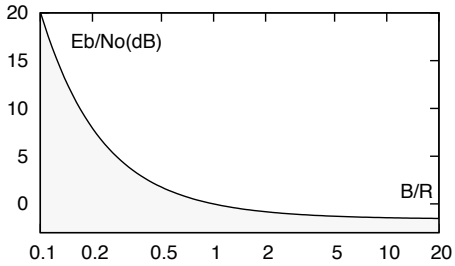
in cui ln è il logaritmo *naturale* in base e, e si è posto $\frac{P_s}{N_0 B} = \lambda$. Ricordando ora lo sviluppo di Maclaurin $f(x) = f(0) + \sum_{n=1}^{\infty}\left(\left.\frac{\partial^n f(x)}{\partial x^n}\right|_{x=0} \cdot \frac{x^n}{n!}\right)$ e che $\frac{d}{dx}\ln x = \frac{1}{x}$, il termine $\ln(1+\lambda)$ può essere espanso in serie di potenze come $\ln(1+\lambda) = \lambda - \frac{1}{2}\lambda^2 + \frac{1}{3}\lambda^3 + \cdots$; notando infine che per $B \to \infty$ si ha $\lambda \to 0$, e che $\lim_{\lambda\to 0}\frac{\ln(1+\lambda)}{\lambda} = 1$, si giunge in definitiva al risultato (17.20).

17.3.4 Limite inferiore per $\frac{E_b}{N_0}$

Una volta assegnato il tasso informativo $R \leq C$ della sorgente e la banda B del canale, partendo dalla (17.18) si può ottenere[16] una relazione che esprime il valore di $\frac{E_b}{N_0}$ necessario a conseguire una trasmissione senza errori (nel caso ideale):

$$\frac{E_b}{N_0} \geq \frac{B}{R}\left(2^{\frac{R}{B}} - 1\right) \tag{17.21}$$

e che, espressa in dB, è graficata alla 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

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 (17.19) 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.

17.3.5 Confronto con le prestazioni di sistemi di modulazione reali

E' possibile svolgere una verifica sperimentale della relazione (17.21) prendendo in considerazione le tecniche di modulazione numerica discusse ai capitoli precedenti, e che consentono di variare l'occupazione di banda B per trasmettere ad una data velocità $R = f_b$, ad esempio riducendone il rapporto B/R come nelle trasmissione multilivello[17], oppure aumentandolo, come nel caso dell'FSK. In questi casi il valore di $\frac{E_b}{N_0}$ necessario a conseguire una determinata prestazione (P_e) varia in funzione del rapporto B/R, e dunque può essere messo a confronto con i valori minimi di $\frac{E_b}{N_0}$ previsti dalla (17.21), come avviene nella figura 17.2 che mostra i valori di E_b/N_0 in funzione di B/R per le tecniche di modulazione numerica QAM (§ 16.3.1) e FSK ortogonale (pag. 514). Per tracciare la figura si sono ricavati i valori di E_b/N_0 necessari a ciascun metodo per ottenere una P_e pari a 10^{-5} per diversi valori di L, e messi in relazione con l'occupazione spettrale associata $B(L)$ rapportata alla velocità f_b, ossia in relazione all'efficienza spettrale ρ (pag. 498) dei metodi.

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 $\left.\frac{B}{R}\right|_{QAM} = \frac{1}{\log_2 L}$; invece come riportato a pag. 516 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

[16] Riscrivendo la (17.18) 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 $\frac{B}{R}(2^{\frac{C}{B}} - 1) = \frac{P_s}{N_0 R}$. L'uguaglianza individua la circostanza limite in cui $R = C$, mentre se nell'esponente di 2 a primo membro sostituiamo C con R, e $R \leq C$, il primo membro diviene più piccolo, e pertanto $\frac{B}{R}(2^{\frac{R}{B}} - 1) \leq \frac{P_s}{N_0 R}$. Infine, notiamo che $\frac{P_s}{N_0 R} = \frac{E_b}{N_0}$, da cui il risultato mostrato (17.21).

[17] Vedi ad es. il caso di banda base al § 15.4.9 o quello del QAM al § 16.3.1.

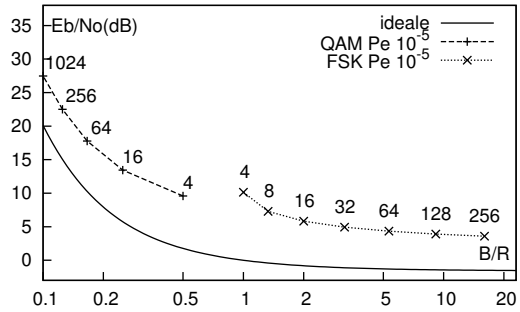


Figura 17.2: Rapporto E_b/N_0 di QAM ed FSK per $P_e = 10^{-5}$ al variare di L, in funzione della efficienza spettrale, confrontato con i valori minimi teorici

$\frac{E_b}{N_0}$ in funzione di $\frac{B}{R}$ ricalchi abbastanza fedelmente quello ideale, a parte una perdita di efficienza, che si riduce per L crescente.

17.4 Codifica di canale

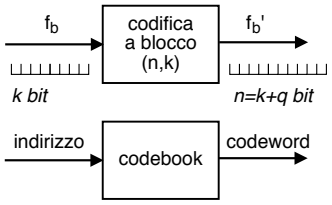
Dopo aver analizzato i risultati che la teoria dell'informazione fornisce a riguardo delle migliori prestazioni ottenibili, proseguiamo il discorso iniziato al § 15.6.2.1 su come aggiungere ridondanza ad un flusso binario a velocità f_b da trasmettere su di un canale numerico, in modo da realizzare una protezione FEC capace di ridurre la probabilità di errore per bit P_e in ricezione. Ricordiamo che al § 17.2 abbiamo mostrato come, aumentando il ritardo di codifica, la probabilità di errore può essere resa piccola a piacere, purché $f_b = R < C$, essendo C la capacità di canale. Mentre una soluzione basata su segnalazione ortogonale (ad esempio, l'FSK del § 16.5.1) determina un aumento asintoticamente *esponenziale* della banda occupata[18], le tecniche illustrate nel seguito consentono di mantenere la relazione tra grado di protezione e banda occupata di tipo *proporzionale*.

Classificazione delle tecniche di codifica di canale Una categoria molto vasta è quella dei codici *a blocco*, che operano suddividendo il messaggio da proteggere in blocchi disgiunti, codificati in modo indipendente; una diversa classe è quella dei codici *convoluzionali*, che invece trattano il messaggio come una sequenza priva di suddivisioni, da cui *calcolare* una nuova sequenza (a velocità maggiore) che ne rappresenta la codifica. Mentre i codici convoluzionali sono descritti al § 17.4.2, la trattazione dei codici a blocco è iniziata al § 15.6.2.1, che si suggerisce di consultare prima di proseguire, riassumendone qui solo alcuni concetti.

[18] Infatti partendo dall'espressione della banda occupata dall'FSK $B \to \frac{f_b}{2} \cdot \frac{L}{\log_2 L}$ (eq. (16.21)) e considerando che $L = 2^M = 2^{f_b T}$ si ottiene $B = \frac{1}{2T} \cdot 2^{f_b T}$ ovvero un aumento esponenziale di B al crescere di T. Un diverso esempio può essere l'uso di forme d'onda ortogonali realizzate come $rect_{T/L}$ posti all'interno del periodo di simbolo T in modo che non si sovrappongano se associati a simboli diversi (vedi fig. 7.8 a pag. 218). Anche qui, aumentando T il numero di simboli $L = 2^M = 2^{f_b T}$ aumenta esponenzialmente, e la durata $T/L = T/2^{f_b T}$ di ogni *rect* tende esponenzialmente a zero se $T \to \infty$, mentre la banda occupata tende ad infinito, sempre con legge esponenziale rispetto a T.

Tasso di codifica, velocità binaria ed E_b/N_0 Riprendiamo la notazione introdotta al § 15.6.2.1 per i *codici a blocco*, in cui ad ogni k bit della sequenza di ingresso (da proteggere)

si genera una *codeword*[19] con lunghezza $n = k + q > k$ bit, in cui sono stati *aggiunti* q bit *di protezione* in funzione dei k del blocco: tale procedura viene indicata come *codice* (n, k), la cui efficienza è misurata dal *tasso di codifica* (o CODE RATE)



$$R_c = \frac{k}{n} < 1$$

che rappresenta la frazione di bit informativi sul totale di quelli trasmessi, nonché *l'inverso* del fattore di *espansione di banda*[20]: la nuova velocità di trasmissione in presenza di codifica vale infatti

$$f_b' = \frac{f_b}{R_c}$$

Osserviamo che all'aumento della velocità di segnalazione (essendo $f_b' > f_b$) corrisponde una eguale diminuzione del rapporto $E_b/N_0 = \frac{\mathcal{P}_x}{N_0 f_b'} = R_c \frac{\mathcal{P}_x}{N_0 f_b}$, e conseguentemente si assiste ad un *peggioramento* della probabilità di errore *grezza* del decisore: pertanto, la capacità correttiva del codice deve essere tale da compensare anche questo aspetto. Per mantenere limitato l'effetto descritto, così come l'espansione di banda, vorremmo trovare codificatori per cui R_c sia il più possibile vicino ad uno.

Distanza di Hamming, distanza minima e capacità correttiva Al § 15.6.2.1 è stata definita $d_H(x_i, x_j)$ come il numero di bit in cui le codeword x_i e x_j differiscono, pari al numero di errori sul bit necessario affinché una si trasformi nell'altra. Valutando la distanza di Hamming d_H tra tutte le possibili coppie di codeword, si definisce la *distanza minima del codice* $d_m = \min_{i \neq j} d_H(x_i, x_j)$, che descrive quanto le codeword siano *vicine* nel caso peggiore.

Esempio Con $k = 4$ e $q = 3$ esistono $2^4 = 16$ codeword su $2^7 = 128$ possibili configurazioni degli $n = k + q$ bit della codeword. Ciò significa che ci sono $2^3 = 8$ *non*-codeword per ogni codeword. Sarebbe bene scegliere queste ultime in modo che risultino *distanti* (nel senso di Hamming) l'una dall'altra!

Come discusso al § 15.6.2.1, la capacità di correzione del codice è direttamente legata alla *minima distanza* d_m, sussistendo le relazioni

- per rivelare 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

[19] Si ricorda che l'insieme delle 2^k codeword costituisce un *codebook*.

[20] Notiamo la differenza tra queste tre grandezze dall'aspetto simile: l'efficienza spettrale $\rho = f_b/B$ indica quanto una tecnica di modulazione numerica faccia buon uso dello spettro, il rapporto $\frac{B}{R}$ esprime l'inverso del grado di utilizzo della banda B di un canale numerico, mentre $R_c = k/n$ misura l'efficienza del codice adottato.

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

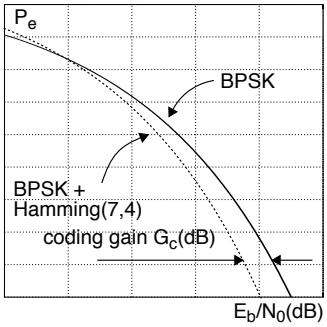
$$d_m \leq q + 1 = n - k + 1$$

che evidenzia la relazione tra d_m e la quantità di bit aggiunti q. L'uguaglianza sussiste solo per una particolare classe di codici[21] tra cui il codice a ripetizione, discusso al § 15.6.2.2, che adottando una dimensione di blocco in ingresso $k = 1$ ha però un tasso di codifica $R_c = {}^{k}/_{n} = {}^{1}/_{n}$ molto inefficiente.

Guadagno di codifica G_c La distanza d_m ed il tasso R_c non possono essere scelti indipendentemente, dato che per aumentare d_m occorre aumentare i bit di protezione q, a cui corrisponde una diminuzione del valore di R_c. Una quantità che tiene conto di entrambi i termini è il *guadagno di codifica asintotico*

$$G_c^a = d_m R_c \tag{17.22}$$

che esprime il fattore di aumento della potenza di segnale che avrebbe prodotto lo stesso miglioramento in termini di P_e dovuto all'adozione del codice.



Esempio Se $G_c = 2$ significa che l'uso del codice porta ad un valore di P_e pari a quello ottenibile con una trasmissione a potenza doppia, ma non codificata.

Tale risultato è valido nel caso di *soft-decoding*[22] (pag. 585), ed è aggettivato come *asintotico* in quanto è valido solo per valori di ${}^{E_b}/_{N_0}$ elevati: infatti all'aumentare della potenza di rumore, per qualunque codice e metodo di decisione il valore di G_c si riduce, fino a divenire inferiore ad uno, quando gli errori sono così numerosi da non poter più essere corretti.

Mostriamo ora soluzioni che consentono di ottenere un adeguato potere di correzione, senza per questo aumentare di molto la velocità di trasmissione del flusso codificato.

17.4.1 Codifica a blocco

Le proprietà di questa classe di codici possono essere meglio analizzate interpretando l'insieme delle possibili codeword da un punto di vista algebrico, ed adottando una notazione matriciale idonea a descrivere la classe di codici *a blocco*, mentre per la sottoclasse dei codici *ciclici* (§ 17.4.1.2) interviene una notazione polinomiale.

Iniziamo ricordando che il *codebook* (§ 15.6.2.1) di un codice a blocco (n, k) è composto da sequenze di n bit indicate come *codeword* $\mathbf{x}$ espresse mediante un *vettore* ad elementi binari

$$\mathbf{x} = (x_1 \quad x_2 \quad \cdots \quad x_n)$$

[21] Indicati come codici MDS, vedi `http://en.wikipedia.org/wiki/Singleton_bound#MDS_codes`.

[22] Nel caso di decisioni *hard* o bit a bit, si ottiene una espressione del tipo $G^a_{c,hard} = R_c\,(t + 1)$, in cui t è il numero di bit per parola che il codice è in grado di correggere.

che può assumere solo 2^k diversi valori tra i 2^n possibili, mentre le ricezione di una delle rimanenti $2^n - 2^k$ combinazioni di bit segnala la presenza di almeno un errore. Ad esempio per un codice a ripetizione 3:1 (§ 15.6.2.2, in cui $k = 1$ ed $n = 3$) vi sono solo due codeword con vettori $x_1x_2x_3$ pari a 000 ed 111, mentre le restanti 8-2=6 configurazioni *non sono* codeword.

Codice lineare Un codebook (che implementa un codice) è detto *lineare* se le sue 2^k codeword costituiscono uno *spazio lineare* (§ 2.4.2), ovvero se comprendono la codeword nulla, e la somma di due codeword è anch'essa una parola di codice. La somma tra due codeword è definita in base alla matematica binaria *modulo due*, ovvero espressa mediante l'operatore di OR *esclusivo* bit a bit $\oplus$ come

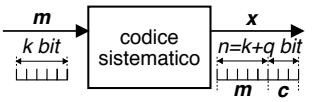
$$\mathbf{x} + \mathbf{y} = (x_1 \oplus y_1 \quad x_2 \oplus y_2 \quad \cdots \quad x_n \oplus y_n) \tag{17.23}$$

Notiamo incidentalmente che, in virtù dell'algebra modulo 2 indotta dall'operatore $\oplus$, la somma tra vettori binari produce un nuovo vettore con elementi pari ad uno nelle posizioni in cui essi differiscono, dunque in numero pari alla d_H tra i vettori.

Distanza d_m per codici lineari Definiamo ora *peso* $w(\mathbf{z})$ di una codeword $\mathbf{z}$ il numero di *uni* in essa contenuti, ovvero la sua distanza di Hamming rispetto alla CW nulla, cioè $w(\mathbf{z}) = d_H(\mathbf{z}, \mathbf{0})$. Per un codice lineare la minima distanza del codice d_m può essere valutata come il *minimo peso* tra tutte le codeword non zero[23], ossia

$$d_m = \min_{\mathbf{z} \neq 0} [w(\mathbf{z})]$$

Codice sistematico e rappresentazione matriciale Con il termine *sistematico* si intende un codice che ottiene gli n bit delle codeword $\mathbf{x}$ concatenando per primi i k bit da proteggere, indicati con $\mathbf{m}$, a cui seguono i $q = n - k$ bit di protezione, indicati con $\mathbf{c}$. Ovvero come abbiamo implicitamente assunto fino ad ora, anche se si tratta di una scelta per nulla scontata. Mostreremo più sotto che un codice sistematico è anche lineare; le sue codeword vengono quindi scritte nella forma



$$\mathbf{x} = \begin{pmatrix} m_1 & m_2 & \cdots & m_k & c_1 & c_2 & \cdots & c_q \end{pmatrix}$$

ovvero come un vettore riga partizionato $\mathbf{x} = (\mathbf{m} \mid \mathbf{c})$, in modo da poterlo calcolare a partire dal vettore $\mathbf{m}$ dei bit da proteggere moltiplicando lo stesso per una *matrice generatrice*[24] $k \times n$ con struttura generale $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$, in cui $\mathbf{I}_k$ è una matrice identità

[23] Infatti dalla definizione di somma tra CW otteniamo che $d_H(\mathbf{x}, \mathbf{y})$ è pari al peso $w(\mathbf{z})$ della codeword $\mathbf{z} = \mathbf{x} + \mathbf{y}$, ossia $\mathbf{z}$ presenta componenti $z_j = 1$ solo in corrispondenza di elementi $x_j \neq y_j$. Ma per la linearità anche $\mathbf{z}$ appartiene al codebook, ovvero sommando tra loro tutte le possibili coppie ottengo l'intero codebook, e dunque la ricerca su tutte le coppie si trasforma in una ricerca su tutte le codeword.

[24] Questa *fantomatica* matrice *generatrice* che cala dall'alto in realtà ha una genesi ben razionale.

Se infatti definiamo una base ortogonale $\{\mathbf{u}_i\}$ per lo spazio $k-$dimensionale descritto da tutti i possibili vettori $\mathbf{m}$ come i k vettori con componenti tutte nulle tranne quella in posizione $i-$esima e pari ad 1, possiamo allora scrivere un generico vettore $\mathbf{m}$ con componenti binarie m_i come una combinazione lineare di vettori della base, ovvero $\mathbf{m} = \sum_{i=1}^{k} m_i \mathbf{u}_i$.

Indicando ora con $\mathbf{g}_i$ la codeword (di n elementi) associata a ciascun vettore $\mathbf{u}_i$, otteniamo che ad un

$k \times k$ e $\mathbf{P}$ è una sotto-matrice di elementi binari $k \times q$. Le codeword si ottengono quindi come $\mathbf{x} = \mathbf{m} \cdot \mathbf{G}$, ovvero

$$\left[\begin{array}{cc} m_1 \cdots m_k & c_1 \cdots c_q \end{array}\right] = \left[\begin{array}{c} m_1 \cdots m_k \end{array}\right] \cdot \left[\begin{array}{cccccc} 1 & \cdots & 0 & p_{11} & \cdots & p_{1q} \\ \vdots & 1 & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & p_{k1} & \cdots & p_{kq} \end{array}\right] \tag{17.24}$$

in modo che $\mathbf{P}$ produca i q bit di protezione come $\mathbf{c} = \mathbf{m} \cdot \mathbf{P}$, in cui sono valide le normali regole di moltiplicazione tra matrici, tranne per l'accortezza di usare la *somma modulo due* anziché quella convenzionale. Il valore della (generica) j-esima ($j = 1, 2, \cdots, q$) componente di $\mathbf{c}$ si calcola pertanto come

$$c_j = m_1 \cdot p_{1j} \oplus m_2 \cdot p_{2j} \oplus \cdots \oplus m_k \cdot p_{kj}$$

in cui il prodotto tra cifre binarie equivale all'operatore logico di AND.

Esempio Se poniamo $\boldsymbol{m} = (0\ 1\ 1\ 0\ 1)$ e $\boldsymbol{p} = (1\ 1\ 0\ 1\ 1)$ il relativo prodotto interno vale $\boldsymbol{m} \cdot \boldsymbol{p} = 0 \cdot 1 \oplus 1 \cdot 1 \oplus 1 \cdot 0 \oplus 0 \cdot 1 \oplus 1 \cdot 1 = 0 \oplus 1 \oplus 0 \oplus 0 \oplus 1 = 0$.

In definitiva ciascuna colonna di $\mathbf{P}$ individua un sotto-insieme di elementi di $\mathbf{m}$ su cui calcolare una *somma di parità*, fornendo il motivo per cui questo sotto-blocco di matrice $\mathbf{G}$ è rappresentato dalla lettera $\mathbf{P}$. Ma non è ancora stato detto 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* (§ 17.4.1.1) ci fornisce una possibile soluzione.

Linearità di un codice sistematico La linearità di un codebook sistematico può essere verificata se riusciamo a mostrare che la somma di due qualunque codeword è ancora una parola del codebook. A tale scopo, dato che la somma modulo due tra vettori binari $\mathbf{x}_1 \oplus \mathbf{x}_2$ avviene bit per bit (eq. (17.23)), consideriamo le due parti $\boldsymbol{m}$ e $\boldsymbol{c}$ di una codeword $\mathbf{x}$ in modo indipendente. Dato che $\boldsymbol{m}$ può assumere una qualunque delle 2^k configurazioni possibili, è sempre vero che $\boldsymbol{m}_3 = \boldsymbol{m}_1 \oplus \boldsymbol{m}_2$ appartiene allo stesso insieme. Per quanto riguarda i q bit di protezione $\boldsymbol{c}$, osserviamo che

$$\boldsymbol{c}_3 = \boldsymbol{c}_1 \oplus \boldsymbol{c}_2 = \boldsymbol{m}_1\boldsymbol{P} \oplus \boldsymbol{m}_2\boldsymbol{P} = (\boldsymbol{m}_1 \oplus \boldsymbol{m}_2)\,\boldsymbol{P}$$

e dunque $\boldsymbol{c}_3$ corrisponde alla protezione di $\boldsymbol{m}_3$, ovvero $\mathbf{x}_3 = (\mathbf{m}_3 \mid \mathbf{c}_3)$ è una codeword esistente.

17.4.1.1 Codice di Hamming

E' un codice a blocco (n, k) sistematico e lineare che permette di conseguire un tasso di codifica elevato pur mantenendo un buon potere correttivo. Aggiunge $q \geq 3$ bit di controllo ai k bit informativi per formare codeword di lunghezza complessiva

$$n = 2^q - 1$$

ottenendo un tasso di codifica pari a

$$R_c = \frac{k}{n} = \frac{n-q}{n} = 1 - \frac{q}{2^q - 1}$$

generico vettore $\mathbf{m}$ è associata la codeword $\mathbf{x} = \sum_{i=1}^{k} m_i \mathbf{g}_i = \mathbf{m} \cdot \mathbf{G}$, in cui $\mathbf{G}$ è la nostra *matrice generatrice* di dimensione $k \times n$ le cui righe sono pari alle codeword $\mathbf{g}_i$ con $i = 1, 2, \cdots, k$ associate ai vettori della base $\mathbf{u}_i$.

che aumenta con il crescere di q, come mostrato in tabella. Le sue codeword si individuano ponendo *le k righe* della sottomatrice $\mathbf{P}$ pari a tutte le parole di q bit con *due o più* uni, in qualsiasi ordine. Ma la cosa *ancora più simpatica* è che per un codebook siffatto si ottiene una distanza di Hamming minima $d_m = 3$, indipendentemente dalla scelta di q.

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

Esempio: codice di Hamming $(7, 4)$. Corrisponde al caso più semplice di scegliere $q = 3$, e dunque $n = 2^3 - 1 = 7$ e $k = 7 - 3 = 4$. Una possibile matrice generatrice è pari a

$$\mathbf{G} = \left[\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{array}\right]$$

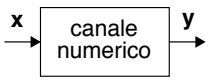
a cui corrispondono le seguenti $2^4 = 16$ codeword, per ognuna delle quali si è evidenziato il peso $w(\mathbf{x})$, confermando che $d_m = 3$.

$\mathbf{m}$	$\mathbf{c}$	$w(\mathbf{x})$
0 0 0 0	0 0 0	0
0 0 0 1	0 1 1	3
0 0 1 0	1 1 0	3
0 0 1 1	1 0 1	4
0 1 0 0	1 1 1	4
0 1 0 1	1 0 0	3
0 1 1 0	0 0 1	3
0 1 1 1	0 1 0	4

$\mathbf{m}$	$\mathbf{c}$	$w(\mathbf{x})$
1 0 0 0	1 0 1	3
1 0 0 1	1 1 0	4
1 0 1 0	0 1 1	4
1 0 1 1	0 0 0	3
1 1 0 0	0 1 0	3
1 1 0 1	0 0 1	4
1 1 1 0	1 0 0	4
1 1 1 1	1 1 1	7

Dato che sia $\mathbf{m}$ che $\mathbf{c}$ assumono tutte le configurazioni possibili, è verificata la linearità del codice (pag. 570), ossia che la somma di una qualunque coppia di codeword corrisponde ad una terza codeword. Notiamo infine che ciascuna della $2^q = 8$ triplette di protezione $\mathbf{c}$ viene usata nel codebook per due volte, ma associata a coppie di sequenze $\mathbf{m}$ da proteggere con valori di distanza di Hamming almeno pari a tre.

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

$\mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \hline \mathbf{I}_q \end{bmatrix}$ 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[25] che, se moltiplicata per una qualunque codeword valida, fornisce un vettore *nullo* di dimensione q, ossia

$$\mathbf{x} \cdot \mathbf{H} = (0 \quad 0 \quad \cdots \quad 0) \tag{17.25}$$

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 $\mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \hline \mathbf{I}_q \end{bmatrix}$ è mostrata a lato. E' 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 $\mathbf{s} = \mathbf{y} \cdot \mathbf{H} = [100]$.

$$\mathbf{H} = \left[\begin{array}{ccc} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]$$

Per quanto riguarda la *correzione*, iniziamo scrivendo 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 fornisce allora

$$\mathbf{s} = \mathbf{y} \cdot \mathbf{H} = (\mathbf{x} \oplus \mathbf{e}) \cdot \mathbf{H} = \mathbf{x} \cdot \mathbf{H} \oplus \mathbf{e} \cdot \mathbf{H} = \mathbf{e} \cdot \mathbf{H}$$

visto che come espresso dalla (17.25), la sindrome delle codeword è nulla. Ma dato 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 ad $\mathbf{e}$. Osserviamo però che, riprendendo i risultati esposti al § 15.6.1.1, la probabilità $P(m, n)$ che si siano verificati m errori su n bit decresce al crescere di m, 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:

$$\hat{\mathbf{e}} = \underset{\mathbf{e} \,:\, \mathbf{e} \cdot \mathbf{H} = \mathbf{s}}{\operatorname{argmin}} \{w(\mathbf{e})\}$$

Osserviamo inoltre che il vettore di errore *più probabile* e con *peso minimo* è quello con *un solo* bit diverso da zero; indichiamo tali n possibili vettori di errore come $\mathbf{e}_i$ ($i = 1, 2, \cdots, n$), qualora solo l'i–esimo bit (di n) sia pari ad uno. Accade poi che (per costruzione) la sindrome $\mathbf{s}_i$ associata ad $\mathbf{e}_i$, calcolata come $\mathbf{s}_i = \mathbf{e}_i \cdot \mathbf{H}$, corrisponda esattamente all'i–esima riga tra le n righe di $\mathbf{H}$. Pertanto l'indice i della riga che corrisponde alla sindrome calcolata, identifica l'indice del bit che ha subito errore.

Notiamo infine 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 com-

[25] In effetti la simpatia c'entra ben poco, ed $\mathbf{H}$ è costruita in modo che le sue q colonne siano *ortogonali* a tutte le k righe di $\mathbf{G}$ (oltre che tra loro), individuando così una base di rappresentazione per il *sottospazio* di dimensione 2^q *complemento ortogonale* di quello di dimensione 2^k descritto dalle codeword di lunghezza $n = k + q$. La dimostrazione che ho trovato (per provare che per codici sistematici il risultato è quello mostrato nel testo) contiene un errore, e non la cito.

plessivo può risultare ancora più elevato. Nel caso del codice di Hamming in cui $d_m = 3$ si può correggere un solo errore per codeword, in accordo alla procedura discussa. Se invece **e** contiene *due* errori, la sua moltiplicazione per **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. 572 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 **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.

Esercizio Un flusso binario con codifica di Hamming $(31, 26)$ è affetto da $P_e = 10^{-4}$. Determinare la probabilità *residua* di errore *sul bit* (pag. 475) dopo decodifica. **Risposta** La presenza di un solo errore nella codeword viene corretta, mentre con due errori la correzione basata sulla sindrome ne sbaglia tre; il caso con più di due errori si considera improbabile, vedi § 15.6.1.1. La prob. di 2 errori su 31 è (eq. (15.27)) pari a $P(2, 31) \simeq \frac{31 \cdot 30}{2} \left(P_e\right)^2 = 4.65 \cdot 10^{-6}$, e l'evento di errore comporta 3 bit errati su 26 decodificati, dunque per la P_e^{bit} residua si ottiene come $\frac{3}{26} \cdot 4.65 \cdot 10^{-6} = 5.36 \cdot 10^{-7}$.

17.4.1.2 Codice ciclico

Anche questo appartiene alla famiglia dei codici lineari a blocco, con la condizione aggiuntiva che se $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ è una codeword lo sono anche tutti i suoi *scorrimenti ciclici*[26], ovvero le codeword

$$\mathbf{x}^{(1)} = (x_2, x_3, \cdots, x_1), \quad \mathbf{x}^{(2)} = (x_3, x_4, \cdots, x_2), \quad \cdots, \quad \mathbf{x}^{(n)} = (x_n, x_1, \cdots, x_{n-1})$$

In tal caso sussistono delle proprietà algebriche aggiuntive che si basano sulla teoria dei *campi di Galois*[27], i cui elementi sono polinomi

$$x(p) = x_1 p^{n-1} + x_2 p^{n-2} + \cdots + x_{n-1} p + x_n$$

nella variabile p, con coefficienti binari definiti a partire dagli elementi delle codeword **x**; per tale motivo, i codici ciclici sono detti anche codici *polinomiali*[28]: senza volerci addentrare nei particolari della teoria[29], citiamo direttamente i principali risultati.

Un codice ciclico (n, k) è completamente definito a partire da un *polinomio generatore*

$$g(p) = p^{n-k} + g_2 p^{n-k-1} + \cdots + g_{n-k} p + 1$$

di grado $n - k$ che deve essere un divisore di $p^n + 1$, ovvero tale che $^{p^n+1}/_{g(p)} = q(p)$ con resto nullo[30]. Una volta definito $g(p)$ è possibile ottenere la matrice generatrice **G** del codice in forma *sistematica* $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$, calcolando le k righe di **P** come i coefficienti del polinomio *resto* della divisione tra p^i e $g(p)$, con $i = n-1, n-2, \cdots, n-k$.

[26] Ad esempio, il codice $[000, 110, 101, 011]$ è un codice ciclico, mentre $[000, 010, 101, 111]$ no. Notiamo che gli scorrimenti della codeword 000, sono la codeword stessa.

[27] `http://it.wikipedia.org/wiki/Campo_finito`

[28] Si veda `http://en.wikipedia.org/wiki/Polynomial_code`

[29] `http://en.wikipedia.org/wiki/Cyclic_code`

[30] Si applicano le regole di divisione tra polinomi `http://it.wikipedia.org/wiki/Divisione_dei_polinomi`

Esempio Troviamo la matrice generatrice in forma sistematica per un codice ciclico $(7, 4)$. Osserviamo innanzitutto che p^n+1 si fattorizza come $p^7+1 = (p+1)\left(p^3+p^2+1\right)\left(p^3+p+1\right)$, 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 la matrice generatrice risulta $\mathbf{G} = \left[\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{array}\right]$. Osserviamo che, a parte una permutazione, le righe di $\mathbf{P}$ sono le stesse di quelle che definiscono il codice di Hamming di pag. 572: infatti, Hamming rientra anche 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 si possono ottenere non solo utilizzando $\mathbf{G}$ come descritto dalla (17.24), ma anche in base alla seguente proprietà: indicando con $m(p) = m_1 p^{k-1} + m_2 p^{k-2} + \cdots + m_{k-1} p + m_k$ 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 calcolando la *convoluzione discreta*[31] tra i coefficienti di $m(p)$ e $g(p)$: $x_i = \sum_{j=1}^{k} m_j g_{i-j+1}$ per $i = 1, 2, \cdots, n$, 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 § 15.6.3.3 a proposito del CRC: i q bit di protezione possono quindi essere anche ottenuti mediante l'utilizzo di un registro a scorrimento controreazionato[32], e lo stesso può essere usato anche per il calcolo della sindrome dal lato ricevente.

17.4.1.3 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 $g(p)$ di grado massimo mt le cui radici sono potenze dell'elemento primitivo α del campo di Galois $GF(2^m)$[33]. 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

[31] Il 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.

[32] Forse ho trovato dove si spiega come faccia un registro a scorrimento controreazionato a svolgere questo genere di compiti: penso che in una prossima edizione sarà interessante aggiungerlo.

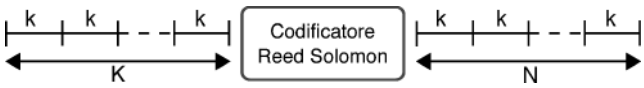
[33] `http://en.wikipedia.org/wiki/BCH_code`

soluzioni circuitali basate su registri a scorrimento controreazionati. Inoltre, è anche possibile applicare l'algoritmo iterativo di *Berlekamp-Massey*[34].

17.4.1.4 Codice di Reed-Solomon

Si tratta di un sottoinsieme dei codici BCH, caratterizzato[35] da parole di codice ad elementi *non binari* ma bensì L-ari, con L che deve essere una *potenza prima*[36]; scegliendo $L = 2^k$ ogni elemento (o simbolo) del codice corrisponde a k bit.

Un codice di *Reed Solomon* (N, K) produce codeword di N simboli $L-$ari ogni K simboli (ognuno di k bit) di informazione prelevati da $\mathbf{m}$; esistono formulazioni in grado di produrre codebook sia di natura sistematica che non. La minima distanza del codice è pari a $d_m = N-K+1$, ed i metodi di decodifica (su cui non ci addentriamo[37]) permettono di correggere fino a $t = \frac{d_m-1}{2} = \frac{N-K}{2}$ simboli $L-$ari per codeword; qualora *la posizione* dei simboli errati sia nota[38] può correggere il doppio dei simboli, ossia $2t$. Dato che opera a livello di simbolo e non di singolo bit, non soffre del problema legato alle sequenze di errore come invece avviene per il codice di Hamming, almeno finché il numero di errori sul bit consecutivi non supera il valore $t \cdot k$, ovvero non coinvolge più di t simboli.



Esempio Scegliendo $k = 8$, $N = 2^k - 1 = 255$, $t = 16$ e $K = N - 2t = 223$ definiamo il codice $RS\,(255, 223)$ che protegge $K \cdot k = 1784$ bit inserendoli in codeword di $N \cdot k = 2040$ bit, di cui $2t \cdot k = 256$ di ridondanza. Tale codice è in grado di correggere fino a 16 simboli ad 8 bit: qualora gli errori sul bit avvengano tutti in simboli differenti si ha il caso *peggiore*, in cui sono correggibili solamente 16 bit su 2040; viceversa nel caso *migliore* in cui gli errori sono consecutivi ed iniziano all'inizio di un simbolo, ne corregge fino a $t \cdot k = 128$. Il tasso di codifica risulta $R_c = {}^{K}/_{N} = 0.937$.

Codice accorciato Può accadere che i bit che costituiscono il flusso f_b di dati da proteggere non sia *omogeneo*, ma presenti strutture sintattiche e quindi delimitazioni che si desidera riflettere nel flusso codificato[39]; oppure, è il sistema di trasmissione ad imporre strutture di trama a dimensione fissa, e non si desidera frammentare una singola codeword a cavallo di time-slot differenti. In tali casi si omette la codifica (e la trasmissione) di alcuni dei K simboli di informazione, idealmente posti a zero, riducendo così la dimensione N della codeword pur mantenendo la stessa quantità di ridondanza $N - K$, accettando di ridurre il tasso di codifica R_c.

[34] `http://en.wikipedia.org/wiki/Berlekamp-Massey_algorithm`

[35] `http://en.wikipedia.org/wiki/Reed-Solomon_error_correction`

[36] Ovvero deve essere della forma $L = \alpha^k$ con α numero primo e k intero positivo. Ciò è necessario affinché gli L simboli corrispondano agli elementi di un campo di Galois $GF\,(L)$.

[37] Ma si veda ad es. `http://scienze-como.uninsubria.it/previtali/Bellini-TeoriaInfoCodiciNote.pdf`

[38] Come nel caso dei codici a cancellazione, vedi ad es. `https://en.wikipedia.org/wiki/Erasure_code`

[39] Come ad esempio avviene nella codifica di sorgenti multimediali (cap. 10)

Esempio Il codice accorciato $RS\,(204, 188)$[40] viene ricavato dal $RS\,(255, 239)$ (con 16 byte di ridondanza) ponendo a zero (e non trasmettendo) 51 dei 239 byte di informazione. Il tasso di codifica passa da 0.937 a 0.921.

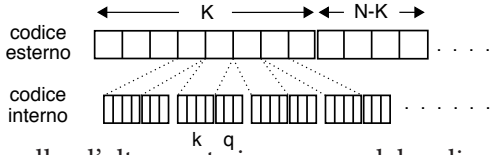
In questi casi la codeword viene calcolata, in formato sistematico, a partire da una sequenza informativa fittizia che contiene anche i simboli posti a zero e poi non trasmessi; qualora in fase di decodifica questi risultino diversi da zero[41], la codeword viene segnalata come errata agli stadi di elaborazione successivi.

17.4.1.5 Codifica concatenata

Come abbiamo fatto notare i codici Reed Solomon riescono a correggere efficacemente errori ripetuti, al contrario di quelli di Hamming, che invece hanno un buon comportamento in presenza di errori singoli. E dato che l'unione fa la forza, è possibile usarli assieme adottando lo schema mostrato in figura 17.3, in cui la prima codifica (di *RS*) è detta *esterna* perché più lontana dal canale, mentre per converso la seconda (di Hamming) è detta *interna*.

Figura 17.3: Schema di codifica concatenata

Ad ogni simbolo di k bit della codeword esterna vengono aggiunti q bit di ridondanza da parte del codice interno, come mostrato a lato. Gli errori *isolati* introdotti da parte del canale sono corretti dal decodificatore interno, e se distanti tra loro per più di $k+q$ bit, il codice esterno neanche si accorge di nulla; d'altra parte in assenza del codice interno, il solo codice di *RS* non sarebbe stato in grado di correggere più di $\frac{N-K}{2}$ errori per codeword, qualora avvenuti ognuno in un simbolo diverso. Al contrario, in presenza di errori ravvicinati è il codice interno di Hamming a non poter fare nulla, anzi ne introduce di ulteriori, ma nello stesso simbolo: in tal caso il codice esterno di *RS* trova gli errori tutti concentrati in pochi simboli, e provvede alla loro correzione.



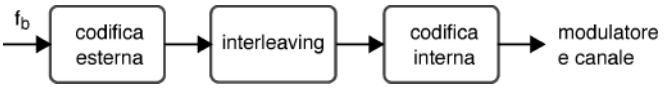
Codice prodotto E' un nome alternativo dato a questo schema di funzionamento[42], e riflette il fatto che il tasso di codifica R_c complessivo è il prodotto dei tassi dei due stadi, dando luogo ad una velocità binaria in ingresso al canale pari a $f'_b = \frac{f_b}{R_c^i \cdot R_c^e}$.

Interleaving Qualora si manifestino sequenze di errore più lunghe di $\frac{N-K}{2} \cdot k$ bit, il numero di simboli del codice esterno che risultano errati diviene maggiore di quanto

[40] Il valore di 188 byte deriva dalla dimensione di una cella ATM (48 byte di dati e 5 di intestazione, § 23.2): infatti $48 \cdot 4 = 192$, ed una codeword si suddivide su 4 celle.

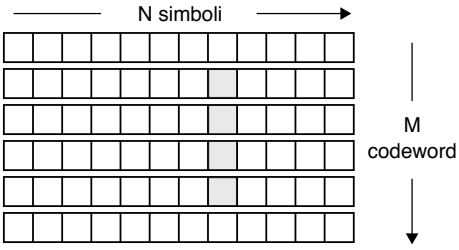
[41] Come osservato a pag. 573 in presenza di un numero di errori superiori al massimo correggibile, se ne verificano ancora di più.

[42] O meglio, ad una versione dello schema in cui la ridondanza interna viene calcolata *sulla totalità* dei simboli nella stessa posizione di M codeword esterne, come avviene per la tecnica esposta al § 15.6.3.2.



esso non possa correggere, e lo schema precedente non funziona più. Per evitare questa situazione, tra i due stadi di codifica (e di decodifica) si frappone un blocco di *interleaving* (ed il suo inverso) (vedi § 15.6.2.3), come mostrato sopra.

La matrice di interleaving viene scritta per righe, inserendovi per intero M codeword del codice esterno, ed è letta per colonne dalla codifica interna, che vi aggiunge ulteriore ridondanza; dal lato ricevente la stessa matrice viene scritta *per colonne* con l'uscita della decodifica interna, e letta *per righe* da parte della decodifica esterna.

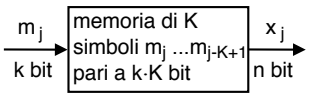


Le aree grigie in figura rappresentano una lunga sequenza di errore, che il codice interno non riesce a correggere: finché M (detto *fattore di interleaving*) è maggiore della più lunga sequenza di simboli errati previsti, questi ultimi vanno a finire in codeword (esterne) differenti, e possono dunque essere ancora corretti. Mentre in assenza di interleaver, i simboli errati sarebbero potuti finire tutti dentro la stessa codeword esterna.

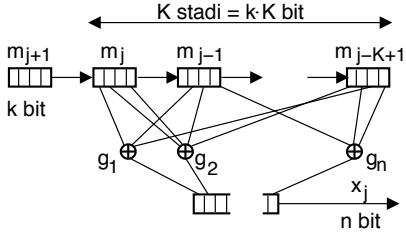
Resta una ultima *sventura* possibile, ovvero la presenza di un disturbo *periodico* che si presenta alla stessa cadenza (o con un suo multiplo) con cui sono scritte le colonne da parte della decodifica interna. Ciò determina che tutti i simboli di una medesima codeword esterna siano errati, rendendone impossibile la correzione. Per rimediare anche a questa evenienza occorre adottare un metodo di interlevaing detto *convoluzionale*, il cui funzionamento non viene approfondito[43], precisando solamente che quello discusso sopra è invece detto interleaver *a blocco*.

17.4.2 Codifica convoluzionale

A differenza della codifica a blocco, questa tecnica produce una sequenza binaria i cui valori dipendono da gruppi di bit di ingresso *temporalmente sovrapposti*, in analogia formale a quanto avviene con l'integrale di convoluzione, che calcola valori di uscita che dipendono da *intervalli* di quelli in ingresso, pesati dai valori della risposta impulsiva. Un generico codice convoluzionale è indicato con la notazione $CC\,(n, k, K)$, che lo descrive come capace di generare gruppi di n bit di uscita (sequenza $\{x\}$) in base alla conoscenza di K simboli di ingresso (sequenza $\{m\}$), ognuno composto da k bit.



Lo schema strutturale del codificatore, raffigurato a lato, ospita i K simboli della sequenza di ingresso (di k bit ciascuno) in un registro a scorrimento in cui per ogni nuovo simbolo m_j che entra da sinistra, i precedenti scorrono a destra, ed il più "vecchio" viene dimenticato. Ognuno



[43] Ma vedi ad es. `https://en.wikipedia.org/wiki/Burst_error-correcting_code#Convolutional_interleaver`

degli n bit di uscita $x_j\,(i)$, $i = 1, 2, ..., n$ è calcolato eseguendo una somma modulo 2 tra alcuni dei $k \cdot K$ bit di ingresso, individuati da un vettore *generatore* g_i $(i = 1, 2, ..., n)$ costituito da una parola binaria di $k \cdot K$ bit, zero od uno a seconda se l'*i-esimo* sommatore modulo due sia connesso (o meno) al corrispondente bit della *finestra* di ingresso.

Il numero $L = k \cdot K$ di bit che contribuiscono al calcolo della parola di uscita viene indicato come *lunghezza del vincolo*, mentre la quantità $\nu = (K - 1) \cdot k$ è indicata come *memoria* del codice per il motivo che vedremo presto. Resta poi valido il concetto di *coding rate* $R_c = \frac{k}{n}$ che rappresenta il rapporto tra quanti *nuovi* bit di informazione sono necessari in ingresso per ogni gruppo di n bit in uscita dal codificatore.

Automa e diagramma di transizione Il numero di possibili configurazioni dei bit contenuti nello shift register è finito e pari a $2^L = 2^{K \cdot k}$. In base alla scelta degli n vettori g_i, ad ogni configurazione corrisponde un unico valore dell'uscita, e ciò comporta che lo schema può essere descritto da una *tabella della verità*, ovvero da un *automa* a stati finiti, con annesso diagramma di transizione.

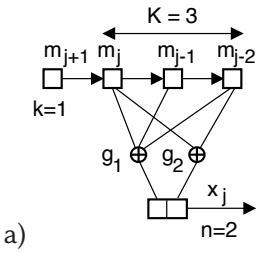
Osserviamo ora che il passaggio da uno stato all'altro non è qualsiasi, ma è stabilito dall'ultimo simbolo in ingresso m_j, e pertanto il *diagramma di transizione* si costruisce individuando i 2^ν *stati* S associati alle possibili combinazioni di bit dei precedenti $K - 1$ simboli di ingresso. Ad ogni stato competono 2^k transizioni, una per ogni possibile m_j di ingresso, diretta verso lo stato individuato dalla nuova configurazione di shift-register che si è determinata. Infine, ad ogni transizione è associato in modo univoco il gruppo di n bit $x_j\,(m_j, S_j)$ da emettere in uscita[44]. Ma prima che il discorso diventi troppo confuso, procediamo con un esempio pratico.

Codice CC(2,1,3) Consideriamo il codice convoluzionale il cui schema è mostrato in fig. 17.4-a), che produce due bit di uscita per ogni bit di ingresso in funzione degli ultimi tre, ovvero $CC\,(n, k, K) = CC\,(2, 1, 3)$, caratterizzato da un tasso di codifica $R_c = \frac{k}{n} = \frac{1}{2}$, da una lunghezza del vincolo $L = K \cdot k = 3$, una memoria $\nu = (K - 1) \cdot k = 2$ ed un numero di stati $2^\nu = 4$, da ognuno dei quali si dipartono $2^k = 2$ transizioni. Se scegliamo come vettori generatori le parole[45] $g_1 = (1\ 1\ 1)$ e $g_2 = (1\ 0\ 1)$ si ottiene la tabella della verità di fig. 17.4-b), che mostra i valori di uscita $x_j\,(i)$ ottenuti eseguendo gli XOR descritti dai vettori g_i sulla parola di tre bit costituita dall'ultimo ingresso m_j e dai due precedenti ingressi, indicati come S_j, e che rappresentano appunto *la memoria del passato* all'istante j. Infine la fig. 17.4-c) rappresenta l'automa ed il diagramma di transizione corrispondente[46], in cui le transizioni tra stati sono disegnate con linee tratteggiate o continue a seconda se il valore m_j dell'ultimo ingresso sia pari a zero o ad uno, e sono etichettate con la coppia di bit in uscita x_j riportati nella tabella della verità. Come si può verificare, ogni stato ha *solo due* transizioni, e dunque per ogni

[44] Lo stesso valore di x_j potrebbe essere prodotto da più di una delle $2^{k \cdot K}$ diverse memorie del codificatore.

[45] Il valore di un vettore generatore viene spesso espresso in notazione ottale, dunque nel nostro caso avremmo $g_1 = 7_8$ e $g_2 = 5_8$.

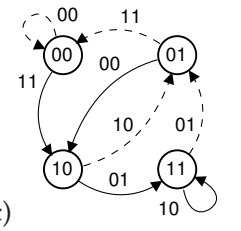
[46] In accordo allo schema `https://it.wikipedia.org/wiki/Macchina_di_Mealy`



a)

b)

m_j	S_j	$x_j(1)$	$x_j(2)$
0	00	0	0
0	01	1	1
0	10	1	0
0	11	0	1
1	00	1	1
1	01	0	0
1	10	0	1
1	11	1	0



c)

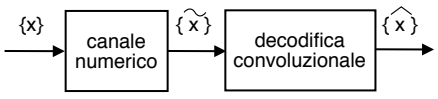
Figura 17.4: a) - architettura del codice convoluzionale $CC(2,1,3)$; b) - tabella della verità; c) - diagramma di transizione; le linee tratteggiate indicano uno zero in ingresso

ingresso sono possibili solo *due valori* dei quattro 4 che sarebbero possibili con due bit; inoltre questi valori differiscono in *entrambi i bit*[47].

Esempio Con una sequenza di ingresso $\{m\} = \{1010\}$ si osserva che, seguendo il diagramma di transizione a partire dallo stato iniziale 00, la sequenza di stati risulta $\{S\} = \{00, 10, 01, 10, 01\}$, mentre quella di uscita è $\{x\} = \{11, 10, 00, 10\}$. E' d'altra parte possibile anche il procedimento inverso, ossia conoscendo $\{x\}$ e quindi $\{S\}$ si può risalire ad $\{m\}$, sempre percorrendo le transizioni etichettate con i simboli x_j. In definitiva, osserviamo come ad ogni coppia di sequenze $(\{m\}, \{x\})$ sia biunivocamente associata una sequenza di stati $\{S\}$.

Diagramma a traliccio Per meglio visualizzare le possibili sequenze di stati si adotta una rappresentazione detta *diagramma a traliccio* (TRELLIS) del codificatore, ottenuta a partire dalla *riorganizzazione* grafica del diagramma di transizione (fig. 17.4-c)) come mostrato a sin. in fig. 17.5-a). Il traliccio è quello al centro, costituito da *nodi* disposti su tante righe quanti sono gli stati dell'automa, e tante colonne quanti sono gli istanti temporali che vogliamo considerare. I collegamenti tra colonne del traliccio corrispondono alle transizioni dell'automa: ponendo uno stato iniziale $S_{j=0} = 00$, si costruisce la colonna $j = 1$ riportando le due possibili transizioni, tratteggiata o continua a seconda che sia entrato uno zero od un uno, e si etichettano le transizioni con la parola x_j emessa in uscita. Il processo si ripete per tutti gli istanti temporali. In basso in figura è riportata una possibile sequenza codificata $\{x\}$ *trasmessa*, in corrispondenza della quale *l'effettivo* percorso attraversato nel traliccio, e dunque la successione di stati $\{S\}$ associata, è rappresentato dalle linee *blu* e più spesse.

17.4.2.1 Criterio di decodifica



Come già osservato non tutte le sequenze di stati (e di uscite) sono ammissibili; indichiamo con $\mathcal{X}$ il loro insieme. Consideriamo quindi la sequenza $\{\tilde{x}\}$ osservata all'uscita di un canale rumoroso, e contenente *errori*. In virtù dei vincoli, nel caso di errori sufficientemente isolati non esiste una sequenza di stati

[47] La dipendenza di x_j da (m_j, S_j) è legata alla scelta dei generatori g_i. Nel caso in cui un valore $x_j(i)$ sia sempre uguale ad uno dei k bit di m_j, il codice diviene *sistematico*.

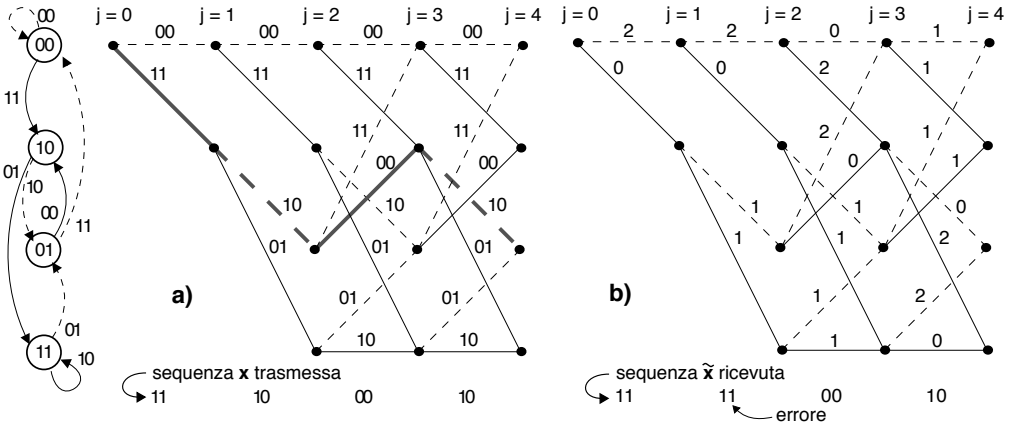


Figura 17.5: a) - diagramma a traliccio e sequenza $\boldsymbol{x}$ trasmessa; b) - costi d_H per la sequenza ricevuta; le linee tratteggiate indicano uno zero in ingresso

$\{S\}$ in grado di produrre *esattamente* la sequenza ricevuta $\{\tilde{x}\}$, e la correzione degli errori consiste nel trovare la sequenza $\{\hat{S}\}$ *ammissibile* e tale che la corrispondente uscita $\{\hat{x}\}$ sia la più *simile* a $\{\tilde{x}\}$. La metrica adottata è la *distanza di Hamming*[48] $d_H(\widehat{x}, \tilde{x})$ tra le due sequenze, ossia il numero di bit diversi tra le due, ed il criterio di decodifica viene definito come

$$\{\hat{x}\} = \underset{\{x\}\in\mathcal{X}}{\arg\min}\{d_H(x, \tilde{x})\} \tag{17.26}$$

Decodifica di Viterbi E' il nome della tecnica basata sui principi della programmazione dinamica[49] che permette di risolvere il problema (17.26) senza dover enumerare[50] tutte le sequenze ammissibili $\{x\} \in \mathcal{X}$.

A tale scopo gli archi del traliccio vengono etichettati come in fig. 17.5-b), con le $d_H(x_j, \tilde{x}_j)$ tra il simbolo x_j associato alla transizione e quello $\tilde{x}_j$ ricevuto all'istante j. Per ogni particolare sequenza di stati $\{S\}$ il *costo* $d_H(x, \tilde{x})$ tra la sequenza ricevuta $\tilde{x}$ e quella x associata ad S è pari alla somma delle $d_H(x_j, \tilde{x}_j)$ che etichettano gli archi attraversati, ossia[51]

$$d_H(x, \tilde{x}) = \textstyle\sum_j d_H(x(S_j/S_{j-1}), \tilde{x}_j) \tag{17.27}$$

Il criterio (17.26) è dunque equivalente a quello di trovare il percorso *di minimo costo*[52] per l'attraversamento di un grafo valutato. Ciò avviene esaminando il traliccio

[48] Questo caso viene indicato con il termine *hard-decision decoding* in quanto il ricevitore *ha già* operato una decisione (quantizzazione) rispetto a $\tilde{x}$. Al contrario, se i valori ricevuti sono passati *come sono* al decodificatore di Viterbi, questo può correttamente valutare le probabilità $p(\tilde{x}/\widehat{x})$ ed operare in modalità *soft decoding*, conseguendo prestazioni migliori.

[49] Vedi as es. `https://it.wikipedia.org/wiki/Programmazione_dinamica`

[50] Dato che da ogni stato si dipartono 2^k archi, ad ogni istante j il numero di percorsi alternativi aumenta di un fattore 2^k, crescendo molto velocemente all'aumentare di j.

[51] Ad esempio, con riferimento alla fig. 17.5, la $\{S\} = \{00, 10, 11, 01, 10\}$ ha un *costo* pari a 3.

[52] Qualora la distanza tra $\tilde{x}_j$ ed un possibile x_j sia invece espressa da una probabilità condizionata $p(\tilde{x}_j/x_j)$, il processo di decodifica è detto di massima verosimiglianza e la decodifica è detta soffice (soft), vedi § 17.4.2.3.

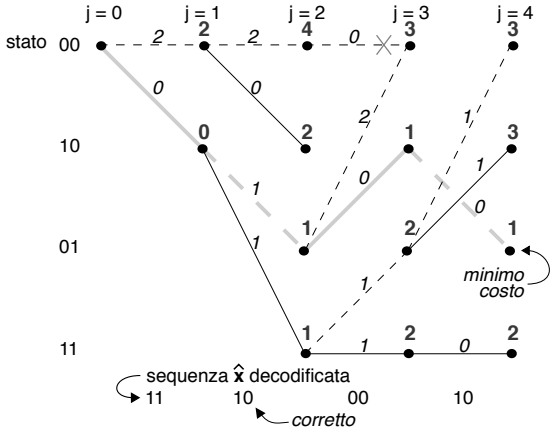


Figura 17.6: Decodifica di Viterbi come percorso di minimo costo

per colonne da sinistra a destra, e valutando per ciascun nodo (riga) il *miglior* costo *parziale*[53] tra i diversi percorsi che lo raggiungono.

Esempio Applichiamo l'algoritmo descritto al caso in questione con l'aiuto della figura 17.6 (ottenuta a partire dalla 17.5-b)) che mostra sopra ad ogni nodo il risultato del calcolo del *costo parziale* del *miglior* percorso che lo raggiunge, ottenuto sommando i costi di attraversamento degli archi (fig. 17.5-b)) che lo compongono, mostrati in corsivo. Ad esempio, lo stato 00 all'istante $j = 3$ potrebbe essere raggiunto tramite il percorso tutto *orizzontale* che rimane nello stato 00, e che assomma un costo parziale di 4. A questo si preferisce il percorso proveniente dallo stato 01, che ha accumulato (per $j = 2$) un costo parziale di 1, a cui sommare $d_H(x(S_{00}/S_{01}), \tilde{x}_3) = d_H(11, 00) = 2$, per un nuovo costo parziale di 3.

All'estremità destra di figura 17.6 viene indicato il percorso di minimo costo, e la corrispondente $d_H(x, \tilde{x})$ minima. In effetti l'algoritmo calcola solamente la d_H minima, e per risalire alla sequenza di stati (e dunque di uscite) che l'ha prodotta, occorre (per ogni nodo del traliccio) memorizzare l'indice del nodo nella colonna precedente da cui parte l'arco che ha determinato il minor costo locale. Una volta individuato (all'ultima colonna) il nodo in cui termina il percorso di minor costo, svolgendo all'indietro la catena dei puntatori si trova la sequenza di stati $\{S\}$ ottima (percorso *verde* a tratto spesso), e da questa in base alla fig. 17.5-a) sia la $\{m\}$ che la $\{\tilde{x}\}$, che come si vede *è quella esatta.*

In fig. 17.7 è riportato un esempio di *pseudo-codice* che implementa l'algoritmo a partire da tre tabelle $m(p, q)$, $x(p, q)$ e $A(p, q)$, di dimensione $2^\nu \times 2^\nu$, che contengono rispettivamente i valori di ingresso m e di uscita x in corrispondenza ad una transizione da p a q, ed un valore pari a 1 o 0 a seconda se la transizione esiste o meno. Il codice opera su di una sequenza di ingresso $\tilde{\boldsymbol{x}}$ di J elementi, e ne produce una $\boldsymbol{m}$ di uguale lunghezza,

[53] La scelta del miglior percorso parziale è l'applicazione del principio di programmazione dinamica, secondo il quale "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".

```
TABELLE:
  m(p,q)                          # ingresso m sull'arco da S_p a S_q
  x(p,q)                          # uscita x sull'arco da S_p a S_q
  A(p,q)                          # presenza di transizione da S_p a S_q
INIZIALIZZA:
per tutti gli stati p = 1,2,··· 2^ν
  CJ(p) = 0
ITERAZIONE:
per tutti gli istanti j = 1,2,··· ,J
  per tutti gli stati q = 1,2,··· 2^ν
    CM(q) = CJ(q)                 # miglior costo parziale al tempo j − 1
    CJ(q) = ∞                     # miglior costo fino a (q, j)
    per tutti gli stati p = 1,2,··· ,2^υ
      se A(p,q) == 1              # esiste transizione da p a q
        c = d_H(x(p,q), x̃_j) + CM(p)  # ipotesi da valutare
        se c < CJ(q)
          CJ(q) = c
          BP(q,j) = p             # backpointer al vincitore
DECODIFICA:
best = ∞
per tutti gli stati q = 1,2,··· 2^ν
  se CJ(q) < best
    best = CJ(q)
    w = q                         # indice stato terminale
per tutti gli istanti j = J,··· ,2,1
  p = BP(w,j)                     # miglior predecessore
  emette m(p,w)                   # messaggio m scritto al contrario
  w = p
```

Figura 17.7: Pseudo codice dell'algoritmo di Viterbi

ma temporalmente invertita. Ovviamente si possono pensare implementazioni molto più efficienti, ma l'esempio ha il solo scopo dimostrativo.

Considerazioni

- l'esempio adottato si mostra in grado di correggere un errore pur impiegando un coding rate pari ad $\frac{1}{2}$, migliore (ad es.) di quello $\left(\frac{1}{3}\right)$ del codice a ripetizione;
- la d_H del miglior percorso corrisponde al numero di bit errati (nel caso in cui siano stati corretti) nella $\tilde{x}$ *ricevuta*, il che permette al decodificatore di *stimare* la qualità del collegamento;
- si verifica errore se esiste una $\hat{x} \neq x$ *ammissibile* e tale che $d_H(\widehat{x}, \tilde{x})$ è *minore* di $d_H(x, \tilde{x})$. In tal caso la sequenza erroneamente decodificata $\hat{x}$ contiene errori *a pacchetto*;
- le capacità di correzione del codice migliorano aumentando la d_H tra le possibili sequenze $\{x\}$. La minima distanza d_m tra sequenze codificate è denominata *distanza libera*, e può essere trovata come la d_H tra una $\{x^0\}$ tutta nulla ($\{x^0\} = \{000000000\}$) e quella con il minor numero di uni, che si diparte e ritorna (nel traliccio) dallo/allo stato 00: nell'esempio di fig. 17.5a) si ottiene $d_m = 5$. Il

decodificatore è in grado di correggere fino a $\frac{d_m-1}{2}$ errori che si presentano in un intervallo pari al vincolo del codice L. Il codice del nostro esempio può dunque correggere 2 bit errati su 5;

- la distanza libera d_m aumenta con il rapporto $\frac{K}{k}$, in quanto la matrice di transizione tra stati diviene più sparsa, ed i valori di $\{x\}$ 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 $R_c = \frac{k}{n}$;
- in presenza di un flusso di dati continuo non ha senso attendere un istante finale (che non esiste) per poter individuare il percorso di minimo costo. In tal caso la decodifica parziale avviene con riferimento ad una colonna del traliccio temporalmente abbastanza lontana dall'ultimo istante j di ingresso, ad es. cinque volte la lunghezza del vincolo L, liberando di conseguenza la memoria occupata dai puntatori precedenti;
- dato che ad ogni istante-colonna sono scartati i percorsi che si incontrano con uno migliore, il numero di percorsi *sopravvissuti* è sempre pari al numero di stati 2^ν;
- all'aumentare del numero 2^ν di stati aumenta la complessità e l'occupazione di memoria dell'algoritmo di Viterbi, che può essere sostituito da tecniche di ricerca a fascio (*beam search*) che estendono solo i percorsi parziali più promettenti, ossia con minor costo parziale.

17.4.2.2 Tail biting

Letteralmente traducibile come *mordendo la coda*, è un metodo per rendere un codice convoluzionale simile ad uno *a blocchi*. Il bitstream di ingresso viene suddiviso in segmenti di m bit, gli ultimi L dei quali sono utilizzati (in ordine inverso) per *inizializzare* lo shift-register del codificatore, che quindi inizia ad elaborare, uno alla volta e dall'inizio, i bit del segmento. Al suo termine, ovvero quando sarà entrato l'$m-$esimo bit del segmento, lo stato del codificatore sarà identico a quello con cui ha iniziato: ciò significa che se venisse di nuovo inserito lo stesso segmento (ovvero in forma periodica), si otterrebbe la stessa uscita, che può dunque essere riguardata nel suo insieme come *la codeword* associata al segmento.

Una variante della tecnica (detta *zerotail*), anziché inizializzare il codificatore ne azzera lo stato, ed aggiunge L bit pari a zero in coda al segmento, ottenendo lo stesso risultato, a spese di un peggioramento di R_c. Con la tecnica *zerotail* la decodifica *sa* che il percorso nel traliccio deve iniziare e terminare con lo stato nullo, e dunque al termine della *digestione* della codeword associata al segmento si può iniziare il backtraking dei puntatori da li, oppure segnalare la presenza di eccessivi errori, qualora non sia quello il percorso di minimo costo.

Con l'inizializzazione dello stato, invece, il traliccio diviene periodico (alcuni lo definiscono *circolare*), e l'algoritmo di Viterbi viene fatto lavorare su di una periodiz-

zazione della codeword ricevuta; dopo alcuni periodi si individua lo stato di minimo costo, e recuperati i puntatori (per un periodo-codeword) a partire da quello.

17.4.2.3 Decodifica a decisione soffice

Fino ad ora la decodifica di canale è stata applicata *a valle* del processo di decisione[54], in modo da individuare le codeword come quelle con la minima *distanza di Hamming* rispetto al risultato prodotto dal decisore. Se invece la decodifica di linea non esegue nessuna decisione, ma inoltra interi blocchi di n campioni $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ prelevati dal segnale ricevuto, la decodifica di canale può individuare le codeword trasmesse come quelle con la minima *distanza euclidea* d_E rispetto a quanto ricevuto[55], ottenendo prestazioni migliori a spese di un maggior onere di calcolo. Questa tecnica valuta la distanza come

$$d_E(\mathbf{y}, \mathbf{x}^i) = \sum_{j=1}^{n} \left(y_j - x_j^i\right)^2 \tag{17.28}$$

tra i valori (continui) y_j ricevuti e quelli (binari) che si sarebbero dovuti ricevere nell'ipotesi che fosse stata trasmessa la $i-$esima delle possibili codeword $\mathbf{x}^i = \left(x_1^i, x_2^i, \cdots, x_n^i\right)$, decidendo quindi per la codeword $\hat{\mathbf{x}}$ che rende $d_E(\mathbf{y}, \hat{\mathbf{x}})$ minima.

Decodifica di massima verosimiglianza Di fatto minimizzare la distanza (17.28) equivale a massimizzare la verosimiglianza logaritmica $\ln\left(p_n\left(\mathbf{y}/\mathbf{x}^i\right)\right)$ dell'ipotesi che sia stata trasmessa la codeword $\mathbf{x}^i$ avendo ricevuto la v.a. $n-$dimensionale $\mathbf{y}$ affetta da rumore gaussiano n di varianza σ^2. Infatti considerando i bit incorrelati[56], $p_n\left(\mathbf{y}/\mathbf{x}^i\right)$ è pari al prodotto delle $p_n\left(y_j/x_j^i\right)$ dei singoli bit, e $\ln\left(p_n\left(\mathbf{y}/\mathbf{x}^i\right)\right)$ diviene una somma, i cui termini[57]

$$\ln\left(p_n\left(y_j/x_j^i\right)\right) = -\ln\left(\sqrt{2\pi}\sigma\right) - \frac{\left(y_j - x_j^i\right)^2}{2\sigma^2} \tag{17.29}$$

sono legati a quelli (cambiati di segno) che compaiono nella (17.28), dato che ne il primo termine di (17.29) ne il denominatore del secondo intervengono nella minimizzazione di (17.28).

Mentre per un codice a blocchi appare problematico valutare la (17.28) per tutte le possibili $\mathbf{x}^i$, le considerazioni ora svolte sono invece particolarmente adatte al contesto della codifica convoluzionale, dato che basta sostituire nella (17.27) la d_H con la d_E per il calcolo dei costi associati ai percorsi nel traliccio esplorato dall'algoritmo di Viterbi. In tal caso si assiste ad un miglioramento di prestazioni (rispetto all'approccio *hard)* equivalente ad un aumento di E_b/N_0 di circa 2 dB.

Prestazioni La figura che segue riporta i valori della P_e^{bit} *residua* per la decodifica di Viterbi di un *CC* con $R_c = 1/2$ al variare della lunghezza del vincolo L, il cui bitstream viene trasmesso mediante modulazione QPSK, in funzione del rapporto E_b/N_0 in ricezione.

[54] Per questo detta *hard decision decoding* in quanto opera su decisioni già prese.

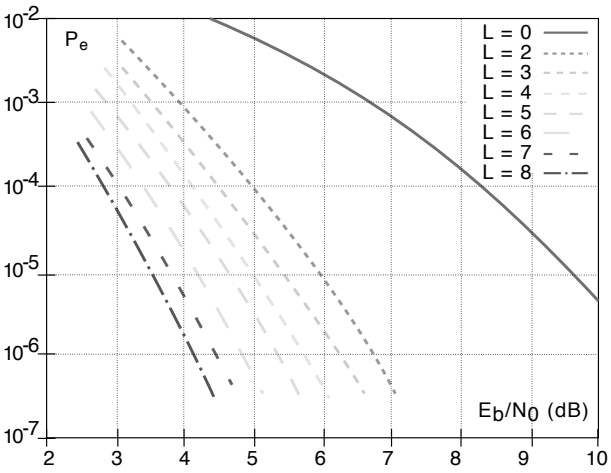
[55] Tale possibilità è indicata come *soft* in quanto richiede operazioni in virgola mobile; si veda ad es. la spiegazione fornita presso
`http://www.gaussianwaves.com/2009/12/hard-and-soft-decision-decoding-2/`.

[56] Anche se la codifica introduce correlazione, l'ipotesi è troppo comoda per poter arrivare al risultato!

[57] Vedere come si è proceduto a pag. 635.

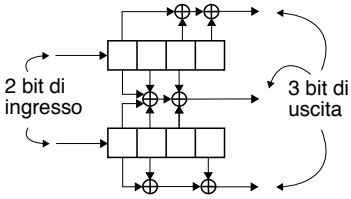
Osserviamo il conseguimento di un guadagno di codifica (pag. 569) che per una $P_e = 10^{-5}$ va da circa 3.5 a 6 dB per le diverse scelte di L, che corrispondono a tralicci da 4 a 256 stati.

Oppure, visto nell'altro senso, per un E_b/N_0 di 5 dB otteniamo una P_e migliore tra 100 e più di 10000 volte rispetto al caso non codificato.

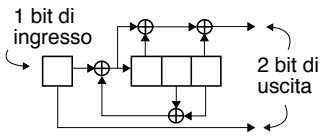


17.4.2.4 Altri schemi di codifica convoluzionale

Lo schema riportato nell'esempio di fig. 17.4 non è l'unico possibile. Anche se permette di variare il tasso di codifica $R_c = \frac{k}{n}$ variando i valori di k ed n, questi impattano anche su complessità del traliccio, lunghezza del vincolo L e distanza libera d_m; per garantirsi la massima libertà di azione sui parametri della codifica, sono state definite anche architetture differenti. Lo schema riportato a lato ad esempio dispone i $k = 2$ bit di ingresso su due diversi rami, ed impiegando tre generatori, consegue un tasso $R_c = \frac{k}{n} = \frac{2}{3}$ con 64 stati, $K = 8$ ed $L = 8$.



Anche se non è stata affrontata la parte di teoria che descrive i *CC* in termini di risposta impulsiva e risposta in frequenza, non può sfuggire la similitudine tra le architetture illustrate e quelle dei filtri FIR, pur se in logica modulo due. Ma esiste anche la possibilità di realizzare un *CC ricorsivo*, in cui cioè sono presenti operazioni XOR *all'indietro* e non solo in avanti, come nello schema mostrato in figura[58] che consegue $R_c = 1/2$, e che è anche *sistematico* in quanto un bit di uscita su due replica quello di ingresso.



17.4.2.5 Codice perforato

Modificare completamente l'architettura del codificatore non sembra la scelta migliore per poter variare R_c, dato che ciò comporta la totale modifica del traliccio su cui è basata la decodifica. Una diversa opzione è quella di *omettere* del tutto la trasmissione di alcuni dei bit presenti nella sequenza codificata, operazione nota come perforazione (*puncturing*) del codice: ad esempio eliminare un bit ogni tre dall'uscita di un *CC* con $R_c = 1/2$ determina un nuovo $R_c = 3/4$, dato che servono ora tre bit di ingresso per produrne quatto di uscita, anziché sei. Ovviamente in questo

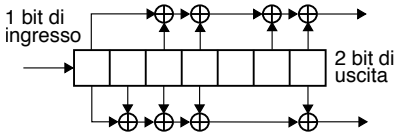
[58] In particolare, lo schema illustrato viene impiegato nella telefonia LTE (ETSI TS 136 212) nell'ambito della codifica *turbo*, vedi § 17.5.1.

caso le capacità correttive peggiorano, ma il processo di decodifica può aver luogo comunque dato che il ricevitore conosce le posizioni non trasmesse, e per esse considera una distanza *neutra* pari a 0.5, come esemplificato alla tabella a lato, dove con δ_i si indica la distanza (hard o soft) tra l'$i-$esimo bit nella sequenza ricostruita, e le corrispondenti ipotesi espresse dal traliccio.

sequenza di ingresso	1	0	1	1	0	0	0
sequenza codificata, $R_c = 0.5$	11	10	00	01	01	11	00
perforazione al tasso $R_c = 0.75$	11	$_1$0	0$_0$	01	$_0$1	1$_1$	00
bit trasmessi dal modulatore	11	00		01	11		00
ricostruzione per la decodifica	11	x0	0x	01	x1	1x	00
metrica	$\delta_1\delta_2$	$\frac{1}{2}\delta_4$	$\delta_5\frac{1}{2}$	$\delta_7\delta_8$	$\frac{1}{2}\delta_{10}$	$\delta_{11}\frac{1}{2}$	$\delta_{13}\delta_{14}$

La percentuale di perforatura può essere resa variabile nel corso della trasmissione in funzione del tasso di errore rilevato, in modo da ridurre la ridondanza nel caso di una buona qualità di ricezione, o mantenerla elevata in caso di peggioramento.

Esempio Lo standard di trasmissione DVB utilizza il *CC* con due generatori mostrato in figura, con 64 stati, vincolo di lunghezza L=7, $d_m = 10$ e $R_c = 1/2$. Le uscite di entrambi i rami possono essere perforate in accordo allo schema in tabella, che mostra anche la variazione del tasso di codifica e della distanza minima: la posizione *degli zeri* nella matrice di perforazione indica i bit di cui è omessa la trasmissione.

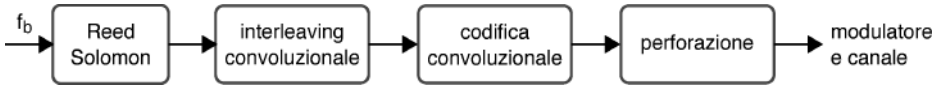


R_c	1/2	2/3	3/4	5/6	7/8
matrice di	1	10	101	10101	1000101
perforazione	1	11	110	11010	1111010
d_m	10	6	5	4	3

17.4.2.6 Concatenazione Solomon-Viterbi

La strategia illustrata al § 17.4.1.5 di collegare in cascata un codice *esterno* in grado di correggere errori a pacchetto (come il Reed-Solomon) con uno *interno* particolarmente idoneo a correggere errori isolati viene impiegata con vantaggio utilizzando come secondo un codice convoluzionale, che come abbiamo illustrato produce errori a pacchetto qualora si eccedano le sue capacità correttive.

E' precisamente la soluzione adottata per la codifica del segnale DVB[59], in cui il *CC* dell'esempio precedente è affiancato (tramite un interleaver convoluzionale di profondità 12) da un codice *RS* accorciato $(214, 188)$ derivato dal $(255, 239)$ (pag. 576) che, con la sua capacità di poter correggere 8 simboli per codeword, porta il tasso di errore *residuo* ad una P_e di 10^{-11} pur in presenza di un E_b/N_0 di 3.2 dB.



17.4.2.7 Viterbi con uscite soffici

Quando la decodifica di canale viene fattorizzata su due blocchi in cascata diviene globalmente vantaggioso che il primo dei due possa alimentare il secondo con valori *continui*, in modo da esprimere *l'affidabilità* della decisione presa.

[59] Vedi le specifiche ufficiali presso
`https://www.etsi.org/deliver/etsi_en/300400_300499/300421/01.01.02_60/en_300421v010102p.pdf`

Al § 17.4.2.3 abbiamo notato come l'algoritmo di Viterbi alimentato con ingressi soffici di n campioni $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ pervenga ad una decisione di *massima verosimiglianza* a riguardo di una codeword $\hat{\boldsymbol{x}}$ ovvero tale che $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\mathbf{y}/\mathbf{x})$, da cui risalire al messaggio $\hat{\boldsymbol{m}} = (m_1, m_2, \cdots, m_k)$ associato alla codeword $\hat{\boldsymbol{x}}$. Analizziamo ora una sua *variante* capace di produrre anche una *probabilità a posteriori* $p(m_j/\boldsymbol{y})$ per ognuno dei bit m_j $j = 1, 2 \cdots, k$ decodificati, o quantomeno una misura di affidabilità della decisione. Tale variante è denominata SOVA[60] (*soft output Viterbi algorithm*), ed in associazione ad un ingresso soffice, realizza una decodifica detta SISO (*soft input, soft output*).

L'uscita soffice si ottiene associando ad ogni $m_j = \pm 1$ decodificato anche un valore p_j che misura la probabilità che la decisione sia *errata*, in modo che l'affidabilità possa essere espressa come

$$L_j = \ln \frac{1 - p_j}{p_j} \geq 0 \tag{17.30}$$

e l'uscita soffice come $\Lambda_j = m_j L_j$ in cui il segno è la decisione *hard*, e la verosimiglianza logaritmica (17.30) ne esprime il modulo. Vediamo quindi come determinare p_j.

Valutazione della prob. di decisione errata Adottiamo per semplicità un *CC* sui cui nodi del traliccio convergono due sole transizioni, e prendiamo in considerazione un istante i in cui per ognuno dei 2^ν stati[61] s_i occorre scegliere tra *due* ipotesi di percorso parziale, che indicizziamo con $h = 1, 2$. Indicando con δ un istante del passato per il quale tutti i 2^ν percorsi superstiti condividono con elevata probabilità un medesimo *progenitore* (vedi fig. 17.8) la decisione all'istante i avviene confrontando i costi parziali

$$C_h = \sum_{j=i-\delta}^{i} d_E\left(\mathbf{y}_j, \mathbf{x}_j^{(h)}\right) \quad h = 1, 2 \tag{17.31}$$

in cui $\mathbf{y}_j$ è la parola (soft) di n bit ricevuta all'istante j ed $\mathbf{x}_j^{(h)}$ è la parola (con valori ± 1) emessa allo stesso istante dal codice in corrispondenza del $h-$esimo percorso. Ricordando quanto discusso al § 17.4.2.3, la (17.31) è legata in modo diretto alla verosimiglianza logaritmica $-\ln\left(p\left(\mathbf{y}/\mathbf{x}^{(h)}\right)\right)$ delle ipotesi $\boldsymbol{x}^{(h)}$ avendo osservato $\boldsymbol{y}$; pertanto è lecito

[60] Illustrato per esteso in J.Hagenauer, P.Hoeher, *A Viterbi algorithm with soft-decision outputs and its applications*, 1989 IEEE, a cui si ispira questa parte, e reperibile ad es. presso
`http://shannon.ece.ufl.edu/eel6550/lit/SOVA.pdf`

[61] Si è volutamente mantenuta la notazione introdotta al § 17.4.2.

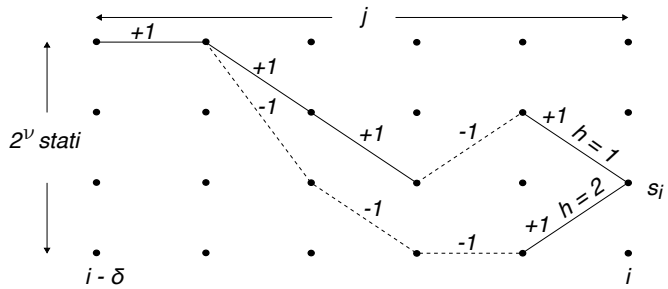


Figura 17.8: Traliccio con cui illustrare il funzionamento di SOVA. Due percorsi convergono su uno stesso stato s_i, e le sequenze informative associate differiscono in due istanti

assumere che

$$Prob\{percorso\ h\} \approx e^{-C_h} \quad h = 1, 2$$

indicando con $\approx$ una qualunque relazione monotona crescente. Assumendo ora $h = 1$ l'indice del miglior percorso e $h = 2$ di quello scartato (e dunque $C_1 < C_2$), la probabilità di aver scelto il percorso *sbagliato* è pari a

$$p_{s_i} = \frac{e^{-C_2}}{e^{-C_1} + e^{-C_2}} = \frac{1}{1 + e^{C_2 - C_1}} = \frac{1}{1 + e^{\Delta}} \tag{17.32}$$

con $\Delta = C_2 - C_1 \geq 0$. Dunque p_{s_i} è pari a 0.5 quando $C_2 = C_1$ mentre tende a zero qualora $C_2 \gg C_1$; in altre parole quanto più i due costi sono simili, tanto più la decisione potrebbe essere errata.

La misura di affidabilità (17.32) ci avverte che, qualora la decisione sia errata, allora (con probabilità p_{s_i}) sono errate anche le decisioni nei q istanti in cui i bit di informazione m nei due percorsi differiscono, ossia

$$m_j^{(1)} \neq m_j^{(2)} \quad j = j_1, \cdots, j_q \tag{17.33}$$

evidenziati in fig. 17.8 da un differente tratteggio, mentre gli istanti j in cui $m_j^{(1)} = m_j^{(2)}$ non sono danneggiati dalla decisione errata. Se abbiamo preventivamente salvato le prob. di errore $\hat{p}_j$ per gli indici espressi dalla (17.33), queste vengono aggiornate come

$$\hat{p}_j = \hat{p}_j \left(1 - p_{s_i}\right) + \left(1 - \hat{p}_j\right) p_{s_i} \quad j = j_1, \cdots, j_q \tag{17.34}$$

ovvero rimangono uguali con prob. $1 - p_{s_i}$, o si aggiornano al nuovo valore p_{s_i} con prob. $1 - \hat{p}_j$ di non aver sbagliato all'istante j.

Differenze rispetto a Viterbi classico In sostanza l'algoritmo ricalca quello in § 17.4.2.1 e produce le medesime decisioni *hard*, ma ognuna corredata dalla verosimiglianza logaritmica $L_j = \ln \frac{1-\hat{p}_j}{\hat{p}_j}$ calcolata a partire dai valori ottenuti tramite la (17.34). Per arrivare a questo risultato occorre memorizzare per ogni istante e per ogni stato, oltre al puntatore al miglior predecessore, anche le differenze Δ tra i costi, da cui ottenere i valori $\hat{p}_j$ (eqq. (17.32) - (17.34)) e L_j (17.30) in fase di decodifica.

Applicazioni Il metodo esposto trova impiego nella demodulazione di segnali con memoria come la modulazione *a traliccio* e CPM (§ 16.10), nella equalizzazione di canali con memoria (§ 18.4.5), e come stadio di decodifica interno di una codifica concatenata (§§ 17.4.1.5 e 17.4.2.6). Quest'ultimo caso si presta particolarmente bene all'utilizzo di una tecnica convoluzionale anche per il codice esterno, operante ovviamente in modalità soft, e separato da quello interno da un adeguato stadio di interleaving. Lo stadio esterno potrà quindi eseguire eseguire una decodifica di massima verosimiglianza a partire dalla sequenza $\Lambda_j = \hat{m}_j L_j$ ($\hat{m}_j = \pm 1$, $L_j \geq 0$) proveniente dallo stadio SOVA interno, decidendo per il messaggio $\boldsymbol{m}^{dec}$ tale che

$$\boldsymbol{m}^{dec} = \arg\max_{\boldsymbol{m}} \left\{\sum_j m_j \Lambda_j\right\} \tag{17.35}$$

Mentre il caso di un codice esterno a blocchi non permette di risolvere agevolmente la (17.35), consideriamo il caso di un semplice controllo di parità a bit singolo (§ 15.6.3.1), e poniamoci nella condizione che lo stesso rilevi la presenza di un bit errato. Anziché invocare una ritrasmissione, può procedere modificando la decisione per il bit a cui corrisponde la massima probabilità di errore, ovvero a cui corrisponde il valore L_j più piccolo.

17.5 Verso il limite di Shannon

I codici a blocco e convoluzionali (e loro combinazioni) esaminati nella precedente sezione sono utilizzati in innumerevoli sistemi[62] anche grazie ai progressi avvenuti nel frattempo dal punto di vista delle capacità di calcolo e memorizzazione, arrivando a conseguire prestazioni che si discostano di circa 3 dB da quelle *limite* previste dalla teoria di Shannon (§ 17.3.2). Sembrava che non si riuscisse a fare di meglio, quando nei primi anni '90 sono stati definiti i TURBO-CODICI[63] (*Berrou, Glavieux*), e poco dopo rivalutati i codici LDPC[64], inizialmente proposti nel '62 da *Gallager*. Mentre il primo metodo prende le mosse da un nuovo modo di applicare la codifica convoluzionale, il secondo è un codice lineare a blocchi *non sistematico*, con n molto grande (decine di migliaia). In entrambi i casi è determinante l'adozione di una decodifica *soffice*, a cui si affianca la novità dell'adozione di un algoritmo *iterativo*, in modo da arrivare *per gradi* alla decodifica del messaggio. Il risultato è che (sempre grazie all'evoluzione della tecnologia) si è riusciti ad approssimare ancor più da vicino il limite dei -1.6 dB per E_b/N_0 (§ 17.3.4) rispetto al quale mancano solo da 0.7 a 0.5 dB per le due tecniche, determinando la loro adozione da parte dei sistemi più recenti[65]. Vediamo dunque di che si tratta.

17.5.1 Codifica turbo

L'aggettivo *turbo* deve la sua origine al funzionamento iterativo dell'algoritmo di decodifica, che fa uso dei risultati parziali ottenuti al passo precedente[66].

Codifica La versione di turbo codice più studiata è il *codice convoluzionale concatenato parallelo* (PCCC) che consiste nel codificatore *sistematico* mostrato in figura, in cui

[62] In particolare sono stati adottati nell'ambito della telefonia GSM, GPRS, EDGE, e 3G e 3GPP, della diffusione televisiva DVB-S e DVB-T, del WIFI (802.11a-g) e delle missioni spaziali, per non parlare dei supporti di memorizzazione come CD audio, DVD, unità RAID. Per una narrazione di questa evoluzione, oltre che degli argomenti che stiamo trattando, si veda `http://www.crit.rai.it/eletel/LeMiniSerie/MS1.pdf`

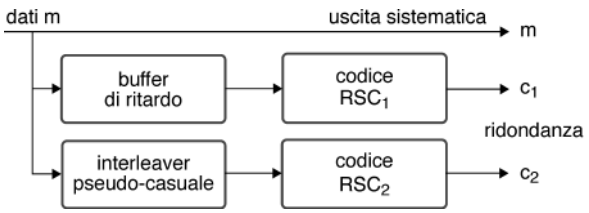
[63] Vedi ad es. `https://en.wikipedia.org/wiki/Turbo_code`

[64] Vedi ad es. `https://en.wikipedia.org/wiki/Low-density_parity-check_code`

[65] Come ad esempio DVB-2, telefonia UMTS ed LTE, 10GBASE-T Ethernet, WIFI 802.11*n* e *ab*, WiMAX 802.16, nonché le missioni spaziali più recenti.

[66] Evidenziamo tra breve la presenza di un vero e proprio percorso di *retroazione*, ma l'aggettivo *turbo* è nato in analogia a quanto avviene in campo automobilistico con i *motori turbo*, una novità tecnica introdotta negli stessi anni in cui è stato definito questo metodo di decodifica.

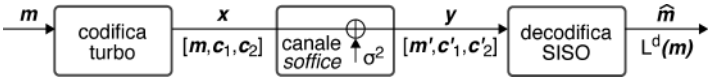
due *CC* ricorrenti (RSC[67]) uguali[68] elaborano il medesimo bitstream $\boldsymbol{m}$ di ingresso producendo i bit di protezione $\boldsymbol{c}_1$ e $\boldsymbol{c}_2$, tranne che il secondo encoder riceve una copia *rimescolata* da parte dell'interleaver, la cui definizione[69] è fondamentale per la riuscita di un turbo codice: grazie ad esso infatti gli ingressi ai due rami divengono istante per istante *incorrelati*, così come le rispettive uscite.



Codificatore turbo parallelo

La fase di codifica suddivide il bitstream di ingresso $\boldsymbol{m}$ in segmenti di dimensione k uguale a quella dell'interleaver, ed applica la tecnica del *tail biting* (§ 17.4.2.2) in modo da ottenere per ogni ramo una sequenza $\boldsymbol{c} = (c_1, \cdots, c_k)$ di k bit *di protezione* del segmento. Alla fine viene quindi trasmessa una parola $\boldsymbol{x} = (\boldsymbol{m}, \boldsymbol{c}_1, \boldsymbol{c}_2)$ di $n = 3k$ bit ottenendo un tasso complessivo $R_c = 1/3$, eventualmente *aumentato* mediante una successiva operazione di perforazione.

Trasmissione e decodifica soft La figura sotto ha il duplice scopo di riepilogare la notazione adottata e di evidenziare la presenza di un canale *soffice* che, pur includendo codifica di linea, modem e canale AWGN (§ 17.3), non prevede la presenza di un decisore, e dunque fornisce al decodificatore la sequenza $\boldsymbol{y}$ di valori analogici (*soft*) necessaria per la decodifica turbo, con elementi $y_j = x_j + \varepsilon$ in cui ε è una v.a. gaussiana a media nulla, varianza σ^2 e valori indipendenti. Il processo di decodifica è anch'esso di tipo soft e per questo detto SISO, ed il suo esito viene espresso come un valore di *verosimiglianza logaritmica a posteriori* L^p per ognuno dei k bit m_i del messaggio informativo, in modo da poter applicare un criterio di massima prob. a posteriori o MAP (§ 17.1.2) ossia decidere che $\hat{m}_j = 1$ o 0 a seconda se $L^p(m_j) \gtrless 0$.



Verosimiglianza logaritmica ed informazione estrinseca Finché non si attua la decodifica, l'osservazione del valore y_j in uscita dal canale permette il calcolo della verosimiglianza logaritmica a posteriori (nel seguito LLR ossia *log likelihood ratio*) per

[67] L'acronimo sta per *Recursive Sistematic Convolutional*, ed al § 17.4.2.4 ne è raffigurato un possibile schema architetturale. Il motivo di questa scelta è triplice: da un lato un RSC è simile ad uno *scrambler* pseudo random, e la teoria di Shannon basa la sua dimostrazione (vedi nota 14 a pag. 564) su codeword casuali; inoltre, un RSC ha prestazioni migliori dei *CC* classici per bassi valori di E_b/N_0. Infine, solo poche sequenze di lunghezza finita in ingresso ne producono di lunghezza finita in uscita, indice di una *elevata ridondanza*.

[68] In realtà possono anche essere diversi e con un diverso tasso R_c, ma non si desidera appesantire la trattazione.

[69] Ad esempio, l'interleaver può essere implementato mediante una sequenza di numeri pseudo casuali da utilizzare ciclicamente come indice di scrittura in un array dove si memorizzano gli elementi della sequenza di ingresso, e la cui lettura avviene poi in modo sequenziale.

ciascun bit x_j in ingresso al canale soffice come

$$L\left(x_j\right) = \ln \frac{Pr\left(x_j=1/y_j\right)}{Pr\left(x_j=0/y_j\right)} = \ln \frac{Pr\left(y_j/x_j=1\right)Pr\left(x_j=1\right)}{Pr\left(y_j/x_j=0\right)Pr\left(x_j=0\right)} = \ln \frac{Pr\left(y_j/x_j=1\right)}{Pr\left(y_j/x_j=0\right)} + \ln \frac{Pr\left(x_j=1\right)}{Pr\left(x_j=0\right)}$$
$$= L^c\left(y_j\right) + L^a\left(x_j\right) \tag{17.36}$$

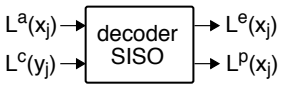
in cui alla seconda eguaglianza si applica il teorema di Bayes, ed il risultato si interpreta osservando che $L\left(x_j\right)$ è somma di due termini: il primo $L^c\left(y_j\right)$ è dovuto al canale e dipende solo dal valore y_j ricevuto, mentre il secondo $L^a\left(x_j\right)$ è legato alla prob. *a priori* di x_j ed è nullo se 0 ed 1 sono equiprobabili.

A differenza della (17.36), il valore $L^p\left(x_j\right) = \ln \frac{Pr\left(x_j=1/\boldsymbol{y}\right)}{Pr\left(x_j=0/\boldsymbol{y}\right)}$ della LLR a posteriori in uscita dal decoder è ottenuto a partire da *tutti* i bit ricevuti $\boldsymbol{y}$, compresi quelli di ridondanza $\boldsymbol{c}_1$ e $\boldsymbol{c}_2$, e per questo la decisione MAP presa in base ai valori di $L^p\left(x_j\right)$ contiene *meno errori*. La variazione della LLR $L^p\left(x_j\right)$ rispetto alla (17.36) vale

$$L^e\left(x_j\right) = L^p\left(x_j\right) - L\left(x_j\right) = L^p\left(x_j\right) - L^c\left(y_j\right) - L^a\left(x_j\right) \tag{17.37}$$

e prende il nome di *informazione estrinseca*[70] in quanto aggiunta da parte del decoder; si può dimostrare che questa *non dipende* dai valori di $\boldsymbol{m}$, ma solo *da quelli della ridondanza* $\boldsymbol{c}_1$ e $\boldsymbol{c}_2$.

Soft input soft output a quattro porte L'algoritmo SISO adottato nella decodifica turbo può essere schematizzato come nella figura a fianco, che lo mostra accettare in ingresso le componenti della LLR (17.36) ovvero i contributi delle singole osservazioni $L^c\left(y_j\right)$ e quelli a priori $L^a\left(x_j\right)$, e ottenere in uscita sia il valore di LLR a posteriori $L^p\left(x_j\right)$ (ottenuta applicando il metodo di decodifica) sia quello della informazione estrinseca ottenuta applicando la (17.37).



Valutazione della LLR di ingresso e di uscita al SISO Consideriamo di effettuare una trasmissione binaria antipodale, in cui i valori del segnale agli istanti di simbolo sono espressi come $x_j = \left(2m_j - 1\right)$ ossia assumono i valori ± 1 quando $m_i = 1$ oppure 0, e lo stesso dicasi per i bit di protezione c_{1j} e c_{2j}: in tal caso il termine $L^c\left(y_j\right) = \ln \frac{Pr\left(y_j/x_j=1\right)}{Pr\left(y_j/x_j=0\right)}$ (eq. (17.36)) assume il valore[71]

$$L^c\left(y_j\right) = \frac{2}{\sigma^2} y_j \tag{17.38}$$

Per quanto riguarda invece $L^p\left(x_i\right) = \ln \frac{Pr\left(x_j=1/\boldsymbol{y}\right)}{Pr\left(x_j=0/\boldsymbol{y}\right)}$ diciamo che nel contesto dei codici RSC usati nel caso in esame il relativo decodificatore opera su di un traliccio analogo a quello visto nella decodifica di Viterbi, ed il valore $L^p\left(x_j\right)$ può essere *approssimato*

[70] Questo è il nome attribuito a tale quantità dalla comunità che ha lavorato alla definizione dei turbocodici. In effetti, essendo la LLR un logaritmo di probabilità può a tutto diritto essere chiamata *informazione*, ma espressa in *nat* anziché in bit, avendo adottato un logaritmo in base e.

[71] Infatti ora y_i è una v.a. gaussiana con media $x_i = \pm 1$ e varianza σ^2, e dunque

$$\ln \frac{Pr(y/x=1)}{Pr(y/x=0)} = \ln \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)} = -\frac{(y-1)^2}{2\sigma^2} + \frac{(y+1)^2}{2\sigma^2} = \frac{-y^2+2y-1+y^2+2y+1}{2\sigma^2} = \frac{4y}{2\sigma^2} = \frac{2}{\sigma^2} y$$

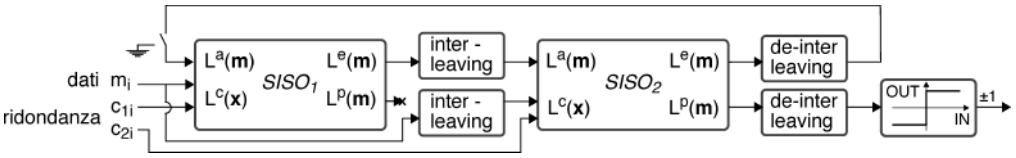


Figura 17.9: Schema di decodifica turbo per un codificatore RSC parallelo

ricorrendo ad una decodifica SOVA (§ 17.4.2.7)[72], ma per ottenere il suo valore esatto (e dunque le migliori prestazioni) occorre ricorrere ad una modifica dell'algoritmo BCJR[73], le cui problematiche numeriche spingono però ad utilizzare metodi sub-ottimi e che operano direttamente nel dominio logaritmico[74].

Decodifica Siamo finalmente in grado di illustrare l'operatività della tecnica, con l'aiuto di fig. 17.9, in cui (come nel seguito) si adotta la notazione $L(\boldsymbol{m})$ per indicare tutti i valori $L(m_j)$ per $j = 1, 2, \cdots, k$:

1. una prima decodifica SISO_1 considera nulla la verosimiglianza *a priori* ovvero $L^a(\boldsymbol{m}) = 0$ (switch *a massa*), e in base a L_c in uscita dal canale, relativa ai bit di $\boldsymbol{m}$ e di $\boldsymbol{c}_1$ (prodotti dal primo RSC di codifica) ottiene la verosimiglianza *a posteriori* $L_1^p(\boldsymbol{m})$ e da questa l'informazione estrinseca $L_1^e(\boldsymbol{m}) = L_1^p - L^c$ che *dipende solo* dai valori di $\boldsymbol{c}_1$;
2. il blocco SISO_2 adotta $L_1^e(\boldsymbol{m})$ come valore della LLR *a priori* $L_2^a(\boldsymbol{m})$ per i bit di messaggio $\boldsymbol{m}$, dopo che l'interleaver ne ha posto i valori nello stesso ordine con cui si sono presentati a RSC_2. Dato che L_1^e dipende dai valori di $\boldsymbol{c}_1$ a cui SISO_2 non ha accesso, rappresenta effettivamente qualcosa in *più*. SISO_2 esegue l'algoritmo di decodifica ottenendo $L_2^p(\boldsymbol{m})$ a partire da $L^c(\boldsymbol{m})$, $L^c(\boldsymbol{c}_2)$ e $L_2^a(\boldsymbol{m})$, e valuta $L_2^e = L_2^p - L^c - L_2^a$, che dipende solamente da $\boldsymbol{c}_2$;
3. dopo essere stata di nuovo riordinata temporalmente, l'informazione $L_2^e(\boldsymbol{m})$ viene fornita al blocco SISO_1 sull'ingresso *a priori* $L_1^a(\boldsymbol{m})$, in quanto anch'essa rappresenta qualcosa che SISO_1 non può calcolare per suo conto. Ecco così attuato il principio di *controreazione!* Ciò consente di ottenere dei nuovi valori per L_1^p e $L_1^e = L_1^p - L^c - L_1^a$;
4. se i valori di L_1^p e L_2^p sono abbastanza simili per tutti gli indici $j = 1, 2, \cdots, k$ i due rami sono collaborativamente addivenuti alla stessa conclusione, e la decodifica

[72]Per poter utilizzare anche le prob. a priori in SOVA occorre che nella (17.29) venga sommato anche un termine $\ln(p(x_j))$.

[73]L. BAHL; J. COCKE; F. JELINEK; J. RAVIV, *Optimal decoding of linear codes for minimizing symbol error rate*, in IEEE Trans. on Inf. Theory, March 1974, per come modificato in C BERROU, A GLAVIEUX, *Near optimum error correcting coding and decoding: Turbo-codes*, IEEE Trans. on Comm., Oct. 1996.

In parole povere, il traliccio è esaminato oltre che da sinistra a destra, anche da destra a sinistra, permettendo il calcolo per ogni istante i della probabilità *congiunta* di trovarsi in uno stato, e che sia stato trasmesso un valore x_i, da cui saturando sugli stati ottenere i valori $Pr(x_i = 1/\boldsymbol{y})$ e $Pr(x_i = 0)/\boldsymbol{y})$. Tale procedura fu poi adottata nel contesto della stima di parametro dei *modelli di Markov nascosti* (HMM) utilizzati per il riconoscimento del parlato, ma quella *è un'altra storia.*

[74]Vedi ad es. P. ROBERTSON; E. VILLEBRUN; P. HOEHER, *A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain*, Proc. IEEE ICC '95

finale per tutti gli $\hat{m}_j$ si ottiene dall'uscita L^p di uno dei due SISO valutando se $L^p \gtrless 0$; altrimenti, si torna al passo 2. La tecnica descritta si è mostrata capace di convergere nel giro di una decina di iterazioni.

Utilizzi I turbo codici sono utilizzati, oltre che nei sistemi UMTS ed LTE, dagli standard DVB-RCS, WiMax, e da missioni spaziali. Il principio della codifica turbo si applica non solo al caso accennato degli RSC paralleli, ma può essere adottato anche per schemi seriali, e per *codici prodotto*. Il blocco di codifica interno può altresì essere sostituito da un modulatore-demodulatore *con memoria*, come ad es. il TCM (pag. 541). In base alla stessa logica anche l'equalizzazione MLSD di un canale con memoria (§ 18.4.5) può beneficiare di uno schema turbo, in cui equalizzatore e decodificatore si scambiano iterativamente informazione estrinseca per addivenire ad una decisione condivisa[75].

17.5.2 Codifica a bassa densità di controllo parità

Questo approccio si basa su di un codice lineare *a blocchi* caratterizzato da una una *matrice di controllo* $\boldsymbol{H}$ con una *bassa densità di uni*, da cui l'acronimo LDPC (*low-density parity-check*).

Riprendendo i concetti espressi al § 17.4.1, la moltiplicazione con somma modulo due $\oplus$ (pag. 571) $\boldsymbol{x} = \boldsymbol{m} \cdot \boldsymbol{G}$ tra il vettore *riga* $\boldsymbol{m}$ dei k bit di messaggio e la matrice binaria $\boldsymbol{G}$ (detta *generatrice*) con k righe ed n colonne produce una codeword $\boldsymbol{x}$ di n elementi. Se $\boldsymbol{G}$ è posta nella forma $\boldsymbol{G} = [\boldsymbol{I}_k \,|\boldsymbol{P}]$ con $\boldsymbol{I}_k$ matrice identità con k righe e colonne e $\boldsymbol{P}$ matrice *di parità* di k righe per $n-k$ colonne, il codice è detto *sistematico* e le codeword possono esser scritte come $\boldsymbol{x} = \left[\begin{array}{ccc|ccc} m_1 & \cdots & m_k & c_1 & \cdots & c_{n-k} \end{array}\right]$ in cui[76] $c_j = \sum_{i=1}^{k} m_i p_{ij}$ valuta la parità sui bit m_i per i quali $p_{ij} = 1$.

Un codice LDPC *non è* definito a partire da $\boldsymbol{G}$ bensì dalla matrice $\boldsymbol{H}$ *di controllo* di dimensione $(n-k) \times n$ e che nel caso sistematico ha la forma[77] $\boldsymbol{H} = \left[\boldsymbol{P}^T \,|\boldsymbol{I}_{n-k}\right]$, ed in generale soddisfa la condizione (valida anche per un codice *non* sistematico)

$$\boldsymbol{G} \cdot \boldsymbol{H}^T = \boldsymbol{0}_{k \times (n-k)}$$

in quanto ciascuna riga di $\boldsymbol{H}$ *è ortogonale*[78] ad ogni riga di $\boldsymbol{G}$; pertanto risulta[79] $\boldsymbol{H} \cdot \boldsymbol{x}^T = \boldsymbol{0}_{n-k}^T$ se e solo se $\boldsymbol{x}$ è una codeword; mentre in presenza di errori il vettore ricevuto è $\boldsymbol{y} \neq \boldsymbol{x}$, e se il codice è sistematico il prodotto $\boldsymbol{H} \cdot \boldsymbol{y}^T \neq \boldsymbol{0}$ è detto *sindrome* e viene usato per individuare i bit errati.

Precisiamo ora che le codeword $\boldsymbol{x}$ di un codice LDPC non fanno distinzione tra bit di messaggio m e di parità c, e sebbene il codice possa essere di tipo sistematico, è di gran lunga preferibile che non lo sia, per i motivi presto illustrati; questo fa si che la decodifica basata sulla sindrome non sia applicabile. Un modo per descrivere il

[75] Vedi ad es. `https://en.wikipedia.org/wiki/Turbo_equalizer`

[76] Si adotta il simbolo $\sum_{\oplus}$ per intendere una somma *modulo due*.

[77] Ci si discosta dalla notazione adottata a pag. 572 in quanto la $\boldsymbol{H}$ definita qui è *la trasposta* di quella definita in tale sede.

[78] Vedi ad es. S.Lin, D.J.Costello, *Error control coding*, Prentice-Hall 1983

[79] Infatti $\boldsymbol{H} \cdot \boldsymbol{x}^T = \boldsymbol{H} \cdot (\boldsymbol{m} \cdot \boldsymbol{G})^T = \boldsymbol{H} \cdot \boldsymbol{G}^T \cdot \boldsymbol{m}^T = (\boldsymbol{G} \cdot \boldsymbol{H}^T)^T \cdot \boldsymbol{m}^T = \boldsymbol{0}_{(n-k)\times k} \cdot \boldsymbol{m}^T = \boldsymbol{0}_{n-k}^T$

funzionamento di un LDPC è pensare che ogni riga i di $\boldsymbol{H}$ rappresenti il vincolo imposto sulle codeword da una tra $n-k$ *equazioni* di parità del tipo

$$\sum_{\oplus}^{n}{}_{j=1} x_j h_{ij} = 0$$

equivalente riga per riga dell'espressione $\boldsymbol{H} \cdot \boldsymbol{x}^T = \boldsymbol{0}$.

Esempio La matrice di controllo $\boldsymbol{H}$ riportata sotto corrisponde alle quattro equazioni di vincolo scritte a fianco, che devono essere soddisfatte dai bit x_j delle codeword esenti da errore. Il codice risultante è descritto dai parametri $n, k = 8, 4$ e da un tasso $R_c = 1/2$.

$$\boldsymbol{H} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \qquad \begin{cases} x_2 \oplus x_4 \oplus x_5 \oplus x_8 = 0 \\ x_1 \oplus x_2 \oplus x_3 \oplus x_6 = 0 \\ x_3 \oplus x_6 \oplus x_7 \oplus x_8 = 0 \\ x_1 \oplus x_4 \oplus x_5 \oplus x_7 = 0 \end{cases}$$

A parte *il piccolo dettaglio* di come poter effettuare la codifica[80] $\boldsymbol{x} = \boldsymbol{m} \cdot \boldsymbol{G}$, per la matrice dell'esempio possiamo osservare che ogni bit x_i compare in due equazioni, ed ogni equazione si applica a quattro bit. In generale un codice LDPC si dice *regolare* se presenta esattamente $w_c \ll n-k$ elementi pari ad uno per ogni colonna, e $w_r = w_c \frac{n}{n-k}$ uni per ogni riga, a cui corrisponde un tasso $R_c = k/n = 1 - w_c/w_r$.

Grafo di Tanner E' il nome dato al grafo di cui $\boldsymbol{H}$ è la matrice *di adiacenza*, e che risulta essere tipo *bipartito* ovvero i cui vertici si dividono in due insiemi, tra gli elementi dei quali non sono presenti archi. Si traccia (fig. 17.10) riportando *sotto* i nodi (detti *variabile*) associati agli n bit ricevuti[81] x_j, e *sopra* quelli (*di controllo*) c_i che verificano il rispetto delle equazioni di vincolo; tra questi nodi si traccia un arco tra x_j e c_i se è presente un *uno* tra la riga i e la colonna j di $\boldsymbol{H}$, ovvero se $h_{ij} = 1$.

In altre parole, i w_c *uni* nelle n colonne rappresentano le connessioni dei nodi x_j (ed infatti ne troviamo due) mentre i w_r uni sulle $n-k$ righe indicano le connessioni dei nodi c_i (e ne troviamo quattro per ciascuno).

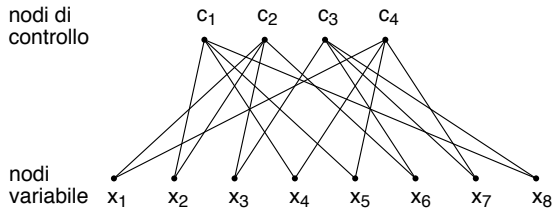


Figura 17.10: Grafo di Tanner per la matrice dell'esempio precedente

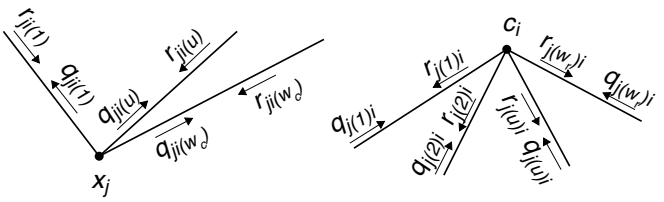
17.5.2.1 Decodifica iterativa

La particolarità più rilevante di un LDPC è quella di svolgere la decodifica in modo *iterativo* basandosi su di un ripetuto scambio di messaggi di natura probabilistica tra

[80] In linea di principio per trovare una matrice generatrice $\boldsymbol{G}_{k \times n}$ tale che $\boldsymbol{G} \cdot \boldsymbol{H}^T = \boldsymbol{0}$ si può procedere trasformando prima $\boldsymbol{H}$ nella forma *canonica* di un codice sistematico, modificandone le righe applicando il metodo di Gauss; ciò determina però una $\boldsymbol{G}$ per nulla sparsa, ed una eccessiva complessità di codifica per n elevato. Fortunatamente hanno escogitato metodi più efficienti, anche ricorrendo a codici LDPC *non regolari*; per un approfondimento si può vedere W.E.RAYAN, *An introduction to LDPC code*, Univ. of Arizona 2003, ed es. presso `http://tuk88.free.fr/LDPC/ldpcchap.pdf`.

[81] La nomenclatura adottata in letteratura indica i nodi-variabile come *v-nodes* e li rappresenta con la lettera c, mentre quelli di controllo (*check-nodes* o *nodi-fattore*) sono rappresentati dalla lettera f. Preferisco qui attenermi alla notazione dell'attuale contesto espositivo.

nodi-variabile e nodi di controllo, realizzando una applicazione di *propagazione della credenza*[82] nota anche come *algoritmo somma-prodotto*. Lo scopo dell'algoritmo è individuare la codeword $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}/\boldsymbol{y})$ che rende massima la prob. a posteriori (PAP) una volta noto il vettore $\boldsymbol{y}$ in uscita dal canale. La ricerca è condotta attraverso un *raffinamento successivo* di ipotesi, a partire dalla conoscenza dei valori y_j indipendenti che fornisce una $p(\boldsymbol{x}/\boldsymbol{y}) = \prod_{j=1}^{n} p(x_j/y_j)$ *di partenza*.



Ad ogni iterazione ciascun nodo-variabile j invia a tutti i w_c nodi di controllo $i(u)$, $u = 1, \cdots, w_c$ a cui è connesso un messaggio q_{ji} (pensiamo stia per *query*) in cui comunica la *sua percezione* della probabilità p_j di essere *pari ad uno*. Ricevuti i messaggi q_{ji}, ogni nodo di controllo c_i invia a ciascun nodo variabile $j(u)$, $u = 1, \cdots, w_r$ a cui è connesso un messaggio r_{ji} in cui *risponde* con nuove stime di p_j ottenute combinando le opinioni q_{ji} ricevute da tutti i nodi *tranne quello a cui è diretta la risposta*. A questo punto i nodi-variabile generano nuovi messaggi q_{ji} integrando la propria opinione di partenza con quella r_{ji} ricevuta dai nodi di controllo, omettendo di includere l'informazione ricevuta da quello verso cui è diretto il messaggio. Il senso di omettere l'informazione proveniente dal destinatario è quello di attingere unicamente all'informazione *estrinseca*, ossia non ricavabile autonomamente a destinazione, come avviene per i codici turbo.

L'opinione *di partenza* sulla PAP p_j che il bit x_j sia pari ad 1 è ottenuta (per ogni istante $j = 1, \cdots, n$) a partire dal valore y_j in uscita dal canale come[83]

$$p_j = p(x = 1/y) = \frac{p(y/x = 1)\, p(x = 1)}{p(y)} = K \cdot p(y/x = 1)$$

in cui $K = \frac{p(x=1)}{p(y)} = \frac{1}{p(y/x=1)+p(y/x=0)}$[84]. A seconda se in presenza di un BSC (§ 17.1.1) oppure di un canale *soffice* (o AWGN, pag. 591)

- BSC: il canale compie una decisione *hard* ed emette un valore y pari a zero od uno, con prob. condizionata *in avanti* $p(y/x)$ di valore $p(1/1) = p(0/0) = 1 - p_e$, $p(0/1) = p(1/0) = p_e$. Dunque $K = 1$[85] sia per $y = 1$ che per $y = 0$, e $p(x = 1/y) = p_e$ se $y = 0$ oppure $1 - p_e$ quando $y = 1$;
- AWGN: in funzione del valore binario di $x \in \{0, 1\}$ il canale emette il valore continuo $y = 2x - 1 + \varepsilon$ che è una v.a. gaussiana a valori indipendenti, media ± 1 a seconda se $x = 1$ o 0, e varianza σ^2; si ha quindi $p(y/x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y\pm 1)^2}{2\sigma^2}\right\}$.

[82] Dall'inglese *belief propagation*, vedi ad es. `https://en.wikipedia.org/wiki/Belief_propagation`

[83] Applicando il solito teorema di Bayes, ed omettendo il pedice j per estetica e spazio.

[84] Questo perché $p(x = 1)$ è la prob. *a priori* considerata pari a 1/2, e quindi $\frac{p(x=1)}{p(y)} = \frac{p(x=0)}{p(y)}$. Imponendo ora $p(x = 1/y) + p(x = 0/y) = 1$ si ottiene $(p(y/1) + p(y/1))\, K = 1$, e dunque il risultato.

[85] In quanto $p(1/1) + p(1/0) = 1 - p_e + p_e = 1$

In entrambi i casi, ogni nodo variabile j pone il messaggio iniziale $q_{ji} = p_j$ uguale per tutti gli i.

Calcolo ai nodi di controllo Consideriamo ora un nodo c_i che riceve più di un q_{ji}, e che *sa* che tra i nodi-variabile a lui collegati ci deve essere un numero *pari* di uni. Per ottenere il valore del messaggio r_{ji} da inviare indietro, c_i *somma* le stime di probabilità ricevute.

Esempio Il nodo c_1 dell'esempio di fig. 17.10 deve far valere il vincolo $x_2 \oplus x_4 \oplus x_5 \oplus x_8 = 0$ ovvero nei quattro bit ci devono essere 4 uni, oppure due, oppure nessuno. Genera quindi il messaggio r_{21} diretto a x_2 tenendo conto delle probabilità q_{j1} ricevute da x_4, x_5 e x_6 e, considerandole *statisticamente indipendenti*, stima

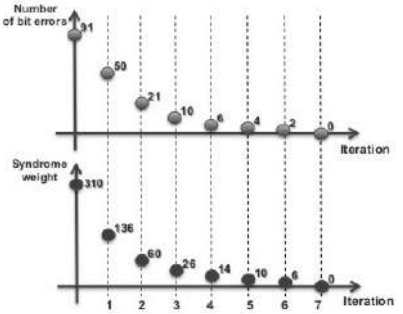
$$r_{21} = \hat{p}_2 = q_{41}\,(1-q_{51})\,(1-q_{61}) + (1-q_{41})q_{51}\,(1-q_{61}) + (1-q_{41})\,(1-q_{51})q_{61} + q_{41}q_{51}q_{61}$$

ossia pari a quella che ci sia un altro uno, oppure tre. Calcola quindi in modo analogo i messaggi r_{41} , r_{51} e r_{61} omettendo ogni volta di considerare l'informazione originata dal nodo destinazione.

Calcolo ai nodi-variabile A questo punto ogni nodo-variabile j ha ricevuto w_c messaggi r_{ji}, da cui ne calcola altrettanti da rispedire indietro, considerando oltre all'informazione $p_j^{(0)}$ proveniente dal canale anche quella r_{ji} proveniente dai nodi c_i, tranne quello di destinazione. Questa volta il calcolo di q_{ji} comporta *il prodotto* dei valori di probabilità ricevuti.

Esempio Il nodo x_1 ha ricevuto r_{12} e r_{14}, e valuta $q_{12} = p_1^{(1)} = k_2 p_1^{(0)} r_{14}$ da inviare a c_2 in cui[86] $k_2 = \frac{1}{\left(1-p_1^{(0)}\right)\left(1-r_{14}\right)+p_1^{(0)} r_{14}}$ serve per *normalizzare* la stima, dato che se k_2 non fosse presente $p_2^{(1)}$ risulterebbe *più piccolo* di tutti i valori ricevuti. In modo simile, il nodo x_1 calcola poi q_{14} da inviare a c_4 omettendo di usare r_{14}.

Arresto Ad ogni ciclo $v = 1, 2, \cdots$ si perviene ad una stima di probabilità $\hat{p}_j^{(v)}$ che ogni bit x_j sia pari ad uno utilizzando *tutte* le fonti informative[87], e da questa si ottiene una *ipotesi* di codeword $\tilde{\boldsymbol{x}}$ operando per ogni bit una decisione *hard* mediante una soglia di probabilità pari a 1/2. Se $\tilde{\boldsymbol{x}}$ soddisfa la condizione $\boldsymbol{H} \cdot \tilde{\boldsymbol{x}}^T = \boldsymbol{0}$ allora è una codeword ammissibile, e la decodifica è terminata: in figura si mostra l'andamento del numero di errori sul bit e del *peso* della sindrome al



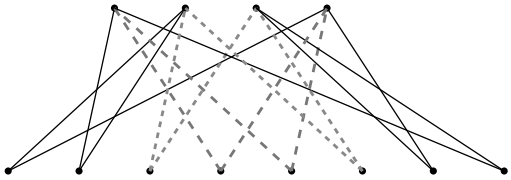
[86] Il risultato si ottiene imponendo che la stessa normalizzazione valga anche per l'evento complementare, ovvero $1 - q_{21} = k_2\left(1 - p_1^{(0)}\right)\left(1 - r_{14}\right)$, ma dall'equazione *sopra* abbiamo anche $1 - q_{21} = 1 - k_2 p_1^{(0)} r_{14}$, ed eguagliando le due espressioni si consegue lo scopo.

[87] Infatti questa stima non deve essere re-inviata a nessuno, per cui nel caso dell'esempio il nodo x_1 calcola $\tilde{p}_1^{(v)} = k_2 p_1^{(0)} r_{12} r_{14}$ con $k_2 = \frac{1}{\left(1-p_1^{(0)}\right)\left(1-r_{12}\right)\left(1-r_{14}\right)+p_1^{(v)} r_{12} r_{14}}$.

progredire delle iterazioni. Se invece anche dopo un loro ragionevole numero[88] la condizione non è mai verificata, significa che una eccessiva quantità di errori impedisce la decodifica corretta, e tale evenienza può essere segnalata agli stadi di elaborazione seguenti (particolare che non avviene per la decodifica *turbo*).

17.5.2.2 Attenti a quel ciclo

Il calcolo svolto sia dai nodi di controllo che da quelli -variabile implica che gli eventi a cui si riferiscono i messaggi ricevuti siano statisticamente indipendenti. Ma se il grafo associato alla matrice $\boldsymbol{H}$ presenta cicli di lunghezza ν, dopo l'iterazione numero $\nu/2$ l'ipotesi perde di validità, in quanto le stime di probabilità divengono dipendenti anche da quelle inviate dal nodo che le riceve.



In figura si evidenziano due cicli di lunghezza 4 presenti nel grafo di fig. 17.10, associati a quattro *uni* disposti agli angoli di una sottomatrice rettangolare di $\boldsymbol{H}$. In fase di progetto della matrice di controllo tale circostanza va evitata, e dato che non è possibile non avere cicli, è bene vincolarne il numero e la lunghezza minima ad valore ritenuto adeguato a non degradare le prestazioni.

17.5.2.3 Implementazione Min-Sum

Il metodo esposto al § 17.5.2.1 comporta difficoltà legate a dover moltiplicare molti valori di probabilità, determinando instabilità numerica per dimensioni n anche di decine di migliaia. Si preferisce allora lavorare nel dominio della verosimiglianza logaritmica (LLR), con un algoritmo del tutto simile a quello esposto, ma con alcune particolarità. Iniziamo con il definire la LLR della PAP di un bit x_j in perfetta analogia con la (17.36), ovvero

$$L\left(x_j/y_j\right) = L^c\left(y_j\right) + L\left(x_j\right)$$

La decodifica iterativa ha ora l'obiettivo di aumentare *il modulo* dalla LLR *a priori* $L\left(x_j\right) = \ln \frac{p(x_j=1)}{p(x_j=0)}$, inizialmente nullo, grazie all'apporto dell'informazione estrinseca proveniente dagli altri nodi.

Per quanto riguarda il contributo del canale $L^c\left(y_j\right) = \ln \frac{p(y_j/x_j=1)}{p(y_j/x_j=0)}$

- nel caso AWGN con mapping $y = 2x - 1 + \varepsilon$, in cui $x \in \{0, 1\}$ e $\varepsilon \in N(0, \sigma)$, si ha (vedi eq. (17.38)) $L^c\left(y_j\right) = \frac{2}{\sigma^2} y_j$;
- nel caso BSC con $P_e = p$ risulta $L^c\left(y_j\right) = \ln \frac{1-p}{p}$ se $y_j = 1$ ed $L^c\left(y_j\right) = \ln \frac{p}{1-p}$ quando $y_j = 0$.

In entrambi i casi si usa $L^c\left(y_j\right)$ per inizializzare i messaggi verso i nodi di controllo, che ora non valgono più q_{ji} ma $L\left(q_{ji}\right) = \ln \frac{Pr\{x_j=1\}}{1-Pr\{x_j=1\}}$.

[88] Tipicamente, tra dieci e trenta. Una simpatica animazione dell'evoluzione della decodifica può essere trovata presso `http://www.inference.org.uk/mackay/codes/gifs/demo2.html` sia per il caso BSC che AWGN.

Min Per il calcolo della LLR dei messaggi *di risposta* r_{ji}, il nodo c_i si auspica che la probabilità $p_j = r_{ji} = Pr\left\{x_j = 1\right\}$ sia uguale a quella che gli altri bit che partecipano al controllo svolto da c_i presentino un numero *dispari* di uni, ossia

$$L(r_{ji}) = L\left(Pr\left\{ \underset{\oplus,\, j' \neq j}{\sum_{j'=1}^{n}} x_{j'} h_{ij'} = 1 \right\}\right) = \ln \frac{Pr\left\{r_{ji} = 1\right\}}{Pr\left\{r_{ji} = 0\right\}}$$

Dopo una serie di sviluppi analitici di cui tralasciamo l'approfondimento[89], sotto l'ipotesi di indipendenza statistica si arriva ad esprimere $L(r_{ji})$ in funzione approssimata della LLR $L\left(q_{ji}\right)$ dei q_{ji} da cui dipende, come

$$L(r_{ji}) \approx (-1)\left\{\prod_{j' \neq j} \text{sgn}\left(L\left(q_{j'i}\right)\right)\right\} \cdot \min_{j' \neq j}\left\{\left|L\left(q_{j'i}\right)\right|\right\}$$

Il risultato si interpreta notando che il modulo (l'affidabilità) della LLR risultante $L(r_{ji})$ è determinato dal *più piccolo* dei moduli dei contributi $\left|L\left(q_{j'i}\right)\right|$ (il meno affidabile), da cui l'appellativo di *Min* a questo passaggio. Il segno positivo di $L(r_{ji})$ indica poi che $Pr\left\{r_{ji} = 1\right\} > Pr\left\{r_{ji} = 0\right\}$ (o viceversa se negativo), e si ottiene come prodotto dei segni di $L\left(q_{j'i}\right)$, indicando così una parità dispari o pari.

Sum I nodi-variabile aggiornano le stime di $L\left(x_j\right)$ come

$$L\left(x_j\right) = L^c\left(y_j\right) + \textstyle\sum_i L\left(r_{ji}\right)$$

da cui ottenere una ipotesi di codeword $\tilde{\boldsymbol{x}}$ con elementi 1 o 0 a seconda se $L\left(x_j\right) \gtrless 0$. Qualora $\boldsymbol{H} \cdot \tilde{\boldsymbol{x}}^T = \boldsymbol{0}$ la decodifica è terminata; altrimenti si calcolano le

$$L\left(q_{ji}\right) = L^c\left(y_j\right) + \textstyle\sum_{i' \neq i} L\left(r_{ji'}\right)$$

(passo *Sum*) e si torna al passo *Min*.

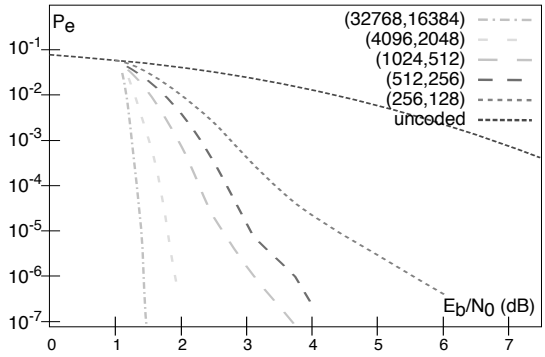
17.5.2.4 Prestazioni

La natura probabilistica del metodo di decodifica non consente di ottenere una espressione in forma chiusa della P_e in funzione di E_b/N_0, il cui grafico deve essere ottenuto mediante simulazione al computer ottenuta mediando su un gran numero di vettori ricevuti $\boldsymbol{y}$: accade infatti che, sebbene il codice si comporti generalmente bene, per alcune configurazioni di partenza l'algoritmo di decodifica non riesca a convergere.

Di seguito sono riportate le prestazioni ottenute da un codice LDPC regolare con $w_c = 3$ e $w_r = 6$, $R_c = 1/2$, per una segnalazione antipodale su canale AWGN, con diverse scelte[90] della lunghezza del blocco n da 256 a 32768.

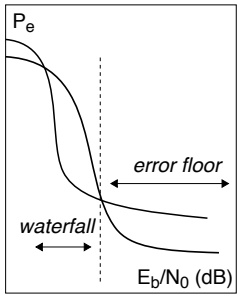
[89] Che può essere svolto incrociando le infomazioni presenti oltre che nel già citato W.E.Rayan, *An introduction to LDPC code*, anche in T.Strutz, *Low-Density Parity-Check codes - An introduction* presso `http://www1.hft-leipzig.de/strutz/Kanalcodierung/ldpc_introduction.pdf`, con la modifica di B.Sklar, *A Primer on Turbo Code Concepts*, IEEE Comm. Mag. 1998 ad es. presso `http://wireless.ece.ufl.edu/eel6550/lit/sklar_primer.pdf`

[90] Figura tratta dal già citato T.Strutz, ottenuta con il software di R. M. Neal disponibile presso `https://www.cs.toronto.edu/~radford/ftp/LDPC-2012-02-11/index.html`



A parte l'evidente guadagno rispetto alle prestazioni in assenza di codifica, i risultati possono essere confrontati con quelli di pag. 586 relativi al codice di Viterbi, e con il *limite di Shannon* (§ 17.3.4) che per un tasso $R_c = 1/2$ fissa un requisito minimo pari a $E_b/N_0 \geq 0.188$ dB, *mancato* dal miglior codice esaminato solamente per un dB e mezzo.

Una alternativa per la matrice $\boldsymbol{H}$ è quella che dà luogo ad un codice *irregolare*, contraddistinto da un numero di uni per riga $w_r(i)$ e per colonna $w_c(j)$ non costanti, e che in generale consegue prestazioni *migliori* di un codice regolare. In questo caso i bit x_j con $w_c(j)$ più grande sono coinvolti in un maggior numero di vincoli e dunque la stima della loro probabilità diviene più affidabile; il miglioramento di verosimiglianza è quindi *distribuito* in modo più *diffuso* da parte dei nodi c_i con $w_r(i)$ maggiore, in quanto collegati ad un maggior numero di nodi-variable. Tra i codici irregolari si menzionano quelli *quasi-ciclici* e quelli *protografici*[91], le cui matrici $\boldsymbol{H}$ presentano una qualche struttura interna, che riduce la complessità del processo di co-decodifica.



Nel grafico di $P_e(E_b/N_0)$ ottenuto dalle simulazioni si può distinguere la presenza di due regioni, la prima cosidetta di *waterfall* (cascata) in cui oltre un certo valore di E_b/N_0 la P_e decade piuttosto bruscamente, a cui fa seguito una regione *piattaforma* (error floor) in cui la riduzione di P_e è molto più graduale, se non nulla. Questo comportamento è attribuibile a *quasi-codeword* ovvero sequenze $\tilde{\boldsymbol{x}}$ la cui sindrome $\boldsymbol{H} \cdot \tilde{\boldsymbol{x}}^T$ presenta un numero ridotto di uni, e che determina una situazione di *minimo locale* per il processo di decodifica. In genere l'error floor si manifesta *prima* per i codici con andamento *più ripido* nella regione di waterfall (come quelli irregolari), sussistendo una situazione di compromesso tra le due esigenze.

Il numero di iterazioni necessario per arrivare alla decodifica corretta diminuisce all'aumentare di E_b/N_0, della dimensione del blocco n, e del tasso R_c, ed è possibile tenerne conto nel fissare il numero massimo di iterazioni prima di dichiarare un fallimento.

Rispetto ai turbo codici gli LDPC hanno il vantaggio che

- la decodifica può essere parallelizzata;
- sono più adatti alle velocità di trasmissione elevate;
- l'error floor si presenta per valori di P_e inferiori;
- resistono meglio agli errori a pacchetto;
- non è necessario alcun interleaver;

[91] Il cui grafo corrispondente è costruito a partire da *prototipi* di sottografo.

- uno stesso codice LDPC è adatto per diversi tipi di canale.

Tra gli svantaggi si citano

- una maggior complessità del codificatore;
- la realizzazione hardware può essere grande e ingombrante;
- un turbo codice si comporta meglio per lunghezze di blocco n più brevi e per tassi R_c minori.

17.5.2.5 Adozione

Il successo della codifica LDPC ha portato alla sua adozione nelle ultime generazioni di standard: dopo la tv satellitare DVB-S2 (2005) viene adottata anche per la diffusione terrestre (DVB-T2) e via cavo (DVB-C), secondo un schema concatenato con LDPC come codice interno e BCH esterno in modo da poter gestire il fenomeno dell'error floor.

E' inoltre adottata per i collegamenti a microonde Wi-MAX 802.16, per le reti wireless WiFi 802.11n, per collegamenti Ethernet 10GBase-T su cavo ritorto, per reti domestiche G.hn con distribuzione su linee elettriche, telefoniche e coassiali fino a 1 Gbit/s (ITU G.9960, 2009), nonchè nel sistema televisivo terrestre DTMB della repubblica popolare cinese, e nella telefonia 5G.[92]

[92] Vedi ad es. *An overview of channel coding for 5G NR cellular communications* presso `doi:10.1017/ATSIP.2019.10`

Capitolo 18

Caratterizzazione circuitale, rumore ed equalizzazione dati

Dedichiamo questo capitolo al completamento della discussione di tre aspetti associati al contesto reale in cui avviene la trasmissione dei segnali, che sono propedeutici allo studio dei mezzi trasmissivi del cap. 19, e profondamente collegati all'aspetto fisico che caratterizza, appunto, la trasmissione. Può essere di aiuto far precedere la lettura di questo capitolo da un ripasso di quanto esposto al cap. 8; viceversa, è possibile saltare direttamente al capitolo successivo, e tornare qui seguendo qualche riferimento ad uno dei temi seguenti.

Il primo aspetto è la descrizione dell'attraversamento di un circuito elettrico da parte di un segnale, di immediata applicabilità nel caso dei collegamenti in cavo, ma che rappresenta un passaggio obbligato per caratterizzare qualunque interfaccia elettrica, inclusi quindi i collegamenti radio. Viene poi ripreso il discorso relativo al rumore termico (§ 8.4.2.1) allargandone le cause anche alla rumorosità introdotta dai circuiti attraversati, individuando come tenere conto dei diversi contributi al rumore in maniera unitaria, compreso il caso dei collegamenti composti da più tratte collegate da ripetitori, che come vedremo segnano un nuovo punto a favore delle trasmissioni numeriche. L'ultimo aspetto affrontato è relativo alle tecniche di equalizzazione attuabili per il caso di una modulazione numerica trasmessa su di un canale che presenta distorsione lineare (§ 8.2).

18.1 Modello circuitale dei segnali

Fino ad ora un *segnale* è stato trattato nelle diverse forme di espressione analitica, sequenza simbolica, grandezza aleatoria, forma d'onda nel tempo, e densità spettrale. E' giunto il tempo di confrontarci con la corrispettiva grandezza elettrica (o elettromagnetica) che lo veicola.

Potenza di segnale e grandezze elettriche La caratterizzazione energetica dei segnali si è finora svolta *a prescindere* dalla natura fisica degli stessi, in quanto non è mai stato specificato se si trattasse di tensioni o correnti, né si sono indicate le impedenze

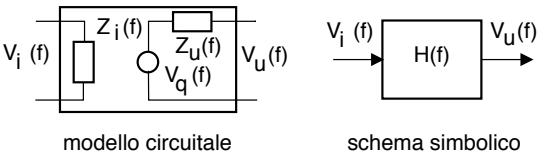
in gioco. Trattando ora di grandezze elettriche le potenze di segnale, di tensione o di corrente, saranno misurate in $(Volt)^2$ o in $(Ampere)^2$ rispettivamente.

Esempio Sia $x(t)$ un segnale di tensione. La sua potenza $\mathcal{P}_x$ ha unità di misura $\left[V^2\right]$, mentre la relativa densità di potenza $\mathcal{P}_x(f)$ si esprime in $\left[V^2/Hz\right]$.

Eseguiamo quindi una distinzione relativa al ruolo che il circuito ha nei confronti del segnale, tradizionalmente basata sul *numero di porte* del circuito.

Numero di porte Le coppie di morsetti a cui applicare o da cui prelevare un segnale vengono denominate *porte*. In questo senso un *generatore* che appunto produce il segnale, ed una *impedenza* di carico che ne assorbe la potenza, costituiscono reti ad *una* porta. Al contrario, l'oggetto che abbiamo fin qui indicato come filtro, o canale, da un punto di vista circuitale è un sistema fisico dotato di una relazione ingresso-uscita, e per questo indicato come 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 suo *schema simbolico*, in cui sono mostrate solamente 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 § 1.6 per i sistemi fisici, possono essere verificate o meno nelle reti due porte. D'altra parte, alcune relazioni e grandezze che nella *teoria dei circuiti* sono definite per segnali puramente sinusoidali, come per la *corrente alternata*, nella *teoria dei segnali* devono essere ridefinite in modo da tenere nel giusto conto dell'intera *densità spettrale* dei segnali con un contenuto informativo.

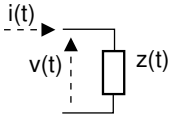
18.1.1 Bipoli

Distinguiamo tra quelli tipo attivo o di tipo passivo.

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:



$$Z(f) = R(f) + jX(f)$$

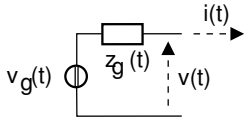
in cui $R(f)$ (*resistenza*) è una funzione *pari* di f (e sempre positiva), mentre $X(f)$ (*reattanza*) è dispari: pertanto, $Z(f) = Z^*(-f)$ e quindi $z(t)$ è reale. Allo stesso tempo

è definita *l'ammettenza*

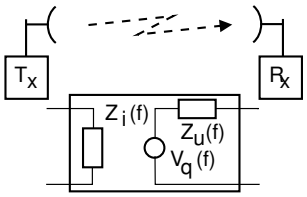
$$Y(f) = \frac{1}{Z(f)} = \frac{1}{R(f) + jX(f)} = \frac{R(f) - jX(f)}{|Z(f)|^2}$$

e la corrispondente $y(t) = \mathcal{F}^{-1}\{Y(f)\}$, permettendo di scrivere $i(t) = v(t) * y(t)$, in cui il ruolo di causa ed effetto per $i(t)$ e $v(t)$ è invertito.

Attivi Sono bipoli al cui interno è presente un generatore. Per il teorema di THÉVENIN,[1] 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).

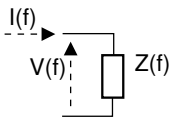


Esempio Una antenna trasmittente (§ 20.1) è schematizzabile come un bipolo passivo di impedenza pari all'impedenza di ingresso dell'antenna che assorbe la potenza erogata dal trasmettitore. Una antenna ricevente è schematizzabile come un generatore di tensione con in serie la propria impedenza di uscita, e trasferisce allo stadio di ingresso del ricevitore la potenza ricevuta per via elettromagnetica.



18.1.1.1 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 (o assorbita), indicata come $\mathcal{W}_z$ per distinguerla da quella *di segnale* $\mathcal{P}_v$, ha densità



$$\mathcal{W}_z(f) = \mathcal{P}_v(f) \cdot \Re\{Y(f)\} = \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} \qquad \left[\frac{V^2}{\Omega \cdot Hz}\right] = \left[\frac{Watt}{Hz}\right] \tag{18.1}$$

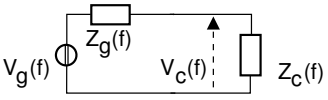
La dimostrazione di tale relazione è fornita al § 18.5.1. Osserviamo come la dipendenza di $Y(f)$ dalla frequenza svolga *una azione filtrante*, e la potenza totale assorbita (o dissipata) su $Z(f)$ vale

$$\mathcal{W}_z = \int_{-\infty}^{\infty} \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} df \qquad [Watt]$$

18.1.1.2 Connessione tra generatore e carico

La tensione ai capi del carico è valutabile applicando la *regola del partitore*[2]:

$$V_c(f) = V_g(f) \frac{Z_c(f)}{Z_c(f) + Z_g(f)}$$



ossia $V_c(f) = V_g(f) H(f)$ con $H(f) = \frac{Z_c(f)}{Z_c(f)+Z_g(f)}$. La densità di potenza di segnale ai

[1] Vedi ad es. `http://it.wikipedia.org/wiki/Teorema_di_Thévenin`

[2] `http://it.wikipedia.org/wiki/Partitore_di_tensione`

capi del carico vale quindi $\mathcal{P}_{v_c}(f) = \mathcal{P}_{v_g}(f)\,|H(f)|^2$, e la densità di potenza *dissipata* su $Z_c(f)$ risulta

$$\begin{aligned}\mathcal{W}_{z_c}(f) &= \mathcal{P}_{v_c}(f)\,\frac{R_c(f)}{|Z_c(f)|^2} = \mathcal{P}_{v_g}(f)\left|\frac{Z_c(f)}{Z_c(f)+Z_g(f)}\right|^2 \frac{R_c(f)}{|Z_c(f)|^2} = \\ &= \mathcal{P}_{v_g}(f)\,\frac{R_c(f)}{\left|Z_c(f)+Z_g(f)\right|^2} \qquad (18.2)\end{aligned}$$

Osserviamo dunque che la densità di potenza assorbita 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, indicata come *potenza disponibile*.

18.1.1.3 Potenza disponibile e massimo trasferimento di potenza

La $\mathcal{W}_{z_c}(f)$ espressa da (18.2) risulta massima quando il denominatore viene reso minimo, ed al § 18.5.2 si mostra che ciò avviene qualora risulti $R_c(f) = R_g(f)$ e $X_c(f) = -X_g(f)$, ovvero $Z_c(f) = Z_g^*(f)$, in modo da poter enunciare

$$\begin{aligned} &se \qquad Z_c(f) = Z_g^*(f) \qquad (18.3)\\ &allora \qquad \mathcal{W}_{z_c}(f) = \max_{Z_c(f)}\left\{\mathcal{W}_{z_c}(f)\right\} = \frac{\mathcal{P}_{v_g}(f)}{4R_g(f)} = \mathcal{W}_{d_g}(f)\end{aligned}$$

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*[3].

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

$$\alpha(f) = \frac{4R_g(f)\,R_c(f)}{\left|Z_g(f)+Z_c(f)\right|^2}$$

(vedi § 18.5.3) e quindi in generale si ha $\mathcal{W}_{z_c}(f) = \alpha(f)\,\mathcal{W}_{d_g}(f)$.

18.1.1.4 Adattamento di impedenza per assenza di distorsione lineare

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* (pag. 231), ovvero risulti $|H(f)| = cost$ e $\arg\{H(f)\} = -2\pi f\tau$: tali condizioni sono anche indicate come *assenza di distorsione lineare*. Il risultato cercato si ottiene qualora si ponga

$$Z_c(f) = \alpha Z_g(f) \quad \text{con } \alpha \text{ reale}$$

[3] E' 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 $Z_g(f) \to 0$.

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 $\alpha = 1$.

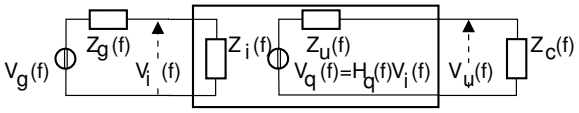
$Z_g(f)$ reale Notiamo che massimo trasferimento di potenza ed assenza di distorsione lineare possono sussistere *congiuntamente,* a patto che $Z_g(f) = R_g$, ovvero quando sia il generatore che il carico sono caratterizzati da una impedenza reale.

18.1.2 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.*

18.1.2.1 Modello circuitale

In figura è mostrato un possibile modello circuitale[4] 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

$$V_u(f) = V_q(f) H_u(f)$$

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 risposta in frequenza complessiva $H(f)$ risulta:

$$V_u(f) = V_g(f) H_i(f) H_q(f) H_u(f) = V_g(f) H(f) \tag{18.4}$$

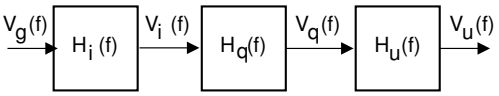
[4]Sono chiaramente possibili modelli diversi, basati su topologie e relazioni differenti. Esistono infatti circuiti a T, ad L, a scala, a traliccio, a pigreco; le relazioni tra le grandezze di ingresso ed uscita possono essere espresse mediante modelli definiti in termini di impedenze, ammettenze, e 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.

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.

18.1.2.2 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 soggiacente, e di rendere del tutto evidente come la risposta in frequenza complessiva abbia origine dal prodotto di tre termini, di cui solo uno ($H_q(f)$) rappresenta la rete due porte in senso stretto.



18.1.2.3 Trasferimento energetico

Applicando ora la (18.1) alla potenza ceduta dal generatore controllato $V_q(f)$ al carico $Z_c(f)$, e tenendo conto della (18.4), si ottiene

$$\mathcal{W}_c(f) = \mathcal{P}_{v_u}(f)\frac{R_c(f)}{|Z_c(f)|^2} = \mathcal{P}_{v_g}(f)\,|H(f)|^2\frac{R_c(f)}{|Z_c(f)|^2}$$

che dipende anche da $Z_c(f)$. Proseguiamo l'analisi nel tentativo di individuare una relazione di trasferimento energetico che rappresenti caratteristiche *della sola 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 all'uscita della rete.

Guadagno di potenza E' il rapporto tra la potenza ceduta al carico e quella assorbita all'ingresso della rete:

$$\begin{aligned} G_{\mathcal{W}}(f) &= \frac{\mathcal{W}_c(f)}{\mathcal{W}_i(f)} = \mathcal{P}_{v_g}(f)\,|H(f)|^2\frac{R_c(f)}{|Z_c(f)|^2}\cdot\frac{1}{\mathcal{P}_{v_g}(f)}\frac{\left|Z_g(f)+Z_i(f)\right|^2}{R_i(f)} = \\ &= |H(f)|^2\frac{R_c(f)}{R_i(f)}\cdot\left|\frac{Z_g(f)+Z_i(f)}{Z_c(f)}\right|^2 = \\ &= \left|H_q(f)\right|^2\cdot\frac{R_c(f)}{R_i(f)}\cdot\left|\frac{Z_i(f)}{Z_u(f)+Z_c(f)}\right|^2 \end{aligned}$$

ed evidentemente è ancora funzione di $Z_c(f)$ ([5]). Notiamo ora che, qualora il carico sia adattato per il massimo trasferimento di potenza ($Z_c(f) = Z_u^*(f)$), la potenza ceduta

[5] L'ultimo passaggio tiene conto che (omettendo la dipendenza da f):

$$|H|^2\cdot\left|\frac{Z_g+Z_i}{Z_c}\right|^2 = \left|\frac{Z_i}{Z_i+Z_g}H_q\frac{Z_c}{Z_c+Z_u}\right|^2\left|\frac{Z_g+Z_i}{Z_c}\right|^2 = \left|H_q\right|^2\cdot\left|\frac{Z_i}{Z_u+Z_c}\right|^2$$

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

$$G_{\mathcal{W}_{Max}}(f) = \left|H_q(f)\right|^2 \cdot \frac{|Z_i(f)|^2}{4R_i(f)\,R_u(f)} \tag{18.5}$$

Guadagno disponibile Il rapporto tra la potenza disponibile di uscita e quella disponibile del generatore posto in ingresso della rete, indipendentemente dal fatto se l'ingresso della rete presenti o meno le condizioni per il massimo trasferimento di potenza, è detto *guadagno disponibile* e vale

$$\begin{aligned} G_d(f) &= \frac{\mathcal{W}_{d_u}(f)}{\mathcal{W}_{d_g}(f)} = \frac{\mathcal{P}_{v_q}(f)}{4R_u(f)} \cdot \frac{4R_g(f)}{\mathcal{P}_{v_g}(f)} = \\ &= \mathcal{P}_{v_g}(f)\,|H_i(f)|^2 \left|H_q(f)\right|^2 \cdot \frac{R_g(f)}{R_u(f)} \cdot \frac{1}{\mathcal{P}_{v_g}(f)} = \\ &= |H_i(f)|^2 \left|H_q(f)\right|^2 \cdot \frac{R_g(f)}{R_u(f)} \end{aligned} \tag{18.6}$$

La relazione trovata mostra la dipendenza di $G_d(f)$ dalle condizioni di chiusura in ingresso. Quando l'impedenza di ingresso $Z_i(f)$ è tale da permettere il conseguimento del massimo trasferimento di potenza, ovvero $Z_g(f) = Z_i^*(f)$, 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 (18.6) diviene

$$G_{d_{Max}}(f) = \left|H_q(f)\right|^2 \cdot \frac{|Z_i(f)|^2}{4R_u(f)\,R_i(f)} \tag{18.7}$$

Quest'ultima quantità è chiamata *guadagno disponibile* DELLA RETE DUE PORTE ed è quella che appunto dipende solo dai parametri della rete stessa. Confrontando (18.7) con (18.5) notiamo che $G_{d_{Max}}(f)$ coincide con $G_{\mathcal{W}_{Max}}(f)$. Confrontando (18.7) con (18.6), 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 $|H_i(f)|^2 \frac{1}{|Z_i(f)|^2} = \left|\frac{Z_i(f)}{Z_i(f)+Z_g(f)}\right|^2 \frac{1}{|Z_i(f)|^2} = \frac{1}{\left|Z_i(f)+Z_g(f)\right|^2}$, otteniamo

$$G_d(f) = \frac{4R_g(f)\,R_i(f)}{\left|Z_g(f)+Z_i(f)\right|^2} \cdot G_{d_{Max}}(f)$$

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

  $$\mathcal{W}_{d_u}(f) = G_{d_{Max}}(f)\,\mathcal{W}_{d_g}(f)$$

  e dunque l'uscita della rete due porte si comporta come un generatore equivalente, caratterizzato da una nuova $\mathcal{W}_{d_u}(f)$ ed una diversa impedenza interna $Z_u(f)$;
- se in ingresso non si verifica 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 in uscita risulta

$$\mathcal{W}_{d_u}(f) = \beta(f) \cdot G_{d_{Max}}(f)\,\mathcal{W}_{d_g}(f) = \frac{4R_g(f)\,R_i(f)}{|Z_g(f) + Z_i(f)|^2} \cdot G_{d_{Max}}(f)\,\mathcal{W}_{d_g}(f)$$

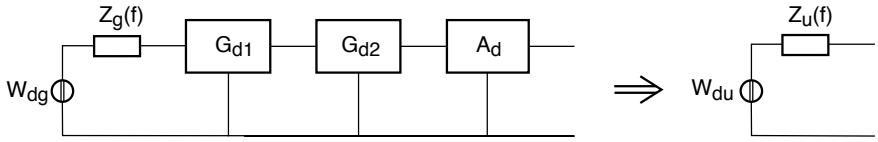
- se infine il carico $Z_c(f)$ in uscita alla rete non è adattato, assorbe una potenza inferiore a $\mathcal{W}_{d_u}(f)$ e pari a (vedi Appendice 18.5.3)

$$\mathcal{W}_c(f) = \alpha(f) \cdot \mathcal{W}_{d_u}(f) = \frac{4R_u(f)\,R_c(f)}{|Z_u(f) + Z_c(f)|^2} \cdot \mathcal{W}_{d_u}(f)$$

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 $A_d(f)\ [\text{dB}] = -G_d(f)\ [\text{dB}]$.

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 disponibile complessivo è il prodotto dei singoli guadagni disponibili: $G_{d_{Tot}} = G_{d1} \cdot G_{d2} \cdot \ldots \cdot G_{dN}$.

Esempio In figura è mostrato un generatore con potenza disponibile 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 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[6]:

$$W_{du} = W_{dg} \cdot G_{d1} \cdot G_{d2} \cdot \frac{1}{A_d} \cdot \frac{1}{A_s}$$

che può essere egualmente valutato operando in decibel, come

$$W_{du}\ [dBW] = W_{dg}\ [dBW] + G_{d1}\ [dB] + G_{d2}\ [dB] - A_d\ [dB] - A_s\ [dB]$$

in cui ovviamente, qualora W_{dg} fosse espresso in dBm anziché dBW, lo stesso accadrebbe per W_{du}.

Collegamento radio Con riferimento al circuito equivalente per una coppia di antenne di pag. 605, precisiamo che la potenza trasmessa è quella assorbita dall'impedenza di ingresso dell'antenna trasmittente, mentre quella ricevuta è quella ceduta dal generatore equivalente dell'antenna ricevente, all'impedenza di ingresso del ricevitore.

[6] L'attenuazione *supplementare* (pag. 643) può esprimere il peggioramento dovuto al mancato verificarsi delle condizioni di massimo trasferimento tra le impedenze di uscita e di ingresso delle reti affacciate.

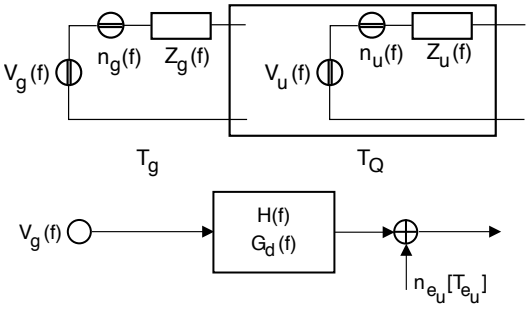
18.2 Rumore nelle reti due porte

Al § 8.4.2.1 abbiamo mostrato come in un bipolo attivo, una volta ridotto[7] nella sua forma canonica come un generatore di segnale (a vuoto) $V_g(f)$ con in serie una impedenza interna $Z_g(f)$ a temperatura T_g, quest'ultima contenga al suo interno *anche* un generatore di rumore *termico*, con densità di potenza di segnale $\mathcal{P}_n(f) = 2kT_gR(f)$ Volt^2, generato dalla sola parte reale $R(f) = \mathcal{R}\{Z_g(f)\}$ dell'impedenza. Qualora tale generatore sia chiuso su di un carico $Z_c(f) = Z_g^*(f)$ adattato per il massimo trasferimento di potenza, su $Z_c(f)$ si dissipa sia la potenza *disponibile* di segnale $\mathcal{W}_{dg}(f) = \frac{\mathcal{P}_g(f)}{4R(f)}$, sia quella (sempre disponibile) di rumore $\mathcal{W}_{d_n}(f) = \frac{\mathcal{P}_n(f)}{4R(f)} = \frac{1}{2}kT_g\ \left[\frac{\text{Watt}}{\text{Hz}}\right]$: pertanto, il generatore nasce *già di per se* rumoroso, con un

$$SNR_g(f) = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g}$$

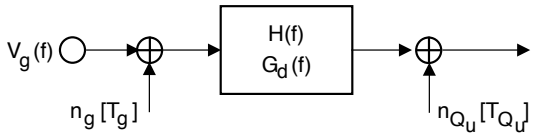
Qualora tra il generatore ed il carico siano invece interposte una (o più) reti due porte, occorre investigare su come tenere conto dell'ulteriore rumore introdotto da queste ultime.

Temperatura equivalente di uscita Collegando un generatore rumoroso a temperatura T_g all'ingresso di una rete due porte a temperatura T_Q (vedi fig. a lato), in uscita della rete troviamo un processo di rumore dipendente sia dal generatore che dalla rete, e la cui potenza disponibile $\mathcal{W}_{dn_u}(f)$ può essere espressa in funzione di una temperatura *equivalente* di uscita $T_{e_u}(f)$ (seconda figura), tale che



$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}kT_{e_u}(f)$$

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* (fig. a lato); scriviamo dunque

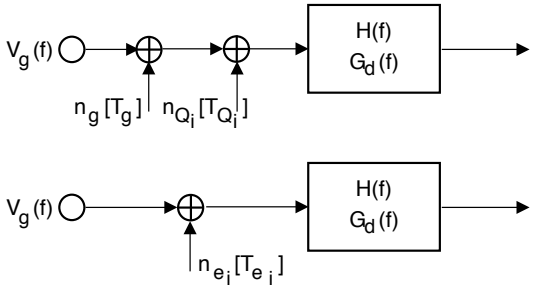


$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}k \cdot \left[T_g(f)\, G_d(f) + T_{Q_u}(f)\right]$$

in cui la potenza disponibile in ingresso alla rete (che ha guadagno disponibile $G_d(f)$) è riportata in uscita, moltiplicata per $G_d(f)$.

Temperatura di sistema Se ora riportiamo in ingresso alla rete il contributo di

[7] Applicando il *teorema di Thévenin*, vedi `https://it.wikipedia.org/wiki/Teorema_di_Thévenin`



rumore dovuto a T_{Q_u} otteniamo

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}kG_d(f)\cdot\left[T_g(f) + T_{Q_i}(f)\right]$$

(in cui $T_{Q_i}(f) = \frac{T_{Q_u}(f)}{G_d}$), ovvero

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}kG_d(f)\,T_{e_i}(f)$$

dove

$$T_{e_i}(f) = T_g(f) + T_{Q_i}(f) = T_g(f) + \frac{T_{Q_u}(f)}{G_d(f)} \tag{18.8}$$

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}?

18.2.1 Reti passive

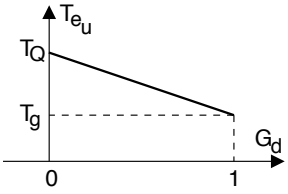
Nel caso in cui la rete due porte non contenga elementi attivi, e considerando tutti i componenti passivi della rete alla stessa temperatura T_Q, si può mostrare che risulta

$$\begin{cases} T_{Q_u}(f) &= [1 - G_d(f)]\,T_Q \\ T_{Q_i}(f) &= \frac{T_{Q_u}(f)}{G_d(f)} = [A_d(f) - 1]\,T_Q \end{cases}$$

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 $0 \le G_d \le 1$), 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.



18.2.1.1 Rapporto SNR in uscita

La valutazione del rapporto segnale rumore in uscita alla rete porta a

$$SNR_u(f) = \frac{\mathcal{W}_{dg}(f)\,G_d(f)}{\frac{1}{2}kT_{e_i}(f)\,G_d(f)} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}k\cdot\left[T_g(f) + [A_d(f) - 1]\,T_Q\right]}$$

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 come

$$\begin{aligned} \frac{SNR_i(f)}{SNR_u(f)} &= \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)} \cdot \frac{\frac{1}{2}k\cdot\left[T_g(f) + [A_d(f) - 1]\,T_Q\right]}{\mathcal{W}_{dg}(f)} = \\ &= 1 + \frac{T_Q}{T_g(f)} \cdot [A_d(f) - 1] \end{aligned} \tag{18.9}$$

18.2.1.2 Fattore di rumore per reti passive

Il rapporto (18.9) $F(f) = 1 + \frac{T_Q}{T_g(f)} \cdot [A_d(f) - 1] \geq 1$ è chiamato *fattore di rumore*[8] 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 presenta 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.

18.2.2 Reti attive

In questo caso il rumore introdotto dalla rete non ha origine *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 $G_d > 1$. Per le reti attive si può quindi esprimere l'SNR in uscita come

$$SNR_u(f) = \frac{\mathcal{W}_{dg}(f)\, G_d(f)}{\frac{1}{2}k\left[G_d(f)\, T_g(f) + T_{Q_u}(f)\right]} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}k \cdot \left[T_g(f) + T_{Q_i}(f)\right]}$$

ed il peggioramento individuato in (18.9) come

$$\frac{SNR_i(f)}{SNR_u(f)} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)} \cdot \frac{\frac{1}{2}k \cdot \left[T_g(f) + T_{Q_i}(f)\right]}{\mathcal{W}_{dg}(f)} \tag{18.10}$$

$$= 1 + \frac{T_{Q_i}(f)}{T_g(f)} = F\left(f, T_g\right) \tag{18.11}$$

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

18.2.2.1 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 $T_0 = 290\ ^oK = 17\ ^oC$. 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. Il valore di $F(f)$ misurato ci permette dunque il calcolo di $T_{Q_i}(f) = T_0\,[F(f) - 1]$ che, sostituito nella (18.8), permette finalmente di valutare la temperatura di sistema come

$$T_{e_i}(f) = T_g(f) + T_0\,[F(f) - 1] \tag{18.12}$$

[8] A 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.

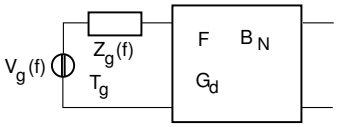
mentre dalla (18.11) si determina il peggioramento dell'SNR come

$$\frac{SNR_i\,(f)}{SNR_u\,(f)} = 1 + \frac{T_0}{T_g\,(f)}\,[F\,(f) - 1]$$

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 $T_0 = 290$ ^{o}K = 17 ^{o}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 è a temperatura $T_Q = T_0$ 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

$$\mathcal{W}_{dn_u}\,(f) = \frac{1}{2}kT_{e_i}G_d = \frac{1}{2}k \cdot \left[T_g + T_{Q_i}\right] \cdot G_d$$

in generale $F = 1 + \frac{T_{Q_i}}{T_0}$ e quindi $T_{Q_i} = T_0\,(F-1)$, dunque

$$\mathcal{W}_{dn_u}\,(f) = \frac{1}{2}k \cdot \left[T_g + T_0\,(F-1)\right] \cdot G_d$$

Pertanto, la potenza disponibile di rumore si ottiene integrando la densità sulla banda di rumore

$$\mathcal{W}_{dn_u} = k \cdot \left[T_g + T_0\,(F-1)\right] \cdot G_d B_N$$

che, nel caso in cui $T_g = T_0$, si riduce a $\mathcal{W}_{dn_u} = kT_0FG_dB_N$. 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

$$SNR_u = \frac{SNR_i}{F} = \frac{\mathcal{W}_{dg}}{kT_0FB_N} = \frac{\mathcal{W}_{dg}}{kT_{e_i}B_N}$$

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:

$$\begin{aligned} SNR_u &= \frac{SNR_i}{F\,(T_g)} = \frac{\mathcal{W}_{dg}}{kT_gB_N} \cdot \frac{1}{1 + \frac{T_{Q_i}}{T_g}} = \frac{\mathcal{W}_{dg}}{kT_gB_N} \cdot \frac{1}{1 + \frac{T_0(F-1)}{T_g}} = \\ &= \frac{\mathcal{W}_{dg}}{k\left[T_g + T_0\,(F-1)\right]B_N} = \frac{\mathcal{W}_{dg}}{kT_{e_i}B_N} \end{aligned}$$

Esercizio Un trasmettitore con potenza di 50 mW e portante 30 MHz modula AM BLD PS un segnale con banda $\pm B = \pm$ 10 KHz, prodotto da un generatore a temperatura T_0. 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. *N.B.: l'esercizio può essere affrontato con le ulteriori conoscenze dei §§ 19.1 e 20.2.*

Svolgimento Assumendo che si verifichino le condizioni di massimo trasferimento di potenza, il valore desiderato $SNR = SNR_0 = \frac{W_R}{W_N}$ può essere ottenuto se $W_R = W_N \cdot SNR = 2B \cdot \mathcal{W}_{dN}(f) \cdot SNR = B \cdot kT_0F \cdot SNR$, e quindi occorre ricevere un potenza

$$W_{R_{min}}(\text{dBm}) = 10\log_{10}10^4(\text{Hz}) - 174(\text{dBm/Hz}) + F_{dB} + SNR_{dB} =$$
$$= 40 - 174 + 10 + 25 = -99 \text{ dBm}.$$

Il guadagno di sistema (pag. 643) 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 applicando la (20.6) di pag. 672 scriviamo

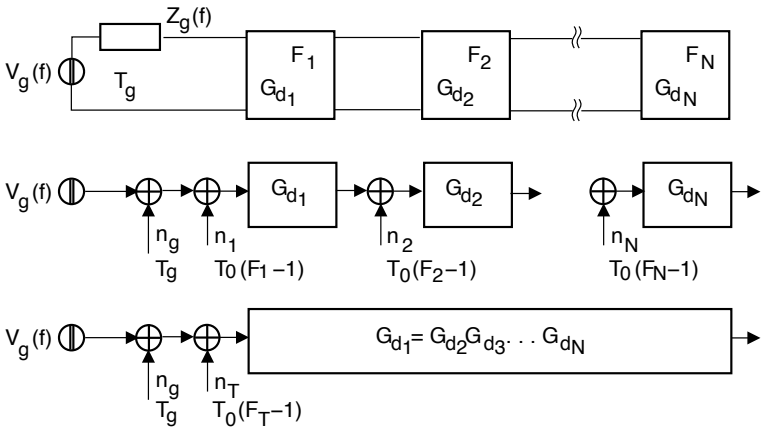
$$A_d = 116 = 32.4 + 20\log_{10} f(\text{MHz}) + 20\log_{10} d(\text{Km}) =$$
$$= 32.4 + 29.5 + 20\log_{10} d(\text{Km})$$

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.

18.2.3 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 $T_{Q_i}^{(n)} = T_0\left(F^{(n)} - 1\right)$



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

$$T_{Q_i}^{(T)} = T_{Q_i}^{(1)} + T_{Q_i}^{(2)} \frac{1}{G_{d_1}} + T_{Q_i}^{(3)} \frac{1}{G_{d_1} G_{d_2}} + \cdots + T_{Q_i}^{(N)} \frac{1}{\prod_{n=1}^{N-1} G_{d_n}}$$

in cui, sostituendo le espressioni per i $T_{Q_i}^{(n)}$ si ottiene

$$T_{Q_i}^{(T)} = T_0 \cdot \left[F_1 - 1 + \frac{F_2 - 1}{G_{d_1}} + \frac{F_3 - 1}{G_{d_1} G_{d_2}} + \cdots + \frac{F_N - 1}{\prod_{n=1}^{N-1} G_{d_n}} \right]$$

Applicando la definizione $F^{(T)} = 1 + \frac{T_{Q_i}^{(T)}}{T_0}$, si ottiene

$$F^{(T)} = F_1 + \frac{F_2 - 1}{G_{d_1}} + \frac{F_3 - 1}{G_{d_1} G_{d_2}} + \cdots + \frac{F_N - 1}{\prod_{n=1}^{N-1} G_{d_n}}$$

che costituisce proprio l'espressione cercata:

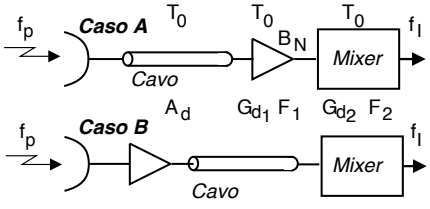
$$F^{(T)} = F_1 + \sum_{i=2}^{N} \frac{F_i - 1}{\prod_{j=1}^{i-1} G_{d_j}}$$

Il risultato, noto come *formula di Friis*[9], si presta alle seguenti considerazioni:

- la *prima* rete due porte deve avere F *più piccolo possibile*, dato che 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*, dato che 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 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 ϕ = 1.2/4.4 mm lunga 50 metri, un filtro-amplificatore con guadagno disponibile G_{d_1} = 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_{d_2} = 0 dB e F_2 = 10 dB. Tutti i componenti a valle dell'antenna si trovano alla stessa temperatura T_0 = 290 ^{o}K. Calcolare:

[9] Vedi `https://it.wikipedia.org/wiki/Formula_di_Friis`, ma da non confondere con la (20.6), anche se si tratta... della stessa persona!

1) La minima potenza disponibile W_{d_R} che occorre ricevere per ottenere SNR_0 = 50 dB nei due casi. Ripetere il calcolo supponendo l'antenna ricevente a temperatura T_a = 10 ^{o}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 ≃ 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_{d_1}	F_2	G_{d_2}
dB	11.85	.4	20	10	0
lineare	15.3	1.1	100	10	1

1) -

A) Il fattore di rumore complessivo risulta

$$F^A = F_{cavo} + A_d\,(F_1 - 1) + \frac{A_d}{G_{d_1}}\,(F_2 - 1) = 15.3 + 15.3 \cdot (.1) + \frac{15.3}{100}\,(9) = 18.2$$

ovvero pari a 12.6 dB. Dato che per la trasmissione televisiva AM-BLU si ha $SNR = SNR_0$, scriviamo

$$W_{d_R} = SNR \cdot W_{d_N} = SNR_0 \cdot F^A \cdot B_N \cdot kT_0 \text{ e quindi}$$

$$W_{d_R}(\text{dBm}) = SNR_0(\text{dB}) + F^A(\text{dB}) + B_N(\text{dBMHz}) + KT_0(\text{dBm/MHz}) =$$

$$= 50 + 12.6 + 8.45 - 114 = -43 \text{ dBm}$$

B) Il fattore di rumore complessivo risulta ora

$$F^B = F_1 + \frac{(F_{cavo} - 1)}{G_{d_1}} + \frac{A_d}{G_{d_1}}\,(F_2 - 1) = 1.1 + \frac{14.3}{100} + \frac{15.3}{100}\,(9) = 2.26$$

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_{d_R} = -52.1 dBm.

Nel caso in cui T_a = 10 ^{o}K ≠ T_0, non si ottiene più $T_{e_i} = FT_0$, ma occorre introdurre la T_{Q_i} della rete riportata al suo ingresso, e considerare la rete non rumorosa in modo da scrivere $T_{e_i} = T_g + T_{Q_i} = T_A + T_0\,(F - 1)$. Ripetiamo i calcoli per i due casi **A** e **B**:

A) $W_{dR_W} = SNR \cdot W_{d_N} = SNR \cdot B_N \cdot k \cdot \left(T_a + T_{Q_i}\right) =$

$$= SNR \cdot B_N \cdot k \cdot \left(T_a + T_0\left(F^A - 1\right)\right), \text{ che espresso in dB fornisce}$$

$$W_{dR_{dBW}} = SNR_{dB} + 10\log_{10} 7 \cdot 10^6 + 10\log_{10}\left(1.38 \cdot 10^{-23}\,(10 + 290 \cdot 17.2)\right)$$

$$= 50 + 68.5 - 191.61 = -73.11 \text{ dBW} = -43.11 \text{ dBm}$$

B) $W_{d_R}(\text{dBW}) = 50 + 68.5 + 10\log_{10}\left(1.38 \cdot 10^{-23}\,(10 + 290 \cdot 1.26)\right) = -84.3$ dBW = -54.3 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_{Q_i} prodotto dal cavo.

2) In un collegamento radio terrestre si assume $T_a = 290\ ^oK$. 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) $W_{d_T} = W_{d_R} + A_d = -43.11 + 72.4 = 29{,}29$ dBm = 850 mW

B) $W_{dT} = W_{d_R} + A_d = -54.3 + 72.4 = 18.1$ dBm = 66 mWatt

Per il downlink si ha $d = 36.000$ Km, mentre $T_a = 10\ ^oK$. 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_{d_R} ottenuto per il caso **B**, otteniamo

$$W_{d_T} = W_{d_R} + A_d = -54.3 + 109.5 = 55.2 \text{ dBm} = 25.2 \text{ dBW} \longrightarrow 331 \text{ Watt}$$

3) Nel caso di adattamento, la potenza ceduta all'antenna T_x è proprio quella disponibile del generatore, e quindi si ha $W_{d_T} = \frac{\sigma_g^2}{4R}$, da cui

$$\sigma_g = \sqrt{W_{d_T} 4R} = \sqrt{66 \cdot 10^{-3} \cdot 4 \cdot 50} = 3.63 \text{ Volt.}$$

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 (18.1))

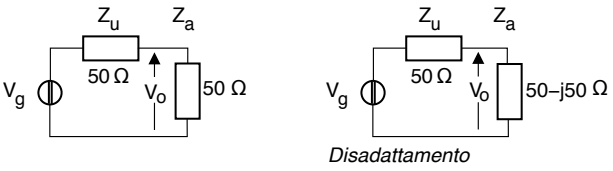
$$\mathcal{W}_T = \mathcal{P}_{v_o} \frac{R_a}{|Z_a|^2} = P_{v_o} \frac{50}{50^2+50^2} = P_{v_o} \cdot 10^{-2}$$

e quindi $P_{v_o} \simeq 6.6$ (Volt^2). 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, $P_{v_g} = \frac{P_{v_o}}{0.4} = \frac{6.6}{0.4} = 16.5$ Volt^2, ovvero $V_{g_{eff}} = \sqrt{16.5} \simeq 4$ Volt.

Evidentemente, il disadattamento produce un innalzamento del valore efficace, se si vuol mantenere la stessa potenza di uscita.



Disadattamento

18.3 Rumore nei ripetitori

Come illustreremo al cap. 19, la propagazione del segnale attraverso un mezzo trasmissivo ne determina una *attenuazione* la cui entità aumenta con la distanza. La realizzazione di un collegamento molto lungo mediante un'unica *tratta* è pertanto praticamente impossibile, sia a causa del livello troppo ridotto del segnale che sarebbe

ricevuto, sia (per un collegamento radio) per la mancanza di condizioni di visibilità. Occorre pertanto suddividere il collegamento in più *tratte*, intervallate da stadi di amplificazione (o *ripetitori*) progettati in modo da compensare l'attenuazione del segmento appena attraversato.

18.3.1 Ripetitore trasparente

L'aggettivo *trasparente* si riferisce al fatto che, oltre al segnale, viene amplificato anche il rumore presente in ingresso, e nel caso di una trasmissione analogica, questo è l'unico modo di procedere. Le trasmissioni numeriche invece adottano un ripetitore *rigenerativo* che produce un *nuovo* segnale, privo di rumore ma con qualche bit errato in più: affronteremo questo caso al § 18.3.2.

Analizziamo ora la questione con riferimento ad un collegamento radio, anche se la trattazione può essere estesa ad altre tecniche trasmissive, come il cavo o le fibre ottiche, e consideriamo una successione di M tratte come mostrato in fig. 18.1.

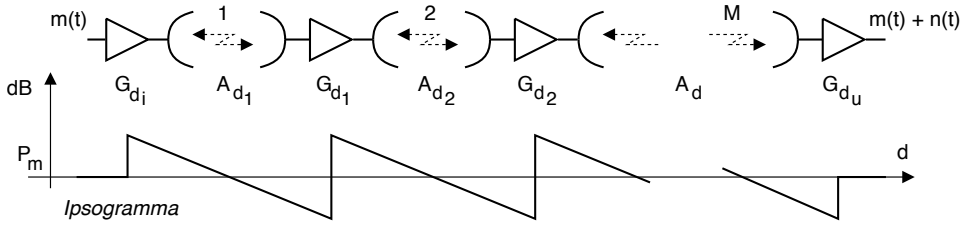


Figura 18.1: Suddivisione di un collegamento in M tratte, e relativo diagramma del livello di segnale in dB o *ipsogramma*

Il ripetitore (il *triangolo*) interposto tra ogni coppia di tratte (le *saette*) amplifica il segnale (ed il rumore) di una quantità pari al proprio guadagno disponibile G_{d_i}, reso uguale all'inverso dell'attenuazione disponibile della tratta precedente, ovvero $G_{d_i} = 1/A_{d_i}$. Il rumore termico accumulato alla fine del collegamento può essere calcolato con i metodi già discussi, ma considerando che il livello di segnale di uscita è lo stesso per tutti i ripetitori si ritrova il risultato ottenuto al § 8.4.1, come andiamo ad illustrare. Al § 18.3.1.2 valuteremo poi come le distorsioni *di non linearità* (§ 8.3) degli amplificatori possano intervenire nel progetto.

18.3.1.1 Rumore termico accumulato

In base all'uscita dell'ultimo ripetitore indicata come $m(t)+n(t)$ si può definire un SNR complessivo come $SNR_T = \mathcal{P}_m/\mathcal{P}_n$. D'altra parte il rumore $n(t)$ è dovuto ai contributi di rumore $n_i(t)$ introdotti nelle singole tratte; essendo tali contributi statisticamente indipendenti tra loro la potenza di rumore accumulata è *la somma* delle singole potenze:

$$\mathcal{P}_n = \sigma_n^2 = E\left\{n^2(t)\right\} = E\left\{\left(\sum_i n_i(t)\right)^2\right\} = \sum_i E\left\{n_i^2(t)\right\} = \sum_{i=1}^{M} \mathcal{P}_{n_i}$$

Deriviamo i singoli termini $\mathcal{P}_{n_i}$ osservando che all'uscita di ogni ripetitore è definito un $SNR_i = \mathcal{P}_{m_i}/\mathcal{P}_{n_i}$ *locale* da cui ottenere $\mathcal{P}_{n_i} = \mathcal{P}_{m_i}/SNR_i$, consentendoci di scrivere

$$SNR_T = \frac{\mathcal{P}_m}{\sum_i \mathcal{P}_{n_i}} = \frac{\mathcal{P}_m}{\sum_i \mathcal{P}_{m_i}/SNR_i}$$

A questo punto notiamo che, essendo il livello di segnale $\mathcal{P}_{m_i}$ in ingresso a ciascun ripetitore lo stesso, ovvero $\mathcal{P}_{m_i} = \mathcal{P}_m$ per $\forall i$, i singoli contributi $\mathcal{P}_{n_i}$ al rumore complessivo possono essere espressi nei termini di uno stesso livello di segnale, ovvero $\mathcal{P}_{n_i} = \frac{\mathcal{P}_m}{SNR_i}$, e dunque per l'$SNR$ complessivo si ottiene

$$SNR_T = \frac{\mathcal{P}_m}{\mathcal{P}_m \sum_i \frac{1}{SNR_i}} = \frac{1}{\sum_i \frac{1}{SNR_i}}$$

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, come illustrato anche al § 8.4.1.

Proseguiamo l'analisi ipotizzando ora che tutte le tratte siano *uguali tra loro*, ovvero con eguali A_d e G_d, uguali temperature di rumore, ed uguali SNR_i. In tal caso si ottiene

$$SNR_T = \frac{1}{\frac{M}{SNR_i}} = \frac{SNR_i}{M}$$

ovvero un $M-$esimo dell'SNR_i locale di ciascun ripetitore[10]. Sembrerebbe dunque che per migliorare l'SNR complessivo sia sufficiente aumentare la potenza 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.

18.3.1.2 Compromesso tra rumore termico e distorsione

Ai §§ 8.3 e 13.3 si è studiato come per un segnale modulato la presenza di un elemento a comportamento non lineare (tipicamente l'amplificatore di potenza del ripetitore) produce *interferenza in banda*, la cui potenza $\mathcal{P}_{NL}$ dipende con legge cubica dalla potenza del segnale trasmesso. Indichiamo quindi con $SNR_{NL} = \frac{\mathcal{P}_m}{\mathcal{P}_{NL}}$ il rapporto SNR complessivo del collegamento dovuto a cause di non linearità, ed osserviamo che questo *diminuisce* all'aumentare della potenza trasmessa da ogni ripetitore. L'SNR complessivo deve tener conto sia del rumore termico che della distorsione non lineare, e dato che questi sono *statisticamente indipendenti*, può essere espresso come "il parallelo" di entrambi, ossia

[10] 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 (cap. 14).

$$SNR = \frac{1}{\frac{1}{SNR_T} + \frac{1}{SNR_{NL}}}$$

Dato che all'aumentare della potenza di trasmissione SNR_T diminuisce mentre SNR_{NL} 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.

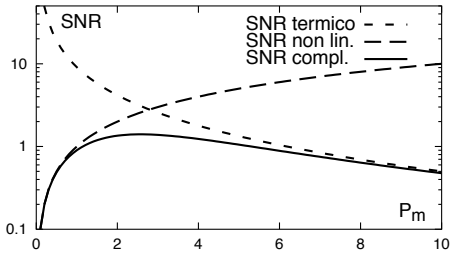
Esempio La figura a lato mostra l'andamento di

$$SNR = \frac{1}{\frac{1}{SNR_T} + \frac{1}{SNR_I}}$$

dovuto ai due termini

. $SNR_T = \mathcal{P}_m$ e $SNR_{NL} = \frac{\mathcal{P}_m}{.1 \cdot \mathcal{P}_m^2 + .01 \cdot \mathcal{P}_m^3}$

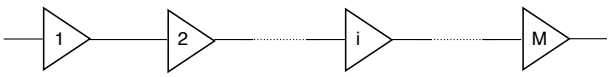
Come si vede, con questi valori SNR presenta un massimo per $\mathcal{P}_m \simeq 2.5$.



18.3.2 Ripetitore rigenerativo

Affrontiamo ora il caso in cui la trasmissione sia di natura numerica (§ 16): in tal caso i dispositivi intermedi di amplificazione anziché essere di tipo *trasparente*, vengono detti *rigenerativi*, ovvero mentre dal lato ricevente svolgono l'intero processo di decisione (§ 15.4) sui simboli del messaggio, dal lato di trasmissione generano un *nuovo* segnale in cui è *assente* il rumore presente in ricezione.

La valutazione delle prestazioni *complessive* nel caso si adottino M ripetitori rigenerativi posti in serie lungo un medesimo collegamento è piuttosto semplice, qualora siano equispaziati, ed ognuno caratterizzato dalla stessa $p = P_e^{bit}$. In tale circostanza si ricade infatti nel caso delle *prove ripetute*, permettendo di esprimere la probabilità che lo stesso bit sia sbagliato da n ripetitori (su M) mediante la distribuzione di Bernoulli (§ 22.1) $p(n) = \binom{M}{n} p^n (1-p)^{M-n}$. Ovviamente il bit risulterà sbagliato solo se avrà subito un numero *dispari* di errori, e dunque complessivamente possiamo scrivere



$$P_e^T = \sum_{\substack{n=1 \\ n\ dispari}}^{M} p(n) = Mp(1-p)^{M-1} + \frac{M(M-1)(M-2)}{3!} p^3 (1-p)^{M-3} + \cdots$$

Nel caso, come spesso accade, che $p \ll 1$, l'espressione indicata è bene approssimata dal solo primo termine, e dunque si può affermare che $P_e^T \simeq MP_e^{bit}$, ovvero che la prob. di errore sul bit è proporzionale al numero di tratte.

Esempio con un valore di $p = 10^{-6}$ occorrono 10 tratte per arrivare a $P_e^T = 10^{-5}$.

Ripetitore trasparente nelle trasmissioni numeriche In questo caso ogni ripetitore *trasparente* amplifica oltre al segnale anche il rumore già presente, e ne aggiunge di suo: come mostrato al § 18.3.1.1 nel caso di tratte identiche all'estremo di destinazione del collegamento si osserva un $SNR_T = \frac{SNR_i}{M}$, pari cioè a quello di ogni singola tratta,

diviso per il numero delle tratte. In tal caso l'E_b/N_0 del ricevitore a destinazione (l'unico ora che esegue il processo di decisione) risulta ridotto dello stesso fattore M (vedi eq. (15.16) pag. 459), producendo un peggioramento delle sue condizioni operative di $10 \log_{10} M$ dB rispetto a quelle delle singole tratte.

Esercizio Una trasmissione numerica 16-QAM (§ 16.3) è realizzata mediante un collegamento suddiviso in $M = 16$ tratte uguali. Qualora si desideri una P_e^T complessiva pari a 10^{-5}, valutare il valore di E_b/N_0 necessario in ingresso a ciascuna tratta, nei due casi **a)** ripetitori rigenerativi oppure **b)** ripetitori trasparenti. Indicare quindi l'entità del rapporto tra le potenze di segnale in ingresso ai ripetitori nei due casi.

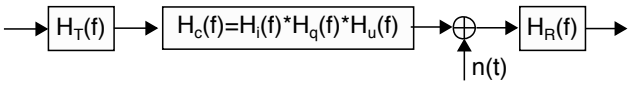
Svolgimento Nel caso **a)** risulta $P_e^T \simeq M P_e$, dunque per ogni tratta si deve ottenere una $P_e = P_e^T/M = 10^{-5}/16 = 6.25 \cdot 10^{-7}$, a cui (per un 16-QAM) corrisponde (fig. 16.14) un valore $E_b/N_0 \simeq 14.5$ dB. Nel caso **b)** la decisione ha luogo solamente nell'ultima tratta, e per ottenere una $P_e^T = 10^{-5}$, la modulazione 16-QAM necessita di un valore $E_b/N_0 \simeq 13.5$ dB. D'altra parte, per ottenere un tale E_b/N_0 in ingresso al decisore, *ogni* tratta che lo precede deve presentare un E_b/N_0 in ingresso allo stadio di amplificazione *maggiore* (rispetto a quello del decisore) di una quantità pari a $\log_2 M = 4$ dB, ossia $E_b/N_0 = 13.5 + 4 = 17.5$ dB, ovvero 3 dB *in più* rispetto al caso **a)**.

Essendo le tratte identiche per ognuna di esse il valore $N_0 = 1/2 k T_{ei}$ è lo stesso, in quanto (vedi eq. 18.12) $T_{ei} = T_g + T_0 (F - 1)$, e sia T_g che F sono gli stessi per tutte le tratte. Pertanto la differenza di 3 dB tra i valori di E_b/N_0 nei casi **a)** e **b)** si riflette in un *raddoppio* della potenza in ingresso ad ogni ripetitore trasparente, in confronto a quello rigenerativo.

18.4 Equalizzazione numerica

Occupiamoci ora di un argomento del tutto differente, ovvero come *correggere* una risposta in frequenza complessiva del mezzo trasmissivo e delle sue chiusure $H_c(f) = H_i(f) H_q(f) H_u(f)$ che presenta distorsione lineare (§ 8.2) e dunque non soddisfa la condizione di canale perfetto (pag. 231), specializzando il discorso al caso di una trasmissione dati[11] per la quale $H_c(f)$ può anche non essere nota a priori.

Andando con ordine, al § 15.3 abbiamo già osservato come l'equalizzazione può essere perseguita inserendo elementi filtranti sia presso il trasmettitore ($H_T(f)$) che al ricevitore ($H_R(f)$), realizzati in modo da ottenere



$$H_T(f) H_c(f) H_R(f) = a e^{-j2\pi f \tau}$$

annullando così l'effetto di $H_c(f)$. Scegliendo di utilizzare uno solo dei due filtri si può delegare tutto il compito dell'equalizzazione a $H_T(f)$, riservando ad $H_R(f)$ il ruolo di filtrare il rumore esterno alla banda di segnale, come fatto ai cap. 14 e 19.

[11] I casi di segnale di banda base e di segnale modulato sono trattati nel seguito in modo unitario, facendo riferimento solo ai primi, ovvero alle c.a. di b.f. nel caso di segnale modulato.

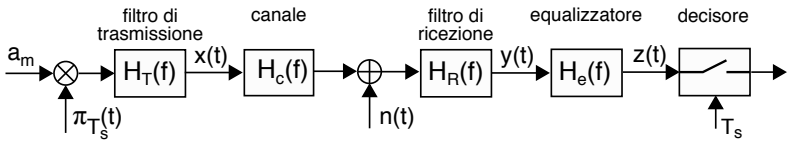


Figura 18.2: Equalizzazione di un canale con distorsione lineare

Quando la $H_c(f)$ del canale è *nota a priori* la sua inversione[12] può essere approssimata sintetizzando $H_T(f)$ mediante un metodo di progetto filtri analogici (§ 5.1).

D'altra parte se l'equalizzazione è svolta solamente *al lato ricevente* si possono applicare tecniche volte a *stimare* la $H_c(f)$ qualora non sia nota[13], e quindi *sintetizzare* una realizzazione numerica di $H_R(f)$ tale da invertire $H_c(f)$. In tal caso si determina la *colorazione* del rumore in ingresso al ricevitore (vedi pag. 625), che viene esaltato anche di molto nelle bande per le quali $|H_c(f)|$ è particolarmente piccolo, come ad es. per un canale con echi di ampiezza simile, vedi pag. 238.

Nel caso di una *trasmissione numerica* è già prevista la presenza di un filtro $H_T(f)$ con la funzione di sagomatore di impulso (come componente del codificatore di linea, § 15.1.2), mentre la teoria del filtro adattato (§ 7.6) stabilisce che realizzando[14] $H_R(f) = H_T^*(f)$ si ottiene il massimo SNR all'istante di decisione. Lo schema di elaborazione complessivo è dunque quello rappresentato in fig. 18.2, che adottiamo come riferimento, ed in cui $H_R(f)\,H_T^*(f) = G(f)$, in modo che *assieme* realizzino un filtro di Nyquist[15], e l'equalizzazione di $H_c(f)$ sia demandata per intero al ricevitore, per mezzo di $H_e(f)$. In ingresso ad $H_e(f)$ si presenta dunque il segnale

$$y(t) = \sum_{m=-\infty}^{\infty} a_m \tilde{g}(t - mT_s) + \nu(t) \tag{18.13}$$

che dipende oltre che dai simboli a_m della sequenza informativa, anche dalla risposta impulsiva complessiva $\tilde{g}(t) = h_T(t) * h_c(t) * h_R(t)$[16], mentre $\nu(t)$ tiene conto dell'effetto di $H_R(f)$ su $n(t)$. Scegliamo di realizzare $H_e(f)$ nella forma di un filtro FIR (§ 5.2.1, vedi fig. 18.3) con $2N+1$ coefficienti c_n, in modo che alla sua uscita si trovi il segnale[17]

$$z(t) = \sum_{n=-N}^{N} c_n y(t - n\tau) \tag{18.14}$$

[12] Detta anche operazione di *deconvoluzione*, vedi `https://en.wikipedia.org/wiki/Deconvolution`.

[13] Operazione necessaria se il canale *cambia* da un collegamento all'altro, come nel transito in una rete commutata, o nel caso di una comunicazione radiomobile.

[14] Si veda l'eq. (7.23) a pag. 213, tralasciando il termine di fase lineare necessario a rendere $h_R(t)$ causale.

[15] Sia che si abbia $H_R(f) = H_T(f) = \sqrt{G(f)}$ come per il caso del ricevitore ottimo (§ 15.5), sia nel caso in cui $H_T(f) = G(f)$, e $H_R(f)$ ha il solo scopo di filtrare il rumore. C'è poi l'aspetto trattato al § 15.5.1 relativo all'equalizzazione *distribuita*, che prevede di ripartire il compito in parti uguali tra i due lati, e che necessita di un canale *di ritorno* per comunicare al trasmettitore le stime operate a destinazione.

[16] Che equivale all'impulso $\tilde{g}(t)$ di eq. (15.1) a pag. 441

[17] La (18.14) 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 pari a $N\tau$, trascurato per semplicità di notazione nel seguito.

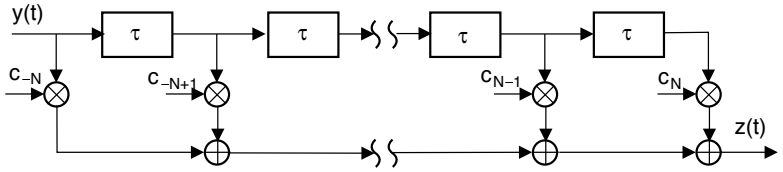


Figura 18.3: Filtro FIR di equalizzazione

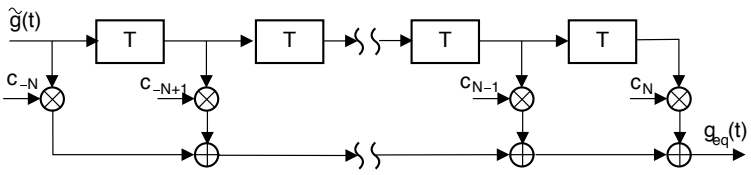
Osserviamo ora che se $H_e(f)$ riesce nel suo intento l'ISI introdotta dal canale è completamente rimossa e campionando l'uscita dell'equalizzatore (18.14) in corrispondenza degli istanti di simbolo si ottiene un valore $z(mT_s)$ che dipende da un solo a_m, oltre che dal rumore filtrato via $H_R(f)\,H_e(f)$; pertanto i coefficienti c_n possono essere scelti direttamente con l'obbiettivo di annullare l'ISI, anziché con quello di invertire $H_c(f)$.

Illustriamo per prima una tecnica che non tiene in particolare conto la presenza del rumore, che viene invece considerato al § 18.4.2 e seguenti.

18.4.1 Equalizzatore zero forcing

Questa tecnica prevede che gli impulsi $\tilde{g}(t)$ ricevuti distorti attraversino un filtro trasversale i cui coefficienti sono calcolati in modo da (quasi) ripristinare le condizioni di Nyquist (nel dominio del tempo) per l'impulso $g_{eq}(t)$ in uscita dal filtro.

Per poter funzionare, si fa precedere il messaggio informativo dalla trasmissione di un impulso isolato *di apprendimento*, la cui forma d'onda $\tilde{g}(t)$ ricevuta distorta presenta un picco a $t = 0$ ed ISI su entrambi i lati (Fig. 18.4a). Tale impulso $\tilde{g}(t)$ è posto in ingresso al filtro trasversale $h_{eq}(t)$ con $2N + 1$ coefficienti che ridisegnamo sotto con il ritardo τ tra gli ingressi pari a T, per un ritardo totale di attraversamento $2NT$. In uscita da $h_{eq}(t)$ si ottiene l'impulso



$$g_{eq}(t) = \sum_{n=-N}^{N} c_n \tilde{g}(t - nT - NT) \tag{18.15}$$

che campionato agli istanti $t_k = kT + NT$ fornisce

$$g_{eq}(t_k) = \sum_{n=-N}^{N} c_n \tilde{g}(kT - nT) = \sum_{n=-N}^{N} c_n \tilde{g}_{k-n} \tag{18.16}$$

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

$$g_{eq}(t_k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, \cdots, \pm N \end{cases} \tag{18.17}$$

assicurando così l'assenza di ISI *almeno* nei confronti degli N simboli precedenti

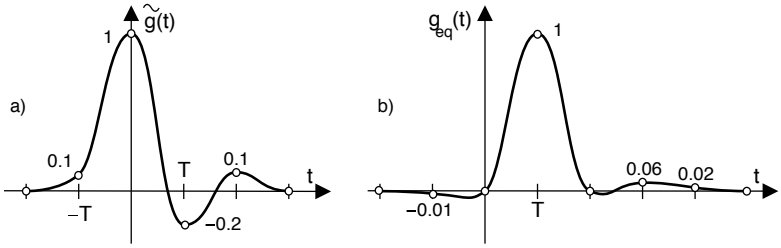


Figura 18.4: Impulso con ISI prima e dopo equalizzazione, caso di $N = 1$

e successivi. Il valore dei coefficienti c_n in grado di soddisfare queste condizioni si ottengono risolvendo un sistema di $2N+1$ equazioni nelle $2N+1$ incognite c_n, impostate a partire dalle relazioni (18.16) e (18.17), e rappresentato in forma matriciale come

$$\begin{bmatrix} \tilde{g}_0 & \cdots & \tilde{g}_{-2N} \\ \vdots & & \vdots \\ \tilde{g}_{N-1} & \cdots & \tilde{g}_{-N-1} \\ \tilde{g}_N & \cdots & \tilde{g}_{-N} \\ \tilde{g}_{N+1} & \cdots & \tilde{g}_{-N+1} \\ \vdots & & \vdots \\ \tilde{g}_{2N} & \cdots & \tilde{g}_0 \end{bmatrix} \begin{bmatrix} c_{-N} \\ \vdots \\ c_{-1} \\ c_0 \\ c_1 \\ \vdots \\ c_N \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{18.18}$$

in cui la matrice dei termini noti è costituita dai $4N + 1$ campioni di $\tilde{g}(t)$ prelevati agli istanti di simbolo $-2N, \ldots, 2N$ 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 se questa non è eccessiva tale peggioramento è più che compensato dal miglioramento dell'ISI.

Esempio Vogliamo realizzare un equalizzatore del terzo ordine, applicando quanto esposto all'impulso con ISI mostrato alla fig. 18.4a. Inserendo i valori di $\tilde{g}_k$ mostrati nella matrice dei coefficienti (18.18) 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 $c_{-1} = -0.096 \quad c_0 = 0.96 \quad c_1 = 0.2$. Inserendo ora 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. 18.4b assieme ad una curva interpolata. Notiamo che sebbene il risultato desiderato (di azzerare $g_{eq}(t)$ ai due istanti di simbolo ai lati del picco) sia stato ottenuto, risulta essere ancora $g_{eq} \neq 0$ ad istanti più remoti.

Effetto del rumore dopo equalizzazione Come anticipato l'approccio dello *zero forcing* trascura la presenza del rumore in ricezione. Poniamo che riesca nel suo intento di realizzare una $H_e(f) = 1/H_c(f)$ come desiderato: la densità di potenza del rumore in uscita sarà allora pari a

$$\mathcal{P}_n(f) = \frac{N_0}{2} |H_R(f)|^2 |H_e(f)|^2 = \frac{N_0}{2} \frac{|H_R(f)|^2}{|H_c(f)|^2}$$

che dunque subisce una maggiore amplificazione proprio alle frequenze per le quali $|H_c(f)|^2$ è particolarmente ridotta. Illustriamo dunque nel seguito alcune tecniche che invece tengono conto della presenza del rumore.

18.4.2 Equalizzatore MMSE e filtro di Wiener

Con questa tecnica si sceglie di realizzare il filtro FIR $H_e(f)$ in forma totalmente numerica (§ 4.6.1) e di campionare $y(t)$ con periodo $\tau = T_s/2$[18], ovvero facendo lavorare il filtro alla velocità di due campioni per simbolo. I coefficienti $\mathbf{c}$ del filtro FIR $H_e(f)$ di fig. 18.3 sono individuati come

$$\tilde{\mathbf{c}} = \arg\min \left\{ \sigma_e^2(\mathbf{c}) \right\}$$

ossia tali da rendere *minimo*[19] l'*errore quadratico medio* $\sigma_e^2 = E\left\{e_m^2\right\}$, in cui

$$e_m = z(mT_s) - a_m \tag{18.19}$$

è l'errore tra i campioni $z(mT_s)$ della (18.14) in uscita da $H_e(f)$, ed i corrispondenti valori a_m della sequenza informativa[20] che è stata trasmessa. Tale criterio è noto come MMSE (MINIMUM MEAN SQUARE ERROR o *minimo errore quadratico medio*), e determina un risultato noto anche come *filtro di* WIENER[21]. In presenza di processi ergodici a media nulla l'*errore quadratico medio* σ_e^2 corrisponde alla *potenza* dell'errore, che (applicando la (18.14) alla (18.19)) può essere espressa come

$$\sigma_e^2 = E\left\{(z(mT_s) - a_m)^2\right\} = E\left\{\left(\sum_{n=-N}^{N} c_n y(mT_s - n\tau) - a_m\right)^2\right\} \tag{18.20}$$

evidenziandone la dipendenza dai coefficienti c_n. La minimizzazione di σ_e^2 si ottiene calcolando le sue derivate rispetto ai coefficienti c_n, ed eguagliando l'espressione delle stesse a zero[22], ovvero

$$\begin{aligned} \frac{\partial}{\partial c_k}\sigma_e^2 &= 2E\left\{\left[\sum_{n=-N}^{N} c_n y(mT_s - n\tau) - a_m\right] y(mT_s - k\tau)\right\} = \\ &= 2\left(\sum_{n=-N}^{N} c_n E\left\{y(mT_s - n\tau)\, y(mT_s - k\tau)\right\} - E\left\{a_m y(mT_s - k\tau)\right\}\right) = 0 \end{aligned} \tag{18.21}$$

nella quale riconosciamo che i *valori attesi* che vi compaiono corrispondono rispettiva-

[18] Dato 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 $f_c = 2f_s$, ovvero $\tau = T_s/2$. 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.

[19] Piuttosto che *l'azzeramento* dell'errore, come nell'approccio *zero forcing*.

[20] Tale errore dipende sia dal rumore presente in ingresso che dall'ISI introdotta dal canale, le cui potenze sono quindi minimizzate in forma congiunta.

[21] Vedi `http://en.wikipedia.org/wiki/Wiener_filter`.

[22] In realtà la minimizzazione ha successo solamente se σ_e^2 è una funzione *convessa* (vedi `https://it.wikipedia.org/wiki/Funzione_convessa`) dei c_n, ma più avanti (pag. 629) troveremo che questo è proprio il nostro caso.

mente all'*autocorrelazione* (§ 7.1.4)

$$\mathcal{R}_Y(n-k) = E\{y(mT_s - n\tau)\, y(mT_s - k\tau)\} \tag{18.22}$$

tra coppie di campioni del segnale ricevuto distanti $|n-k|\,\tau$, ed all'*intercorrelazione*

$$\mathcal{R}_{YA}(k) = E\{y(mT_s - k\tau)\, a_m\} \tag{18.23}$$

tra i campioni di $y(t)$ distanti $|k|\,\tau$ dal generico simbolo a_m ed il valore del simbolo stesso, dove il valore atteso è calcolato rispetto alla variabilità della sequenza a_m e del rumore. Dato che entrambi gli indici n e k variano tra $-N$ e N, le eq. (18.21) si riscrivono come

$$\sum_{n=-N}^{N} c_n \mathcal{R}_Y(n-k) = \mathcal{R}_{YA}(k) \qquad k = -N, -N+1, \cdots, 0, \cdots, N-1, N \tag{18.24}$$

note come equazioni di *Wiener-Hopf*, ovvero un sistema di $2N+1$ equazioni nelle altrettante incognite c_n, la cui matrice dei coefficienti e vettore dei termini noti sono costituiti rispettivamente dalle (18.22) e (18.23).

Soluzione delle equazioni di Wiener-Hopf Adottiamo innanzitutto per la (18.24) una notazione matriciale del tipo

$$\mathbf{Bc} = \mathbf{d} \tag{18.25}$$

dove $\mathbf{B} = [b_{nk}] = [\mathcal{R}_Y(n-k)]$ è una matrice $(2N+1)\times(2N+1)$, $\mathbf{c} = [c_n]$ è un vettore colonna i cui elementi sono i coefficienti del filtro, e $\mathbf{d}$ è il vettore dei termini noti $\mathbf{d} = [d_k] = [\mathcal{R}_{YA}(k)]$[23]. Con tale notazione i coefficienti *ottimi* $\tilde{\mathbf{c}}$ si ottengono come

$$\tilde{\mathbf{c}} = \mathbf{B}^{-1}\mathbf{d}$$

Osserviamo che la *parità* di $\mathcal{R}_Y(n)$ rispetto ad n (vedi schema a lato) rende $\mathbf{B}$ una matrice di *Toepliz*[24], per la quale esistono metodi che semplificano in parte la soluzione diretta del sistema di equazioni, vedi anche la nota 14 a pag. 297.

$$\mathbf{B} = \begin{bmatrix} \mathcal{R}_Y(0) & \mathcal{R}_Y(1) & \mathcal{R}_Y(2) & \cdots & \mathcal{R}_Y(2N) \\ \mathcal{R}_Y(1) & \mathcal{R}_Y(0) & \mathcal{R}_Y(1) & \cdots & \mathcal{R}_Y(2N-1) \\ \mathcal{R}_Y(2) & \mathcal{R}_Y(1) & \mathcal{R}_Y(0) & \cdots & \mathcal{R}_Y(2N-2) \\ \mathcal{R}_Y(3) & \mathcal{R}_Y(2) & \mathcal{R}_Y(1) & \cdots & \mathcal{R}_Y(2N-3) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathcal{R}_Y(2N) & \mathcal{R}_Y(2N-1) & \cdots & \cdots & \mathcal{R}_Y(0) \end{bmatrix}$$

Risposta in frequenza del filtro di Wiener Affrontiamo una prima digressione per derivare l'espressione assunta da $H_e(f)$ nel caso ottimo, ovvero quando $\mathbf{c} = \tilde{\mathbf{c}}$ e σ_e^2 è minimo. Conviene procedere direttamente nel dominio delle sequenze, ed osservare che la risposta impulsiva campionata (con periodo τ) del filtro FIR $H_e(f)$ di

[23] 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, possono essere *stimate* a partire da una fase iniziale *di apprendimento* durante la quale viene trasmesso un segnale di test associato ad una sequenza $\{a_k\}$ nota al ricevitore, in base alla quale quest'ultimo calcola le medie temporali

$$\hat{\mathcal{R}}_Y(n) = \frac{1}{K}\sum_{k=1}^{K} y(kT_s - n\tau)\, y(kT_s) \qquad \text{e} \qquad \hat{\mathcal{R}}_{YA}(n) = \frac{1}{K}\sum_{k=1}^{K} y(kT_s - n\tau)\, a_k$$

che quindi utilizza al posto delle medie di insieme nella (18.24).

[24] `http://en.wikipedia.org/wiki/Toeplitz_matrix`

fig. 18.3 è esattamente pari ai coefficienti $\mathbf{c}$, ovvero $h_e(n) = c_{n-N}$ con $n = 0, 1, 2 \cdots, 2N$. Pertanto la (18.24) può essere letta come l'espressione della sequenza $\mathcal{R}_{YA}(k)$ (per $k = 0, \pm 1, \pm 2, \cdots, \pm N$) nei termini di una *convoluzione discreta* tra $h_e(n)$ ed $\mathcal{R}_Y(n)$ anticipata di N:

$$\mathcal{R}_{YA}(k) = \sum_{n=0}^{2N} h_e(n)\, \mathcal{R}_Y(k - n + N) \tag{18.26}$$

Effettuando la DTFT (§ 4.4) di entrambi i membri di (18.26) si ottiene[25] $\mathcal{P}_{YA}(f) = H_e(f)\, \mathcal{P}_Y(f)\, e^{j2\pi f N\tau}$ in cui $\mathcal{P}_Y(f)$ è la densità di potenza della sequenza $y(n)$ in ingresso ad $h_e(t)$, e $\mathcal{P}_{YA}(f)$ è la densità di potenza *mutua*[26] tra $y(n)$ e la sequenza informativa $a(n)$. Otteniamo pertanto

$$H_e(f) = \frac{\mathcal{P}_{YA}(f)}{\mathcal{P}_Y(f)} \mathrm{e}^{-j2\pi f N\tau} \tag{18.27}$$

dove il termine $\mathrm{e}^{-j2\pi f N\tau}$ esprime il ritardo necessario a realizzare un filtro causale, e viene trascurato nel seguito. Le densità di potenza $\mathcal{P}_Y(f)$ e $\mathcal{P}_{YA}(f)$ possono essere valutate eseguendo la DTFT delle corrispondenti correlazioni che figurano nella (18.26), che (per simboli a_m *incorrelati*) risultano[27]

$$\mathcal{R}_Y(n) = \sigma_a^2 \mathcal{R}_H(n) + \mathcal{R}_\nu(n) \quad \mathrm{e}^{28} \quad \mathcal{R}_{YA}(n) = \sigma_a^2 h(-n)$$

in cui $\mathcal{R}_H(n)$ è l'autocorrelazione dei campioni (con periodo τ) delle risposta impulsiva complessiva $h(t) = h_T(t) * h_c(t) * h_R(t)$, $\mathcal{R}_\nu(n)$ è la correlazione tra campioni di $\nu(t)$ a distanza $n\tau$, e $\sigma_a^2 = E\left\{a_k^2\right\}$, sottintendendo gli a_k a media nulla. La (18.27) si riscrive quindi come

$$H_e(f) = \frac{\sigma_a^2 H^*(f)}{\sigma_a^2 |H(f)|^2 + \mathcal{P}_\nu(f)} \tag{18.28}$$

[25] Viene applicata l'equivalenza tra convoluzione nel tempo e prodotto in frequenza, nonché quella tra trasformata della correlazione e densità di potenza.

[26] Analogamente all'energia mutua (§ 3.2), $\mathcal{P}_{YA}(f)$ esprime una *similitudine* tra due segnali e/o processi, frequenza per frequenza. Laddove risulti $\mathcal{P}_{YA}(f) = 0$, i segnali sono *ortogonali* in tale regione di frequenze.

[27] Ricordando che $y(t) = \sum_k a_k h(t - kT_s) + \nu(t)$ in cui $h(t)$ è la risposta impulsiva complessiva $h(t) = h_T(t) * h_c(t) * h_R(t)$ (vedi fig. 18.2) e $\nu(t)$ rappresenta l'effetto di $H_R(f)$ su $n(t)$, dalla (18.22) scriviamo $\mathcal{R}_Y(n) = E\{y(mT_s)\, y(mT_s + n\tau)\}$ e quindi esprimendo $h(mT_s - kT_s)$ come $h((m-k)T_s)$ si ottiene

$$\begin{aligned} \mathcal{R}_Y(n) = \quad & E\{(\textstyle\sum_k a_k h((m-k)T_s) + \nu(mT_s))\,(\sum_i a_h h((m-i)T_s + n\tau) + \nu(mT_s + n\tau))\} = \\ & E\{\textstyle\sum_k \sum_i a_k a_i h((m-k)T_s)\, h((m-i)T_s + n\tau) + \nu(mT_s)\, \nu(mT_s + n\tau) + \\ & + \nu(mT_s + n\tau) \textstyle\sum_k a_k h((m-k)T_s) + \nu(mT_s) \sum_h a_h h((m-i)T_s + n\tau)\} \end{aligned}$$

i cui termini dell'ultima riga risultano nulli se i simboli a_m sono statisticamente indipendenti dai campioni di rumore, ed almeno uno dei due processi è a media nulla. Essendo inoltre $E\{a_k a_i\} = 0$ per $i \neq k$, il termine con la doppia sommatoria si riduce a

$$E\left\{a_k^2\right\} \textstyle\sum_k h((m-k)T_s)\, h((m-k)T_s + n\tau) = \mathcal{R}_A(0)\, \mathcal{R}_H(n) = \sigma_a^2 \mathcal{R}_H(n)$$

avendo considerato gli a_k a media nulla. Infine, il termine $E\{\nu(mT_s)\, \nu(mT_s + n\tau)\}$ risulta pari alla correlazione $\mathcal{R}_{\nu\nu}(n)$ del processo di rumore.

[28] Dalla (18.23) scriviamo $\mathcal{R}_{YA}(n) = E\{y(mT_s - n\tau)\, a_m\}$ e quindi

$$\begin{aligned} \mathcal{R}_{YA}(n) = \quad & E\{(\textstyle\sum_k a_k h((m-k)T_s - n\tau) + \nu(mT_s - n\tau))\, a_m\} = \\ = \quad & \textstyle\sum_k E\{a_k a_m\}\, h((m-k)T_s - n\tau) + E\{\nu(mT_s - n\tau)\, a_m\} \end{aligned}$$

e come prima troviamo $E\{\nu(mT_s - n\tau)\, a_m\} = 0$, mentre della sommatoria rimane il solo termine $k = m$.

Possiamo osservare che qualora $\mathcal{P}_\nu(f) = 0$ si ottiene $H_e(f) = \frac{1}{H(f)}$, che pertanto assolve in pieno il compito di equalizzazione, come avviene per l'approccio *zero forcing* (§ 18.4.1). D'altra parte, dividendo numeratore e denominatore di $H_e(f)$ per $\sigma_a^2 |H(f)|^2$ si ottiene

$$H_e(f) = \frac{1}{H(f)} \frac{1}{1 + \frac{1}{SNR(f)}}$$

che evidenzia come l'effetto del rumore sia quello *attenuare* la risposta in frequenza $H_e(f)$ nelle regioni in cui $SNR(f)$ (non in dB) è più piccolo, ossia dove c'è meno segnale, e/o più rumore. Infine, nel caso di rumore elevato la (18.28) diviene $H_e(f) = H^*(f) \frac{\sigma_a^2}{\mathcal{P}_\nu(f)}$ (o, in caso di rumore *bianco*, $H_e(f) = H^*(f) \frac{\sigma_a^2}{\sigma_\upsilon^2}$) come per un *filtro adattato* (§ 7.6); essendo il termine $H^*(f)$ comunque presente, si può scegliere di realizzare $H_R(f)$ (vedi fig. 18.2) con il solo scopo di limitare la potenza di rumore[29].

Qualora i simboli informativi a_m siano tra loro *correlati* la (18.28) diviene[30]

$$H_e(f) = \frac{\mathcal{P}_A(f) H^*(f)}{\mathcal{P}_A(f) |H(f)|^2 + \mathcal{P}_\nu(f)}$$

in cui $\mathcal{P}_A(f)$ è la DTFT della correlazione dei simboli $\mathcal{R}_A(n) = E\{a_m a_{m+n}\}$ o *spettro del codice* (pag. 226). La stessa espressione è valida per un filtro di Wiener di tipo più generale, che esegue ad esempio la *deriverberazione* di un segnale audio con densità di potenza $\mathcal{P}_A(f)$.

Analisi della potenza di errore Sviluppiamo ora una seconda digressione sul significato *geometrico* della minimizzazione di σ_e^2, che può essere considerato come una funzione dei coefficienti $\mathbf{c}$, e di cui si cerca il minimo (ovvero $\tilde{\mathbf{c}} = \arg\min\left\{\sigma_e^2(\mathbf{c})\right\}$), che mostriamo essere unico. Infatti, l'espressione di $\sigma_e^2(\mathbf{c}) = E\left\{e_m^2\right\}$ si ottiene dalla (18.20) come

$$\begin{aligned}\sigma_e^2(\mathbf{c}) &= E\left\{\left(\sum_{n=-N}^{N} c_n y(mT_s - n\tau) - a_m\right)\left(\sum_{k=-N}^{N} c_k y(mT_s - k\tau) - a_m\right)\right\} = \\ &= \sum_{n=-N}^{N}\sum_{k=-N}^{N} c_n c_k \mathcal{R}_Y(n-k) - 2\sum_{k=-N}^{N} c_k \mathcal{R}_{YA}(k) + E\left\{a_m^2\right\}\end{aligned} \tag{18.29}$$

[29] In particolare, scegliendo $H_R(f) = rect_{2f_s}(f)$ si ottiene che i campioni di rumore presi con intervallo $\tau = 1/2T_s$ sono incorrelati e dunque statisticamente indipendenti perché gaussiani. Infatti $\mathcal{P}_\nu(f) = \frac{N_0}{2}|H_R(f)|^2$ e dunque $\mathcal{R}_\nu(\tau) = \frac{N_0}{2} 2f_s sinc(2f_s\tau)$, che quindi si azzera per $\tau = 1/2T_s$, vedi § 7.2.4. In tal caso la (18.28) diviene $H_e(f) = \frac{\sigma_a^2 H^*(f)}{\sigma_a^2 |H(f)|^2 + \sigma_\upsilon^2}$.

[30] Per quanto il risultato sembri banale, la dimostrazione non è troppo diretta, e si basa sullo scomporre il termine

$$E\left\{\sum_k \sum_i a_k a_i h((m-k)T_s) h((m-i)T_s + n\tau)\right\} = \sum_k \sum_i \mathcal{R}_A(i-k) h((m-k)T_s) h((m-i)T_s + n\tau)$$

tenendo conto della stazionarietà di a_m e dunque della simmetria di $\mathcal{R}_A(n)$, in una somma di termini

$$\sum_p \mathcal{R}_A(p) \sum_i h((m-i)T_s) h((m-i-p)T_s + n\tau) = \sum_p \mathcal{R}_A(p) \mathcal{R}_H(n+2p)$$

in cui si tiene conto che $T_s = 2\tau$, mentre gli a_m sono prodotti uno per simbolo; introducendo $\mathcal{R}'_A(n) = \begin{cases}\mathcal{R}_A\left(\frac{n}{2}\right) & n\ pari \\ 0 & altrimenti\end{cases}$ e con qualche cambio di variabile si ottiene infine $\sum_q \mathcal{R}'_A(q) \mathcal{R}_H(n-q)$ che ha l'aspetto rassicurante della convoluzione, da cui scaturisce il risultato.

che è un *polinomio completo* di secondo grado nelle $2N+1$ variabili c_n, ed essendo $[\mathcal{R}_Y(n-k)]$ una matrice *definita positiva*, $\sigma_e^2(\mathbf{c})$ presenta *un unico* minimo globale[31], e la sua espressione può essere riscritta in forma matriciale come[32]

$$\sigma_e^2(\mathbf{c}) = \mathbf{c}^t\mathbf{B}\mathbf{c} - 2\mathbf{c}^t\mathbf{d} + \sigma_a^2 \tag{18.30}$$

In corrispondenza della soluzione $\tilde{\mathbf{c}} = \mathbf{B}^{-1}\mathbf{d}$ della (18.25) si ottiene la *minima* varianza di errore, pari a[33]

$$\sigma_{e_{min}}^2 = \sigma_e^2(\tilde{\mathbf{c}}) = \sigma_a^2 - \mathbf{d}^t\mathbf{B}^{-1}\mathbf{d} = \sigma_a^2 - \tilde{\mathbf{c}}^t\mathbf{B}\tilde{\mathbf{c}} = \sigma_a^2 - \sigma_z^2 \tag{18.31}$$

avendo notato che il termine $\sigma_z^2 = \tilde{\mathbf{c}}^t\mathbf{B}\tilde{\mathbf{c}}$ corrisponde[34] alla potenza dell'uscita $z = z_m$, ovvero $\sigma_z^2 = E\{z^2\}$. Accade ora che la (18.30) può essere espressa anche come[35] $\sigma_e^2(\mathbf{c}) = (\mathbf{B}\mathbf{c}-\mathbf{d})^t\mathbf{B}^{-1}(\mathbf{B}\mathbf{c}-\mathbf{d}) - \mathbf{d}^t\mathbf{B}^{-1}\mathbf{d} + \sigma_a^2$, e tenendo conto della prima delle (18.31) si ottiene

$$\sigma_e^2(\mathbf{c}) = \sigma_{e_{min}}^2 + (\mathbf{B}\mathbf{c}-\mathbf{d})^t\mathbf{B}^{-1}(\mathbf{B}\mathbf{c}-\mathbf{d}) \tag{18.32}$$

evidentemente pari a $\sigma_{e_{min}}^2$ qualora $\mathbf{c} = \tilde{\mathbf{c}}$ e dunque $\mathbf{B}\tilde{\mathbf{c}} = \mathbf{d}$. Essendo $\mathbf{B}$ una matrice definita positiva lo è anche la sua inversa $\mathbf{B}^{-1}$, ed esiste (vedi § 6.7.3) una matrice unitaria $\mathbf{\Gamma}$ con colonne pari agli autovettori di $\mathbf{B}^{-1}$ tale da poter scrivere $\mathbf{B}^{-1} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^t$, dove $\mathbf{\Lambda}$ è una matrice diagonale con elementi pari all'inverso degli autovalori λ_i di $\mathbf{B}$; indicando allora con $\mathbf{c}_\delta$ il vettore $\mathbf{B}\mathbf{c}-\mathbf{d} = \mathbf{B}(\mathbf{c}-\tilde{\mathbf{c}})$ legato alla *differenza* tra i coefficienti generici e quelli ottimi, e con $\mathbf{u} = \mathbf{\Gamma}^t\mathbf{c}_\delta$ la sua versione trasformata nella base individuata dagli autovettori, la (18.32) diviene

$$\sigma_e^2(\mathbf{u}) = \sigma_{e_{min}}^2 + \mathbf{u}^t\mathbf{\Lambda}\mathbf{u} = \sigma_{e_{min}}^2 + \sum_{i=0}^{2N}\frac{u_i^2}{\lambda_i}$$

che rappresenta l'equazione di un iperparaboloide ellittico, con assi principali pari agli autovettori di $\mathbf{B}^{-1}$.

---

[31] Ciò deriva dal fatto che $[\mathcal{R}_Y(n-k)] = \mathbf{B}$ è legata alla *matrice di covarianza* $\mathbf{\Sigma}_Y$ (§ 6.7.3) dalla relazione $[\mathcal{R}_Y(n-k)] = \mathbf{\Sigma}_Y + m_Y^2$ (vedi eq. (7.3) a pag. 192 in condizioni stazionarie), e dato che $\mathbf{\Sigma}_Y$ è una *matrice definita positiva* lo è anche $[\mathcal{R}_Y(n-k)]$, non decadendo la proprietà in seguito alla somma per una quantità positiva (m_Y^2). Pertanto la *forma quadratica* $\sum_{n=-N}^{N}\sum_{k=-N}^{N} c_n c_k \mathcal{R}_Y(n-k)$ è ovunque convessa (vedi § 6.7.3), e dotata di un unico minimo globale. Questa proprietà permane anche considerando gli altri due termini della (18.29) che concorrono al valore di $\sigma_e^2(\mathbf{c})$, essendo questi rispettivamente lineari in $\mathbf{c}$ e costanti, e che dunque non ne modificano la convessità.

[32] Considerando la sequenza informativa a_m a media nulla, si ha $E\{a_m^2\} = \sigma_a^2$.

[33]
$$\begin{aligned}\sigma_e^2(\tilde{\mathbf{c}}) &= \tilde{\mathbf{c}}^t\mathbf{B}\tilde{\mathbf{c}} - 2\tilde{\mathbf{c}}^t\mathbf{d} + \sigma_a^2 = \tilde{\mathbf{c}}^t\mathbf{B}\tilde{\mathbf{c}} - 2\tilde{\mathbf{c}}^t\mathbf{B}\tilde{\mathbf{c}} + \sigma_a^2 = \\ &= \sigma_a^2 - \tilde{\mathbf{c}}^t\mathbf{B}\tilde{\mathbf{c}} = \sigma_a^2 - \tilde{\mathbf{c}}^t\mathbf{d} = \sigma_a^2 - \mathbf{d}^t\mathbf{B}^{-1}\mathbf{d}\end{aligned}$$

avendo prima sostituito $\mathbf{d} = \mathbf{B}\tilde{\mathbf{c}}$ e poi il contrario, e valutato $\tilde{\mathbf{c}}^t = (\mathbf{B}^{-1}\mathbf{d})^t = \mathbf{d}^t(\mathbf{B}^{-1})^t = \mathbf{d}^t\mathbf{B}^{-1}$ in quanto $\mathbf{B}$ è una matrice simmetrica, così come la sua inversa.

[34] Considerando z a media nulla, si ha $\sigma_z^2 = E\{z^2\} = E\{\tilde{\mathbf{c}}^t\mathbf{y}\mathbf{y}^t\tilde{\mathbf{c}}\} = \tilde{\mathbf{c}}^t E\{\mathbf{y}\mathbf{y}^t\}\tilde{\mathbf{c}} = \tilde{\mathbf{c}}^t\mathbf{B}\tilde{\mathbf{c}}$.

[35] Infatti

$$\begin{aligned}(\mathbf{B}\mathbf{c}-\mathbf{d})^t\mathbf{B}^{-1}(\mathbf{B}\mathbf{c}-\mathbf{d}) &= \mathbf{c}^t\mathbf{B}^t\mathbf{B}^{-1}\mathbf{B}\mathbf{c} - \mathbf{c}^t\mathbf{B}^t\mathbf{B}^{-1}\mathbf{d} - \mathbf{d}^t\mathbf{B}^{-1}\mathbf{B}\mathbf{c} + \mathbf{d}^t\mathbf{B}^{-1}\mathbf{d} = \\ &= \mathbf{c}^t\mathbf{B}\mathbf{c} - \mathbf{c}^t\mathbf{d} - \mathbf{d}^t\mathbf{c} + \mathbf{d}^t\mathbf{B}^{-1}\mathbf{d} = \mathbf{c}^t\mathbf{B}\mathbf{c} - 2\mathbf{c}^t\mathbf{d} + \mathbf{d}^t\mathbf{B}^{-1}\mathbf{d}\end{aligned}$$

18.4.3 Metodo del gradiente

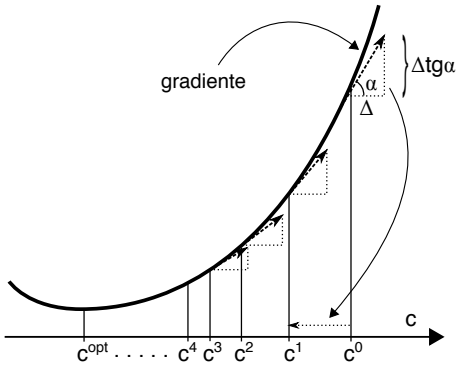
Benché si sia pervenuti alla (18.32) che esprime $\sigma_e^2(\mathbf{c})$ in funzione del vettore di coefficienti $\mathbf{c}$ di $H_e(f)$, la sua minimizzazione *diretta* comporta la soluzione del sistema (18.25). Ciò non è necessario adottando un metodo *iterativo* noto come *discesa del gradiente*[36] che si basa su di un approccio per così dire *altimetrico* e che considera $\sigma_e^2(\mathbf{c})$ come una (iper)superficie in uno spazio $2N+1$ dimensionale, e partendo da un vettore iniziale $\mathbf{c}^0$ di coefficienti scelti casualmente e dall'associata *quota* $\sigma_e^2(\mathbf{c}^0)$, ad ogni iterazione calcola nuovi coefficienti $\mathbf{c}^i$ tali che $\sigma_e^2(\mathbf{c}^{i+1}) \leq \sigma_e^2(\mathbf{c}^i)$, di fatto *scendendo* lungo il crinale[37] della superficie $\sigma_e^2(\mathbf{c})$.

A questo scopo ci si basa sulla conoscenza del *gradiente* di $\sigma_e^2(\mathbf{c})$ che indichiamo come $\mathbf{g}(\mathbf{c})$, cioè (pag. 9.7.1) del vettore le cui componenti g_n sono le derivate parziali di σ_e^2 rispetto ai coefficienti che compongono $\mathbf{c}$, ovvero $g_n = \frac{\partial \sigma_e^2(\mathbf{c})}{\partial c_n}$: esse sono tanto maggiori quanto più è ripida la pendenza di σ_e^2 nella direzione di c_n, e si azzerano tutte nel punto di minimo $\tilde{\mathbf{c}}$. Pertanto *l'orientamento* del vettore $\mathbf{g}(\mathring{\mathbf{c}})$ per ogni vettore $\mathring{\mathbf{c}} \neq \tilde{\mathbf{c}}$ indica *la direzione* verso la quale $\sigma_e^2(\mathring{\mathbf{c}})$ cresce di più.

Essendo anche il gradiente una funzione di $\mathbf{c}$ il suo valore può essere calcolato a partire dalla (18.30) ottenendo $\mathbf{g}(\mathbf{c}) = 2\mathbf{B}\mathbf{c} - 2\mathbf{d}$, che infatti si azzera per $\tilde{\mathbf{c}} = \mathbf{B}^{-1}\mathbf{d}$. Scegliendo casualmente un vettore iniziale di coefficienti $\mathbf{c}^0$, ad ogni iterazione i nuovi valori sono calcolati come

$$\mathbf{c}^{i+1} = \mathbf{c}^i - \Delta \mathbf{g}(\mathbf{c}^i) \tag{18.33}$$

ovvero *spostando* ogni coefficiente c_n in direzione *opposta* alla relativa componente del gradiente $\mathbf{g}^i = \mathbf{g}(\mathbf{c}^i)$ calcolato in quel punto, di una quantità ad esso proporzionata dal parametro Δ, quest'ultimo pari ad piccolo numero positivo che determina la velocità di convergenza dell'algoritmo[38]. Viene quindi calcolato un nuovo valore $\mathbf{g}^{i+1} = \mathbf{g}(\mathbf{c}^{i+1})$, ed il procedimento iterato fino a quando non si verifica che $|\mathbf{g}^i| < \varepsilon$, nel qual caso si ritiene raggiunto l'ottimo $\tilde{\mathbf{c}}$. La figura precedente rappresenta un esempio unidimensionale di applicazione del metodo, in cui il gradiente è pari alla derivata dy/dx, ovvero alla tangente dell'angolo α tra la funzione e l'asse orizzontale, e mostra i diversi valori c^i ottenuti nelle iterazioni.



18.4.3.1 Equalizzazione adattiva Least Mean Square (LMS)

Il metodo ora esposto si basa sul calcolo ripetuto di $\mathbf{g}^i = 2(\mathbf{B}\mathbf{c}^i - \mathbf{d})$, ma... se la $H(f)$ del canale varia nel tempo, lo stesso avviene per $\mathbf{B}$ e per $\mathbf{d}$, così come dovrebbe avvenire

[36] Vedi `http://it.wikipedia.org/wiki/Discesa_del_gradiente`

[37] Coma abbiamo discusso, il minimo è unico, dunque raggiungibile da dovunque si parta.

[38] 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 dare anche luogo ad instabilità della soluzione, mentre valori più piccoli rallentano la convergenza, ma possono produrre errori finali minori.

per il valore dei coefficienti $\tilde{\mathbf{c}}$! Per fronteggiare la contingenza ed allo stesso tempo evitare di stimare $\mathcal{R}_Y(n)$ e $\mathcal{R}_{YA}(n)$ si può adottare un diverso metodo di discesa del gradiente, detto ora *stocastico* (o LMS[39]) perché opera a partire da una *stima* del gradiente, aggiornata ad ogni periodo di simbolo, così come ad ogni simbolo sono aggiornati i coefficienti $\mathbf{c}$, utilizzati *allo stesso tempo* per eseguire l'equalizzazione.

Tutto inizia considerando che in base alla (18.21) il gradiente può essere valutato (a meno di un fattore due) per i valori $\mathbf{c}_m$ utilizzati all'istante $t = mT_s$ anche come[40]

$$\mathbf{g}_m = E\{e_m \mathbf{y}_m\} \tag{18.34}$$

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 ricezione[41]) ed $\mathbf{y}_m$ è il vettore dei $2N+1$ campioni $y_m(n)$ memorizzati *nello stato* del filtro $h_e(t)$ al momento di produrre l'uscita $z(mT_s)$. La (18.34) viene poi *approssimata* tralasciando di effettuare la media di insieme, ossia valutando la stima del gradiente come[42] $\hat{\mathbf{g}}_m = e_m \mathbf{y}_m$, da usare poi nella formula di aggiornamento dei coefficienti c_n derivata dalla (18.33):

$$\mathbf{c}_{m+1} = \mathbf{c}_m - \Delta e_m \mathbf{y}_m \tag{18.35}$$

La (18.35) è improntata ad un principio di *correzione dai propri errori*, ovvero i coefficienti ricevono *un colpetto* in su o in giù, in modo da ridurre l'errore commesso[43]. Il

[39] Vedi anche `http://en.wikipedia.org/wiki/Least_mean_squares_filter`

[40] Un interessante modo di interpretare la (18.34) si basa sull'osservare che la minimizzazione di (18.20) ovvero l'azzeramento del gradiente (18.21) $\mathbf{g}_m = \mathbf{0}$ comporta che ciascun valore di errore e_m deve essere *ortogonale* (in senso statistico, vedi § 7.7.2) a tutte le osservazioni $y_{m-N}, \cdots, y_{m+N}$ da cui dipende la stima $z(mT)$, ossia le $2N+1$ coppie di sequenze devono essere *incorrelate*, e quindi prive di legami di tipo lineare.

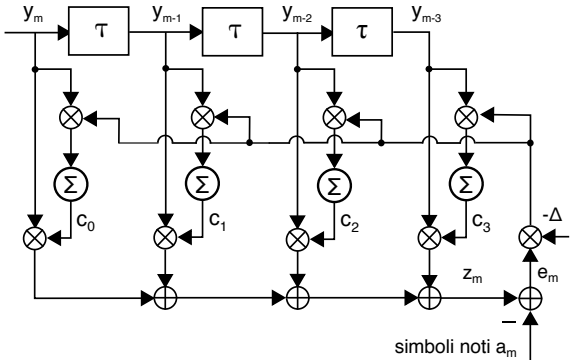
[41] La sequenza dei simboli trasmessi *è nota* all'inizio della trasmissione, e spesso generata nella forma di una sequenza *pseudo-noise*. Durante la fase *di apprendimento* l'algoritmo viene lasciato iterare arrivando alla stima dei coefficienti $\mathbf{c}$; terminata questa fase ha inizio la trasmissione vera e propria, e l'equalizzatore commuta su di una modalità *dipendente dalle decisioni*, in cui l'errore e_m è valutato rispetto ai valori $\widetilde{a}_m$ 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.

[42] Poniamo di eseguire l'aggiornamento (18.35) dei coefficienti $\mathbf{c}$ ogni K periodi di simbolo: otterremmo $\mathbf{c}_{m+K} = \mathbf{c}_m - \Delta \sum_{k=0}^{K-1} e_{m+k} \mathbf{y}_{m+k}$, in cui il termine sommatoria risulta in effetti proporzionale alla stima di $E\{e_m \mathbf{y}_m\}$, almeno più di quanto non appaia il termine $e_m \mathbf{y}_m$ che compare isolato nella (18.35).

[43] Possiamo infatti notare che alla m–esima iterazione della (18.35) il coefficiente c_n viene *aumentato* se e_m e y_n sono di segno opposto, e *diminuito* se concorde. Ovvero: se l'uscita è *piccola* allora $e_m = z(mT_s) - a_m < 0$, e il contributo c_n dell'n-esima memoria viene aumentato se $y_m(n)$ è positivo, o diminuito se $y_m(n)$ è negativo; viceversa, se $e_m = z(mT_s) - a_m > 0$ (uscita *troppo grande*) il valore di c_n viene aumentato se $y_m(n)$ è negativo, e diminuito nel caso opposto. Inoltre, se c_n e $y_m(n)$ hanno lo stesso segno contribuiscono a z_m come un termine positivo, oppure come un termine negativo se di segno diverso. Infine, la colonna δ esprime la variazione del contributo di $c_n y_m(n)$ a z_m, frutto della variazione di c_n e dei segni di c_n e di $y_m(n)$. Come si può verificare, δ è sempre tale da contribuire alla riduzione di e_m.

e_m	$y_m(n)$	c_n	c_n	$c_n y_m(n)$	δ
+	+	↓	+	+	↓
+	+	↓	−	−	↓
+	−	↑	+	−	↓
+	−	↑	−	+	↓
−	+	↑	+	+	↑
−	+	↑	−	−	↑
−	−	↓	+	−	↑
−	−	↓	−	+	↑

metodo del gradiente stocastico trova applicazione in molteplici ambiti, ivi comprese le reti neurali.



La figura a lato mostra uno schema di architettura per il filtro adattivo, che opera come equalizzatore con periodo di campionamento $\tau = {}^{T_s}/_{2}$, e ad ogni periodo di simbolo esegue l'aggiornamento dei propri coefficienti, in accordo alla (18.35), mediante l'operatore di *accumulo* simbolizzato dalla $\sum$ cerchiata.

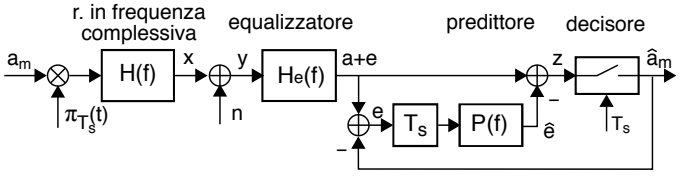
18.4.4 Equalizzatore a reazione

Questo approccio realizza una sorta di *equalizzazione dell'equalizzazione*, vediamo in che senso. Intanto osserviamo che i valori e_m che compongono la sequenza di errore $e_m = z(mT_s) - a_m$ (la differenza tra l'uscita z_m di $H_e(f)$ ed il simbolo trasmesso a_m) sono tra loro *correlati*, dato che

- il rumore bianco $n(t)$ (da cui $z(mT_s)$ anche dipende) si è *colorato* attraversando $h_R(t) * h_e(t)$;
- se l'ISI non è completamente rimossa significa che $H_e(f) \neq {}^{1}/_{H(f)}$ e quindi non si verificano le condizioni di Nyquist, ed i simboli *ricevuti* a_m divengono statisticamente dipendenti.

La tecnica del *decision feedback equalizer* o DFE sfrutta questa correlazione per *predire* i valori e_m a partire dagli errori agli istanti di simbolo precedenti, mediante un secondo filtro FIR che stima $\hat{e}_m = \sum_{n=1}^{N} p_n e_{m-n}$, e contribuisce ad una grandezza di decisione $z_m = a_m + e_m - \hat{e}_m$ affetta da un errore ancora più piccolo.

Il filtro di predizione è indicato come $P(f)$ nello schema di figura, nel quale si suppone assenza di errori *di decisione*, ovvero $\hat{a}_m = a_m$, ed i coefficienti p_n *ottimizzati* applicando ad es. il criterio MMSE all'errore sull'errore, ossia $E\left\{(e_m - \hat{e}_m)^2\right\} = \min$. Ma lo stesso schema può essere *ricablato* come appare in fig. 18.5-a), in cui $P(f)$ ed il sommatore in uscita sono *sdoppiati*: ciò consente di riunificare $H_e(f)$ e $P(f)$ in un unico *filtro in avanti* o di *smoothing*, la cui uscita viene ora fatta dipendere dal segnale ricevuto nei soli istanti di simbolo *passati*[44].



Lo schema di equalizzazione diviene quindi quello di fig. 18.5-b), in cui il filtro all'indietro *o di controreazione* è indicato come *di feedback*, opera alla frequenza di

[44] Ovvero il filtro FIR ha solo i ritardi relativi ai coefficienti c_n con $n \geq 0$, così come la sommatoria (18.14).

simbolo f_s, e la sua uscita $\hat{e}_m$ dipende solamente dalle decisioni precedenti $\hat{a}_m$[45], una cui combinazione lineare è usata per *correggere* i valori emessi dal primo filtro: la grandezza in ingresso al decisore può quindi essere ora espressa come

$$z_m = \sum_{n=0}^{N_1} c_n y\left(mT_s - n\tau\right) - \sum_{n=1}^{N_2} d_n \hat{a}_{m-n} \tag{18.36}$$

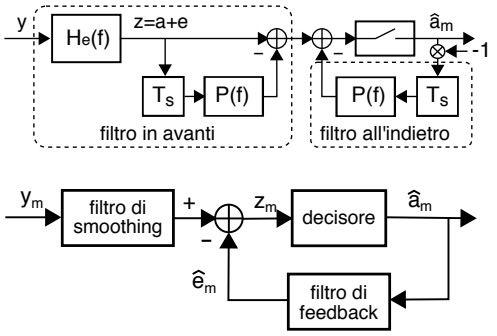


Figura 18.5: Schema simbolico di un decision feedback equalizer

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 18.6, in cui la grandezza che *guida* l'aggiornamento dei d_n è la stessa misura dell'errore e_m usata per l'aggiornamento dei c_n del filtro in avanti, portando ad adattare la (18.35) come $\mathbf{d}_{m+1} = \mathbf{d}_m - \Delta e_m \mathbf{a}_m$, in cui $\mathbf{a}_m$ è il vettore dei simboli precedentemente decisi $a_{m-1}, a_{m-2}, \cdots, a_{m-N_2}$ e presenti nella memoria del filtro di feedback.

Dato che il filtro di feedback opera sui simboli già decisi e per questo privi di rumore, non introduce nuovo rumore, che anzi è in generale ridotto grazie alla minore lunghezza del filtro in avanti. D'altra parte la presenza del decisore nell'anello di reazione rende il metodo *non lineare*, e per questo l'effetto *moltiplicativo* di eventuali errori di decisione non può essere analizzato dal punto di vista teorico, ma solo essere studiato mediante simulazione.

[45] Anche qui, nella 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.

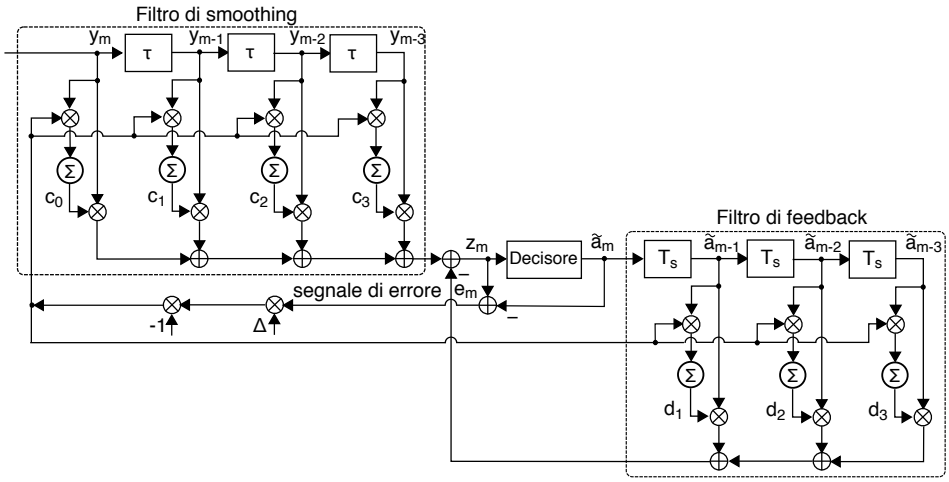


Figura 18.6: Architettura di un equalizzatore DFE basato sul gradiente stocastico

18.4.5 Equalizzazione come sequenza di massima verosimiglianza

Le tecniche finora illustrate effettuano la decisione simbolo per simbolo, mentre il metodo che stiamo per analizzare 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. In questo caso non viene utilizzato nessun filtro aggiuntivo, e la funzione di equalizzazione è assolta dal decisore che opera in base all'algoritmo *di Viterbi* (pag. 581), con costi dei percorsi calcolati in modo simile a quelli del *soft decision decoding* di pag. 585.

Per illustrare il funzionamento di MLSD iniziamo con il *riconsiderare* i campioni $y_m = y(mT_s)$ del segnale ricevuto (18.13) come

$$y_m = \sum_{n=0}^{N} a_n h(mT_s - nT_s) + \nu_m = h_0 a_m + \sum_{n=0,\, n\neq m}^{N} a_n h_{m-n} + \nu_m \tag{18.37}$$

in cui $h(t) = h_T(t) * h_c(t) * h_R(t)$, evidenziando così le componenti di errore dovuto all'ISI (la $\sum_{n\neq m}$) ed al rumore ν_m. Il problema viene quindi impostato come quello della stima della sequenza $\mathbf{a} = \{a_m\}$ a partire dalla conoscenza della sola sequenza di osservazione $\mathbf{y} = \{y_m\}$, e per il quale la soluzione *di massima verosimiglianza* individua $\hat{\mathbf{a}}$ in modo tale che $p(\mathbf{y}/\mathbf{a})|_{\mathbf{a}=\hat{\mathbf{a}}}$ è massima.

La (18.37) evidenzia inoltre come la componente aleatoria di $\mathbf{y}$ sia tutta riconducibile a quella della sequenza di rumore ν, 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 indipendenti[46] 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

$$\hat{\mathbf{a}} = \arg\min \sum_{m=0}^{M} -\ln\left(p_\nu(y_m/\mathbf{a})\right) \tag{18.38}$$

Quest'ultimo passaggio trova motivazione nella espressione della d.d.p. gaussiana $p_\nu(y) = \frac{1}{\sqrt{2\pi}\sigma_\nu} \exp\left\{-\frac{(y-\bar{y})^2}{2\sigma_\nu^2}\right\}$ per cui $\ln(p_\nu(y)) = -\ln\left(\sqrt{2\pi}\sigma_\nu\right) - \frac{(y-\bar{y})^2}{2\sigma_\nu^2}$, e dato che σ_ν è la stessa per tutti gli istanti m, il suo valore non contribuisce alla minimizzazione (18.38), che diviene pertanto

$$\hat{\mathbf{a}} = \arg\min \sum_{m=0}^{M} \left(y_m - \bar{y}_m(\mathbf{a})\right)^2$$

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. 581) riconducendo il problema a quello di individuare un *percorso di minimo costo* nell'ambito del *traliccio* che rappresenta tutte le possibili sequenze $\mathbf{a}$.

[46] A questo proposito, è opportuno che $H_R(f)$ sia realizzato come un passa basso e non come un filtro adattato, o nel caso di rumore colorato in ingresso, che si effettui una operazione di *sbiancamento*.

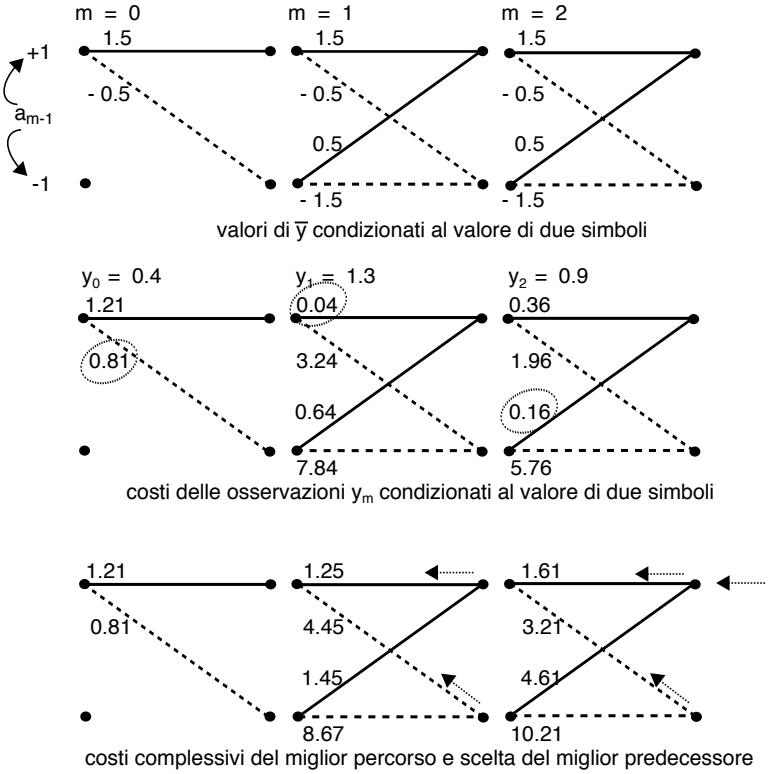


Figura 18.7: Costruzione del traliccio per un decodificatore MLSD

Esempio Ipotizziamo una $h_n = \begin{cases} 1 & con\ n = 0 \\ 0.5 & con\ n = 1 \end{cases}$ ed una trasmissione antipodale binaria con $a = \{1, -1\}$; la (18.37) 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 appare in alto in fig. 18.7, 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 centrale della figura gli archi sono invece etichettati con i costi $\left(y_m - \bar{y}_m(\mathbf{a})\right)^2$ associati all'osservazione della sequenza $\mathbf{y} = 0.4, 1.3, 0.9$, e sono evidenziati i costi *minimi* simbolo per simbolo, che porterebbero a decidere per $-1, 1, 1$. Al contrario, la figura in basso mostra il costo totale del *miglior percorso* che raggiunge le righe del traliccio per ogni istante di simbolo, assieme alle frecce che indicano il miglior predecessore. L'applicazione del criterio MLSD determina quindi la decisione per $1, 1, 1$, con un bit di differenza.

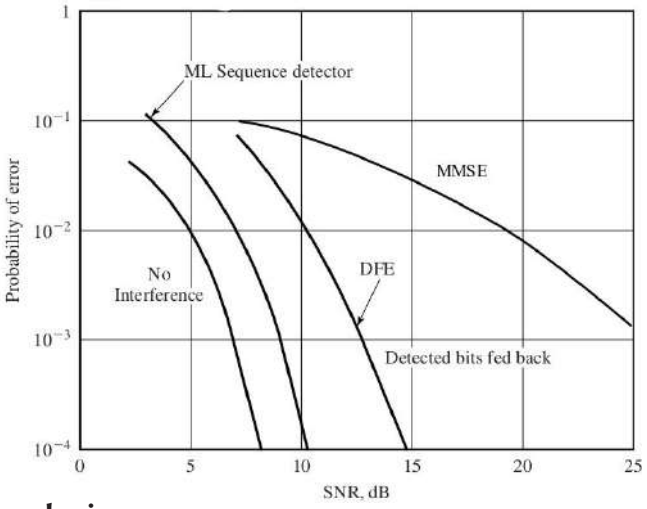
$a_m a_{m-1}$	$\bar{y}_m$
1,1	1.5
1,-1	0.5
-1,1	-0.5
-1,-1	-1.5

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 al § 17.4.2.1 e non viene qui ripetuto.

18.4.6 Confronto delle prestazioni di equalizzazione

La figura che segue[47] confronta P_e vs. SNR per una modulazione binaria antipodale nei casi di assenza di ISI oppure di un canale con $h_0 h_1 h_2 = 0.4, 0.8, 0.4$, i cui effetti sono compensati mediante equalizzazione realizzata in base alle tecniche fin qui 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 bande in cui il canale ha guadagno nullo[48]; 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 al ritardo da attendere per una decisione su intere sequenze, nonché le difficoltà a rendere adattativo l'algoritmo nel caso di condizioni non stazionarie.



18.4.7 Considerazioni conclusive

A volte l'operazione di equalizzazione non si svolge mediante un filtro, ma in un modo peculiare a seconda del tipo di modulazione. Abbiamo osservato al § 16.8.7 come nel caso dell'OFDM sia sufficiente moltiplicare il simbolo $\underline{a}_n$ ricevuto su ogni portante per l'inverso della risposta in frequenza $\underline{H}_n$ del canale equivalente passa basso, ovvero $\tilde{\underline{a}}_n = \underline{a}_n \cdot \frac{1}{\underline{H}_n}$. Nel caso di una trasmissione a spettro espanso, la stima delle caratteristiche dei cammini multipli viene usata da un ricevitore Rake (§ 20.5.2) per attuare l'equalizzazione di canale mediante ritardi e prodotti complessi; qualora però il rumore additivo sia colorato (o generato da interferenti) il ricevitore RAKE può essere *generalizzato* per tenerne conto.

Citiamo infine la soluzione dell'equalizzazione *turbo*[49], in cui lo stadio di equalizzazione e quello di decodifica si scambiano *decisioni soffici* ovvero verosimiglianze logaritmiche che rappresentano l'*informazione estrinseca* (pag. 591) relativa ai simboli ricevuti, iterando i rispettivi compiti finché non pervengono alle stesse decisioni.

[47] Tratta da J.G. Proakis, M. Salehi, *Communication systems engineering*, 2^{nd} Ed., 2002 Prentice-Hall

[48] Nonostante MMSE tenti di minimizzare σ_e^2 tenendo conto sia dell'ISI che del rumore, le sue prestazioni degradano in presenza di canali che manifestano profonde attenuazioni per alcune frequenze. D'altra parte per canali meno problematici, le sue prestazioni possono avvicinarsi a quelle dell'MLSD.

[49] Vedi ad es. `https://en.wikipedia.org/wiki/Turbo_equalizer`

18.5 Appendici

18.5.1 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

$$\begin{aligned}
\mathcal{W}_z &= \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} w(t)\, dt = \\
&= \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} v(t) \left[\int_{-\infty}^{\infty} v(t-\tau)\, y(\tau)\, d\tau \right] dt = \\
&= \int_{-\infty}^{\infty} y(\tau) \left[\lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} v(t)\, v(t-\tau)\, dt \right] d\tau \\
&= \int_{-\infty}^{\infty} y(\tau)\, \mathcal{R}_v(-\tau)\, d\tau = \int_{-\infty}^{\infty} Y(f)\, \mathcal{P}_v(f)\, df = \\
&= \int_{-\infty}^{\infty} \left[\Re\{Y(f)\} + j\Im\{Y(f)\} \right] \mathcal{P}_v(f)\, df = \\
&= \int_{-\infty}^{\infty} \mathcal{P}_v(f)\, \frac{R(f)}{|Z(f)|^2} df
\end{aligned}$$

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 $\mathcal{W}_z = \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} w(t)\, dt$, 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 omesso dalla definizione di *potenza attiva* $\mathcal{W}_z(f)$.

18.5.2 Condizioni per il massimo trasferimento di potenza

Svolgiamo per intero la dimostrazione delle (18.3). Verifichiamo subito che, mantenendo $Z_g(f)$ fisso e per qualunque valore di $R_c(f)$, la potenza ceduta ad un carico espressa dalla (18.2): $\mathcal{W}_{z_c}(f) = \mathcal{P}_{v_g}(f)\, \frac{R_c(f)}{|Z_c(f)+Z_g(f)|^2}$ 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

$$\mathcal{W}_{z_c}(f) = \mathcal{P}_{v_g}(f)\, \frac{R_c(f)}{\big(R_c(f) + R_g(f)\big)^2} = \frac{\mathcal{P}_{v_g}(f)}{R_c(f) + 2R_g(f) + \big(R_g(f)\big)^2/R_c(f)} \tag{18.39}$$

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:

$$\frac{d}{dR_c}\left(R_c + 2R_g + \frac{R_g^2}{R_c}\right) = 1 - \left(\frac{R_g}{R_c}\right)^2 = 0 \tag{18.40}$$

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 (18.3). Volendo verificare che la (18.40) individui effettivamente un minimo e non un massimo del denominatore di (18.39), 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.

18.5.3 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

$$\mathcal{W}_{z_c}(f) = \mathcal{P}_v(f)\, \frac{R_c(f)}{\big|Z_g(f) + Z_c(f)\big|^2}$$

Il rapporto tra la densità di potenza effettivamente ceduta a $Z_c(f)$, e quella che le sarebbe ceduta se $Z_c(f)$ fosse adattato per il massimo trasferimento di potenza, fornisce

la perdita di potenza subita:

$$\frac{\mathcal{W}_{z_c}(f)}{\mathcal{W}_d(f)} = \mathcal{W}_d(f)\, 4R_g \frac{R_c(f)}{\left|Z_g(f) + Z_c(f)\right|^2} \cdot \frac{1}{\mathcal{W}_d(f)} = \frac{4R_g(f)\, R_c(f)}{\left|Z_g(f) + Z_c(f)\right|^2} = \alpha(f)$$

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 50 Ω, e con densità di potenza disponibile (a frequenze positive)

$$\mathcal{W}_d^+(f) = \frac{\mathcal{W}_d}{4W} rect_{2W}(f - f_0)$$

in cui $\mathcal{W}_d$ = 1 Watt è la potenza disponibile totale, distribuita uniformemente in una banda $2W$ = 10 KHz centrata a frequenza $f_0 = 1MHz$. Il generatore è collegato ad un carico

$$Z_c(f) = R_c(f) + jX_c(f)$$

con $R_c(f)$ = 50 Ω ed $X_c(f) = 2\pi f L$ = 50 Ω 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

$$\alpha = \frac{4R_g R_c}{\left|Z_g + Z_c\right|^2} = \frac{4 \cdot 50 \cdot 50}{\left|50 + 50 + j50\right|^2} = \frac{10000}{12500} = 0.8$$

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).

Capitolo 19

Collegamento in cavo e fibra ottica

La trasmissione vera e propria dei segnali, che siano di banda base o modulati, di natura analogica o resi tali dopo codifica di linea, avviene per il tramite di un *mezzo trasmissivo* a cui corrisponde una *entità fisica*, la cui analisi permette di descrivere il peggioramento (o distorsione) causato dal transito nel mezzo trasmissivo, nei termini fin qui adottati. I mezzi *convenzionali* sono *cavo*, *canale radio* e *fibra ottica*, ed in questo capitolo vengono analizzate le relazioni tra i parametri fisici che caratterizzano il primo e l'ultimo, assieme ai fenomeni che possono manifestarsi, consentendo di giungere ad una descrizione dei mezzi nei termini della rappresentazione tempo-frequenza di un canale di comunicazione. Al canale radio è invece dedicato il cap. 20.

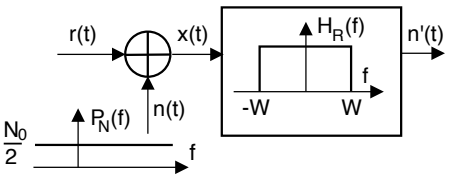
Per ciascuno dei mezzi presi in considerazione l'analisi perviene ad una caratterizzazione del tutto peculiare. Per il caso dei collegamenti *in cavo* si fa ampio uso della teoria dei circuiti esposta al § 18.1, specializzando ulteriormente i risultati per i casi di *linea aerea*, *coppia ritorta* e *cavo coassiale*. Anche se nel caso della *fibra ottica* il segnale non è più di natura elettrica ma luminosa, l'aspetto *dissipativo* nei confronti dell'energia in transito la accomuna alla trasmissione via cavo. Ma per la fibra sono i fenomeni di *dispersione temporale* a caratterizzare il canale, imponendo vincoli di *banda massima* e dunque di velocità di segnalazione. L'evoluzione della tecnologia trasmissiva in fibra porta quindi a nuove possibilità di intervento, come l'amplificazione *ottica* e la *multiplazione a divisione di lunghezza d'onda* (WDM), nonché la possibilità di realizzare una sorta di equalizzazione (o meglio *compensazione*) della dispersione cromatica. Si affrontano quindi *aspetti di rete*, essendo la fibra storicamente legata all'infrastruttura di trasporto, illustrando architetture e dispositivi in grado di attuare un principio di *commutazione* di lunghezza d'onda, e permettere l'*accesso in fibra* anche da parte dei dispositivi di utente.

L'aspetto che accomuna tutti i mezzi trasmissivi è l'analisi della attenuazione subita dalla comunicazione in transito, o meglio di come questa incida sulla distanza (o *portata*) che è possibile coprire mediante una unica tratta, ovvero sulla potenza con cui è necessario trasmettere: tali valutazioni prendono il nome di *bilancio di collegamento*, ed è la prima cosa di cui ci occupiamo.

19.1 Bilancio di collegamento

Come un bilancio economico individua l'equilibrio tra entrate ed uscite, così il bilancio *di collegamento* mette in relazione i parametri *energetici* che lo descrivono, ovvero la potenza disponibile del trasmettitore W_{dT}, la *sensibilità* del ricevitore ovvero la minima potenza $W_{R_{Min}}$ che occorre ricevere, e l'attenuazione disponibile A_d *del mezzo trasmissivo* che si intende utilizzare.

Mentre W_{dT} e A_d dipendono dal trasmettitore e dal canale, la sensibilità $W_{R_{Min}}$ è legata alla qualità (SNR o P_e) desiderata per il collegamento e dunque alla potenza di rumore in ingresso al ricevitore, ovvero alla *banda occupata* W dal *segnale trasmesso*, come illustrato ai § 15.4.1 e 14.1.



Determinazione della sensibilità Si ottiene in base alla conoscenza del livello di rumore $\frac{N_0}{2}$ in ingresso al ricevitore (vedi cap. 8.4.2.1) e dell'SNR (cap. 14) o della P_e (vedi § 15.21 e cap. 16) che si intende conseguire. Nel caso di trasmissioni analogiche, se si desidera ottenere un valore $SNR = \alpha SNR_0 = \alpha \frac{W_R}{N_0 W}$ occorre ricevere una potenza[1]

$$W_{R_{Min}} = N_0 W \cdot \frac{SNR}{\alpha}$$

mentre per trasmissioni numeriche, il vincolo ad ottenere un valore di P_e^{bit} prefissato consente di determinare il valore minimo di $\frac{E_b}{N_0} = \frac{W_R}{N_0 f_b}$ e quindi

$$W_{R_{Min}} = N_0 f_b \cdot \frac{E_b}{N_0}$$

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 eq. (8.13) a pag. 247). Infatti

$$\frac{W_{dR}}{W_{dN}} = \frac{P_R}{4R_g} \frac{4R_g}{P_N} = \frac{P_R}{P_N}$$

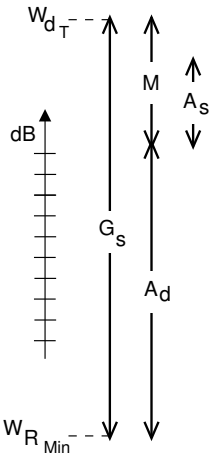
così come l'SNR non varia se, anziché le potenze *disponibili*, si considerano quelle assorbite da un carico (lo stadio di ingresso del ricevitore), dato che segnale e rumore subiscono il medesimo rapporto di partizione (vedi § 18.1.1.2).

A partire dalla sensibilità può essere tracciato il diagramma della figura seguente, con l'aiuto del quale interpretare le grandezze definite appresso.

[1] Come definito al § 14.1.4, SNR_0 dipende solo dalle caratteristiche del collegamento, mentre il coefficiente α rappresenta la dipendenza dal tipo di modulazione adottata, e differisce da 1 nei casi di modulazione FM, AM-PI e AM-PPS.

Guadagno di sistema Individua il rapporto[2]

$$G_s = \frac{W_{dT}}{W_{R_{Min}}}$$



e rappresenta il massimo valore di attenuazione *disponibile* A_d che è possibile superare. La differenza in decibel

$$G_{s_{dB}} = W_{dT}\ [dBW] - W_{R_{Min}}\ [dBW]$$

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

$$A_{d_{dB}} \leq G_{s_{dB}}$$

Margine di sistema La differenza tra $G_{s_{dB}}$ ed $A_{d_{dB}}$, 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:

$$M_{dB} = G_{s_{dB}} - A_{d_{dB}}$$

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) nascere da un mancato massimo trasferimento di potenza, o nel caso di una trasmissione radio essere causate da perturbazioni atmosferiche, cammini multipli e mobilità, mentre nel caso multiutente occorre considerare i fenomeni di interferenza, o ancora nelle fibre ottiche possono dipendere da perdite nei giunti ...

Grado di servizio Esprime un indicatore di qualità del collegamento, valutabile considerando il valore delle attenuazioni supplementari come determinazioni di variabili aleatorie. Il *grado di servizio* esprime infatti la probabilità che l'attenuazione supplementare *non superi* il valore del margine a disposizione, perché altrimenti la potenza ricevuta si riduce sotto la minima $W_{R_{Min}}$, ed il collegamento "va *fuori specifiche*". Tale probabilità individua pertanto la percentuale di tempo per la quale si mantiene $W_R > W_{R_{Min}}$, ed è definita come

$$Grado\ di\ Servizio = Pr\left\{\sum A_{s_{dB}} < M_{dB}\right\}$$

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 $Pr\left\{\sum A_{s_{dB}} > M_{dB}\right\} = 10^{-4}$.

[2]Notiamo che G_s è definito come ingresso/uscita, contrariamente agli altri guadagni. Infatti, non è una *grandezza* del collegamento, bensì una *potenzialità* dello stesso.

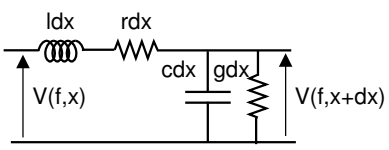
I concetti ora introdotti sono utilizzati nelle sezioni che seguono, specializzandoli al caso del mezzo trasmissivo in esame.

19.2 Collegamenti in cavo

Iniziamo l'analisi dei mezzi trasmissivi con la descrizione delle caratteristiche e delle prestazioni dei cavi in rame, utilizzati fin dall'inizio della storia della trasmissione allo scopo di recapitare a distanza i segnali in forma elettrica. Il risultato più rilevante 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.

19.2.1 Costanti distribuite, grandezze derivate, e condizioni generali

Un conduttore elettrico uniforme e di lunghezza infinita è descritto in base ad un modello cosiddetto a *costanti distribuite* in quanto espresso nei termini delle costanti *primarie* resistenza r, conduttanza g, capacità c e induttanza l *per unità di lunghezza*, di cui in figura è fornita una rappresentazione nei termini della rete due porte (§ 18.1.2) corrispondente ad una sezione di *lunghezza infinitesima* del cavo. A partire dalle costanti primarie sono quindi definite due grandezze derivate, l'*impedenza caratteristica* $Z_0(f)$ e la *costante di propagazione* $\gamma(f)$, da cui ottenere un modello di rete due porte equivalente *all'intero* collegamento in cavo.



Impedenza caratteristica Rappresenta il rapporto tra $V(f)$ ed $I(f)$ in un generico punto del cavo, ed è definita come

$$Z_0(f) = R_0(f) + jX_0(f) = \sqrt{\frac{r + j2\pi fl}{g + j2\pi fc}} \tag{19.1}$$

permettendo di scrivere

$$I(f) = \frac{V(f)}{Z_0(f)}$$

Costante di propagazione Il rapporto tra valori di tensione presenti in due punti distanti d di un cavo di lunghezza infinita viene espresso come

$$V(f, x + d) = \mathrm{e}^{-\gamma(f)d} V(f, x)$$

in cui la dipendenza da f dovuta al cavo

$$\gamma(f) = \alpha(f) + j\beta(f) = \sqrt{(r + j2\pi fl)(g + j2\pi fc)} \tag{19.2}$$

è indicata come *costante di propagazione*.

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:

$$r_g(f) = \frac{Z_g(f) - Z_0(f)}{Z_g(f) + Z_0(f)} \quad \text{e} \quad r_c(f) = \frac{Z_c(f) - Z_0(f)}{Z_c(f) + Z_0(f)} \tag{19.3}$$

Osserviamo che qualora $Z_g(f) = Z_c(f) = Z_0(f)$ si ottiene $r_g(f) = r_c(f) = 0$.

Quadripolo equivalente L'impedenza vista dai morsetti di *ingresso* e di *uscita* di un cavo interposto tra generatore e carico vale rispettivamente

$$Z_i(f) = Z_0(f)\frac{1 + r_c(f)\cdot \mathrm{e}^{-2d\gamma(f)}}{1 - r_c(f)\cdot \mathrm{e}^{-2d\gamma(f)}} \quad \text{e} \quad Z_u(f) = Z_0(f)\frac{1 + r_g(f)\cdot \mathrm{e}^{-2d\gamma(f)}}{1 - r_g(f)\cdot \mathrm{e}^{-2d\gamma(f)}} \tag{19.4}$$

Allo stesso tempo la funzione di trasferimento *intrinseca* (§ 18.1.2) risulta

$$H_q(f) = 2\frac{\mathrm{e}^{-d\gamma(f)}}{1 - r_g(f)\cdot r_c(f)\cdot \mathrm{e}^{-2d\gamma(f)}} \tag{19.5}$$

Condizioni di adattamento Corrispondono ad avere

$$Z_g(f) = Z_c(f) = Z_0(f) \tag{19.6}$$

a cui corrisponde *assenza* di distorsione lineare (§ 8.2), dato che in tal caso le (19.3) forniscono $r_g(f) = r_c(f) = 0$, e dunque dalle (19.4) si ottiene che

$$Z_i(f) = Z_u(f) = Z_0(f)$$

e la (19.5) diviene

$$H_q(f) = \frac{V_q(f)}{V_i(f)} = 2\mathrm{e}^{-d\gamma(f)}$$

Pertanto il verificarsi delle condizioni di adattamento (19.6) implica che il cavo *si comporta* come se avesse lunghezza *infinita*. Come ulteriore conseguenza troviamo che $H_i(f) = \frac{Z_i(f)}{Z_g(f)+Z_i(f)} = \frac{1}{2}$ (vedi § 18.1.2) e dato che $R_g(f) = R_u(f)$, per il guadagno disponibile (eq. (18.6) pag. 609) si ottiene

$$G_d(f) = |H_i(f)|^2 \left|H_q(f)\right|^2 \frac{R_g(f)}{R_u(f)} = \frac{1}{4}\left|2\mathrm{e}^{-d[\alpha(f)+j\beta(f)]}\right|^2 = \mathrm{e}^{-2d\alpha(f)} \tag{19.7}$$

Condizione di Heaviside Qualora i valori delle costanti primarie verifichino la relazione $r\cdot c = l\cdot g$ (nota come *condizione di Heaviside*[3]) le (19.1) e (19.2) si semplificano, implicando

$$\gamma(f) = \sqrt{rg} + j2\pi f\sqrt{lc} \qquad e \qquad Z_0(f) = \sqrt{\frac{r}{g}} = \sqrt{\frac{l}{c}} = R_0$$

Pertanto le parti reale $\alpha(f)$ ed immaginaria $\beta(f)$ di $\gamma(f)$ divengono rispettivamente costante e linearmente crescente con la frequenza, realizzando così le condizioni di un canale perfetto (pag. 231) per il termine $\mathrm{e}^{-d\gamma(f)}$ che compare in $H_q(f)$; dato inoltre che l'impedenza caratteristica $Z_0(f) = R_0$ è ora solamente resistiva ed indipendente dalla frequenza, diviene semplice realizzare la condizione di adattamento (19.6), così come quella $Z_c(f) = Z_g^*(f)$ di massimo trasferimento di potenza (§ 18.1.1.3), oltre ad

[3] Pe una breve biografia ed il link agli scritti, vedi `https://it.wikipedia.org/wiki/Oliver_Heaviside`

implicare $r_g(f) = r_c(f) = 0$, 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 risposta in frequenza complessiva per questo caso vale

$$H(f) = H_i(f)\,H_q(f)\,H_u(f) = \frac{1}{2}2e^{-d\sqrt{rg}}e^{-jd2\pi f\sqrt{lc}}\frac{1}{2} = \frac{1}{2}e^{-d\sqrt{rg}}e^{-jd2\pi f\sqrt{lc}}$$

dunque equivalente 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[4]

$$A_d(f) = 1/G_d(f) = e^{2d\sqrt{rg}}$$

19.2.2 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 (19.4) e (19.5).

Cavo molto lungo Se il cavo è sufficientemente lungo da poter considerare

$$\left|e^{-2d\gamma(f)}\right| = e^{-2d\alpha(f)} \ll 1$$

le (19.4) divengono $Z_i(f) = Z_u(f) \simeq Z_0(f)$, mentre la (19.5) si semplifica in $H_q(f) = 2e^{-d\gamma(f)}$; nel caso generale risulta pertanto

$$G_d(f) = \left|H_q(f)\right|^2 \cdot |H_i(f)|^2 \cdot \frac{R_g(f)}{R_u(f)} = 4 \cdot e^{-2d\alpha(f)} \cdot |H_i(f)|^2 \cdot \frac{R_g(f)}{R_u(f)}$$

che evidenzia due cause di distorsione lineare, ossia quella intrinseca legata ad $\alpha(f)$, e quella che dipende dal disadattamento di impedenze in ingresso ed uscita. La seconda causa può essere rimossa qualora si realizzi la condizione di adattamento $Z_g(f) = Z_c(f) = Z_0(f)$, ottenendo (eq. (19.7))

$$A_d(f) = \frac{1}{G_d(f)} = e^{2d\alpha(f)}$$

che circoscrive la causa della distorsione lineare al solo comportamento *non perfetto* di $H_q(f) = 2e^{-d\gamma(f)}$, che può essere *neutralizzato* solo nel caso in cui le costanti primarie soddisfino le condizioni di Heaviside.

> Ma in pratica il risultato è diverso, perché.... le "costanti primarie" *non sono costanti !!!*

Effetto pelle Si tratta di un fenomeno[5] legato all'addensamento del moto degli elettroni presso la superficie del cavo, fenomeno sempre più marcato al crescere della frequenza, come riportato in fig. 19.1. Per questo motivo la superficie del conduttore realmente attraversata da corrente elettrica è sempre più ridotta all'aumentare di f, ed a questo corrisponde un aumento della resistenza per unità di lunghezza r. Si può mostrare che per frequenze maggiori di 50-100 KHz il valore di r aumenta proporzio-

[4] Vedi l'eq. (19.7) con $R_g(f) = R_u(f) = R_0$.

[5] Vedi ad es. `https://en.wikipedia.org/wiki/Skin_effect`

nalmente 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

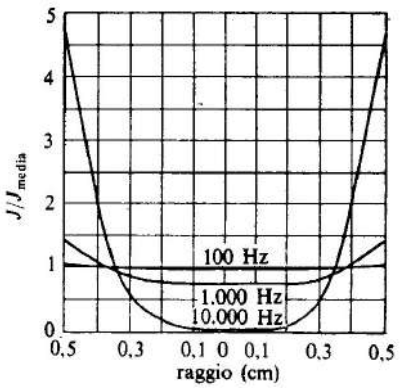
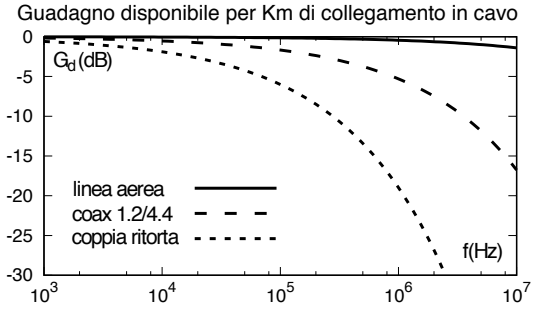


Figura 19.1: Distribuzione della corrente in un conduttore a sezione circolare per tre diverse frequenze

$$A_d(f)|_{dB} = A_0(f_R)\cdot d_{Km}\cdot\sqrt{\frac{f}{f_R}} \tag{19.8}$$

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 lunghezza[6].

La figura a lato mostra l'andamento di $G_d(f)|_{dB} = -A_d(f)|_{dB}$ che si ottiene adottando i valori di A_0 e f_R riportati al § 19.2.3 per alcune tipologie di cavo.



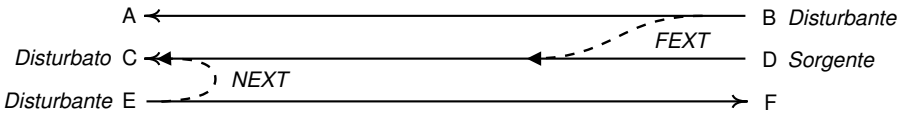
Equalizzazione In presenza di effetto pelle la funzione di trasferimento intrinseca $H_q(f) = 2e^{-d\gamma(f)}$ presenta una dipendenza da f tutt'altro che perfetta, causando distorsione lineare sui segnali in transito qualora questi contengano frequenze oltre la banda audio. Un problema analogo insorge anche in assenza di effetto pelle, qualora si manifesti un disadattamento di impedenze ed il cavo non sia sufficientemente lungo (vedi pag. 649).

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. D'altra parte, una volta stimata la $H(f)$ da equalizzare, la natura

[6] Questa circostanza è comune con le trasmissioni in fibra ottica (vedi fig. 19.4 a pag. 657), ed è legato alla presenza nel mezzo di una componente *dissipativa*, in questo caso la resistenza.

statica del collegamento permette di evitare tecniche di equalizzazione marcatamente adattative.

Diafonia La diafonia, indicata in inglese con il termine di *crosstalk*, consiste nel fenomeno di *interferenza* tra segnali che transitano 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 comunicazioni[7]; 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 tra **D** e **C** su cui gravano due cause di interferenza: 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 paradifonia*

$$\Delta A_{EC}|_{dB} = A_{EC}|_{dB} - A_{DC}|_{dB}$$

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 **E** ed osservato al punto **C** risulta quindi pari a[8]

$$W_E^{next} = W_E - A_{EC} = W_D - A_{EC} = W_C + A_{DC} - A_{EC} = W_C - \Delta A_{EC}$$

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 pari a $W_B^{fext} = W_C - \Delta A_{BC}$, in cui lo *scarto di telediafonia* ΔA_{BC} ha il valore

$$\Delta A_{BC}|_{dB} = A_{BC}|_{dB} - A_{DC}|_{dB}$$

[7] ... le famose *interferenze* telefoniche, praticamente scomparse con l'avvento della telefonia numerica (PCM), da non confondere con ... *le intercettazioni*.

[8] Omettiamo di indicare di operare in dB per compattezza di notazione.

19.2.2.1 Casi limite

Completiamo l'analisi considerando i seguenti due casi particolari.

Cavo a basse perdite E' un modello applicabile per tutte quelle frequenze tali da verificare $r \ll 2\pi f l$ e $g \ll 2\pi f c$. In tal caso le (19.1) e (19.2) forniscono

$$Z_0(f) = R_0 = \sqrt{\frac{l}{c}} \text{ reale} \quad \text{e} \quad \gamma(f) = j2\pi f\sqrt{lc}$$

Di conseguenza, è facile realizzare $Z_g = Z_c = R_0$ in modo da ottenere

$$H_q(f) = 2\mathrm{e}^{-jd2\pi f\sqrt{lc}}$$

e quindi il cavo non presenta distorsione 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

$$\mathrm{e}^{-d\gamma(f)} = \mathrm{e}^{-d\alpha(f)}\mathrm{e}^{-jd\beta(f)} \simeq \mathrm{e}^{-jd\beta(f)}$$

in quanto $\mathrm{e}^{-d\alpha(f)} \simeq 1$.

Qualora si verifichi un disadattamento di impedenze i coefficienti di riflessione $r_g(f)$ e $r_c(f)$ risultano diversi da zero, rendendo

$$H_q(f) = 2\frac{\mathrm{e}^{-jd\beta(f)}}{1 - r_g(f) \cdot r_c(f) \cdot \mathrm{e}^{-j2d\beta(f)}}$$

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. (19.3) mostra come rispettivamente risulti $r_c(f) = \pm 1$, per cui la prima delle (19.4) diviene

$$Z_i(f) = Z_0(f)\,\frac{1 \pm \mathrm{e}^{-j2d\beta(f)}}{1 \mp \mathrm{e}^{-j2d\beta(f)}}$$

e pertanto per i valori (ricorrenti) di frequenza f (o di distanza d) che rendono $\mathrm{e}^{-j2d\beta(f)} = \pm 1$ ([9]), l'impedenza di ingresso del cavo può risultare infinita o nulla.

Evidentemente la distorsione lineare prodotta in questo caso ha un andamento del tutto dipendente dalle particolari condizioni operative, e dunque la sua equalizzazione deve prevedere componenti in grado di adattarsi alla $H_q(f)$ del caso[10]. D'altra 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 collegamento[11].

[9] Ovvero, tali che $2d\beta(f) = k\pi$ con $k = 0, 1, 2, \ldots$

[10] Può ad esempio rendersi necessario "tarare" un trasmettitore radio, la prima volta che lo si collega all'antenna.

[11] E' questa la fase in cui il modem *anni 90* che si usava per collegarsi al provider Internet emetteva

19.2.3 Tipologie di cavi per le telecomunicazioni

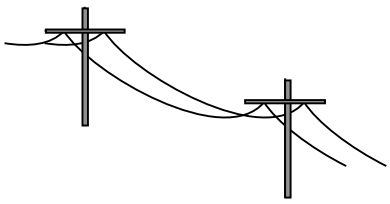
Descriviamo i principali tipi di cavo utilizzati, per i quali forniamo in tabella i valori tipici delle grandezze essenziali, nelle condizioni illustrate nel testo che segue.

Tipo di cavo	A_0 $[dB/Km]$	Z_0 $[\Omega]$	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, polietilene	89, $1.88{\cdot}10^{-7}$, $.26{\cdot}10^{-6}$, 10^{-10}
" 2.6/9.5 mm	2.3 ad 1 MHz	50, aria	41, ", ", "
" 8.4/38 mm	.88 ad 1 MHz	$\frac{138}{\sqrt{\varepsilon_r}}\log_{10}\frac{D}{d}$	1.45, ", ", "

19.2.3.1 Coppia simmetrica

In questo caso il cavo è costituito da due conduttori uguali, la distanza tra i quali permette di definire due sottoclassi: linea aerea e coppia ritorta.

Linea aerea Si tratta 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 meno sviluppati.



I valori riportati in tabella sono riferiti a conduttori con ϕ = 3 mm[12], 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.

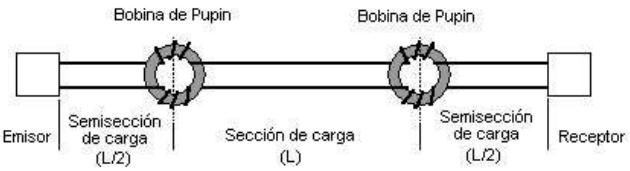
Coppia ritorta E' costituita da una coppia di conduttori in rame con ϕ da 0.4 ad 1.3 mm, rivestiti di materiale isolante, ed avvolti tra loro secondo *eliche* con passo grande rispetto al diametro. Un numero variabile di tali coppie (tra qualche decina e qualche centinaio) sono poi raggruppate assieme, 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,

una serie di orribili suoni... corrispondenti alla ricezione della sequenza *di apprendimento*, vedi anche la nota 23 di pag. 627.

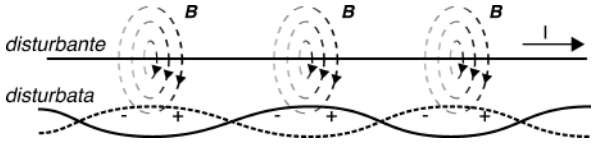
[12]Ovvero, una sezione capace di reggere il peso del cavo lungo una campata.

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*).

I valori riportati in tabella sono riferiti a conduttori con ϕ = .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'espediente di innalzare artificialmente l, collocando ad intervalli regolari una induttanza "concentrata" (le cosiddette bobine *Pupin*[13]), realizzando così nella banda del canale telefonico un comportamento approssimativamente perfetto. Ma al crescere della frequenza le bobine Pupin producono un effetto passa basso, aumentando di molto il valore di attenuazione. In tempi successivi, quando le stesse coppie ritorte sono state utilizzate per la trasmissione di segnali numerici PCM, le bobine Pupin sono state rimosse, ed al loro posto inseriti ripetitori rigenerativi (§ 18.3.2). L'impedenza caratteristica di circa 600 Ω riportata nella tabella di pag. 650 è quella valida a frequenze audio, con cavi di diametro ϕ = .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 ϕ = .7 mm; per diametri di 1.3 mm si ottengono valori circa dimezzati, mentre con ϕ = .4 mm il valore di A_0 risulta maggiore.



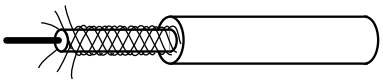
La configurazione a spirale dei conduttori ha infine lo scopo di ridurre i fenomeni di diafonia tra circuiti differenti. Infatti se il passo dell'elica è diverso tra le coppie affasciate in unico cavo, le tensioni e correnti indotte[14] da una coppia su di un'altra non interessano sempre lo stesso conduttore, ma entrambi in modo alternato. L'avvolgimento della coppia disturbante produce inoltre 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.



[13] Vedi `https://it.wikipedia.org/wiki/Pupinizzazione`

[14] L'induzione elettromagnetica è causata dal campo magnetico tempo-variante $\boldsymbol{B}$ generato dalla corrente che scorre nella coppia disturbante, vedi ad es.
`https://it.wikipedia.org/wiki/Induzione_elettromagnetica`.

19.2.3.2 Cavo coassiale



In questo caso è presente un conduttore centrale ricoperto di dielettrico, su cui è avvolto il secondo conduttore, intrecciato a formare una sorta di *calza*, connesso *a massa* ad entrambe le estremità, e racchiuso a sua volta in una guaina isolante. La particolare conformazione del cavo lo rende molto più resistente ai fenomeni di interferenza, svolgendo una funzione di *gabbia di* FARADAY[15].

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. 650. 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 relazioni

$$r = 8.4 \cdot 10^{-8}\sqrt{f}\left(\frac{1}{D}+\frac{1}{\phi}\right)\frac{\Omega}{m} \qquad l = 0.46\log_{10}\frac{D}{\phi}\frac{\mu H}{m}$$

$$g = 152 \cdot 10^{-12}\frac{f\varepsilon_r \tan\delta}{\log_{10}\frac{D}{\phi}}\frac{S}{m} \qquad c = \frac{24.2\cdot\varepsilon_r}{\log_{10}\frac{D}{\phi}}\frac{pF}{m}$$

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÷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

$$W_{dT}^{(n)} = \tfrac{10}{120} = 83.3 \text{ mW, con } n = 1, 2, \ldots, 120.$$

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

$$G_s^{(120)}\Big|_{dB} = 10\log_{10}\frac{W_{dT}^{(120)}}{W_{R_{Min}}} = 10\log_{10}\frac{83.3}{1} = 19.2 \text{ dB},$$

[15] Vedi `https://it.wikipedia.org/wiki/Gabbia_di_Faraday`

essendo $W_{R_{Min}} = 1$ mW come indicato nel testo. Come noto, G_s corrisponde alla massima attenuazione $A_{d_{Tot}}$ che può essere accettata, e pertanto

$$A^{(120)}_{d_{Tot}}\Big|_{dB} = A^{(120)}_{d}\,(dB/Km) \cdot L_{Km} = 19.2 \text{ dB},$$

da cui si ricava per la massima lunghezza

$$L_{Km} = \frac{A^{(120)}_{d_{Tot}}\Big|_{dB}}{A^{(120)}_{d}(dB/Km)} = \frac{19.2}{6.46} = 2.97 \text{ Km},$$

che come vediamo è imposta dal canale più attenuato. Per il primo canale si ha invece $A^{(1)}_{d}\,(dB/Km) = A_0$, e dunque

$$A^{(1)}_{d_{Tot}}\Big|_{dB} = A_0\,(dB/Km) \cdot L_{Km} = 5.3 \cdot 2.97 = 15.74 \text{ dB}.$$

La differenza tra $G_s|_{dB}$ (uguale per tutti i canali) e $A^{(1)}_{d_{Tot}}\Big|_{dB}$ 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 una misura della *distorsione lineare di ampiezza* del cavo nella banda del segnale.

Nel caso si voglia superare una lunghezza doppia anche $A^{(120)}_{d_{Tot}}\Big|_{dB}$ raddoppia, e per mantenere $W_{R_{Min}} = 1$ mW deve raddoppiare anche $G^{(120)}_{s}\Big|_{dB}$. Pertanto la nuova potenza/canale risulta

$W'_{dT}\,(dBm) = W_{R_{min}}\,(dBm) + G'_s\,(dB) = 0 + 2G_s\,(dB)$; quindi

$$W'_{dT}\,(mW) = 10^{\frac{W'_{dT}(dBm)}{10}} = 10^{\frac{2G_s(dB)}{10}} = 10^{\frac{2\cdot 19.2}{10}} = 10^{3.84} = 6918.3$$

milliWatt, cioè 6.91 Watt/canale, e quindi $6.91 \cdot 120 = 830$ Watt complessivi !

19.3 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 e 50 nm dell'ultravioletto (1 nm=10^{-9} metri), che corrispondono a frequenze (ricordando ancora che $f = \frac{c}{\lambda}$) che vanno da $3 \cdot 10^{12}$ fino a $6 \cdot 10^{15}$ Hz. Questi valori individuano una banda 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

Infrarosso	$\rightarrow$	Ultravioletto	
10^{-4}	$\rightarrow$	$50{\cdot}10^{-9}$	λ [metri]
$3{\cdot}10^{12}$	$\rightarrow$	$6{\cdot}10^{15}$	f [Hz]

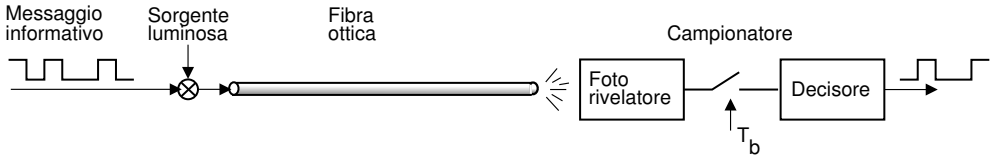


Figura 19.2: Schema di trasmissione in fibra ottica

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!

19.3.1 Trasmissione ottica

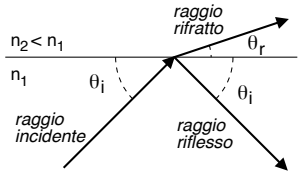
Anche se sono teoricamente possibili schemi di *modulazione analogica*, le fibre ottiche sono usate per trasportare informazione di natura *binaria* secondo lo schema di fig. 19.2, in cui la luce emessa da una sorgente è accesa o spenta, ovvero modulata in ampiezza, con uno schema detto *on-off keying* o OOK[16]. All'altro estremo della fibra un *fotorivelatore* effettua (appunto) 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.

Propagazione luminosa e indice di rifrazione Lo spazio libero è il mezzo in cui la luce viaggia alla sua *massima* velocità, pari a $c = 3 \cdot 10^8$ m/sec; il rapporto n tra c e la velocità di propagazione v in un mezzo trasparente prende il nome di *indice di rifrazione*[17] n del mezzo stesso, ossia $n = c/v$, risultando sempre $n \geq 1$.

Esempio Se $n = 2$ allora la velocità di propagazione della luce nel nuovo mezzo è la metà di quella che avrebbe nello spazio.

Quando un raggio luminoso incontra una superficie di separazione tra mezzi con diverso indice n (ad esempio, da n_1 ad $n_2 < n_1$) una parte di energia *si riflette* con angolo θ_i uguale a quello *incidente*, e la restante parte continua *rifrangendosi* nell'altro mezzo, ma con diverso angolo θ_r. La relazione tra gli angoli è nota come *legge di Snell*

$$\cos\theta_r = \frac{n_1}{n_2}\cos\theta_i \qquad (19.9)$$



[16] Indicata anche come *intensity modulation and direct detection* (IMDD). In realtà è anche possibile adottare tecniche di modulazione numerica come PSK e QAM, che richiedono una detezione *coerente* (vedi ad es. https://doi.org/10.1364/OE.16.000753), ma tali sistemi sono tuttora in fase sperimentale, e l'esposizione prosegue per il caso universalmente adottato.

[17] Vedi http://it.wikipedia.org/wiki/Indice_di_rifrazione, ma anche il video https://www.youtube.com/watch?v=8VZHym6HqVU.

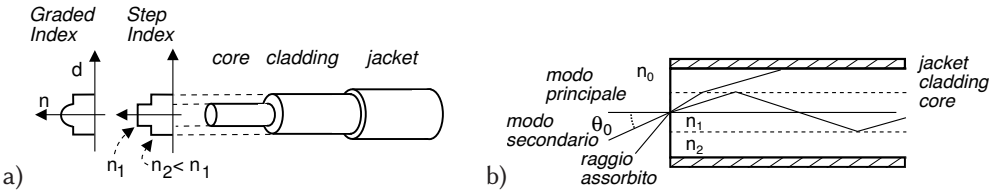


Figura 19.3: Struttura della fibra ottica (a) e modi di propagazione (b)

ed essendo $n_1/n_2 > 1$ il raggio *rifratto* risulta più inclinato nel mezzo con n inferiore (dove viaggia più veloce), ovvero si ha sempre $\theta_r < \theta_i$.

Immaginiamo ora di aumentare un po' alla volta l'inclinazione, a partire da $\theta_i = \pi/2$ (perpendicolare) fino a $\theta_i = 0$: esiste un *valore critico* $0 < \theta_c < \pi/2$ per l'angolo di incidenza θ_i in corrispondenza del quale θ_r si azzera: ciò corrisponde ad avere il termine sinistro della (19.9) pari ad uno, da cui si ottiene $\theta_c = \arccos \frac{n_2}{n_1}$.

Per valori di incidenza $\theta_i < \theta_c$ non si verifica rifrazione, ma *tutto il raggio viene riflesso*. La capacità della fibra ottica di trasportare energia luminosa si fonda proprio su questo fenomeno[18], che a sua volta ne determina la struttura, costituita da un nucleo (*core*) centrale con indice di rifrazione n_1, circondato da un rivestimento (*cladding*) con indice $n_2 < n_1$[19]; entrambi racchiusi in una guaina (*jacket*) di materiale opaco, raffigurati in fig. 19.3a.

Applicando la (19.9) anche all'interfaccia tra sorgente luminosa (con indice di rifrazione $n_0 < n_1$) e fibra, si definisce *apertura numerica*[20] il valore

$$\Delta = \sqrt{n_1^2 - n_2^2} = n_0 \sin \theta_0^{Max}$$

dove θ_0^{Max} è il massimo angolo θ_0 (vedi fig. 19.3b) con cui può *entrare* energia nella fibra, e quindi continuare a propagarsi mediante riflessione totale. Pertanto si ottiene $\theta_0^{Max} = \arcsin \frac{\Delta}{n_0}$ da cui osserviamo che quanto più Δ è piccola (ovvero n_1 ed n_2 sono simili) tanto più θ_0^{Max} si riduce, e dunque si riduce la potenza luminosa che viene immessa nella fibra ottica, ma... si ottiene il beneficio illustrato di seguito.

Quando un raggio luminoso attraversa la fibra, l'energia si propaga mediante diversi *modi di propagazione,* uno per ogni angolo $\theta_0 < \theta_c$ con cui entra la luce incidente[21]. Il *modo principale* è quello che si propaga lungo l'asse rettilineo, mentre i *modi secondari* sono quelli con angolo $\theta_i < \theta_c$ 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.

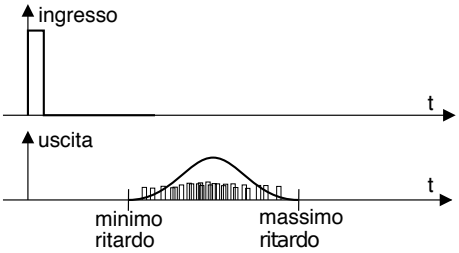
[18] Qui descritto in termini di ottica geometrica, approssimazione valida per un diametro del *core* ben maggiore di quello della λ incidente. Per dimensioni comparabili, occorre invece ricorrere alla *teoria di propagazione delle onde*, in cui non ci avventuriamo.

[19] I diversi valori di n sono ottenuti *drogando* differentemente la sezione della fibra.

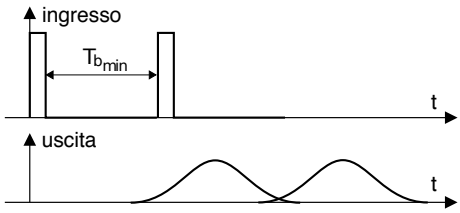
[20] Vedi ad es. `https://en.wikipedia.org/wiki/Numerical_aperture`

[21] In realtà questa interpretazione data in chiave di ottica geometrica è una semplificazione, ed in effetti i *modi di propagazione* sono quelli che risultano dalla applicazione delle *equazioni di Maxwell* alla propagazione in fibra.

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, come esemplificato in figura.



L'entità della *dispersione temporale* (differenza tra ritardo massimo e minimo) è tanto maggiore quanto più il collegamento è lungo, e quanti più modi partecipano alla propagazione: un suo valore tipico è dell'ordine di 10 nsec/Km. La conseguenza di questo fenomeno è la limitazione della *massima frequenza* con cui gli impulsi luminosi possono essere posti in ingresso alla fibra; impulsi temporalmente troppo vicini causano infatti interferenza intersimbolica (ISI) in uscita, rendendo gli impulsi sovrapposti. Pertanto la massima frequenza di segnalazione in una fibra ottica *dipende dalla sua lunghezza.*



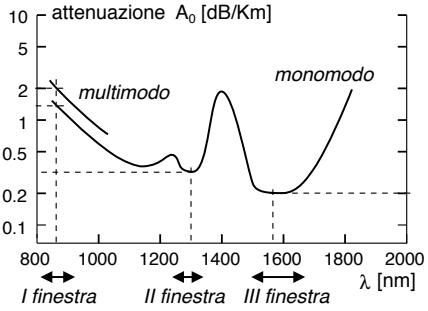
Nelle fibre ottiche *multimodo* sono presenti più modi di propagazione, e vengono distinte (fig. 19.3a) nel tipo STEP INDEX se n cambia in modo brusco, o in quello GRADED INDEX se il *core* ha un indice graduato. Nel secondo caso la dispersione temporale è ridotta; infatti quando i modi secondari attraversano 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 piccolo[22], da consentire alla sorgente luminosa di immettere luce nella fibra solo con angolo di incidenza nullo, e quindi permettere la propagazione del solo modo principale.

Ovviamente le ultime due soluzioni (graded index e fibra monomodo) sono state rese possibili grazie ai progressi nei processi di fabbricazione.

Attenuazione Similmente 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 che si manifestano sono quelli di natura *intrinseca* del materiale silicio, quelli legati allo *scattering* per disomogeneità della densità e del diametro della fibra, e quelli legati alla presenza di *impurità chimiche*[23], che possono ridurre la trasparenza oppure avere dimensioni (a livello molecolare) comparabili con le lunghezze d'onda in gioco.

[22] Si passa dai 50 μm per le fibre multimodo, a circa 8 μm nel caso monomodo.

[23] Ovvero molecole e ioni di altri elementi. Ad esempio, lo ione OH- è quello che determina il picco di assorbimento a 1.39 μm.



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 19.4: Dipendenza della attenuazione chilometrica dalla lunghezza d'onda λ

La caratteristica comune ai fenomeni di assorbimento è una marcata dipendenza da λ, cosicché la loro combinazione determina la caratteristica di attenuazione *chilometrica* A_0 mostrata in fig. 19.4, dove possono essere 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).

Tra le fonti di attenuazione *supplementare* troviamo quella causata dalle *giunzioni* tra tratte in fibra ottica: 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. Inoltre, le fibre si possono *fondere* tra loro, con perdite tra 0,01 e 0,1 dB. Una ulteriore fonte di perdite localizzate può essere costituita *dalle curve* nel percorso, che non devono avere un raggio troppo stretto, altrimenti parte dell'energia non subisce riflessione totale, e viene assorbita dal *jacket*.

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 se in esso 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[24]. La dispersione cromatica *nominale* D_0 della fibra si misura in $\left[\frac{psec}{Km \cdot nm}\right]$, e dà luogo ad una effettiva dispersione temporale

$$D = D_0 \cdot L \cdot \Delta\lambda \quad psec$$

che è direttamente proporzionale alla lunghezza L della fibra ed alla estensione della

[24]Il fenomeno descritto viene detto dispersione *da materiale* o D_M, oltre al quale ne interviene anche un altro detto dispersione di *guida d'onda* o D_W, che dipende da fattori geometrici come la dimensione del *core* e l'apertura numerica.

gamma cromatica $\Delta\lambda$ della sorgente[25]. 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 nominale 15 volte inferiore a 1.3 μm ($\sim 1 \left[\frac{psec}{Km \cdot nm}\right]$) che non a 1.5 μm[26];
- scegliere una sorgente con la *minima* estensione cromatica $\Delta\lambda$;
- adottare una tecnica di *controllo* della dispersione, vedi § 19.3.3.3.

Dispersione del modo di polarizzazione Indicata come PMD, è una conseguenza della non perfetta simmetria cilindrica del *core*, che causa il fenomeno della *birifrangenza*[27]; dato che queste variazioni geometriche sono casuali e disperse su tutta la fibra, ciò determina un continuo scambio di energia tra le componenti a polarizzazione verticale ed orizzontale del segnale in transito, a cui si associa una *dispersione temporale* che dipende dalla radice della lunghezza della fibra, in una proporzione compresa tra 0.1 e 0.01 $ps/\sqrt{Km}$. Normalmente questo fenomeno ha conseguenze trascurabili, ma può incidere sulle prestazioni di collegamenti lunghi ed a velocità elevata.

Effetti non lineari Nascono dall'interazione tra la luce ed il materiale in cui si propaga, e dipendono dalla intensità del fascio ottico, ovvero da quanto questo è *concentrato spazialmente*. Causano perdite di intensità del segnale, rumore, interferenza intercanale nel WDM, e dispersione temporale, ma è anche possibile trarne vantaggio, come nel caso dell'amplificazione ottica, della conversione di lunghezze d'onda e della compensazione di dispersione. Possiamo distinguere due categorie di effetti non lineari:

- diffusione stimolata (o *scattering*) legata alla interazione tra fotoni ed atomi della fibra, come lo *stimulated Brillouin scattering* (SBS) e lo *stimulated Raman scattering* (SRS);
- fenomeni legati all'*effetto Kerr*[28] ed alla dipendenza dell'indice di rifrazione dalla potenza ottica, come la *self phase modulation* (SPM), la *cross phase modulation* (XPM) ed il *four wave mixing* (FWM).

Sono in genere fenomeni di lieve entità, ma il loro effetto si accumula durante la propagazione, e dunque come per la dispersione, dipende dalla lunghezza del tratto percorso. Ne rimandiamo la descrizione a quando saranno citati nel seguito.

19.3.2 Bilancio di collegamento

Nel caso delle fibre ottiche lo schema definito al § 19.1 deve essere particolarizzato al tipo di trasduttori in uso, che assieme alle caratteristiche della fibra influenzano oltre che la lungheza del collegamento, anche la sua banda.

[25] Il fenomeno della dispersione cromatica è l'equivalente ottico della distorsione di fase (o distorsione di ritardo) introdotta al § 8.2 per i segnali elettrici.

[26] D'altra parte, dato che i termini D_M e D_W descritti alla nota 24 hanno una diversa dipendenza da λ, variando i loro contributi a D_0 si è riusciti a realizzare un tipo di fibra detto *dispersion-shifted* (o DS) che presenta il minimo di dispersione in terza finestra, vedi fig. 19.5.

[27] `https://it.wikipedia.org/wiki/Birifrangenza`

[28] `https://it.wikipedia.org/wiki/Effetto_Kerr`

Trasduttori elettro-ottici I primi ad essere usati sono stati gli economici LED (*Light Emitting Diode*), che richiedono una circuiteria di interfaccia semplice, sono poco sensibili alle condizioni ambientali, e risultano affidabili. D'altra parte, i LED raggiungono frequenze di *segnalazione limitate* al centinaio di Mbps, immettono nella fibra una *potenza ridotta*, ed emettono luce su di una *gamma cromatica* $\Delta\lambda > 50$ nm.

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

Tabella 19.1: Caratteristiche delle sorgenti luminose

Per ridurre la dispersione cromatica (e quindi raggiungere frequenze di segnalazione più elevate) occorre ricorrere ai *Diodi Laser* (LD)[29], che forniscono anche una maggiore potenza, e dunque divengono indispensabili per coprire distanze maggiori[30]; d'altra parte i LD sono più costosi, hanno vita media ridotta rispetto ai LED, e richiedono condizioni di lavoro più controllate. Notiamo inoltre come una fibra ottica posta inizialmente in opera mediante sorgenti LED, possa essere potenziata (in termini di banda) semplicemente 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* (vedi fig. 19.5), che presenta il minimo[31] della dispersione cromatica nominale in III finestra anziché in II, e che raggiunge valori migliori di 3.5 psec/Km·nm.

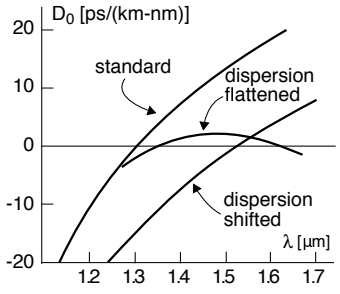


Figura 19.5: Dispersione nominale per tipi di fibra

Prodotto banda-lunghezza 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 $T_b = 1/f_b$,

[29] In particolare, con laser detti *distributed feedback* (DFB) si riesce ad eccitare un solo modo di emissione, producendo una luce di fatto monocromatica, la cui effettiva λ può anche essere variata in tutta la gamma che va dalla II alla III finestra.

[30] La 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.

[31] La presenza di valori di dispersione *negativi* in fig. 19.5 può destare una legittima curiosità. Ma non si tratta di un fenomeno *anticausale*! Come indicato dall'unità di misura $\frac{ps}{Km \cdot nm}$ di D_0, la dispersione cromatica rappresenta la *derivata* di un ritardo rispetto a λ, derivata che dipende essa stessa da λ. Dunque, come i suoi valori positivi indicano che il ritardo *aumenta* con λ, e quindi le frequenze *più basse* (con λ maggiore) arrivano *dopo* di quelle più alte, i valori negativi di D_0 individuano il fenomeno inverso, ovvero che il ritardo aumenta *con il diminuire* di λ, ovvero le frequenze *più alte* arrivano dopo (di quelle basse).

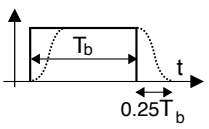
ed imponiamo che la dispersione temporale sia non maggiore di $\frac{1}{4}T_b$, deve risultare

$$D = D_0 \cdot L \cdot \Delta\lambda \leq 0.25 \cdot T_b \tag{19.10}$$

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 (19.10) può essere riscritta in modo da evidenziare il *prodotto della banda per la lunghezza PBL*, che è pari al valore

$$PBL_{NRZ} = f_b \cdot L = \frac{.25}{D_0 \cdot \Delta\lambda} \quad [Tbps \cdot Km] \tag{19.11}$$

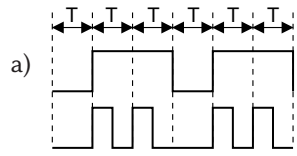
che è una grandezza che dipende dalla coppia fibra-sorgente[32], e che rappresenta la relazione tra f_b ed L necessaria ad ottenere $D = \frac{1}{4}T_b$. Inserendo dunque i valori di $\Delta\lambda$ (della sorgente) e D_0 (della fibra) nella (19.11) si ottiene *una costante* da usare per calcolare la 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 doppio[33] rispetto al caso precedente:

$$PBL_{RZ} = \frac{.5}{D_0 \cdot \Delta\lambda} = 2 \cdot PBL_{NRZ}$$

In fig. 19.6-b) sono mostrati i valori di PBL (per il caso NRZ) associati alle accoppiate fibra-sorgente indicate.



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

Figura 19.6: a) codice di linea; b) prodotto banda-lunghezza per tipiche coppie sorgente-fibra

Esercizio Determinare la lunghezza massima di un collegamento in fibra ottica monomodo, operante con λ=1.3 μm, 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. 19.4 si trova che per λ=1300 nm l'attenuazione chilometrica è di 0,35 dB/Km, che determina una $A_d = 0,35 \cdot L_{Km}$ [dB]. Imponendo ora $A_d = G_s = 42$ dB si ottiene una lunghezza pari a $L = \frac{A_d}{0.35} = \frac{42}{0.35}$ = 120 Km, che identifica il *Limite di Attenuazione* del collegamento. Verifichiamo quindi che non intervenga un limite più

[32] In questo senso, il prodotto *banda-lunghezza* costituisce un parametro di sistema che tiene conto di un concorso di cause. Un po' 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....

[33] Tuttavia il dimezzamento della durata di un bit causa una perdita di potenza di 3 dB, in base alle considerazioni riportate a pag. 661.

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 $L = \frac{PBL}{f_b} = \frac{250.000}{417} = 600$ 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 *più potente* per aumentare il guadagno di sistema, e di conseguenza migliorare il limite di attenuazione. In questo 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 occorre suddividere il collegamento in più segmenti, collegati da *ripetitori rigenerativi* (§ 18.3.2), oppure ripartire la banda su più fibre poste in parallelo; d'altra parte l'affermazione delle tecniche discusse al § 19.3.3 come WDM, *amplificazione ottica* e *controllo della dispersione*, consentono di attuare soluzioni ancora diverse.

Trasduttori ottico-elettrici Sono i dispositivi che effettuano la conversione del segnale luminoso uscente dalla fibra ottica in uno elettrico e per i quali, come per le sorgenti, non entriamo nei dettagli tecnologici. Il trasduttore utilizzato fin dall'inizio, economico ed affidabile, è il diodo P-I-N[34]. Un secondo tipo di trasduttore molto usato è il diodo APD[35] (*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 gli APD sono più delicati, e più sensibili alla temperatura. La tabella 19.2 riporta i valori di sensibilità W_R (ossia la minima potenza che è necessario ricevere) di diversi fotorivelatori, necessaria a conseguire[36] una probabilità di errore per bit $P_e = 10^{-11}$.

Fotorivelatore	λ *[nm]*	W_R *[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 P-I-N	1550	-37	1200
InGaAs APD	1550	-37.5	678

Tabella 19.2: Valori di sensibilità dei fotorivelatori

Sensibilità e frequenza di segnalazione Nella tabella 19.2 è riportato anche il valore della frequenza di segnalazione f_b a cui si riferisce la sensibilità, ma occorre

[34]Che sta per *p-intrinseco-n* riferito al tipo di drogaggio del semiconduttore - vedi `http://it.wikipedia.org/wiki/Diodo_PIN`

[35]`http://it.wikipedia.org/wiki/Fotodiodo_a_valanga`

[36]La 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!

tenere presente che quest'ultima peggiora all'aumentare di f_b. Infatti, le prestazioni conseguite dal decisore che si trova a valle del trasduttore dipendono (pag. 641) 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 come[37]

$$W_R\left(f_b'\right)\ [\text{dBm}] = W_R\left(f_b\right)\ [\text{dBm}] + 10\log\frac{f_b'}{f_b} \tag{19.12}$$

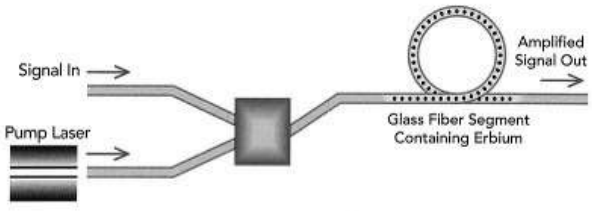
19.3.3 Seconda generazione

Quanto finora esposto può considerarsi una prima generazione[38] di sistemi in fibra ottica (anni '90), e per la quale assumendo un valore $PBL = 200$ (fig. 19.6), si ottiene una lunghezza di tratta di 80 Km a 2,5 Gbps e di soli 20 Km a 10 Gbps. Da allora si sono rese possibili nuove tecniche che consentono di aumentare di molto il PBL, e che sono ora brevemente illustrate.

19.3.3.1 Amplificazione ottica

Consiste nell'aumento della dinamica (e quindi della potenza) del segnale ottico in transito, senza effettuare la conversione in segnale elettrico e viceversa, come invece accade con un ripetitore rigenerativo (§ 18.3.2), la cui realizzazione nel caso dei sistemi WDM (che si stavano affermando nello stesso periodo) è particolarmente complessa. Con l'adozione dell'amplificazione ottica si riescono a realizzare collegamenti con rigeneratori intervallati di circa 500 km, ed amplificatori ogni circa 100 Km. Il funzionamento di questi ultimi si basa sulla *emissione stimolata* di fotoni legata alla λ in transito, prodotta da un *segnale di pompa* elettrico o luminoso, che ne determina il guadagno.

Amplificatore in fibra drogata all'erbio In questo caso il *mezzo attivo* corrisponde ad un tratto di qualche decina o centinaio di metri di fibra (appunto, drogata) in cui vengono miscelati il segnale in transito e quello di pompa. Il drogaggio a base di *erbio* è il tipo più diffuso in terza finestra, in quanto presenta un guadagno massimo in corrispondenza della *banda C* (1525 - 1565 nm) e della *banda L* (1570 - 1610 nm). Il guadagno può raggiungere i 30 dB con un segnale di pompa di 15 mW, e dipende (in modo inverso) anche dalla potenza del segnale in transito, presentando un effetto di *saturazione*; inoltre



[37] Questo metodo di calcolo 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 (19.12).

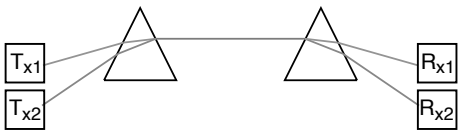
[38] Anche se, relativamente a queste prime fasi, si è soliti distinguere tre generazioni, corrispondenti all'uso delle corrispettive *finestre*, vedi fig. 19.4.

il guadagno può essere *non uniforme* su tutta la banda (in termini di λ), ma questo fenomeno può essere compensato mediante filtri ottici. L'amplificatore EDFA[39] presenta inoltre un fattore di rumore[40] di 4 - 8 dB, che pone un limite al massimo numero di tratte amplificate otticamente, dopodiché occorre intercalare un ripetitore rigenerativo.

Amplificazione a semiconduttore e Raman L'amplificatore ottico *a semiconduttore* (SOA) è di piccole dimensioni, viene pilotato da un segnale di pompa elettrico, è più economico dell'EDFA, ed opera su un ampio intervallo di λ. Di contro, il SOA è più rumoroso, presenta un guadagno inferiore a quello dell'EDFA, ed è affetto da fenomeni non lineari. Viene anche utilizzato come *interruttore ottico* nei dispositivi di multiplazione e conversione di λ.

Anche l'*amplificazione Raman* utilizza un segmento di fibra per mescolare il segnale in transito con quello (ottico) di pompaggio, ma a differenza dell'EDFA, il guadagno non dipende dal drogaggio, ma dal verificarsi dello *scattering di Raman*[41] che richiede un pompaggio maggiore, anche di 0.5 - 1 W, ed una lunghezza maggiore, anche alcuni km. Dato che non è necessario drogare la fibra, il metodo è applicabile ad impianti già in esercizio, ed il guadagno può essere reso uniforme su ampi intervalli di λ.

19.3.3.2 Multiplazione a divisione di lunghezza d'onda - WDM



Il successivo passo verso l'incremento della capacità di trasporto della fibra viene compiuto applicando alle trasmissioni ottiche il principio della multiplazione a divisione di frequenza, ovvero immettendo sulla stessa fibra più di un segnale ottico, ognuno con la sua propria λ. In questo caso si parla di WDM (*Wavelength Division Multiplex*), che viene realizzata mediante lo schema di principio[42] dei *rifrattori prismatici*, realizzando un circuito ottico del tipo illustrato alla figura precedente[43]. I dispositivi di multiplazione WDM sono *passivi* e *reversibili*, dato che non necessitano di alimentazione, ed lo stesso apparato può indifferentemente svolgere una funzione e la sua inversa. Nondimeno, spesso al multiplatore è fatto seguire uno stadio di amplificazione ottica.

Nella figura che segue si illustra come le diverse portanti ottiche vengano disposte nelle regioni a bassa attenuazione[44]. In funzione di quante portanti vengano utilizzate, si distingue tra il caso di *coarse* WDM o CWDM, con al massimo 16 λ, e quello di *dense*

[39] *Erbium doped fiber amplifier.*

[40] La natura del rumore è ottica anziché elettrica, ed è indicato come *emissione spontanea amplificata* (ASE) in quanto ha origine dai fotoni che si producono in modo *spontaneo* (anziché *stimolato* come nei laser), e che poi interagiscono con gli ioni di drogante producendone l'amplificazione.

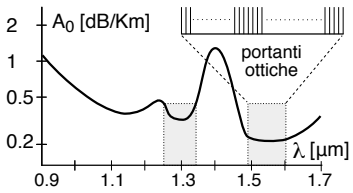
[41] `https://it.wikipedia.org/wiki/Scattering_Raman`

[42] I dispositivi reali basano il loro funzionamento su fenomeni di *diffrazione* e *interferenza*.

[43] Si sfrutta il principio "dell'arcobaleno" (ma che a me ricorda *The dark side of the moon...*), 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.

[44] Sono anche prodotte delle fibre prive dello ione OH responsabile del picco di assorbimento a 1.4 μm, dette *dry fibre*, per le quali è possibile allocare portanti in una regione veramente estesa!

WDM o DWDM. Nel DWDM sono previste 40 portanti spaziate di 100 GHz nella BANDA C[45], oppure 80 portanti spaziate di 50 GHz, su ognuna delle quali inviare un segnale con velocità 10 Gbps, per una capacità complessiva da 400 ad 800 Gbps; capacità che può ulteriormente raddoppiare qualora venga utilizzata allo stesso tempo anche la BANDA L.



I sistemi DWDM necessitano di dispositivi dotati di notevole stabilità in frequenza, dotati di controllo della temperatura, e dato che il loro uso è in pratica *relegato* alle dorsali ad alta velocità, soffrono di un prezzo elevato a causa del mercato ristretto. L'amplificazione dei collegamenti DWDM viene tipicamente svolta mediante EDFA, che a differenza dei SOA non produce effetti di intermodulazione tra canali; d'altra parte, devono esser prese contromisure rispetto alle irregolarità del guadagno tra le diverse portanti[46], e tener presente che l'amplificazione della BANDA L necessita di una lunghezza di fibra maggiore rispetto alla BANDA C, svolta pertanto su due tratte consecutive. Inoltre, livelli eccessivi di potenza (che per il DWDM è moltiplicata per il numero di λ attive) intensificano i fenomeni non lineari (pag. 658) che possono portare ad interferenza tra canali.

Un 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.

19.3.3.3 Controllo della dispersione

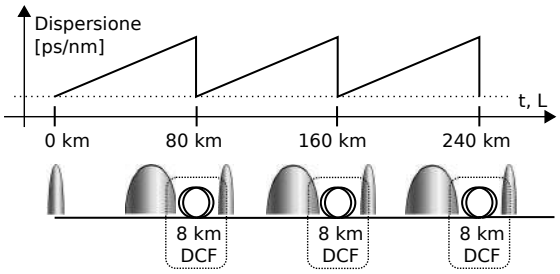
Con l'avvento degli amplificatori ottici la massima lunghezza di un collegamento in fibra non è più limitata dalla sua attenuazione, ma solo dai fenomeni di dispersione temporale, e da quelli non lineari. In realtà l'amplificazione ottica *peggiora* i fenomeni di dispersione, dato che in assenza di uno stadio di rigenerazione, queste degradazioni *si accumulano* di amplificatore in amplificatore; per questo motivo, sono state sviluppate le tecniche di gestione della dispersione. Alcune di queste agiscono al trasmettitore od al ricevitore, rispettivamente in modo da *predistorcere* il segnale, oppure di *equalizzarlo*, facendo ricorso a tecniche di demodulazione coerente, od a tecniche non lineari. In tal modo però non si riesce ad andare oltre un semplice *raddoppio* del PBL.

Fibre compensatrici Il fenomeno della dispersione *cromatica* può essere tenuto sotto controllo anche per collegamenti di migliaia di chilometri inserendo lungo gli

[45] Le portanti sono centrate attorno $f_0 = 193$ THz. Ricordando che $\lambda = v/f_0$ e ponendo $v \simeq c = 3 \cdot 10^8$ m/sec, otteniamo che alla f_0 corrisponde $\lambda = {}^{3 \cdot 10^8}/_{193 \cdot 10^{12}} = 1554$ nm, mentre una spaziatura tre le f_0 di 100 GHz equivale ad un $\Delta\lambda = c\,({}^1/_{f_2} - {}^1/_{f_1}) \simeq 0.8$ nm; pertanto, 40 portanti occupano un intervallo di 32 nm, e dunque entrano perfettamente nei 1565 - 1525 = 40 nm della BANDA C.

[46] Ancor più grave se l'irregolarità si ripete uguale su diverse sezioni consecutive di amplificazione, e che può essere affrontata interponendo filtri ottici progettati in modo da *compensare* le differenze di guadagno.

stessi alcune tratte di fibra con un coefficiente di dispersione D_0 *negativo*, e quindi in grado di *invertire* l'effetto prodotto sulle diverse componenti cromatiche[47]. Tipicamente occorre inserire qualche km di *Dispersion Compensating Fiber* (DCF) ogni cinquantina di km di collegamento, applicando la relazione $D_0 L_0 + D_{DCF} L_{DCF} = 0$ in cui il pedice $_{DCF}$ individua dispersione e lunghezza della fibra compensatrice. E' una soluzione sempre più diffusa, anche in virtù della progressiva riduzione della perdita di potenza che ne caratterizzava le prime realizzazioni.



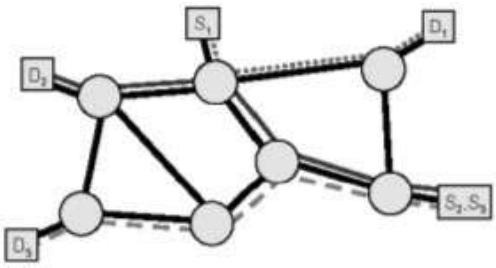
Filtri ottici E' una soluzione che evita di *allungare* il collegamento con le DCF ed opera inserendo filtri interferometrici (o basati su *reticolo*[48]) subito dopo gli amplificatori ottici. Un tale posizionamento, oltre ad avere un vantaggio logistico, permette di compensare le perdite introdotte dai filtri. Questi ultimi possono inoltre svolgere anche una funzione di controllo del rumore e di normalizzazione del guadagno dell'amplificatore ottico.

19.3.4 Sistemi in fibra ottica

Fin qui le fibre ottiche sono state descritte come mezzo trasmissivo per un collegamento punto-punto ad alta velocità, mentre il loro utilizzo si è esteso alla rete di accesso e distribuzione, e sono stati sviluppati dispositivi in grado di interconnettere i nodi di rete e svolgere le operazioni di instradamento operando direttamente a livello ottico, senza dunque dover passare dal dominio elettrico, con evidenti vantaggi e semplificazioni da un punto di vista realizzativo.

19.3.4.1 Dalle fibre ottiche alle reti ottiche

La trasmissione WDM permette di realizzare lo schema di *rete ottica* mostrato in figura e detta *wavelength routed optical network*, in cui ad ogni tributario è assegnata una λ che lo identifica da estremo ad estremo; in realtà ciò che viene realizzato è uno schema di instradamento del tipo a *circuito virtuale* (pag. 785), e l'effettiva λ associata ad un circuito *cambia* di nodo in nodo. A tal fine sono stati sviluppati i

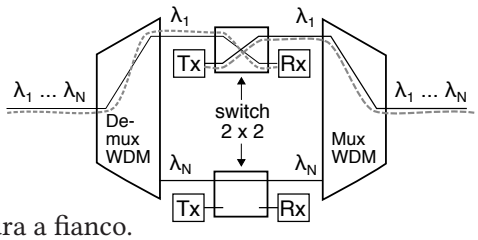


[47] Facendo riferimento alla fig. 19.5, notiamo come per una fibra *normale* D_0 in terza finestra sia positivo, ed *aumenti* con λ. Per invertire questo fenomeno, la fibra compensatrice oltre ad avere un D_0 negativo, deve anche variarne il valore con un andamento complementare a quello della fibra da compensare, in particolar modo nel caso di trasmissione DWDM.

[48] Traduzione di *grating*, con cui si descrive una alterazione periodica di un parametro fisico, vedi ad es. `https://en.wikipedia.org/wiki/Fiber_Bragg_grating`.

seguenti dispositivi, che permettono di realizzare in forma completamente ottica le funzioni svolte da quelli descritti al § 24.6.1.

Multiplatori e demultiplatori passivi che rispettivamente convogliano più λ in unica fibra, oppure le estraggono, oppure ancora che combinati assieme ad un commutatore a due vie permettono la funzionalità *optical add and drop* (vedi § 24.3.4.2) o OADM, come mostrato nella figura a fianco.



Accoppiatori a stella (o *star copulers*) che assemblano le λ provenienti da sorgenti diverse in un unico flusso WDM, che viene quindi inoltrato a molteplici ricevitori mediante altrettante fibre di uscita[49].

Convertitori di lunghezza d'onda basati su effetti non lineari[50], che pur se più costosi di altri componenti, permettono (come discutiamo sotto) di realizzare instradamenti *non bloccanti*.

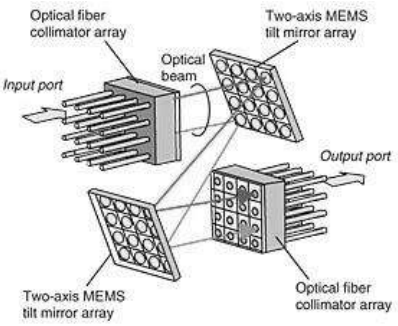
Optical cross-connects (OXC) che svolgono la funzione di commutazione ed instradamento dei segnali ottici, di cui è possibile distinguere, in ordine di complessità e di costo, tra:

- *matrici di commutazione spaziale* che permettono l'interconnessione tra M fibre in ingresso ed altrettante in uscita, e possono essere realizzate mediante dei *micro-specchi* a controllo elettromeccanico, una tecnologia nota come *micro electro-mechanical systems* o MEMS[51].

[49] In questo modo si realizza una rete di tipo *broadcast* (ovvero non *switchata*) qualora ogni nodo terminale emetta su di una sua propria λ, e riceva quelle emesse dagli altri nodi.

[50] Come la *cross gain modulation* che si manifesta nei SOA, il cui il guadagno satura con la potenza in transito. Quando al segnale OOK in arrivo con λ_1 è mescolato quello (debole e continuo) di pompa con λ_2, il guadagno satura nei periodi di bit di λ_1, mentre invece quando λ_1 è *spenta*, λ_2 viene amplificato. Un filtro ottico rimuove λ_1, e la sua informazione è stata trasferita su λ_2, con segno invertito; altri schemi risolvono anche questo aspetto. Altri dispositivi fanno uso dell'effetto FWM, in cui la presenza di λ_1 e $\lambda_p = 2\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}$ (di pompa) produce la comparsa di λ_2 in uscita.

[51] Vedi `https://it.wikipedia.org/wiki/MEMS`. Ogni specchietto ha dimensioni inferiori al mm^2, e riflette o meno la luce a seconda della sua disposizione controllata da micro-attuatori, il tutto realizzato direttamente su dei chip in silicio o polisilicio. Adottando una architettura di commutazione a due stadi è possibile realizzare strutture tridimensionali come quella mostrata in figura, che consente di adottare un numero di specchi pari a $2N$ (dato che un MEMS altera l'indice di riga, e l'altro di colonna) contro gli N^2 relativi al caso di una matrice bidimensionale (i cui flussi entranti ed uscenti sono disposti rispettivamente sui due lati di un MEMS squadrato), e di mantenere le differenze di percorso ottico entro limiti ridotti.



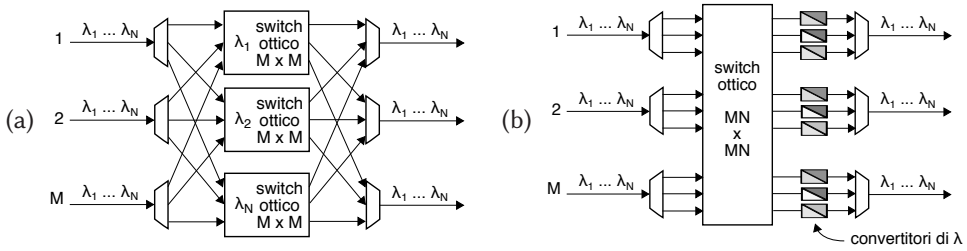


Figura 19.7: (a) commutatore di lunghezza d'onda; (b) wavelength router;

- *commutatori di lunghezza d'onda* che permettono di commutare le singole λ di M flussi WDM verso altrettanti (diversi) flussi. Sono realizzati combinando elementi di commutazione spaziale con moduli di multi-demultiplazione delle λ, come mostrato in fig. 19.7-(a).
- *wavelength selective switch* o WSS, in grado anch'essi di combinare la funzione di demultiplazione spaziale delle λ con il loro direzionamento verso una diversa fibra di uscita per mezzo di celle a cristalli liquidi, eventualmente realizzate su silicio[52], offrendo anche il vantaggio della programmabilità. Ma nel caso in cui flussi entranti differenti ma con λ uguali debbano uscire sulla medesima fibra, si verifica un *fenomeno di blocco* (§ 24.8.2), evitato dal dispositivo che segue;
- *wavelength router* (WR), con la capacità di instradare il segnale trasportato dalle λ sulle porte di ingresso verso una diversa λ in uscita, sulla base di una matrice di routing. In fig. 19.7-(b) ne viene mostrato uno schema realizzativo, in cui sono evidenziati i convertitori di λ necessari a realizzare un comportamento *non bloccante*.

19.3.4.2 Rete ottica di trasporto

Al § 24.4 viene descritta la rete SDH, che offre un servizio di trasporto a divisione di tempo per tributari di diverso tipo: ma l'architettura descritta al § 24.6 prevede la fibra *solo* come mezzo trasmissivo tra dispositivi, che invece operano in modalità elettronica, e necessitano di una conversione elettro-ottica ad ogni porta di I/O. Una rete ottica (OTN) come quella sopra descritta, al contrario, svolge tutte le funzioni direttamente nel dominio ottico, ed a questo fine sono stati definiti gli standard necessari a permettere l'interconnessione dei dispositivi ed il loro controllo. D'altra parte, non si è ancora in grado di evitare del tutto le forme di degradazione legate al rumore introdotto dagli amplificatori ottici e dai fenomeni non lineari; pertanto convivono *isole di trasparenza ottica*, interconnesse tra loro mediante stadi di completa rigenerazione.

L'approfondimento necessario a descrivere l'architettura di una OTN, e le modalità atte ad ospitare traffico eterogeneo (SDH, Ethernet, ATM, IP) travalica i limiti di questo testo, per cui si rimanda ad alcune risorse Internet[53].

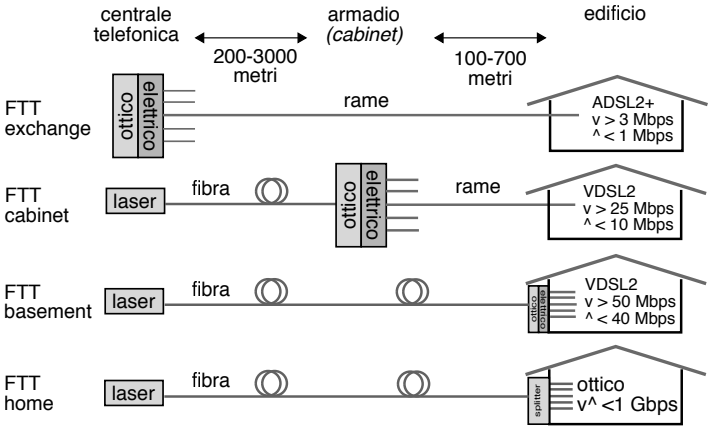
[52] Vedi `https://en.wikipedia.org/wiki/Liquid_crystal_on_silicon`

[53] `https://it.wikipedia.org/wiki/Optical_Transport_Network`
`https://studylibit.com/doc/7555525/verso-una-rete-tutta-ottica`
`https://www.itu.int/ITU-T/studygroups/com15/otn/OTNtutorial.pdf`

19.3.4.3 Rete passiva di distribuzione

Nella maggior parte dei casi il collegamento in fibra termina presso la propria centrale telefonica, dove sono alloggiati i DSLAM (§ 24.9.4) che inviano il segnale ADSL all'utente finale mediante un collegamento in rame, con la velocità consentita da questa tecnologia. Ma attualmente il collegamento in fibra ottica si avvicina sempre più alla residenza dell'utente finale, e viene classificato con una sigla del tipo FTTx, che sta per *fiber to the "x"*, in cui la *x* indica appunto fin dove arriva la fibra. In tal senso, possiamo distinguere tra

- FTT *exchange*: la situazione di base, in cui la fibra si ferma in centrale;
- FTT *cabinet*, o *curb*: viene raggiunto l'armadio tra la centrale e l'utente finale, dove vengono spostati i DSLAM;
- FTT *basement*: sono raggiunte le fondamenta del palazzo;
- FTT *home*: la fibra raggiunge direttamente l'utente finale.



Nell'ultimo caso la fibra ottica entra direttamente in casa; per ridurre complessità e costi quest'ultima tratta è priva di apparati *attivi*[54] e si basa sullo *splitting* del segnale ottico, che raggiunge in *broadcast* tutti gli utenti serviti dalla stessa fibra, i quali si avvalgono poi di meccanismi di indirizzamento e crittografici per recuperare solo ciò che è effettivamente indirizzato loro.

19.3.5 Ridondanza e pericoli naturali

Le fibre vengono normalmente interrate, e per questo sono esposte ai pericoli di essere attaccate da roditori, 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 § 24.6.3), in modo che in caso di interruzione di un collegamento sia possibile deviarne il traffico su di un altro.

[54] Che per questo motivo prende il nome di *passive optical network* o PON, vedi anche `https://it.wikipedia.org/wiki/FTTx`.

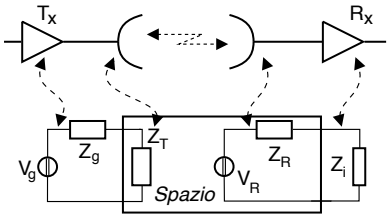
Capitolo 20

Collegamento radio

Concludiamo l'analisi dei mezzi trasmissivi iniziata al cap. 19 con *il canale radio*, su cui in pratica si basa la nostra vita attuale, permettendo ai nostri dispositivi (telefono, computer, radio e tv) di ricevere e trasmettere informazione in modalità *wireless*. Anche in questo caso vogliamo studiare le relazioni tra i parametri fisici del collegamento radio, studiare i fenomeni che possono manifestarsi, e giungere ad una descrizione del mezzo nei termini della rappresentazione tempo-frequenza di un canale di comunicazione.

Il caso del *collegamento radio* è del tutto particolare: basato sul fenomeno di trasduzione elettromagnetica che interessa l'apparato di *antenna*, dopo una disamina relativa ai fenomeni di *propagazione atmosferica* e la definizione del fenomeno dei *cammini multipli* e delle sue conseguenze, viene svolto un approfondimento relativo ai collegamenti *radiomobili* ed alla tipizzazione dell'attenuazione in funzione delle caratteristiche dell'ambiente circostante e dalle condizioni di visibilità (o meno) delle antenne. Ciò porta a descrivere l'intensità del segnale ricevuto sia in termini statistici (fading di *Rayleigh* o di *Rice*), sia in termini spettrali e/o tempo-varianti (banda e tempo *di coerenza*), a derivare nuove espressioni per la probabilità di errore, ed individuare architetture di ricezione *a correlazione* capaci di trarre vantaggio da un canale *dispersivo in frequenza*.

Modello circuitale La trasmissione via onda radio si differenzia da quella via cavo o fibra ottica sotto diversi aspetti, tra cui la condivisione di uno stesso mezzo tra più comunicazioni, e la possibilità di comunicare in movimento. E' resa possibile dalla conversione di un segnale elettrico in radiazione elettromagnetica[1] ad opera dei *dispositivi di antenna*, che fungono (vedi figura) da carico dal lato trasmissione, e da generatore dal lato ricezione. La descrizione



[1]Dato che tale conversione avviene unicamente a seguito delle *variazioni* del segnale, è esclusa la presenza di una componente continua, e per questo (ma non solo) il segnale può unicamente essere di natura *modulata*.

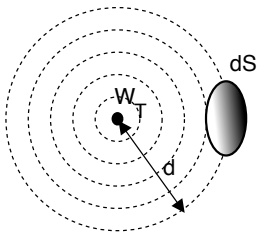
circuitale delle antenne viene poi semplificata dalla circostanza che per il segnale modulato è praticamente sempre vera la condizione di occupare una banda *stretta* attorno alla portante f_0, al punto da poterlo assimilare ad una singola sinusoide. Con una tale approssimazione le condizioni di massimo trasferimento di potenza (§ 18.1.1.3) 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 § 13.1.2.4. 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 effetti Z_T dipende dalla frequenza portante ed in parte dalla geometria dello spazio circostante, mentre Z_g è in genere fissata a 50 Ω; perciò tra stadio di uscita del trasmettitore Tx e cavo di antenna può essere interposto un *adattatore di impedenza*[2].

20.1 Trasduzione elettromagnetica

Senza minimamente affrontare alcuna analisi rigorosa, descriviamo come la potenza W_{dT} ceduta all'antenna di trasmissione si propaghi nello spazio.

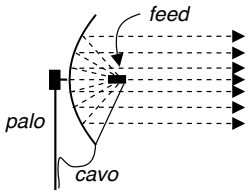
Antenna isotropa Qualora l'antenna trasmittente irradiasse allo stesso modo in tutte le direzioni, W_{dT} si distribuirebbe su di una sfera. Una superficie dS posta a distanza d sarebbe quindi attraversata da una potenza pari a

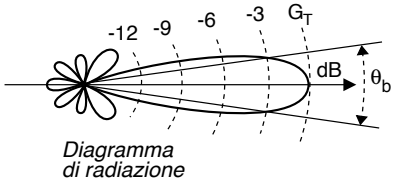
$$dW = W_{dT} \frac{dS}{4\pi d^2} \quad \text{[Watt]} \tag{20.1}$$



Si noti che il denominatore rappresenta la superficie di una sfera di raggio d.

Antenna direttiva Praticamente qualsiasi antenna presenta direzioni *privilegiate* di emissione. Ad esempio le antenne paraboliche dispongono di un *illuminatore* o FEED[3] posto in corrispondenza del *fuoco* della parabola stessa, la cui superficie riflette le onde elettromagnetiche in modo che si propaghino in forma pressoché parallela[4].





La potenza W_{dT} quindi non si distribuisce più con simmetria sferica, e la direzione di propagazione massima presenta un guadagno G_T rispetto all'antenna isotropa, e l'intensità di campo irradiato spazialmente è descritta da un *diagramma di radiazione*.

[2] Vedi ad es. `https://en.wikipedia.org/wiki/Antenna_tuner`

[3] Dall'inglese *to feed* = alimentare.

[4] Il processo di focalizzazione parabolica, comunemente usato ad esempio nei *fanali* degli autoveicoli, era ben noto ad ARCHIMEDE da Siracusa, che lo impiegò negli *specchi ustori*...

Il valore di G_T dipende dal rapporto tra le dimensioni dell'antenna e quelle della lunghezza d'onda λ secondo la relazione

$$G_T = 4\pi \frac{A}{\lambda^2} \tag{20.2}$$

avendo indicato con A l'area dell'antenna. Il guadagno G_T viene spesso espresso in dBi, ovvero dB riferiti all'antenna *isotropa*.

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 privilegiata[5]. Ovviamente minore è θ_b, e maggiore è G_T.

Antenna ricevente Se una antenna identica a quella di trasmissione 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

$$A_e = G_R \frac{\lambda^2}{4\pi} \tag{20.3}$$

legato alla forma e dimensione dell'antenna, a meno di un fattore di efficienza ρ ([6]). Perciò una stessa antenna (A_e fisso) aumenta il suo guadagno (e stringe il *beam*) all'aumentare della frequenza, ovvero al diminuire di $\lambda = \frac{c}{f}$ ([7]).

20.2 Bilancio di collegamento per spazio libero

Applichiamo il modello visto al § 19.1 per il caso di un collegamento in condizioni di *visibilità* tra le antenne.

Potenza ricevuta Usando l'area efficace dell'antenna ricevente (20.3) per intercettare parte della potenza irradiata (20.1) 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

$$A_{sl} = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4\pi d f}{c}\right)^2 \tag{20.4}$$

[5] Si tratta di un concetto del tutto analogo alla "frequenza di taglio a 3 dB", ma applicata ad un dominio spaziale con geometria radiale.

[6] Indicando con A_r l'area *reale* (fisica) dell'antenna, risulta $A_e = \rho A_r$, con $\rho < 1$. La diseguaglianza 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 (20.2), a meno di non usare appunto il valore di area efficace.

[7] La 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.

è chiamato *attenuazione di spazio libero*, e dipende da f^2 oltreché da d^2. Anche se, ai fini del bilancio di collegamento, la dipendenza dalla frequenza si elide con quella relativa al guadagno delle antenne: $G_T = A_e \frac{4\pi}{\lambda^2} = A_e \frac{4\pi f^2}{c^2}$ ([8]).

Attenuazione disponibile Il rapporto

$$A_d = \frac{W_{dT}}{W_{dR}} = \left(\frac{4\pi df}{c}\right)^2 \frac{1}{G_T G_R} \tag{20.5}$$

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

$$A_d\,(\text{dB}) = 32.4 + 20\log_{10} f\,(\text{MHz}) + 20\log_{10} d\,(\text{Km}) - G_T\,(\text{dB}) - G_R\,(\text{dB}) \tag{20.6}$$

nota come *equazione di Friis*. Osserviamo che a differenza della trasmissione in cavo l'attenuazione cresce con il quadrato della distanza, ovvero 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. (19.8)) e fibra ottica. Per un esempio di applicazione della (20.6) si veda l'esercizio a pag. 615.

Il sistema di telecomunicazione che meglio rappresenta le condizioni di spazio libero è quello tra terra e satellite (§ 25.3), per il semplice fatto che non vi sono ostacoli frapposti tra le antenne. D'altra parte i collegamenti radio terrestri, sia fissi che mobili, sono affetti da una serie di ulteriori fenomeni, mentre la (20.6) 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*: 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*[9], in quanto la crosta terrestre agisce 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 dB[10] 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.

20.3 Fenomeni propagativi e atten. supplementare

La dipendenza della propagazione radio dalla geometria del territorio e dalle condizioni atmosferiche causa l'insorgenza di termini di *attenuazione supplementare* A_s (pag. 643)

[8] Mantenendo fissa la dimensione delle antenne si ottiene il risultato che trasmissioni operanti a frequenze più elevate permettono di risparmiare potenza. Purtroppo però, guadagni di antenna superiori a 30-40 dB (corrispondenti a piccoli valori di θ_b) sono controproducenti, per i motivi esposti al §20.3.1.

[9] `http://it.wikipedia.org/wiki/Onda_superficiale`

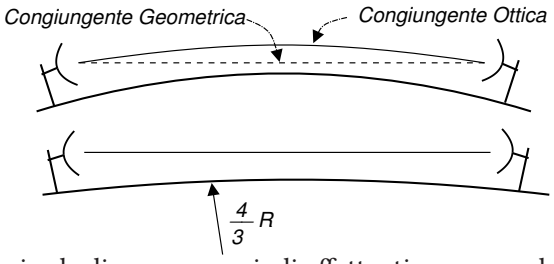
[10] equivalente ad una riduzione di potenza di $10^{2.5} = 316$ volte

da sommare al valore A_d (dB) fornito dalla (20.6) per stabilire la potenza realmente ricevuta. Iniziamo con l'affrontare gli aspetti più generali, riservando quelli legati a cammini multipli e radio mobile ai §§ successivi.

20.3.1 Condizioni di visibilità

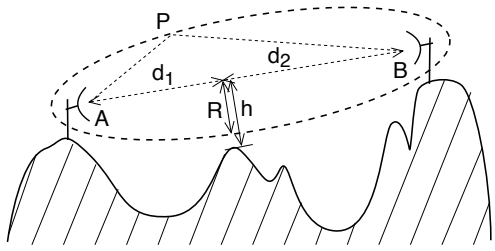
Come spiegato nel commento relativo alla definizione di area efficace (20.3) all'aumentare della frequenza la lunghezza d'onda λ diminuisce con legge reciproca, permettendo di realizzare antenne di dimensioni ridotte e di elevato guadagno. Allo stesso tempo, per evitare l'assorbimento terrestre occorre posizionare l'antenna in alto (in cima ad una torre), e trasmettere per *onda diretta*, condizione detta anche LOS o di *line of sight*.

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 (pag. 19.3.1) *maggiori,* ossia verso terra: i calcoli vengono quindi effettuati supponendo che il raggio terrestre sia per 4/3 maggiore di quello reale. Inoltre l'indice di rifrazione può variare con l'ora e con le condizioni climatiche e quindi (di nuovo) le antenne con guadagno elevato (e molto direttive) possono andare fuori puntamento.

Ellissoide di Fresnel Anche quando le antenne si trovano in condizioni di visibilità occorre comunque tenere conto dei fenomeni di *diffrazione*[11], che *deviano* nella zona *in ombra* le onde radio che transitano in prossimità di ostacoli[12]. La determinazione dell'orizzonte radio deve pertanto prevedere un *margine 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

[11] `http://it.wikipedia.org/wiki/Diffrazione`

[12] Lo 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.

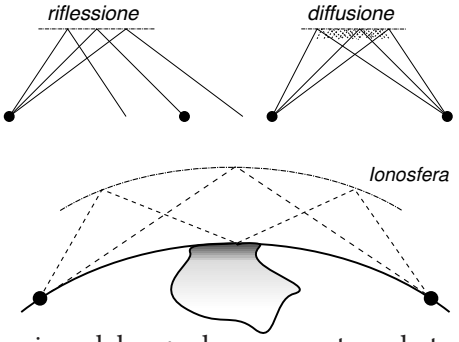
definito come il luogo dei punti P per i quali la somma delle distanze $d(A,P) + d(P,B)$ è pari a $d(A,B) + \frac{\lambda}{2}$, in cui $\lambda = \frac{c}{f}$ è la lunghezza d'onda della trasmissione a frequenza f. Suddividendo la distanza $d(A,B)$ tra i due fuochi A e B in due segmenti d_1 e d_2 individuati 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 h/R, ed è maggiore per gli *spigoli vivi*, fino ad arrivare ad una decina di dB.

20.3.2 Condizionamenti atmosferici

E' il turno di descrivere gli effetti dovuti alla natura dei diversi strati ed alle condizioni climatiche.



Diffusione e riflessione Tra 0,1 e 10 GHz si può verificare il fenomeno della *diffusione troposferica*[13] (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*[14] (la fascia oltre gli 80 Km di quota), dove strati ionizzati causano una *riflessione* del segnale e consentono la trasmissione anche tra luoghi non in visibilità[15], ma con il rischio di cammini multipli. E' questo il caso tipico della propagazione delle *onde corte*, per le quali λ va dai 100 ai 10 metri, corrispondenti ad un banda dai 3 ai 30 MHz..

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 qui può verificarsi la diffusione troposferica, specie *di notte*.

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([16]). Inoltre, il vapor d'acqua (con molecole di dimensioni maggiori) produce

[13] `http://en.wikipedia.org/wiki/Tropospheric_scatter`

[14] `http://en.wikipedia.org/wiki/Skywave`

[15] Anche, ma non solo, in concorso con la riflessione operata da masse d'acqua, come mostrato in figura.

[16] L'elevata attenuazione chilometrica presente a 60 GHz può essere sfruttata nei sistemi di comunicazione allo scopo di *riusare* una stessa banda di frequenze *per altri utenti*, anche a breve distanza.

una attenuazione supplementare di 1-2 dB/Km (al massimo) a 22 GHz[17]. Sotto i 10 GHz non si verifica assorbimento né da ossigeno, né da vapore.

20.3.2.1 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 per l'attenuazione da pioggia viene pertanto posto pari al valore di attenuazione supplementare che viene superato con una probabilità p pari a quella di fuori servizio.

Una formula sperimentale che consente di determinare il valore in dB dell'attenuazione supplementare che viene superato con probabilità p è:

$$A_s\,(r_0,\,d,\,p) = K \cdot r_0^{\alpha} \cdot d \cdot \beta\,(d) \cdot \gamma\,(p) \qquad [\text{dB}]$$

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 riportata a lato.

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 $\gamma\,(p) = 6.534 \cdot 10^{-3} \cdot p^{-\left(.718+.043\cdot\log_{10} p\right)}$ (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. 20.1 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.

Esempio Un ponte radio numerico opera tra due località distanti 50 Km con una portante $f_0 = 15$ GHz. Valutare l'attenuazione supplementare superata per lo 0.1% del tempo, nell'ipotesi che l'intensità di precipitazione superata per lo 0.01% del tempo sia pari a 40 mm/h.

[17] L'assorbimento della potenza di un'onda elettromagnetica a 2.45 GHz da parte delle molecole d'acqua è il principio su cui si basa il funzionamento di un forno a microonde.

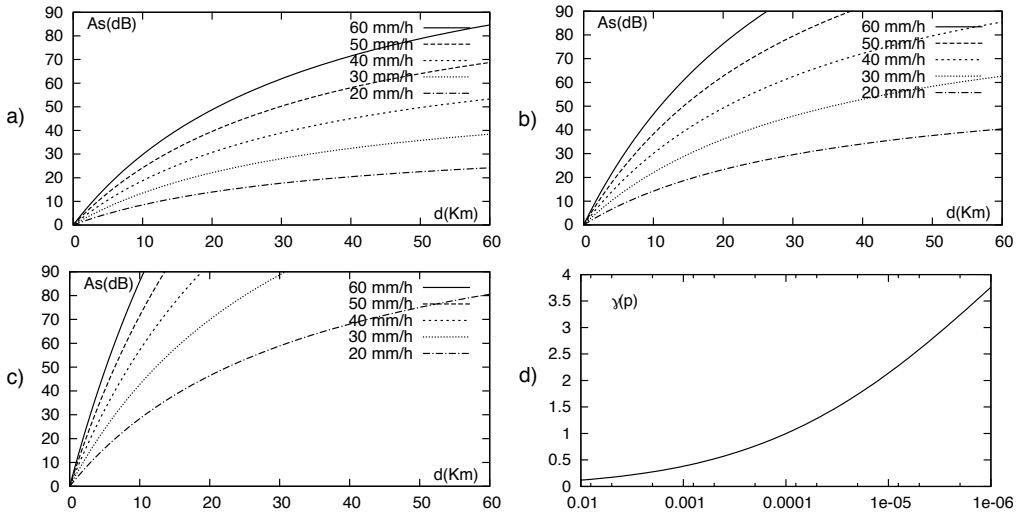


Figura 20.1: a), b), c) - attenuazione supplementare per pioggia superata per lo 0.01% del tempo, rispettivamente a 15, 20 e 30 GHz;
d) - fattore di attenuazione $\gamma(p)$ al variare della probabilità di fuori servizio

Dal primo grafico di fig. 20.1 si ricava un valore di $A_s^{10^{-4}} \geq 50$ dB per lo 0.01% del tempo; considerando invece un grado di servizio 10 volte peggiore, occorre considerare il fattore $\gamma\left(10^{-3}\right) \simeq 0.45$, e dunque $A_s^{10^{-3}} \geq 50 \cdot 0.45 = 22.5$ dB.

20.3.3 Cammini multipli

Dopo aver preso in esame i collegamenti in visibilità ed analizzato i fenomeni legati al territorio ed atmosferici, occupiamoci ora degli aspetti conseguenti la ricezione di più di una replica ritardata di uno stesso segnale trasmesso. Infatti 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 atmosferici[18] 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

$$h(t) = \sum_{n=1}^{N} a_n \delta(t - \tau_n) \tag{20.7}$$

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 § 5.2. La presenza del multipath comporta che la corrispondente risposta in frequenza

$$H(f) = \sum_{n=1}^{N} a_n e^{-j2\pi f \tau_n} \tag{20.8}$$

[18]Nel 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 si propaga come lungo una *guida d'onda*, vedi anche `http://en.wikipedia.org/wiki/Tropospheric_propagation`.

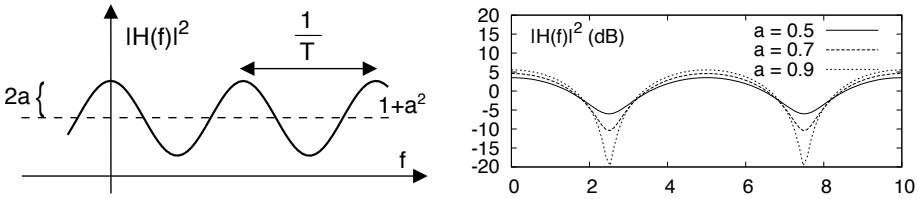
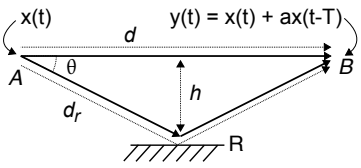


Figura 20.2: Modulo quadro della risposta in frequenza per un collegamento affetto da eco singola. A sinistra in scala lineare, a destra in dB

introduce *distorsione lineare*. Come esempio *semplice* consideriamo la presenza di una *eco singola* con ritardo T, per il quale (vedi § 5.2.3) il modulo quadro della risposta in frequenza risulta

$$|H(f)|^2 = 1 + a^2 + 2a\cos 2\pi fT$$

periodico in frequenza con periodo $f = \frac{1}{T}$, come mostrato in fig. 20.2 per valori lineari ed in dB, e per diverse scelte di a. Osserviamo che per valori $a \simeq 1$ la risposta in frequenza presenta una notevole attenuazione nell'intorno di $f = \frac{2k+1}{2T}$, impedendo di fatto la trasmissione su tali frequenze; inoltre all'aumentare di T le oscillazioni di $|H(f)|^2$ si infittiscono e dunque aumenta la possibilità che $|H(f)|^2$ vari di molto nella banda del segnale[19], causando una distorsione lineare che sarà necessario equalizzare.

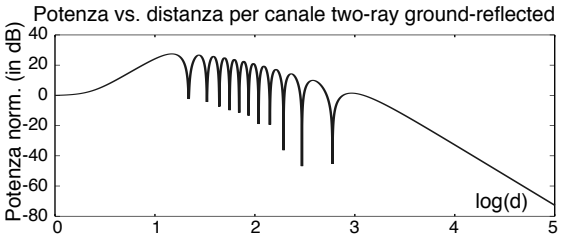


Esempio Consideriamo la geometria descritta in figura, in cui un collegamento di portata d tra A e B subisce un fenomeno di riflessione a metà della sua lunghezza, da parte di una superficie riflettente R che dista h dalla congiungente, e ricaviamo l'espressione del ritardo T. Ricordando che $tempo = \frac{spazio}{velocita'}$ e indicando con d la distanza d_{AB} e con d_r quella 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 $\theta = 11^o31'$, $\cos\theta = 0.98$, e $T = 0,066$ μsecondi. Pertanto $|H(f)|^2$ presenta un periodo (in frequenza) di $\frac{1}{T} = 15.15 \cdot 10^6 = 15.15$ MHz.

Modello two-ray ground-reflected E' il nome attribuito allo schema descritto dall'esempio precedente, esteso ad un caso generale in cui vengono prese in considerazione possibili altezze differenti per le antenne, il cui guadagno viene considerato variabile in funzione dell'angolo di emissione, e sono prese in considerazione le caratteristiche del coefficiente di riflessione al suolo. L'approfondita analisi[20] di tali particolarità porta al risultato che per distanze brevi tra le antenne le onde diretta e

[19] Ad esempio, desiderando $\frac{1}{T} > 1$ MHz, si ottiene $T_{Max} = 1$ μsec; 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.

[20] Vedi ad esempio `https://en.wikipedia.org/wiki/Two-ray_ground-reflection_model`, da cui è tratta l'immagine mostrata. Molto interessante, anche l'applet java disponibile presso `https://www.cdt21.com/technical_tools/wave-propagation-calculation-tool`, che grafica l'andamento della attenuazione del modello al variare di alcuni dei parametri sopra illustrati.



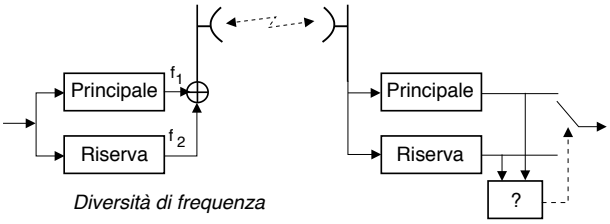
riflessa si sommano costruttivamente, producendo un guadagno anziché ad una attenuazione; aumentando la distanza si assiste ad una attenuazione che cresce con d^2, come per il caso di spazio libero, ma con sovrapposta l'oscillazione illustrata in figura, e che dipende dalla geometria del problema. Oltre una distanza detta *critica*, e che corrisponde alla prima zona di Fresnel, l'attenuazione aumenta con d^4.

Il fading piatto Quando la banda del segnale è sufficientemente piccola rispetto a $\frac{1}{T}$ ed $|H(f)|^2$ si può considerare costante in tale banda (§ 13.1.2.4), l'attenuazione dovuta ai cammini multipli prende il nome di *flat fading* (vedi § 20.4.5). Il 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*. Nel seguito continuiamo a riferirci alle attenuazioni supplementari con il termine più generale di fading.

20.3.3.1 Collegamento in diversità

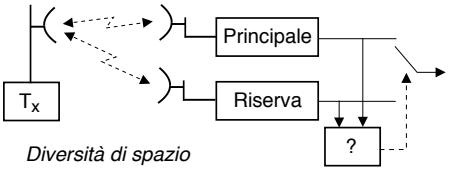
Quando la banda del segnale è sufficientemente estesa da non poter considerare $|H(f)|^2$ costante il *fading* causato dai cammini multipli viene detto *selettivo in frequenza*, potendo le variazioni di $|H(f)|^2$ diventare anche rilevanti quando due repliche del segnale giungono al ricevitore con ampiezze molto simili. Una via per ridurre la probabilità di subire forti attenuazioni a specifiche frequenze è quella di prevedere una *ridondanza* degli apparati, in modo da realizzare vie di collegamento *alternative*. Entrambi gli aspetti illustrati appresso saranno approfonditamente sviluppati ai § successivi.

Diversità di frequenza Ideata per prima in ordine di tempo, consiste nel trasmettere lo stesso messaggio mediante *due* diverse portanti: se una delle due subisce attenuazione, la trasmissione che utilizza l'altra ne è probabilmente esente (o viceversa). Qualora il collegamento tra le antenne sia condiviso tra diverse trasmissioni, una unica *via di riserva* può essere impiegata per fornire una ridondanza $N : 1$. Ad esempio in una trasmissione multiplata FDM (§ 11.1.1.2) la portante di riserva viene assegnata al canale del banco FDM che presenta la maggiore attenuazione.



Diversità di frequenza

Diversità di spazio Trasmettendo invece lo stesso segnale mediante due diverse antenne (riceventi o trasmittenti) collocate in posizioni diverse, le copie del segnale prodotte dai cammini multipli giungono a destinazione con



Diversità di spazio

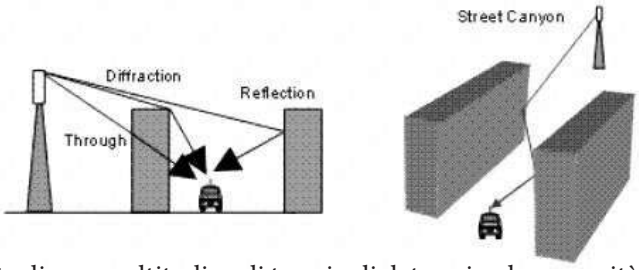
ritardi differenti per le due antenne, e dunque la risposta in frequenza (20.8) è differente nei due casi. Pertanto anche se un ricevitore subisce una attenuazione selettiva ad una determinata frequenza, l'altro ricevitore può esserne esente.

Esempio Utilizzando lo stesso modello di propagazione e gli stessi dati del precedente esempio, valutiamo cosa accade se la riserva viene posta *dieci metri più indietro* dell'antenna principale. In tal caso il nuovo ritardo tra il raggio diretto e quello riflesso diviene pari a $T' = 65.3$ nanosecondi contro i $T = 66.0$ nsec ottenuti per la via principale, e dunque $|H(f)|^2$ per la riserva ha un periodo pari a $1/T' = 15.29$ MHz, una differenza di 140 KHz. Per ottenere che i minimi della $|H(f)|^2$ nei due casi siano distanziati di almeno 3 Mhz, ovvero il 20% del periodo in frequenza, occorre operare con portanti oltre i 300 MHz.

20.4 Collegamenti radiomobili

Le condizioni di propagazione per comunicazioni radiomobili, come nel caso della telefonia cellulare, presentano diversi aspetti particolari che influenzano il fading.

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 Fresnel[21], ed una attenuazione supplementare da assorbimento terrestre. *Inoltre* (specie in ambito urbano) si verifica un elevato numero di cammini multipli e diffrazioni, che per di più 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 differenti[22], e l'attuazione di meccanismi di codifica di canale (§ 17.4) per ridurre gli effetti delle interferenze e del fading variabile[23].



Analizziamo di seguito i fenomeni legati a *posizione* ed *ambiente*, fornendo modelli che descrivono le *attenuazioni supplementari* ed i fenomeni di *multipath variabile*, rimandando la discussione sulle *tecniche di accesso multiplo* ai §§ 16.8.12 e 16.9.2.5.

20.4.1 Le componenti del fading

Al fine di distinguere tra le diverse cause di fading, la rappresentazione grafica del bilancio di collegamento mostrata a pag. 643 può essere re-impostata come illustrato in fig. 20.3, in cui si considera una componente di attenuazione *nominale* A_{pl} indicata come *path loss* (o attenuazione di percorso), e due componenti *aleatorie* di attenuazione

[21] Alla 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.

[22] Vedi ad es. i §§ 11.1.1.3, 16.9.2.5, 16.8.12.

[23] Mentre il fading produce una attenuazione variabile sul segnale, la stessa variabilità delle condizioni di propagazione può portare a livelli di interferenza variabili, causati da altre trasmissioni nella stessa banda. La variabilità temporale della qualità del segnale ricevuto, in particolare quella *veloce* (vedi § 20.4.6), produce errori a *burst*, che possono essere corretti mediante codifica di canale ed interleaving (vedi § 15.6.2.3).

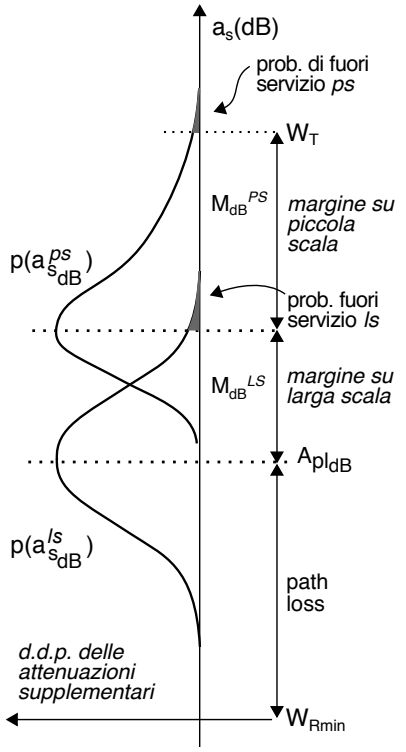


Figura 20.3: Bilancio di collegamento per il caso radiomobile

supplementare legate a posizione e movimento, indicate rispettivamente come *fading* su *larga scala* o *shadowing* (ombreggiatura) a_s^{ls} e fading su *piccola scala* a_s^{ps}.

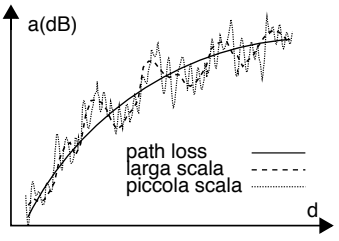
Il valore di attenuazione A_{pl} del *path loss* risulta maggiore di quello A_{sl} di spazio libero (eq. 20.4) a causa delle condizioni di propagazione non ideali, determinando una attenuazione disponibile A_d più elevata, come analizzato al § 20.4.2. L'attenuazione supplementare *su larga scala* a_s^{ls} 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 § 20.4.3 si mostra come il suo valore *in dB* possa considerarsi quello di una v.a. gaussiana a media nulla e varianza σ_{ls}^2, consentendo di determinare il *margine su larga scala* M_{dB}^{ls} come quel valore di $a_{s_{dB}}^{ls}$ che viene superato con probabilità sufficientemente bassa.

La variabilità *su piccola scala* è quella che maggiormente caratterizza il *fading*, e tiene conto degli innumerevoli cammini multipli presenti in ambito urbano ed *indoor*, che possono produrre una attenuazione supplementare $a_{s_{dB}}^{ps}$ 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 $a_{s_{dB}}^{ps}$; 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 § 20.4.4, dove si determina il margine M_{dB}^{ps} necessario a rendere trascurabile la probabilità che $a_{s_{dB}}^{ps} > M_{dB}^{ps}$; mentre ai § 20.4.5 e 20.4.6 si illustrano gli effetti dei fenomeni di variabilità in frequenza e nel tempo.

La fig. 20.3 mostra come queste tre componenti di attenuazione si sommano[24] 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 lato tenta di rappresentare come varia la somma dei tre contributi di attenuazione con la posizione del ricevitore.



[24] Considerando le v.a. statisticamente indipendenti.

20.4.2 Path loss

La dipendenza della attenuazione dal quadrato della distanza espressa dalla (20.4) 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 ricevente[25]. Pertanto il termine $20\log_{10} d\,(Km)$ che compare in (20.6) viene sostituito con $A_{pl} = n \cdot 10\log_{10} d\,(Km) + \alpha$, e quindi in questo caso anziché la (20.6), l'espressione da usare per l'attenuazione disponibile è

$$A_d\,(\text{dB}) = 32.4 + 20\log_{10} f\,(\text{MHz}) + n \cdot 10\log_{10} d\,(Km) + \alpha - G_T\,(\text{dB}) - G_R\,(\text{dB}) \quad (20.9)$$

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.

Esercizio Valutare il path loss per un collegamento a 2 GHz lungo un chilometro, considerando le antenne omnidirezionali, in un ambiente per il quale sono stati stimati i parametri $n = 4$ e $\alpha = 32$.

E' sufficiente applicare la (20.9) utilizzando i valori forniti per i parametri:

$$A_d\,(\text{dB}) = 32.4 + 20\log_{10} 2 \cdot 10^3 + 4 \cdot 10\log_{10} 1 + 32 = 130.4\ \text{dB}.$$

20.4.3 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, misure che in realtà variano spostandosi tra territori diversi, in cui si riscontrano valori di fading diversi, anche per uguali valori di d. Questo fenomeno è 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. Ma anche stando fermi, 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 dB[26] (per questo detto *lognormale*) ed a media nulla, cioè del

[25] Inoltre, la condizione di NLOS introduce una attenuazione supplementare *costante*. Per una rassegna dei diversi modelli di propagazione, si veda ad es.
`http://www.slideshare.net/deepakecrbs/propagation-model`.

[26] La 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'inviluppo complesso $\underline{x}$ del segnale ricevuto è adeguatamente rappresentato da una v.a. di Rice (vedi pag. 429) $\underline{x} = a + \underline{r}$, 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

$$\begin{aligned} a_s(dB) &= 10\log_{10}\tfrac{1}{|a+\underline{r}|^2} = -10\log_{10}\left((a+r_c)^2 + r_s^2\right) = \\ &= -10\log_{10}\tfrac{a^2}{a^2}\left(a^2 + 2ar_c + r_c^2 + r_s^2\right) = 10\left(\log_{10} a^2 + \log_{10}\left(1 + \tfrac{2r_c}{a} + \tfrac{|\underline{r}|^2}{a^2}\right)\right) = \\ &\simeq 10\left(\log_{10} a^2 + \log_{10}\left(1 + \tfrac{2r_c}{a}\right)\right) \simeq 10\left(\log_{10} a^2 + \tfrac{2r_c}{a}\right) = 10\log_{10} a^2 + 20\tfrac{r_c}{a} \end{aligned}$$

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.

tipo

$$p_{A_s}(a_s(dB)) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(a_s(dB))^2}{2\sigma_{ls}^2}\right\}$$

dove σ_{ls} varia tra 6 ed 8 dB per una altezza dell'antenna tra 5 e 15 metri[27]. 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 Una trasmissione LOS per la quale occorre ricevere una potenza di almeno $W_R = -50$ dBm è realizzata mediante un collegamento radio tra antenne omnidirezionali poste a $d = 20$ Km e con portante $f_0 = 27$ MHz. 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 $\sigma_{ls} = 3$ dB. Utilizziamo la (20.6) 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

$$W_T^{slib}(dBm) = W_R(dBm) + A_d(dB) = -50 + 87 = 37\ \text{dBm}$$

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. 154 per cui $0.05 = \frac{1}{2}\text{erfc}\left(\frac{M_{dB}^{ls}}{\sqrt{2}\sigma_{ls}}\right)$, e quindi graficamente si ottiene $\frac{M_{dB}^{ls}}{\sqrt{2}\sigma_{ls}} = 1.5$, da cui $M_{dB}^{ls} = 1.5 \cdot \sqrt{2} \cdot 3 = 1.5 \cdot 1.41 \cdot 3 = 6.3$ dB, che ci consente di calcolare la nuova W_T^{sfad} come $W_T^{sfad}(dBW) = W_T^{slib}(dBW) + M_{dB}^{ls} = 7 + 6.3 = 13.3$ dBW, ovvero $10^{1.33} = 21.4$ Watt.

20.4.4 Fading su piccola scala

Consiste nella fluttuazione di livello del segnale radio osservata durante *il movimento*, e causata dalla variazione dei ritardi con cui i cammini multipli giungono al ricevitore: spostandosi infatti di $\frac{\lambda}{2}$ si può passare[28] da una situazione di somma coerente ad una completa opposizione di fase. Analizziamo ora la situazione dal punto di vista del livello di segnale ricevuto, distinguendo tra i casi di fading *piatto*, di *Rayleigh* e di *Rice*.

Fading piatto Considerando che la (20.7) consente di scrivere il segnale ricevuto come $y(t) = \sum_{n=1}^{N} a_n x(t-\tau_n)$, il relativo inviluppo complesso $\underline{y}(t)$ in presenza di cammini multipli può essere espresso in funzione di quello trasmesso $\underline{x}(t)$ come[29]

[27] Anche se l'aumentare dell'altezza di una antenna ne estende la relativa area di copertura, in ambito urbano questo corrisponde ad una maggiore variabilità delle effettive condizioni operative.

[28] A 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.

[29] La (20.10) discende dal considerare un generico segnale modulato $x(t) = a(t)\cos(2\pi f_0 t + \varphi(t))$ ed il suo inviluppo complesso $\underline{x}(t) = a(t)\,e^{j\varphi(t)}$: per ogni sua replica ritardata $x_n(t) = x(t-\tau_n)$ possiamo scrivere

$$\underline{y}(t) = \sum_{n=1}^{N} a_n \underline{x}(t-\tau_n)\, \mathrm{e}^{-j2\pi f_0 \tau_n} = \sum_{n=1}^{N} a_n \underline{x}(t-\tau_n)\, \mathrm{e}^{-j\varphi_n} \tag{20.10}$$

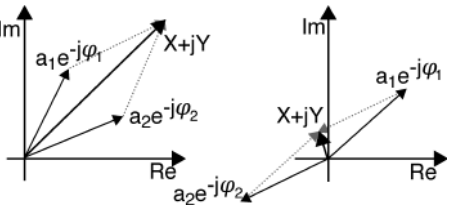
in cui τ_n è il *ritardo* dell'n-esimo cammino, a_n il rispettivo *guadagno*, e $\varphi_n = 2\pi f_0 \tau_n$ la *rotazione* del associata. Se durante il tempo che intercorre tra l'arrivo della prima replica (ritardata di τ_{min}) e l'arrivo dell'ultima (ritardata di τ_{max}) il segnale $\underline{x}(t)$ non varia di molto (e cioè $\underline{x}(t-\tau_{min}) \simeq \underline{x}(t-\tau_n) \simeq \underline{x}(t-\tau_{max})$)[30] il risultato equivale alla moltiplicazione di $\underline{x}(t)$ per un numero complesso, senza quindi produrre distorsione lineare (vedi § 13.1.2.4). Infatti in tal caso la (20.10) può essere riscritta come

$$\begin{aligned} \underline{y}(t) &\simeq \underline{x}(t) \sum_{n=1}^{N} a_n \mathrm{e}^{-j\varphi_n} = \underline{x}(t) \sum_{n=1}^{N} a_n (\cos\varphi_n - j\sin\varphi_n) \\ &= \underline{x}(t) \cdot (X + jY) = \underline{x}(t) \cdot \rho \mathrm{e}^{j\varphi} \end{aligned} \tag{20.11}$$

in cui il valore complesso

$$X + jY = \rho \mathrm{e}^{j\varphi} = \sum_{n=1}^{N} a_n \cos\varphi_n - j \sum_{n=1}^{N} a_n \sin\varphi_n$$

riassume l'effetto delle diverse repliche e costituisce una 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 elevate[31], bastano piccole variazioni di ritardo τ_n per produrre una fase $\varphi_n = 2\pi f_0 \tau_n$ (nota 29) uniformemente distribuita tra 0 e 2π e del tutto indipendente per le diverse repliche. Pertanto 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 (§ 6.7.2), e quindi i valori di X ed Y nella (20.11) possono considerarsi realizzazioni di v.a. indipendenti, gaussiane, a media nulla ed uguale varianza σ^2.



Fading di Rayleigh Consideriamo ora *l'ampiezza* $\left|\underline{y}(t)\right|$ del segnale ricevuto, che dalla (20.11) risulta pari a $\left|\underline{y}(t)\right| = \rho \cdot \left|\underline{x}(t)\right|$, in cui nelle condizioni descritte $\rho = \sqrt{X^2 + Y^2}$ è una v.a. di Rayleigh (pag. 427) la cui d.d.p. ha espressione

$$p_P(\rho) = \frac{\rho}{\sigma^2} \mathrm{e}^{-\frac{\rho^2}{2\sigma^2}} \tag{20.12}$$

$x_n(t) = a(t-\tau_n)\cos\left[2\pi f_0 (t-\tau_n) + \varphi(t-\tau_n)\right] = a(t-\tau_n)\cos\left(2\pi f_0 t - 2\pi f_0 \tau_n + \varphi(t-\tau_n)\right)$
ed il cui inviluppo complesso rispetto ad f_0 può quindi essere espresso come

$$\underline{x}_n(t) = \underbrace{a(t-\tau_n)\, e^{j\varphi(t-\tau_n)}}_{\underline{x}(t-\tau_n)} e^{-j2\pi f_0 \tau_n} = \underline{x}(t-\tau_n)\, e^{-j2\pi f_0 \tau_n}$$

[30] Si consideri che il risultato dell'esempio di pag. 677 valuta i ritardi in gioco dell'ordine di grandezza delle decine di nanosecondi, mentre (ad esempio) ad un segnale $\underline{x}(t)$ limitato in banda a 10 KHz corrisponde un periodo di campionamento $T_c = 50\ \mu$sec.

[31] Se ad esempio i ritardi τ_n sono dell'ordine di 10^{-8}, l'ipotesi è valida per $f_0 > 100$ MHz, quasi 1/10 delle frequenze a cui operano i radiomobili.

con $\rho \geq 0$. Il valore della *potenza istantanea* ricevuta, legata[32] a $\left|\underline{y}(t)\right|^2 = \rho^2 \left|\underline{x}(t)\right|^2$, risulta pertanto variato di una quantità pari a ρ^2, che è una v.a. *esponenziale negativa*[33], descritta dalla d.d.p. (vedi § 22.2.1)

$$p_E\left(\rho^2\right) = \lambda \mathrm{e}^{-\lambda\rho^2} = \frac{1}{2\sigma^2}\mathrm{e}^{-\frac{\rho^2}{2\sigma^2}} \tag{20.13}$$

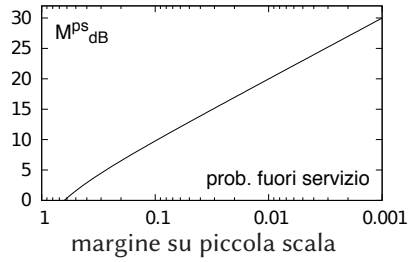
dove si è posto in evidenza il valor medio

$$m_{\rho^2} = E\left\{\rho^2\right\} = 1/\lambda = 2\sigma^2$$

In base alla (20.13) è possibile determinare il margine M_{dB}^{ps} necessario a contrastare un fading di Rayleigh, qualora si desideri una *probabilità di fuori servizio* pari a p[34]:

$$M_{dB}^{ps} = -10\log_{10}\left(-\ln\left(1-p\right)\right) \tag{20.14}$$

il cui andamento è mostrato a lato al variare del grado di servizio.



Qualora trasmettitore, ricevitore ed ambiente siano *statici*, ρ assume un unico valore casuale distribuito come indicato dalla (20.12). Se invece (ad es.) il ricevitore è in movimento i cammini multipli si modificano nel tempo, e la figura 20.4 mostra come varia la potenza in dB del segnale ricevuto, relativamente alle condizioni di ricezione *medie* (ovvero su larga scala, rappresentate dalla condizione di zero dB), per posizioni via via

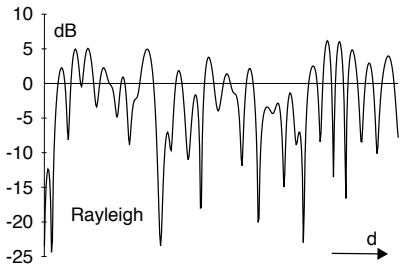
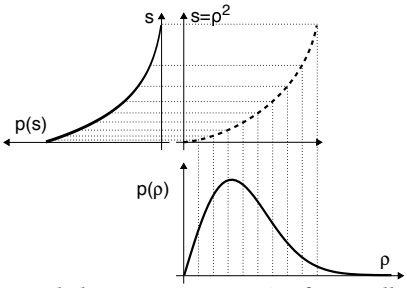


Figura 20.4: Intensità del segnale con fading di Rayleigh

[32] Per semplicità nel seguito consideriamo $\underline{x}(t)$ a potenza unitaria, in modo che ρ^2 sia proprio la potenza istantanea ricevuta.

[33] Impostando il cambiamento di variabile $s = \rho^2$ si possono applicare le regole viste al § 6.4, 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

$$p_S(s) = p_P\left(\sqrt{s}\right)\cdot\frac{d}{ds}\rho(s) = \frac{\sqrt{s}}{\sigma^2}\exp\left(-\frac{\left(\sqrt{s}\right)^2}{2\sigma^2}\right)\cdot\frac{1}{2\sqrt{s}} = \\ = \frac{1}{2\sigma^2}\exp\left(-\frac{s}{2\sigma^2}\right)$$

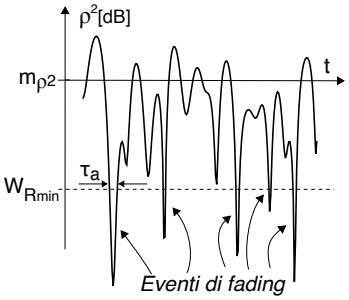


In figura si mostra il processo di costruzione grafica che produce una d.d.p. esponenziale negativa a partire dal quadrato di una d.d.p. di Rayleigh.

[34] A tal fine osserviamo che il collegamento va fuori servizio quando la potenza ricevuta è inferiore alla sensibilità del ricevitore $W_{R_{min}}$, e la probabilità di questo evento si esprime come $p = Pr\left(\rho^2 < W_{R_{min}}\right) = 1 - \exp\left(-\frac{W_{R_{min}}}{m_{\rho^2}}\right)$, essendo appunto ρ^2 una v.a. a d.d.p. esponenziale con media $m_{\rho^2} = 2\sigma^2$, e tenendo conto dell'eq. (22.3) a pag. 771. Al tempo stesso, $m_{\rho^2} = E\left\{\rho^2\right\}$ rappresenta la potenza *media* ricevuta, ovvero lo *zero dB* di fig. 20.4: esprimendo dunque il margine M (non in dB) come il rapporto tra la potenza media ricevuta e la sensibilità del ricevitore $M = \frac{m_{\rho^2}}{W_{R_{min}}}$, si ottiene $p = 1 - \mathrm{e}^{-\frac{1}{M}}$, e quindi $-\frac{1}{M} = \ln\left(1-p\right)$ e, passando ai decibel, $-10\log_{10} M = 10\log_{10}\left(-\ln\left(1-p\right)\right)$, da cui la (20.14).

più distanti: si nota chiaramente come la potenza possa diminuire anche di molto, condizione indicata come *deep fade*.

Frequenza e durata media del fading Se è presente movimento a *velocità costante* la fig. 20.4 rappresenta altrettanto bene l'andamento di ρ^2/m_{ρ^2} (dB) in funzione del tempo. In tal caso è interessante valutare *per quanto tempo* la potenza istantanea ρ^2 del segnale ricevuto *scende sotto* la soglia $W_{R_{min}}$, e dunque valutare quanti bit, ricadendo in tale intervallo temporale, saranno soggetti ad una P_e peggiore di quella desiderata. Come osservato alla nota 34 la probabilità che ρ^2 sia minore di $W_{R_{min}}$ vale



$$p = Pr\left(\rho^2 < W_{R_{min}}\right) = 1 - \exp\left(-\frac{W_{R_{min}}}{m_{\rho^2}}\right) \qquad (20.15)$$

e la durata media $\bar{\tau}_a$ di questo evento si ottiene dividendo p per il *numero medio* N_a di affievolimenti per secondo[35], ovvero $\bar{\tau}_a = \frac{p}{N_a}$. D'altra parte, si può mostrare che risulta

$$N_a = \sqrt{2\pi} f_D \alpha \mathrm{e}^{-\alpha^2} \qquad (20.16)$$

in cui si è posto $\alpha^2 = \frac{W_{R_{min}}}{m_{\rho^2}} = \frac{1}{M^{ps}}$, mentre f_D è la massima *deviazione doppler* (pag. 691) che come vedremo è direttamente legata alla velocità di movimento: infatti, per velocità maggiori aumenta la frequenza dei fenomeni di fading. Combinando le (20.15) e (20.16) si ottiene pertanto

$$\bar{\tau}_a = \frac{p}{N_a} = \frac{1 - \mathrm{e}^{-\alpha^2}}{\sqrt{2\pi} f_D \alpha e^{-\alpha^2}} = \frac{\mathrm{e}^{\alpha^2} - 1}{\sqrt{2\pi} f_D \alpha} \qquad (20.17)$$

il cui andamento *normalizzato* è rappresentato nella figura 20.5 assieme a quello di N_a, al variare di α ovvero di $M^{ps}_{dB} = 10 \log_{10} \frac{1}{\alpha^2} = -20 \log_{10} \alpha$.

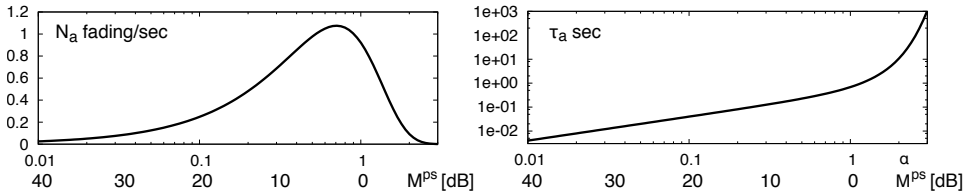


Figura 20.5: Frequenza e durata media del fading di Rayleigh per f_D=1 Hz in funzione di α ovvero di M^{ps}_{dB}

Esercizio Valutare la durata media del fading di Rayleigh in presenza di doppler $f_D = 20$ Hz e di un margine $M^{ps}_{dB} = 20$ dB. Consideriamo quindi errato un bit se durante il suo periodo T_b si verifica un affievolimento che rende la potenza istantanea ricevuta minore

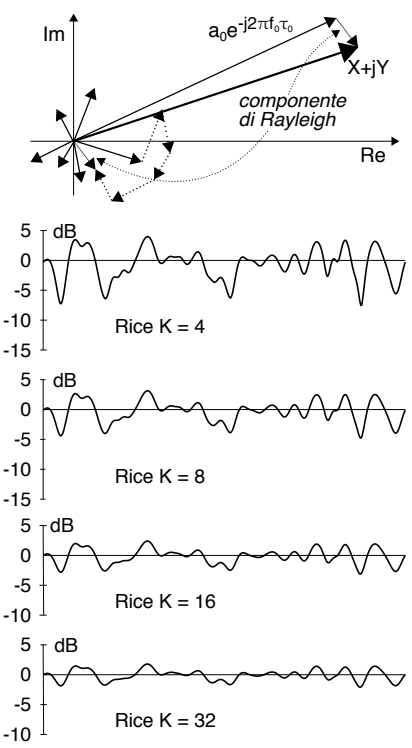
[35] Infatti in tal modo la percentuale di tempo p viene *spalmata* su di un secondo, e suddivisa per il numero (medio) di volte (in un secondo) per cui avviene che $\rho^2 < W_{R_{min}}$. **Esempio** Se $p = 0.1$ ed $N_a = 5$ fading/sec allora $\bar{\tau}_a = 0.1/5 = 0.02$, ossia 20 msec, ripartendo i 100 msec (10% di 1 secondo) sui 5 affievolimenti medi.

di quella media per più di M_{dB}^{ps}. In presenza di una modulazione BPSK a velocità $f_b = 50$ bit/sec, quanti sono in media i bit errati per secondo, e la corrispondente P_e^{bit}?

Ad un $M_{dB}^{ps} = 20$ dB corrisponde $\alpha = 0.1$, ed in base alla (20.17) si ottiene $\bar{\tau}_a = 2$ msec, minore di $T_b = 1/50 = 20$ msec, e quindi l'intervallo temporale per cui il fading è maggiore del margine, interessa un solo bit. Mediante la (20.16) (con $\alpha = 0.1$ e $f_D = 20$ Hz) si ottiene che $N_a = 4.96$ fading/sec, e dunque in un secondo risultano errati quasi 5 bit su 50, ovvero $P_e^{bit} = 5/50 = 0.1$.

Come evidente, ottenere il margine a partire dal % di fuori servizio (eq. (20.14)), e poi dal margine risalire alla P_e, è un procedimento un po' contorto. Un'elegante alternativa che non richiede di passare dal margine viene esposta all'appendice 20.5.1.

Fading di Rice Si verifica nel caso in cui le ampiezze a_n dei diversi percorsi che compaiono nella (20.10) non sono identicamente distribuite, ma ne esiste una (a_0 in figura) che *prevale* su tutte le altre, 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. (14.16) a pag. 429, essendo la risultante $X + Y$ tipicamente ora *vicina* al cammino prevalente $a_0 e^{-j2\pi f_0 \tau_0}$. In particolare il rapporto $K = a_0^2/2\sigma^2$ tra la potenza $a_0^2/2$ dell'onda diretta e quella σ^2 della componente dovuta al multipath prende il nome di *fattore di Rice*, e nella figura a lato si mostra come in presenza di una forte componente diretta la profondità del fading si riduca sensibilmente.

Effettivamente in corrispondenza di un K elevato il fading di Rice può essere descritto nei termini di un fading su larga scala (§ 20.4.3), come discusso alla nota 26. Viceversa qualora la ricezione avvenga principalmente in *assenza di visibilità* i valori del modulo dell'inviluppo complesso del segnale $\rho(t) = \left|\underline{y}(t)\right|$ sono soggetti al fading di Rayleigh precedentemente discusso.

20.4.5 Fading selettivo in frequenza

Individua il caso in cui $H(f)$ non può essere considerata costante, equivalente[36] a rimuovere l'ipotesi fatta a pag. 682 e dunque accettare che nell'intervallo temporale $\Delta\tau = \tau_{max} - \tau_{min}$ tra l'arrivo della prima e dell'ultima replica, detto anche *dispersione temporale*, il segnale possa modificare il suo valore. Per analizzare cosa succede par-

[36] Visto che la causa nota per cui $H(f) \neq cost$ è l'eccessiva banda del segnale, a ciò corrisponde un maggior contenuto di alte frequenze e dunque una maggiore velocità di variazione temporale.

tiamo dalla (20.10) per scrivere l'espressione dell'inviluppo complesso della risposta impulsiva del canale come[37]

$$\underline{h}(t) = \sum_{n=0}^{N-1} a_n \delta(t - \tau_n)\, \mathrm{e}^{-j2\pi f_0 \tau_n} = \sum_{n=0}^{N-1} Z_n \delta(t - \tau_n) \tag{20.18}$$

in cui si è posto $Z_n = a_n \mathrm{e}^{-j2\pi f_0 \tau_n}$. Facciamo quindi l'ipotesi semplificatrice che i ritardi τ_n siano multipli di un comune intervallo T, cioè $\tau_n = nT$, considerando eventualmente nullo qualche valore a_n: in tal modo la (20.18) può essere assimilata all'espressione di un segnale campionato (§ 4.1) $\underline{h}^{\bullet}(t) = \sum_{n=0}^{N-1} Z_n \delta(t - nT)$, interpretando dunque i coefficienti complessi Z_n come campioni di un processo $Z(t)$[38], ovvero $Z_n = Z(nT) = a_n \mathrm{e}^{-j2\pi f_0 nT}$. Ciò consente di esprimere la risposta in frequenza equivalente di b.f. del canale come la DTFT (vedi § 4.4) della sequenza Z_n, ovvero

$$\underline{H}(f) = \sum_{n=0}^{N-1} Z_n \mathrm{e}^{-j2\pi f nT} \tag{20.19}$$

Notiamo ora che i valori di $\underline{H}(f)$ in funzione di f sono variabili aleatorie, dipendendo dalle caratteristiche statistiche dei termini Z_n che per i motivi illustrati a pag. 683 sono v.a. complesse, indipendenti ed a valor medio nullo, e quindi (vedi § 7.5.3)

$$E\left\{Z_n^* Z_{n+m}\right\} = \begin{cases} 0 & se\ m \neq 0 \\ \sigma_{a_n}^2 & altrimenti \end{cases} \tag{20.20}$$

in cui la sequenza di valori $\sigma_{a_n}^2 = E\left\{a_n^2\right\}$ è indicata nel seguito come...

Dispersione potenza-ritardo[39] E' costituita dalla sequenza $\mathcal{P}_n = E\left\{a_n^2\right\}$ e rappresenta la distribuzione temporale (media) della potenza (o energia) delle repliche del segnale. Infatti, trasmettendo un impulso di energia unitaria $\delta(t)$ si ricevono N impulsi di energia $\mathcal{E}_n = a_n^2$, ovvero viene ricevuto l'inviluppo complesso $\underline{h}(t)$ espresso dalla (20.18), la cui energia vale

$$\mathcal{E}_{\underline{h}} = \int_{-\infty}^{\infty} \underline{h}^*(t)\, \underline{h}(t)\, dt = \sum a_n^2 = \sum \mathcal{E}_n$$

ed il cui valore atteso rispetto all'aleatorietà degli a_n risulta $E\left\{\mathcal{E}_{\underline{h}}\right\} = \sum E\left\{a_n^2\right\} = \sum \mathcal{E}_n$.

Misura della dispersione potenza-ritardo Può essere portata a termine con tre diverse tecniche, di cui ci si limita ad accennare i principi operativi:

- un *primo* metodo consiste nel trasmettere una portante modulata in ampiezza da impulsi molto brevi, ottenendo dopo demodulazione la convoluzione tra l'impulso usato in trasmissione e l'$h(t)$ del canale: benché questa soluzione sia molto semplice, è affetta sia dal rumore a larga banda che *entra* nel passa-banda di ricezione, sia dalle interferenze presenti;
- una *seconda* tecnica fa invece uso di una segnale DSSS, il cui despreading in ricezione avviene variando di volta in volta la fase della PN: quando questa risulta allineata

[37] Il cambiamento negli indici della sommatoria è legato a considerare l'origine dei tempi in corrispondenza al primo arrivato dei cammini multipli.

[38] Si sottintende che T sia minore dell'inverso del doppio della banda di $Z(t)$, ovvero $T < 1/2W$.

[39] Libera traduzione del termine POWER DELAY SPREAD.

temporalmente con una delle repliche dovute al multipath a valle del filtro passabasso si rivela *un massimo* con ampiezza legata ad a_n. In tal modo la sensibilità al rumore viene ridotta dal guadagno di processo, ma la misura richiede il tempo necessario a *provare* tutte le fasi della PN;

- l'*ultimo* metodo opera nel dominio della frequenza e si basa su diverse frequenze trasmesse una alla volta, la cui ampiezza e fase viene confrontata con quella ricevuta, come illustrato a pag. 74; i campioni della $H(f)$ così ottenuti sono quindi antitrasformati mediante IDFT, per ottenere i campioni di $h(t)$. Ma per effettuare il confronto, occorre che trasmettitore e ricevitore siano fisicamente vicini, e dunque il metodo è applicabile solo per ambiti *indoor*.

Una volta pervenuti alla misura della sequenza delle ampiezze a_n^2 l'operazione è ripetuta più volte spostandosi di poco alla volta[40], ed alla fine i risultati sono mediati tra loro in modo da ottenere una stima di $\mathcal{P}_n = \sigma_{a_n}^2 = E\left\{a_n^2\right\}$.

Da questa si possono derivare parametri statistici come il *ritardo medio*

$$\bar{\tau} = \frac{\sum_n \mathcal{P}_n \tau_n}{\sum_n \mathcal{P}_n}$$

e la *deviazione standard* dei ritardi

$$\sigma_\tau = \sqrt{\overline{\tau^2} - \bar{\tau}^2}$$

in cui $\overline{\tau^2} = \frac{\sum_n \mathcal{P}_n \tau_n^2}{\sum_n \mathcal{P}_n}$, mentre la *dispersione temporale*

$$\Delta\tau = \tau_{max} - \tau_{min}$$

è definita con riferimento ad una soglia che permette di distinguere le repliche dal rumore.

La figura 20.6 mostra la curva di *dispersione potenza-ritardo* misurata per un ambiente al coperto, per il quale sono calcolate $\bar{\tau}$, σ_τ e $\Delta\tau$ per una soglia di -10 dB. In appendice 20.5.4 sono riportati alcuni valori tipici di questi parametri per diversi contesti ambientali. L'andamento tendenziale rilevato per le $\mathcal{P}_n$ misurate suggerisce l'approssimazione della dispersione potenza-ritardo mediante una densità esponenziale:

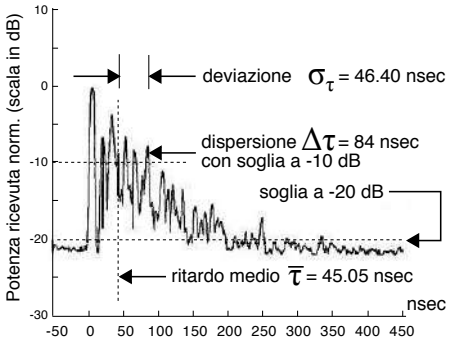


Figura 20.6: Profilo di dispersione potenza-ritardo per ambito indoor

$$\mathcal{P}(\tau) = \frac{1}{\sigma_\tau} \exp\left(-\frac{\tau}{\sigma_\tau}\right) \quad \text{ovvero} \quad \mathcal{P}_n = \frac{1}{\sigma_\tau} \exp\left(-\frac{nT}{\sigma_\tau}\right) \tag{20.21}$$

Banda di coerenza Per poter descrivere la $H(f)$ definita dalla (20.19) ma in cui i termini Z_n sono aleatori, impostiamo l'analisi con lo scopo di valutare per quale intervallo Δf si ottengano coppie di valori della risposta in frequenza $(\underline{H}(f), \underline{H}(f + \Delta f))$ che iniziano a divenire *incorrelati*, dato che in tal caso un segnale che occupa una banda

[40] Tipicamente di 1/4 della lunghezza d'onda relativa alla portante adottata.

comparabile a Δf è affetto da distorsione lineare. A tal fine, partendo dalla (20.19) interpretiamo $\underline{H}(f)$ come un processo ad *aleatorietà parametrica* (§ 6.3.7) in *frequenza*, e dunque ne calcoliamo la funzione di autocorrelazione (appunto, in frequenza):

$$\begin{aligned}\mathcal{R}_{\underline{H}}(\Delta f) &= E\left\{\underline{H}^*(f)\,\underline{H}(f+\Delta f)\right\} = \\ &= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} E\left\{Z_n^* Z_m\right\} \mathrm{e}^{j2\pi f nT}\mathrm{e}^{-j2\pi(f+\Delta f)mT} = \sum_{n=0}^{N-1}\mathcal{P}_n \mathrm{e}^{-j2\pi\Delta f nT} \qquad (20.22)\end{aligned}$$

in cui all'ultimo passaggio si è applicata la (20.20) considerando $m = n$: $\mathcal{R}_{\underline{H}}(\Delta f)$ è dunque pari alla trasformata di Fourier di sequenze (§ 4.4) della dispersione potenza-ritardo $\mathcal{P}_n = \sigma_{a_n}^2 = E\left\{a_n^2\right\}$.

La *banda di coerenza* B_c è quindi definita come l'intervallo di frequenze Δf entro cui $\underline{H}(f)$ si mantiene *correlata*, e può essere fatto corrispondere alla *larghezza di banda* di $\mathcal{R}_{\underline{H}}(\Delta f)$. Pertanto, quanto più la dispersione temporale $\Delta\tau$ (o, più in generale, la deviazione σ_τ) risulta elevata, tanto minore sarà il valore di B_c. Convenzionalmente una sua valutazione approssimata ricade nell'intervallo

$$\frac{1}{50\,\sigma_\tau} \le B_c \le \frac{1}{5\,\sigma_\tau} \qquad (20.23)$$

Esempio Consideriamo un canale radio in un contesto urbano, caratterizzato da una deviazione standard dei ritardi $\sigma_\tau = 5$ μsec, e per il quale si assume valida l'approssimazione del profilo di dispersione potenza-ritardo esponenziale (20.21), ovvero $\mathcal{P}(\tau) = \frac{1}{\sigma_\tau} e^{-\tau/\sigma_\tau}$. L'applicazione della (20.23) porta ad una stima di B_c compresa tra 4 e 40 KHz.

Dato che[41] $\mathcal{R}_{\underline{H}}(\Delta f) = \mathcal{F}\{\mathcal{P}(\tau)\} = \frac{1}{1+j2\pi\sigma_\tau\Delta f}$, osserviamo che questa ha il massimo nell'origine (vedi fig. 4.13 a pag. 108), ed il suo modulo si dimezza[42] per $\Delta f = \frac{1}{3.63\sigma_\tau}$: pertanto la scelta $B_c = 1/5\sigma_\tau = 40$ KHz corrisponde ad una correlazione in frequenza maggiore di 0.5. Lo stesso calcolo mostra che scegliere invece la stima più restrittiva $B_c = 1/50\sigma_\tau = 4$ KHz corrisponde ad una correlazione $\left|\mathcal{R}_{\underline{H}}(\Delta f)\right| > 0.9$ (per l'esattezza, si ottiene $\left|\mathcal{R}_{\underline{H}}(\Delta f = 1/50\sigma_\tau)\right| = 0.94$).

Ricapitolando se la banda W del segnale modulato non eccede B_c ci si trova nelle condizioni di fading *piatto*, mentre se $W > B_c$ 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 che $T_s \simeq \frac{1}{W} > \frac{1}{B_c} > \sigma_\tau$, ovvero la deviazione standard dei ritardi è ben inferiore al periodo di simbolo, limitando gli effetti dell'ISI.

[41]
$$\begin{aligned}\mathcal{F}\{\mathcal{P}(\tau)\} &= \int_{-\infty}^{\infty}\mathcal{P}(\tau)\,e^{-j2\pi f\tau}d\tau = \frac{1}{\sigma_\tau}\int_0^\infty e^{-\frac{\tau}{\sigma_\tau}}e^{-j2\pi f\tau}d\tau = \frac{1}{\sigma_\tau}\int_0^\infty e^{-\left(\frac{1}{\sigma_\tau}+j2\pi f\right)\tau}d\tau = \\ &= \frac{1}{\sigma_\tau}\frac{-1}{\frac{1}{\sigma_\tau}+j2\pi f}\,e^{-\left(\frac{1}{\sigma_\tau}+j2\pi f\right)\tau}\Big|_0^\infty = \frac{1}{\sigma_\tau}\frac{1}{\frac{1}{\sigma_\tau}+j2\pi f} = \frac{1}{1+j2\pi\sigma_\tau f}\end{aligned}$$

[42] Si ha $\left|\mathcal{R}_{\underline{H}}(\Delta f)\right| = \frac{1}{2}$ se $\sqrt{1+(2\pi\sigma_\tau\Delta f)^2} = 2$, dunque $2\pi\sigma_\tau\Delta f = \sqrt{3}$ ovvero $\Delta f = \frac{1.73}{6.28\sigma_\tau} = \frac{1}{3.63\sigma_\tau}$

Conseguenze e rimedi Dato che la correzione degli effetti di distorsione lineare e ISI richiede al ricevitore complesse operazioni di equalizzazione (§ 18.4) 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. Un conveniente *escamotage* è l'adozione di una trasmissione OFDM che suddivide W in tante sotto-bande più piccole, e adotta un $T_s > \sigma_\tau$. Occupare una banda $W > B_c$ è possibile anche ricorrendo alla modulazione DSSS, dato che in tal caso al § 20.5.2 mostreremo come poter evitare uno stadio di equalizzazione *classico* adottando una speciale architettura di ricevitore detta *Rake*. L'adozione infine di *più antenne* in trasmissione e/o ricezione (cap. 21) offre una ulteriore via per rendere i fenomeni di attenuazione selettiva *un vantaggio* del collegamento.

Dimensione di cella e velocità trasmissiva Per celle molto grandi la differenza di percorso tra cammini multipli può essere notevole (vedi § 20.5.4), determinando una B_c ridotta, e quindi una bassa velocità di trasmissione. Riducendo la dimensione di cella è possibile aumentare la velocità, dato che le differenze di ritardo si riducono. Pertanto se celle con raggio di chilometri e $\triangle\tau > 10$ μsec possono richiedere equalizzazione anche per trasmissioni a 64 kbps, al contrario comunicazioni *indoor* con $\triangle\tau < 1$ μsec possono presentare *flat fading* per velocità dell'ordine del 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.

20.4.6 Dispersione spettrale e variabilità temporale

Addentriamoci nella descrizione dell'effetto prodotto dal *movimento*. 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, sia esso di Rayleigh o di Rice; in tal caso il fading (piatto o selettivo) è *costante* nel tempo. Viceversa nel caso in cui ci sia movimento[43] l'inviluppo complesso ricevuto (20.10) si riscrive come

$$\underline{y}(t) = \sum_{n=1}^{N} a_n(t)\, \underline{x}(t - \tau_n(t))\, \mathrm{e}^{-j2\pi f_0 \tau_n(t)}$$

evidenziando come ora sia le ampiezze a_n che i ritardi τ_n dipendono dal tempo.

Allo scopo di analizzare le conseguenze di questa *non stazionarietà*, consideriamo una portante non modulata $x(t) = \cos 2\pi f_0 t$, con inviluppo complesso $\underline{x}(t) = 1$[44], che produce la ricezione di

$$\underline{y}(t) = \sum_{n=1}^{N} a_n(t)\, \mathrm{e}^{-j2\pi f_0 \tau_n(t)} = \sum_{n=1}^{N} a_n(t)\, \mathrm{e}^{-j\alpha_n(t)} \tag{20.24}$$

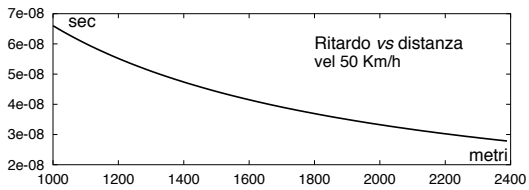
Si ottengono quindi N diversi segnali *modulati* sia in ampiezza che angolarmente, anche se è stata trasmessa una sola frequenza. In generale le ampiezze $a_n(t)$ non

[43] Del ricevitore, del trasmettitore, o degli oggetti riflettenti.

[44] Come evidente dalla eq. (11.3) a pag. 345

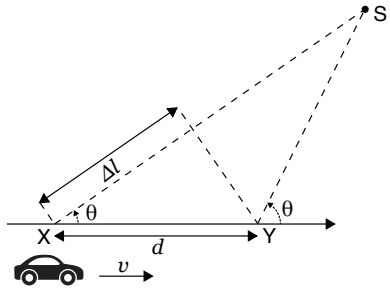
variano di molto con il movimento, mentre come già osservato sono sufficienti piccole variazioni di $\tau_n(t)$ per causarne di grandi per $\alpha_n(t) = 2\pi f_0 \tau_n(t)$: ad esempio, con una $f_0 = 1$ GHz basta la variazione di τ pari ad 1 nsec per produrre una rotazione di 2π.

Esempio Riprendiamo i dati ed il modello usati a pag. 677 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.



Effetto Doppler A lato è raffigurato un mobile che viaggia a velocità costante v ed impiega $\Delta t = d/v$ secondi per spostarsi tra i punti X ed Y distanti d, mentre riceve una portante a frequenza $f_0 = c/\lambda$ dalla sorgente S. La differenza di distanza Δl dalla sorgente nei due punti risulta[45]

$$\Delta l = d \cos\theta = v\Delta t \cos\theta$$



e quindi la differenza di fase nel segnale ricevuto in X e Y vale[46]

$$\Delta\alpha = \frac{2\pi\Delta l}{\lambda} = \frac{2\pi v\Delta t}{\lambda}\cos\theta \tag{20.25}$$

Pertanto durante il tragitto la frequenza ricevuta differisce da f_0 per una quantità[47]

$$f_d = \frac{1}{2\pi}\frac{\Delta\alpha}{\Delta t} = \frac{v}{\lambda}\cos\theta = \frac{c}{c}\frac{v}{\lambda}\cos\theta = f_0\frac{v}{c}\cos\theta \tag{20.26}$$

denominata *scostamento Doppler*[48].

Dispersione Doppler Se al posto di una singola sorgente S sono presenti tutti gli N riflettori che danno origine al multipath l'effetto Doppler si verifica per ciascuno di essi, causando la ricezione di N diverse frequenze $f_n = f_0 \pm f_d^n$, ognuna aumentata (o diminuita) rispetto alla portante f_0 della *frequenza Doppler*

$$f_d^n = f_0\frac{v}{c}\cos\theta_n \tag{20.27}$$

[45] Approssimiamo θ come uguale in X e Y, nell'ipotesi che S sia molto lontana rispetto a d.

[46] Il rapporto $n = \triangle l/\lambda$ indica quanti periodi di portante entrano in $\triangle l$, che moltiplicato per 2π fornisce appunto la differenza tra le fasi di arrivo, nulla se n è intero.

[47] La (20.26) si ottiene applicando alla (20.25) la definizione di deviazione di frequenza f_d come differenza $f_d(t) = f_i(t) - f_0$ in cui la frequenza istantanea f_i è data dalla (12.12) come $f_i(t) = f_0 + \frac{1}{2\pi}\frac{d}{dt}\alpha(t)$

[48] Si tratta dello stesso effetto che produce la variazione del suono della sirena di un mezzo di soccorso, vedi `http://it.wikipedia.org/wiki/Effetto_Doppler`

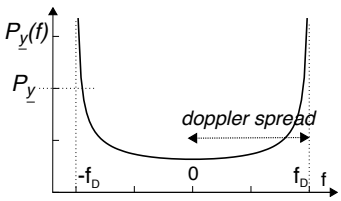
in cui θ_n è l'angolo tra la direzione del moto e la congiungente con il riflettore[49]. Con i dati dell'esempio precedente (relativo ad un moto con $v = 50$ Km/h ovvero 13.8 m/sec) e ponendo $f_0 = 1$ GHz si ottiene una f_d^n *massima* di 46.3 Hz, relativa al caso di $\theta_n = 0$[50]; indichiamo con

$$f_D = \max_n \left\{ f_d^n \right\} = f_0 \frac{v}{c}$$

tale valore. 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*.

Dispersione spettrale e variabilità temporale Considerando il mobile raggiunto da infiniti percorsi con direzione di arrivo distribuita uniformemente (condizione di *scattering isotropo*), si può mostrare[51] che la densità spettrale ricevuta a partire da una singola portante trasmessa è pari a

$$\mathcal{P}_{\underline{y}}(f) = \begin{cases} \dfrac{\mathcal{P}_{\underline{y}}}{\pi f_D \sqrt{1-\left(\frac{f}{f_D}\right)^2}} & \text{con } |f| \le f_D \\ 0 & \text{altrove} \end{cases} \qquad (20.28)$$



mostrata a lato, e del tutto simile a quella di pag. 163.

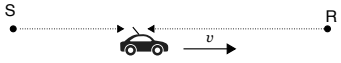
La dispersione Doppler f_D costituisce nella pratica una misura della *velocità di variazione* del canale[52], come già evidenziato in relazione alla frequenza degli affievolimenti di cui all'eq. (20.16). Infatti l'inviluppo complesso ricevuto $\underline{y}(t)$ descritto dalla (20.24) è il risultato 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 rotazione 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 Dettagliamo meglio il legame tra la dispersione Doppler f_D ed una valutazione quantitativa del tempo per cui il canale può essere considerato stabile, di nuovo ricorrendo a considerazioni di tipo statistico. Calcoliamo a tal fine l'antitra-

[49] La stessa analisi è valida anche nel caso di un ricevitore *fermo* ma con i riflettori in movimento, come per la la *riflessione ionosferica*: in tal caso l'espressione si scrive come $f_d^n = f_0 \frac{v_n}{c} \cos\theta_n$, considerando cioè la possibilità che i riflettori abbiano velocità diverse tra loro.

[50] Notiamo che se $\theta_n = 0$ ci stiamo riferendo al caso in cui il moto si realizza lungo la congiungente tra ricevitore e sorgente (o riflettore).

[51] Notiamo che il risultato è diretta conseguenza della condizione di scattering *isotropo*: infatti la (20.27) costituisce un *processo armonico* (pag. 163) quando $-\pi < \theta_n < \pi$ con d.d.p. uniforme, ed al tempo stesso rappresenta la deviazione della frequenza istantanea f_i rispetto ad f_0 (§ 11.2.2), e dunque si verifica l'effetto di conversione AM-FM descritto al § 12.3.3.3. Se viceversa esistono ad es. due soli cammini, il primo diretto (S) e l'altro riflesso (R) con il mobile nel mezzo, $P_{\underline{y}}(f)$ corrisponde a due impulsi in $\pm f_D$.



[52] In questo modo si ottiene una trattazione unificata sia per il caso di un ricevitore 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.

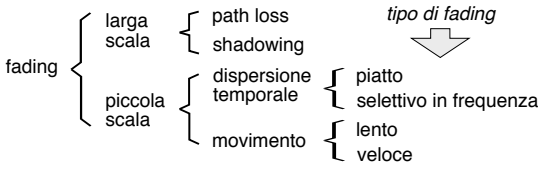
sformata di $\mathcal{P}_{\underline{y}}(f)$, ovvero l'autocorrelazione $\mathcal{R}_{\underline{yy}}(\tau)$ di $\underline{y}(t)$, che nel caso della (20.28) fornisce $\mathcal{R}_{\underline{yy}}(\tau) = J_o(2\pi f_D \tau)$ in cui J_0 è la funzione di Bessel del primo tipo di ordine zero graficata a pag. 387. Sappiamo poi che una correlazione nulla corrisponde a valori non predicibili l'uno dall'altro, e da pag. 387 troviamo che il primo passaggio per zero di $\mathcal{R}_{\underline{yy}}(\tau)$ avviene per $\tau \simeq \frac{0.4}{f_D}$, corrispondente al minimo intervallo di tempo necessario per osservare valori di $\underline{y}(t)$ *incorrelati*; viceversa, un intervallo τ sufficientemente più piccolo trova il canale in condizioni pressoché immutate. Definendo allora

$$T_c = \frac{0.1}{f_D} \tag{20.29}$$

come *tempo di coerenza,* osserviamo che una trasmissione con periodo di simbolo $T_s \geq T_c$ subisce condizioni del canale differenti nell'arco di tempo di un simbolo, ostacolandone la sincronizzazione[53], ed in tal caso il fading viene detto *veloce*. Se invece $T_s \ll T_c$ il canale si mantiene in condizioni pressoché stazionarie per tutto il periodo di simbolo, il fading è detto *lento*, 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.

20.4.7 Tipologia di canale radiomobile

E' univocamente determinata dalla tipologia del fading su piccola scala, denominato in base allo schema rappresentato a lato, e che a sua volta dipende dalla natura del messaggio trasmesso, come ora mostriamo.



Condizione di sottodispersione Notiamo che il verificarsi contemporaneo della assenza di distorsione lineare in quanto $W < B_c$ (fading *piatto*) e della stazionarietà del canale in quanto $T_s < T_c$ (fading *lento*) equivale al verificarsi della condizione di *canale perfetto* (pag. 231). Ciò accade a patto che[54]

$$\begin{cases} W < B_c = \frac{0.1}{\sigma_\tau} & \text{no dist. lin.} \\ T_s < T_c = \frac{0.1}{f_D} & \text{stazionario} \end{cases}$$

Condizioni di slow flat fading

$$f_D \cdot \sigma_\tau < 0.01 \quad \text{ovvero} \quad T_c \cdot B_c > 1 \tag{20.30}$$

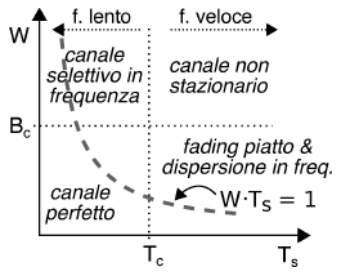
detta condizione di *sottodispersione* (*underspread*). Nella pratica i valori di f_D e σ_τ per i canali in uso nelle telecomunicazioni soddisfano tale condizione.

Classificazione del canale Come anticipato dipende da come i valori della banda W e del periodo di simbolo T_s del messaggio si relazionano rispetto a banda B_c e tempo T_c di coerenza del canale radio in uso[55].

[53] Ciò avviene perché in pratica è come se due simboli consecutivi pervenissero attraverso due differenti canali, e dunque non è possibile eseguire operazioni di media.

[54] Infatti le due condizioni $W < B_c$ e $T_s < T_c$ possono essere riscritte in base alle (20.23) e (20.29) come $W\sigma_\tau < 0.1$ e $f_D T_s < 0.1$, e moltiplicando queste ultime tra loro si ottiene $W\sigma_\tau f_D T_s < 0.01$. Ponendo quindi $W \simeq 1/T_s$ (eq. (15.5)) si ottiene la condizione $f_D \cdot \sigma_\tau < 0.01$, in cui sostituendo $f_D = 0.1/T_c$ e $\sigma_\tau = 0.1/B_c$ (di nuovo in virtù delle (20.23) e (20.29)) si ottiene la seconda relazione.

[55] Ovvero contesto rurale, urbano, indoor, oltre ovviamente ai fenomeni legati al movimento.



A lato viene raffigurato il piano $T_s - W$ in cui è evidenziato il ramo di iperbole corrispondente alla relazione $W \cdot T_s = 1$[56] assieme ad una coppia di possibili valori per B_c e T_c tali da rispettare[57] la condizione $T_c \cdot B_c > 1$. Osserviamo come la (20.30) sia necessaria *ma non* sufficiente ad individuare un canale perfetto: muovendosi infatti lungo l'iperbole (ovvero variando W e T_s) ci si può comunque trovare in condizioni di canale selettivo in frequenza (in alto a sin.) qualora divenga $W > B_c$, oppure (in basso a destra) in condizioni di variabilità temporale quando diventa $T_s > T_c$.

Esempio Dato un canale con assegnati T_c e B_c, determinare la massima velocità per una trasmissione QPSK con impulso a coseno rialzato e $\gamma = 1$, in modo da evitare l'uso di un equalizzatore. Affrontiamo l'analisi fissando la banda occupata $B = f_s\,(1+\gamma)$ pari a B_c, da cui si ottiene una $f_b = f_s \cdot 2 = {}^{B_c}/_2 \cdot 2 = B_c$. In tal caso $T_s = {}^1/_{f_s} = {}^2/_{B_c}$, e se il canale verifica la condizione di sottodispersione (20.30) si ottiene anche $T_s < T_c$, ovvero il canale può essere ritenuto stazionario per la durata di un simbolo.

20.5 Appendici

Sviluppiamo qui alcuni approfondimenti a riguardo del calcolo della P_e in condizioni di fading di Rayleigh, della architettura di ricevitore *Rake*, della allocazione delle frequenze radio, e della caratterizzazione del fenomeno di dispersione temporale.

20.5.1 Probabilità di errore in presenza di fading di Rayleigh

Anche se un collegamento radiomobile presenta fading piatto, con il movimento l'ampiezza del segnale ricevuto subisce fluttuazioni *alla Rayleigh* rispetto al suo valor medio (fig. 20.4) tali da dover aumentare la potenza da trasmettere allo scopo di dotare il collegamento di un margine M^{ps} in grado di assicurare un adeguato grado di servizio (eq. (20.14)). Dato che a pag. 686 abbiamo trovato questo procedimento un po' macchinoso, sviluppiamo ora una discussione su come modificare le formule di calcolo della probabilità di errore in modo da valutare direttamente l'${}^{E_b}/_{N_0}$ necessario *senza* dover passare per il grado di servizio.

La variabilità temporale della potenza istantanea ricevuta può essere tenuta direttamente in conto se l'espressione della $P_e^{bit}\,({}^{E_b}/_{N_0})$ ottenuta al cap. 16 per un canale gaussiano viene considerata come quella di una probabilità *condizionata* $Pr\left[\mathrm{err}/\frac{E_b}{N_0}\right]$ rispetto ad un determinato valore di ${}^{E_b}/_{N_0}$, di cui valutare il *valore atteso* rispetto alla variabilità statistica dei valori di E_b *ricevuto*. Per poter sviluppare i passaggi indichiamo allora con E_b l'energia per bit *media* che si riceverebbe in *assenza* di fading di Rayleigh, e con $E'_b = \rho^2 E_b$ la stessa quantità (istantanea) ricevuta a seguito del fading, in cui ρ è il modulo dell'inviluppo complesso ricevuto, descritto da una v.a. di Rayleigh. La P_e è

[56] Come osservato alla nota precedente $W \simeq {}^1/_{T_s}$, da cui $W \cdot T_s \simeq 1$

[57] Trovandosi il prodotto *sopra* l'iperbole unitaria.

dunque definita come

$$P_{e,Rayleigh}^{bit} = E_\rho \left\{ Pr\left[\text{err}/\frac{\rho^2 E_b}{N_0} \right] \right\} = \int_0^\infty Pr\left[\text{err}/\frac{\rho^2 E_b}{N_0} \right] p(\rho) d\rho \tag{20.31}$$

Prendendo come esemplare il caso della modulazione BPSK[58] (pag. (16.7)) abbiamo $P_e^{bit}\left(E_b/N_0\right) = \frac{1}{2}\text{erfc}\left\{\sqrt{E_b/N_0}\right\}$ e dunque possiamo scrivere

$$Pr\left[\text{err}/\frac{\rho^2 E_b}{N_0} \right] = P_e^{bit}\left(\frac{\rho^2 E_b}{N_0}\right) = \frac{1}{2}\text{erfc}\left\{\rho\sqrt{E_b/N_0}\right\} \tag{20.32}$$

Per applicare la (20.32) alla (20.31) occorre specificare la d.d.p. di ρ pari a $p\left(\rho\right) = \frac{\rho}{\sigma^2}\text{e}^{-\frac{\rho^2}{2\sigma^2}}$ (eq. (20.12)) e ricordare l'espressione di $\text{erfc}\left\{\alpha\right\} = \frac{2}{\sqrt{\pi}}\int_\alpha^\infty \text{e}^{-y^2} dy$ ((§ 6.2.4)), in modo che la (20.31) divenga

$$P_{e,Rayleigh}^{bit} = \frac{1}{\sqrt{\pi}} \int_0^\infty \int_{\rho\sqrt{E_b/N_0}}^\infty \text{e}^{-y^2} dy \frac{\rho}{\sigma^2} \text{e}^{-\frac{\rho^2}{2\sigma^2}} d\rho$$

ed invertendo l'ordine di integrazione[59] otteniamo

$$P_e^{bit} = \frac{1}{\sqrt{\pi}} \int_0^\infty \text{e}^{-y^2} \int_0^{y\sqrt{N_0/E_b}} \frac{\rho}{\sigma^2} \text{e}^{-\frac{\rho^2}{2\sigma^2}} d\rho dy = \frac{1}{\sqrt{\pi}} \int_0^\infty \text{e}^{-y^2} \left[1 - \text{e}^{-\frac{N_0}{E_b}\frac{y^2}{2\sigma^2}} \right] dy \tag{20.33}$$

Allo scopo di semplificare la (20.33) notiamo innanzitutto che (essendo ρ una v.a. di Rayleigh) risulta $E\left\{\rho^2\right\} = 2\sigma^2$ (vedi eq. (14.13)), da cui ne deriva che l'energia per bit $\overline{E}_b'$ *media ricevuta* ha espressione $\overline{E}_b' = E\left\{\rho^2 E_b\right\} = 2\sigma^2 E_b$. Indichiamo quindi con $\Gamma = \frac{2\sigma^2 E_b}{N_0} = \frac{\overline{E}_b'}{N_0}$ l'SNR per bit *medio* che viene *ricevuto* in modo che la (20.33) possa essere riscritta come

$$P_e^{bit} = \frac{1}{\sqrt{\pi}} \int_0^\infty \text{e}^{-y^2} dy - \frac{1}{\sqrt{\pi}} \int_0^\infty \text{e}^{-y^2 - \frac{y^2}{\Gamma}} dy \tag{20.34}$$

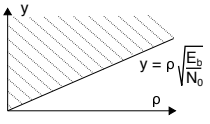
Tenendo ora conto[60] che $\int_{-\infty}^{\infty} \text{e}^{-y^2} dy = \sqrt{\pi}$ il primo termine di (20.34) risulta pari ad $\frac{1}{2}$, mentre dato che $-y^2 - \frac{y^2}{\Gamma} = -\frac{\Gamma+1}{\Gamma}y^2$ e che risulta anche $\int_{-\infty}^{\infty} \text{e}^{-\alpha y^2} dy = \sqrt{\pi/\alpha}$, il secondo termine di (20.34) si riscrive come $\frac{1}{\sqrt{\pi}}\int_0^\infty \text{e}^{-\frac{\Gamma+1}{\Gamma}y^2} dy = \frac{1}{2}\sqrt{\frac{\Gamma}{1+\Gamma}}$, in modo da ottenere per la (20.34) il risultato *in forma chiusa*

$$P_{e,Rayliegh}^{bit} = \frac{1}{2}\left(1 - \sqrt{\frac{\Gamma}{1+\Gamma}}\right) \tag{20.35}$$

---

[58] A causa delle fluttuazioni di ampiezza legate al fading non è possibile ricorrere a modulazioni di tipo QAM, e nel seguito sono prese in considerazione unicamente modulazioni di fase e di frequenza.

[59] Il dominio di integrazione è rappresentato in figura, e anziché muoversi prima lungo y dalla retta $y = \rho\sqrt{E_b/N_0}$ ad infinito ottenendo una funzione di ρ, e quindi integrare con $0 < \rho < \infty$, ci si muove in orizzontale tra $\rho = 0$ e $\rho = y\sqrt{N_0/E_b}$ ottenendo una funzione di y, quindi integrata con $0 < y < \infty$. Verifichiamo quindi che



$$\int_0^{y\sqrt{N_0/E_b}} \frac{\rho}{\sigma^2}\text{e}^{-\frac{\rho^2}{2\sigma^2}} d\rho = -\text{e}^{-\frac{\rho^2}{2\sigma^2}}\Big|_0^{y\sqrt{N_0/E_b}} = -\text{e}^{-\frac{y^2}{2\sigma^2}\frac{N_0}{E_b}} + 1.$$

[60] Vedi ad es. `https://it.wikipedia.org/wiki/Integrale_di_Gauss`)

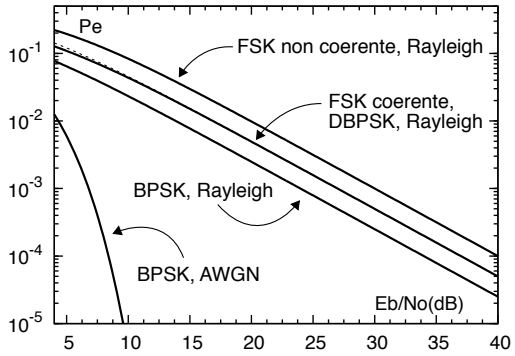


Figura 20.7: Confronto tra probabilità di errore in presenza di fading di Rayleigh per alcuni tipi di modulazione

che confrontato in fig. 20.7 con l'espressione di $P^{bit}_{e,BPSK} = \frac{1}{2}\text{erfc}\left\{\sqrt{E_b/N_0}\right\}$ per un canale AGWN evidenzia come in presenza di fading di Rayleigh la P_e sia sensibilmente peggiore, e diminuisca molto più lentamente all'aumentare di E_b/N_0. Se poi valutiamo la differenza in E_b/N_0 (dB) mostrata in fig. 20.7 per i casi di presenza ed assenza di fading e per una stessa P_e, troviamo valori confrontabili con quelli del margine M^{ps} mostrati in fig. 20.4. Procedendo in modo simile si possono valutare le prestazioni per le altre forme di modulazione numerica, il cui risultato è pure riportato in fig. 20.7 come anche nella tabella seguente, assieme al valore approssimato di P^{bit}_e per grandi valori di Γ.

modulazione	$P^{bit}_{e,Rayleigh}$	$P_e\|_{\Gamma\to\infty}$
BPSK antipodale coerente	$\frac{1}{2} - \frac{1}{2}\sqrt{\frac{\Gamma}{1+\Gamma}}$	$\frac{1}{4\Gamma}$
DBPSK	$\frac{1}{2(1+\Gamma)}$	$\frac{1}{2\Gamma}$
BFSK ortogonale coerente	$\frac{1}{2} - \frac{1}{2}\sqrt{\frac{\Gamma}{2+\Gamma}}$	$\frac{1}{2\Gamma}$
BFSK ortogonale incoerente	$\frac{1}{2+\Gamma}$	$\frac{1}{\Gamma}$

Esempio Determinare l'incremento di potenza necessario a conseguire una $P^{bit}_e = 10^{-4}$ nel caso di una modulazione BPSK affetta da fading di Rayleigh, rispetto alla potenza necessaria su di un canale AWGN.

Dal grafico di fig. 20.7 osserviamo che nel caso AWGN è necessario un E_b/N_0 circa pari a 8 dB, mentre in presenza di fading ne occorrono circa 34, dunque l'incremento di potenza assomma a 26 dB.

20.5.2 Ricevitore Rake

Questa particolare architettura di ricevitore *trae vantaggio* da una modulazione DSSS (§ 16.9.2) che occupa una banda maggiore della banda di coerenza $W_p > B_c$ e per la quale il canale presenta dunque una attenuazione *selettiva in frequenza* (§ 20.4.5) legata alla ricezione di più repliche del segnale trasmesso a causa del fenomeno dei cammini multipli. Se infatti le repliche prodotte dal multipath arrivano con intervalli temporali maggiori del periodo di chip T_p la proprietà di *bassa autocorrelazione* delle sequenze

PN (§ 16.9.1) utilizzate nel DSSS rendono le repliche equivalenti ad una qualsiasi altra interferenza a larga banda, ed il ricevitore svolge la funzione di equalizzazione in modo *del tutto particolare*.

Per capire ciò che accade occorre affrontare un po' di conti. Indichiamo allora con $\{b_i\}$ la sequenza dei *bit* trasmessi e con

$$\underline{\tilde{x}}(t) = \sum_i b_i \, \mathrm{pn}(t - iT_b)$$

l'inviluppo complesso del segnale DSSS, mentre come discusso al § 20.4.5 la risposta impulsiva del canale ha espressione

$$\underline{h}(t) = \sum_{n=0}^{N-1} Z_n \, \delta(t - \tau_n)$$

(eq. (20.18)) in cui i coefficienti $Z_n = a_n \mathrm{e}^{-j2\pi f_0 \tau_n}$ sono i guadagni *complessi* dovuti ai cammini multipli: il *trucco* del ricevitore Rake è quello di *conoscere*[61] i valori di Z_n e τ_n. Trascurando il fattore $1/2$ della convoluzione tra inviluppi complessi (eq. (13.3)), il segnale ricevuto in presenza di multipath ha quindi espressione

$$\underline{\tilde{r}}(t) = \underline{\tilde{x}}(t) * \underline{\tilde{h}}(t) = \sum_{n=0}^{N-1} Z_n \, \underline{\tilde{x}}(t - \tau_n) = \sum_{n=0}^{N-1} Z_n \sum_i b_i \, \mathrm{pn}(t - iT_b - \tau_n)$$

La decodifica (§ 16.9.2.2) del $j-$esimo simbolo b_j *inizia* (ad esempio) moltiplicando $\underline{\tilde{r}}(t)$ per $\mathrm{pn}(t - jT_s - \tau_0)$ (allineata cioè al primo ritardo τ_0) ed integrando il risultato su di un periodo di bit, realizzando così un *correlatore* (pag. 218) alla sequenza PN, ovvero

$$\begin{aligned}
\hat{b}_j &= \frac{1}{T_b} \int_{jT_b+\tau_0}^{(j+1)T_b+\tau_0} \underline{\tilde{r}}(t) \, \mathrm{pn}(t - jT_b - \tau_0) \, dt = \\
&= \frac{1}{T_b} \int_{jT_b+\tau_0}^{(j+1)T_b+\tau_0} \left[\sum_{n=0}^{N-1} Z_n \sum_i b_i \, \mathrm{pn}(t - iT_b - \tau_n) \right] \mathrm{pn}(t - jT_b - \tau_0) \, dt = \\
&= \frac{1}{T_b} \int_0^{T_b} \sum_{n=0}^{N-1} Z_n \, b_j \, \mathrm{pn}(\alpha - \tau'_n) \, \mathrm{pn}(\alpha) \, d\alpha = \\
&= \sum_{n=0}^{N-1} Z_n \, b_j \, \frac{1}{T_b} \int_0^{T_b} \mathrm{pn}(\alpha - \tau'_n) \, \mathrm{pn}(\alpha) \, d\alpha = Z_0 \, b_j + \sum_{n=1}^{N-1} Z_n \, b_j \, R_{PN}(\tau'_n)
\end{aligned}$$

in cui alla terza riga sopravvive solo il termine $j-$esimo della $\sum_i$ in quanto è l'unico entro gli estremi di integrazione, dopodiché si pone $\alpha = t - jT_b - \tau_0$ e $\tau'_n = \tau_n - \tau_0$. Il primo termine del risultato finale per $\hat{b}_j$ è pari (a meno del coeff. complesso Z_0) al simbolo cercato b_j, mentre il secondo rappresenta $N-1$ termini di interferenza associati agli altri percorsi, in cui il prodotto tra i coefficienti $Z_{n\neq 0}$ e lo stesso simbolo b_j viene *ridotto* di una quantità pari all'autocorrelazione $R_{PN}(\tau'_n)$ della sequenza pseudonoise calcolata per slittamenti pari alla differenza τ'_n tra ritardi, in modo che il secondo termine risulta ridotto rispetto al primo del rapporto $R_{PN}(\tau'_n)/R_{PN}(0)$.

Interrompiamo per un attimo l'aspetto analitico per indicare la stima a cui siamo arrivati come $\hat{b}_j^0$ in quanto ottenuta dalla replica con ritardo τ_0, e notare che lo stesso tipo di elaborazione è applicabile fruttuosamente *a tutti* i ritardi τ_n per ottenere, mediante un banco di correlatori[62], altrettante stime $\hat{b}_j^n$ che il ricevitore RAKE (letteralmente,

[61] Ottenuti grazie ad una tecnica di stima di canale.

[62] I correlatori del Rake sono anche detti *fingers*, ovvero dita (del rastrello), ognuno dei quali utilizza

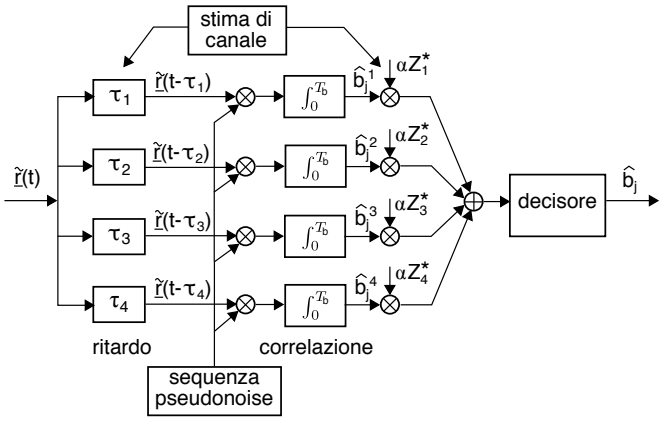


Figura 20.8: Schema di principio del ricevitore Rake

rastrello) ricombina come mostrato alla fig. 20.8.

La figura schematizza i passaggi discussi adottando una notazione lievemente diversa, in quanto anziché ritardare la sequenza *pn* viene ritardato il segnale in arrivo[63]; inoltre, viene evidenziato il ruolo ed i punti di intervento del risultato della stima di canale.

Le uscite dei correlatori sono quindi combinate tra loro con il duplice intento di rendere la somma *coerente* eliminando il contributo di fase dovuto al multipath, e di applicare il principio di *massimo rapporto* (§ 21.3.1.2) per trarre vantaggio dallo schema a diversità. Ciò avviene moltiplicando l'uscita del correlatore $n-$esimo per Z_n^*, e dato che l'uscita stessa contiene il fattore Z_n, tale operazione elimina il contributo di fase, e *pesa* il contributo del ramo con $\left|Z_n\right|^2$, ovvero con l'energia associata al ritardo τ_n[64]. Tali pesi sono infine *scalati* di una quantità $\alpha = \frac{1}{\sum_{n=0}^{N-1} Z_n^2}$ in modo da mantenere la dinamica del risultato entro valori noti.

20.5.3 Allocazione delle frequenze radio

L'assegnazione generale dello spettro radio ai diversi utilizzi è riportata in tabella 20.1, 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 le trasmissioni televisive in *digitale terrestre*, ad eccezione dei canali da 61 a 69, che sono stati assegnati agli operatori di telefonia mobile di 4^a generazione, detta LTE/4G.

una PN con un ritardo pari a quello di uno degli echi del multipath, realizzando uno schema di ricezione a *diversità di tempo*.

[63] In pratica nello schema in figura la pn è allineata al ritardo maggiore τ_n, mentre i ritardi mostrati vanno da zero (per correlare la replica più ritardata) alla massima differenza $\tau_n - \tau_0$.

[64] Ciò riduce il peso dei contributi relativi a rami su cui perviene un segnale di ampiezza ridotta, la cui uscita dipende in misura maggiore dal rumore.

Intervallo	λ	*Sigla*	*Denominazione*	*Uso*
30 - 300 Hz	$10^4 - 10^3$ Km	**ELF**	*Extremely Low*	
.3 - 3 KHz	$10^3 - 10^2$ Km	**VF**	*Voice Frequency*	
3 - 30 KHz	100 – 10 Km	**VLF**	*Very Low*	Radionavigazione a largo raggio. Attività nucleare.
30 - 300 KHz	10 – 1 Km	**LF**	*Low Frequency*	Radiolocalizzazione marittima ed aeronautica
.3 - 3 MHz	.1 – 1 Km	**MF**	*Medium Frequency*	Comunicazioni aeree e marittime. Radionavigazione. Broadcast AM
3 - 30 MHz	10 – 100 metri	**HF**	*High Frequency*	Collegamenti a lunga distanza fissi e mobili. Radioamatori.
30 - 300 MHz	1 – 10 metri	**VHF**	*Very High*	Broadcast FM e TV. Collegamenti in visibilità. Radiomobili civili e militari.
.3 - 3 GHz	.1 – 1 metro	**UHF**	*Ultra High*	Ponti radio e radiomobili terrestri. Broadcast TV. Satelliti meteo e TV.
3 - 30 GHz	10 – 100 mm	**SHF**	*Super High*	Ponti radio terrestri. Satelliti. Radar.
30 - 300 GHz	1 – 10 mm	**EHF**	*Extremely High*	Radar

Tabella 20.1: Allocazione delle frequenze radio

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

Intervallo	*utilizzo*
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

Intervallo Uplink (MHz)	*Intervallo Downlink(MHz)*	*utilizzo*
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

20.5.4 Caratterizzazione della dispersione temporale

Valori tipici per la dispersione temporale da cammini radio multipli

ambiente	f_0 (MHz)	σ_τ	Note
urbano	910	600 ns	New York City, $\overline{\tau}$ = 1.3 μs, $\Delta\tau$ = 3.5 μs
urbano	892	10 - 25 μs	San Francisco, caso *peggiore*
rurale	910	200 - 310 ns	caso tipico medio
rurale	910	1.9 - 2.1 μs	caso *estremo* medio
indoor	1500	10 - 50 ns	ufficio, $\overline{\tau}$ = 25 ns
indoor	850	-	ufficio, $\Delta\tau$ = 270 ns
indoor	1900	-	grattacieli, $\overline{\tau}$ = 70 - 94 ns, $\Delta\tau$ = 1.47 μs

Dispersione potenza-ritardo ETSI-GSM (900 MHz)

	ambito collinare		area urbana	
cammino n.	τ_n [μsec]	a_n^2 [dB]	τ_n [μsec]	a_n^2 [dB]
1	0	-10	0	-4.0
2	0.1	-8	0.1	-3.0
3	0.3	-6	0.3	0.0
4	0.5	-4	0.5	-2.6
5	0.7	0	0.8	-3.0
6	1.0	0	1.1	-5.0
7	1.3	-4	1.3	-7.0
8	15.0	-8	1.7	-5.0
9	15.2	-9	2.3	-6.5
10	15.7	-10	3.1	-8.6
11	17.2	-12	3.2	-11.0
12	20.0	-14	5.0	-10.0

Capitolo 21

Sistemi multiantenna o MIMO

Non appena l'evoluzione tecnologica lo ha permesso i sistemi di trasmissione radio hanno iniziato a dotarsi di più di una antenna, presso una od entrambe le estremità del collegamento. Tale modifica architetturale ha reso possibile lo sviluppo di tecniche trasmissive basate sul *signal processing* del tutto nuove, capaci di migliorare le prestazioni sia nei termini di un maggior SNR in ricezione sia in quelli di una più elevata velocità di trasmissione, conseguendo così una *efficienza spettrale* prima non immaginabile. Ma la cosa più sorprendente è che i miglioramenti avvengono *grazie ai fenomeni di multipath*, che nello scenario ad antenna singola costituivano invece un aspetto peggiorativo.

In condizioni di *fading piatto* ciascun canale radio tra una coppia di antenne è descritto da un unico coefficiente complesso h_{ij}, l'insieme dei quali individua una matrice $\boldsymbol{H}$ di dimensione $n_T \cdot n_R$ la cui conoscenza da parte del ricevitore, del trasmettitore, o di entrambi, costituisce un fattore determinante per la realizzazione di quasi tutte le soluzioni che andremo ad esaminare. Innanzitutto vedremo come molteplici antenne in ricezione consentano di combinare *in modo coerente* i segnali di ciascuna, mentre una molteplicità in trasmissione permette di inviare *più copie* dello stesso segnale. Quindi studieremo come la presenza di n_T antenne in trasmissione ed n_R in ricezione determini la possibilità di trasmettere *fino a* $\min(n_T, n_R)$ diverse comunicazioni indipendenti, il cui numero effettivo dipende dal *rango* della matrice $\boldsymbol{H}$, funzione dal grado di indipendenza tra i suoi elementi, a sua volta funzione dalla spaziatura tra antenne e dalla natura dell'ambiente circostante.

Dopo aver ridefinito il concetto di capacità di canale per adattarlo al nuovo caso di canale *vettoriale* affetto da fading variabile, vengono descritte tecniche lineari subottime che permettono la *separazione* dei diversi flussi ricevuti senza incorrere in una eccessiva complessità computazionale. Tecniche duali possono essere altresì impiegate al lato trasmittente, aprendo la possibilità di adottare la soluzione multiantenna anche per realizzare schemi di trasmissione *uno a molti* tipici delle reti di telefonia mobile e di accesso WIFI e WIMAX. In tali contesti si è affermato l'abbinamento delle tecniche di trasmissione MIMO con quelle di modulazione a portante multipla OFDM, determinando la *moltiplicazione* del numero di canali frutto della molteplicità delle antenne, per il

numero delle sotto-portanti su cui si basa l'OFDM. Una soluzione del tutto simile si è affermata anche per sistemi di diffusione radio televisiva digitali o DVB, in cui un unico dispositivo riceve in modo sincrono lo stesso segnale OFDM trasmesso da molteplici antenne dislocate sul territorio.

21.1 Lo scenario delle possibilità

Iniziamo con un riassunto sommario dei molteplici aspetti da affrontare. Una prima discriminazione terminologica riguarda la *localizzazione* della molteplicità di antenne, specificando che mentre adottarne più di una sia in trasmissione che in ricezione viene indicato con l'acronimo MIMO ovvero *multiple input multiple output,* ed il caso *tradizionale* di una singola antenna da entrambi i lati è detto SISO (*single input single output*), l'utilizzo di antenne multiple da un solo lato del collegamento, in trasmissione od in ricezione, viene rispettivamente indicato con i termini MISO (*multiple input single output*) e SIMO (*single input multiple output*), vedi fig. 21.1a.

Il secondo punto importante è che la presenza di più antenne ad uno o entrambi i lati del collegamento consente l'attuazione di differenti principi genericamente descritti come *diversità*, *multiplazione spaziale* e *beamforming* (vedi fig. 21.1b), con lo scopo di perseguire tre obiettivi complementari e che in generale *non* possono essere conseguiti congiuntamente, ma di cui esiste la possibilità di individuare soluzioni di compromesso. Vediamo di che si tratta.

Diversità spaziale E' la più diretta conseguenza della possibilità di poter trasmettere e/o ricevere una stessa trasmissione mediante antenne diverse. Come riferito al § 20.4.4 se le antenne sono abbastanza distanziate tra loro i *cammini multipli* su cui si sviluppa il collegamento sono differenti, così come la risposta in frequenza $H(f)$ *vista* dalle singole antenne, cosicché la distorsione lineare subita da ciascuna di esse diviene una v.a. statisticamente indipendente. Ciò rende molto improbabile che le antenne subiscano *tutte assieme* una forte attenuazione, cosicché la ricezione risente molto meno della variabilità del canale; come introdotto al (§ 20.3.3.1), questo risultato è frutto dello

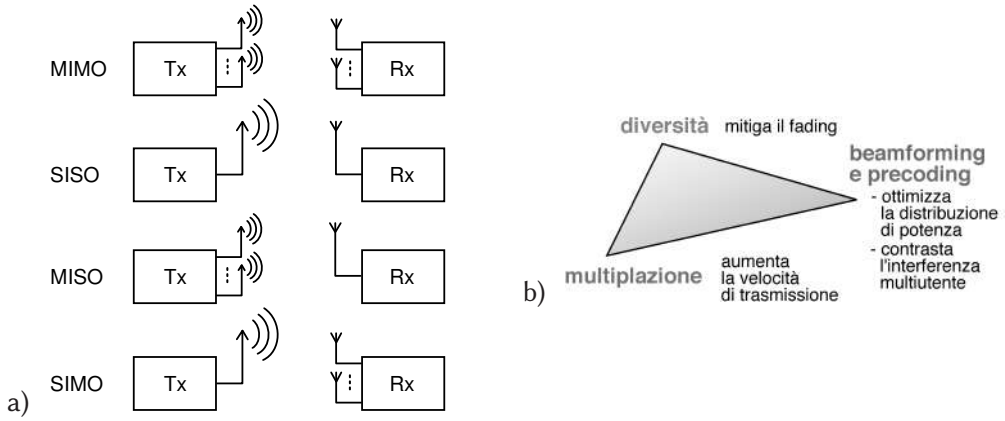


Figura 21.1: MIMO: a) - configurazioni di antenna; b) - tecniche applicabili

sfruttamento della *diversità spaziale*. In tale scenario la velocità di trasmissione non cambia, dato che ogni antenna trasmette lo stesso simbolo, o come approfondiremo al § 21.3.2.2, una versione codificata della stessa sequenza. Due particolarità della tecnica sono che in generale non è necessario conoscere le caratteristiche del canale per le diverse antenne, e che una qualche sua variante può essere attuata anche se solo una delle parti in comunicazione è dotata di più antenne. Il *guadagno di diversità* conseguito può essere impiegato a favore di una maggiore affidabilità del collegamento (ossia per ottenere un miglior E_b/N_0 e dunque P_e), oppure per preservare la stessa affidabilità di un sistema SISO ma diminuire la potenza totale irradiata, a vantaggio della durata delle batterie dei terminali di utente.

Multiplazione spaziale In presenza di più antenne da entrambi i lati del collegamento la multiplazione consiste nel trasmettere flussi numerici *diversi* da parte delle differenti antenne, contemporaneamente, ed alla stessa frequenza. Ciò è possibile a patto che il ricevitore conosca la matrice $\boldsymbol{H}$ che descrive la risposta in frequenza tra coppie di antenne; quando tali valori risultano sufficientemente incorrelati la matrice diviene invertibile, rendendo possibile risalire ai singoli segnali a partire dalle loro diverse combinazioni lineari, ricevute da ciascuna antenna. Il risultato è che la velocità aggregata è *la somma*[1] di quella dei diversi flussi, mentre la banda occupata è quella di un flusso singolo, conseguendo pertanto una elevata *efficienza spettrale* (pag. 498). Qualora anche il trasmettitore conosca $\boldsymbol{H}$ può combinare la multiplazione con il *precoding*, rendendo possibile trasmettere i diversi flussi verso *destinatari differenti*, anche ad antenna singola, realizzando in tal modo un sistema *MultiUtente* detto MU-MIMO basato sullo *space division multiple access* o SDMA.

Compromesso diversità - multiplazione Non è possibile conseguire allo stesso tempo sia un pieno guadagno di diversità che un altrettanto pieno guadagno di multiplazione, semplicemente perché le antenne o trasmettono gli stessi dati, oppure dati differenti. Nei sistemi reali si ricerca una condizione operativa di compromesso, ed in grado di adattarsi in funzione delle condizioni del collegamento: mentre in presenza di un buon SNR può non essere necessario ricorrere alla diversità e dunque è possibile sfruttare al massimo la multiplazione (conseguendo una velocità elevata), per valori di SNR ridotti è opportuno rinunciare a buona parte di multiplazione, a favore della diversità.

Beamforming Il senso più letterale di questo termine si riferisce alla consolidata tecnica di *orientare* verso una determinata direzione l'onda elettromagnetica emessa da una schiera di antenne[2], in modo da aumentare la potenza ricevuta in tale direzione e/o riducendo la potenza interferente per soggetti in posizioni diverse. Nato nel contesto di una comunicazione in visibilità, nel caso di fading di Rayleigh ovvero di prevalenza

[1] In altre parole la velocità aggregata aumenta linearmente con il numero delle antenne, inteso come il numero *minimo* tra quelle di trasmissione e quelle di ricezione.

[2] Tecnica nota anche come *smart antenna* o *phased array*, vedi ad es. `https://en.wikipedia.org/wiki/Phased_array`

dei cammini riflessi tale approccio perde di applicabilità, ma lo stesso termine è usato per intendere anche la tecnica del...

Precoding Nel caso di una molteplicità di antenne di tipo MIMO e della perfetta conoscenza di $\boldsymbol{H}$ da parte del trasmettitore, studieremo come tecniche di signal processing consentano di realizzare il cosiddetto *eigen-beamforming*, ovvero utilizzare il collegamento ripartendo l'informazione da trasmettere nelle *direzioni* individuate dagli autovettori della matrice di canale. Come per il multiplexing anche con il precoding si ottiene un aumento della velocità di trasmissione, sia per quanto riguarda un collegamento punto-punto, sia per la possibilità di servire più utenti in un sistema MU-MIMO. Ma in questo secondo caso la tecnica da adottare si modifica allo scopo di minimizzare l'effetto interferente nei confronti degli utenti non destinatari.

Esaurita l'esposizione delle tecniche principali, il capitolo prosegue illustrando le problematiche correlate, riassunte appresso.

Modalità duplex e canale di ritorno Le due direzioni di trasmissione possono essere distinte nel tempo o in frequenza: mentre nel primo caso la proprietà di reciprocità del canale radio permette di usare la stima di canale svolta in fase di ricezione anche durante quella di trasmissione, nel secondo caso ciò non è possibile, determinando l'esigenza di rendere nota al trasmettitore la matrice $\boldsymbol{H}$ stimata in ricezione. La trasmissione di $\boldsymbol{H}$ inoltre avviene per forza di cose in modo approssimato (quantizzato), comportando errori ed imprecisioni nel precoding, ed impegnando tante più risorse di trasmissione quanto più precisa si desidera la codifica della $\boldsymbol{H}$ da trasmettere.

Trasmissione multiportante Gran parte degli approcci fin qui descritti si basano sul presupposto di avere a che fare con segnali a banda stretta ovvero per i quali si verificano le condizioni di *fading piatto*, impedendo di aumentare troppo la velocità di trasmissione ovvero l'occupazione di banda, a causa dell'aumento di complessità della procedura di equalizzazione che si renderebbe necessaria. Ciò non è più vero nel caso della trasmissione OFDM, per la quale la banda a disposizione è suddivisa in P sotto-canali di banda ridotta su ciascuno dei quali trasmettere una frazione del flusso informativo complessivo, rendendo l'ipotesi di $H(f)$ costante verificata per ciascun sotto-canale. In pratica è come avere a disposizione P diverse trasmissioni MIMO, su ciascuna delle quali applicare le tecniche discusse, con valutazioni di compromesso potenzialmente differenti per ciascuna sottoportante. A questo aspetto sono da aggiungere due ulteriori vantaggi di una trasmissione MIMO-OFDM: il primo deriva dalla semplicità con cui l'ODFM consente di equalizzare la distorsione lineare subita da ciascuna sottoportante, mentre la seconda è la possibilità di assegnare sottoportanti diverse ad utenti differenti per realizzare con facilità un sistema multiutente.

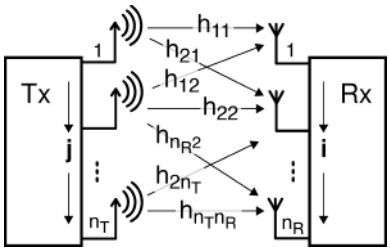
Rete a frequenza singola La trasmissione broadcast attuata dai sistemi di radio e TV digitale si basa su molteplici antenne disseminate sul territorio che trasmettono tutte il medesimo segnale OFDM, allo stesso tempo e nella stessa regione di frequenza, secondo

un approccio[3] noto come *single frequency network* o SFN[4]. Ciò richiede una stringente sincronizzazione sia nel tempo che in frequenza (derivata dal sistema GPS) in modo che il ritardo relativo dei segnali che pervengono ad un medesimo destinatario ricada tutto entro i limiti del tempo di guardia del simbolo OFDM.

Come evidente la natura delle implicazioni della presenza di più antenne investe i più disparati contesti, che in questo capitolo tentiamo di affrontare uno alla volta, a partire dalla definizione formale del *canale MIMO*.

21.2 Il canale MIMO

In presenza di n_T antenne in trasmissione ed n_R in ricezione ogni coppia i, j di esse ($i = 1, 2, \cdots, n_R$; $j = 1, 2, \cdots, n_T$) *vede* un differente canale radio, che in condizioni di fading piatto (pag. 682) ovvero di segnale a banda stretta (§ 13.1.2.4) ossia con banda inferiore alla banda di coerenza (pag. 688), viene descritto da un unico coefficiente *complesso* h_{ij}, pari alla $H(f)$ valutata alla $f = f_0$ di modulazione.



Il valore esatto di h_{ij} dipende dai fenomeni di cammini multipli tipici delle comunicazioni wireless, sia rurali che urbane che indoor, ed in tal senso può variare nel tempo qualora la trasmissione sia da associare ad un contesto di mobilità. Il valore h_{ij} della risposta in frequenza tra le antenne j ed i ad un dato istante viene quindi messo in corrispondenza all'elemento i, j di una matrice $\boldsymbol{H}$ di dimensioni $n_R \times n_T$, indicata anche come CSI ovvero *channel state information*.

La trasmissione da parte della antenna j di un segnale $x(t)$ ottenuto mediante modulazione numerica PSK o QAM a partire da una sequenza simbolica[5] $\left\{s_j^k\right\}$ produce agli istanti di simbolo $t = kT_s$ la ricezione presso l'antenna i di un inviluppo complesso[6]

$$r_i^k = h_{ij} s_j^k + n_i \tag{21.1}$$

con n_i v.a. gaussiana complessa a media nulla, parti reale ed immaginaria incorrelate, e varianza[7] $\sigma_n^2 = E\{nn^*\}$, condizione indicata anche come $n_i \in \mathcal{CN}\left\{0, \sigma_n^2\right\}$. Indicando

[3] Al contrario, le tecniche di broadcast analogico (cap. 25) su scala nazionale prevedono l'uso di regioni di frequenza diverse per lo stesso canale trasmesso in bacini di propagazione differenti, in cui questi ultimi sono definiti dalle condizioni di visibilità legate alla conformazione del territorio.

[4] Vedi ad es. `https://en.wikipedia.org/wiki/Single-frequency_network`

[5] I valori reale e immaginario di s^k rappresentano le coordinate nel piano dell'inviluppo complesso di un punto appartenente alla costellazione $\mathcal{A}$ scelta per la trasmissione, vedi ad es. la figura a pag. 501.

[6] In virtù della (13.3) scriviamo $\underline{y}(t) = \frac{1}{2}\underline{h}(t) * \underline{x}(t) = \frac{1}{2}\underline{h} \cdot \underline{x}(t)$ dato che per fading piatto si ha $\underline{h}(t) = \underline{h} \cdot \delta(t)$; a sua volta (eq. (11.21)) $\underline{h} = \underline{H}(f) = 2H^+(f + f_0) = 2H(f_0)$ e dunque $\underline{y}(t) = \frac{1}{2} \cdot 2H(f_0) \cdot \underline{x}(t)$ ovvero, ponendo $\underline{x}(t_k) = s^k$ si ottiene $r^k = \underline{y}(t_k) = H(f_0) \cdot s^k$.

[7] Essendo n una v.a. complessa $n = n_c + jn_s$, otteniamo $\sigma_n^2 = E\{nn^*\} = E\left\{n_c^2 + n_s^2\right\} = 2\sigma_{n_c}^2 = 2\sigma_{n_s}^2$, in accordo con i risultati del § 14.1.3.

ora con $\boldsymbol{s}$ il vettore n_T-dimensionale trasmesso ad un generico istante di simbolo[8] da tutta la schiera di n_T antenne, presso le n_R antenne in ricezione otterremo al medesimo istante il vettore complesso di dimensione n_R

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n} \tag{21.2}$$

con $\boldsymbol{n}$ vettore gaussiano complesso ad n_R elementi. Come discusso al § 20.4.4 se ciascun valore $h_{ij} = \rho e^{j\varphi}$ è una v.a. gaussiana complessa a media nulla con parti reale ed immaginaria incorrelate e varianza σ_h^2 ([9]) il valore $\rho = |h_{ij}|$ del modulo di h_{ij} ha d.d.p. di Rayleigh

$$p(\rho) = \frac{\rho}{\sigma_h^2} \exp\left(-\frac{\rho^2}{2\sigma_h^2}\right)$$

mentre φ si distribuisce uniformemente tra $\pm\pi$, come avviene per collegamenti in assenza di visibilità e con molteplici cammini multipli statisticamente indipendenti.

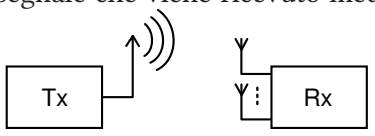
Come scopriremo strada facendo, molti risultati si basano sulla conoscenza dei valori di $\boldsymbol{H}$, stimati al ricevitore (§ 21.7.3.2) e la cui stima si ritiene valida per un periodo comparabile con il *tempo di coerenza* (pag. 692) del collegamento. In base a tali valori scopriremo come le proprietà di $\boldsymbol{H}$ possano condizionare il livello di prestazione conseguibile. Ma iniziamo la trattazione con un caso *relativamente* semplice.

21.3 Diversità spaziale

L'esistenza contemporanea (e sulla stessa banda di frequenze) di più canali radio, differenziati per gli elementi di $\boldsymbol{H}$ (21.2), offre innanzitutto l'opportunità di analizzare come sfruttare la *diversità spaziale* disponibile allo scopo di migliorare le prestazioni del collegamento, ovvero risparmiare potenza o migliorare la P_e, od un compromesso tra i due obiettivi. Discutiamo innanzitutto cosa si può fare quando solo uno dei due estremi del collegamento dispone di più di una antenna.

21.3.1 Ricevitore multi-antenna

Si tratta del caso indicato come SIMO ($n_T = 1$) che studiamo nel contesto di una trasmissione punto-punto oppure broadcast, in cui una unica antenna trasmette un segnale che viene ricevuto mediante $n_R > 1$ antenne, e dunque $\boldsymbol{H}$ è un vettore ad una colonna ed n_R righe, ovvero $\boldsymbol{H} = \left(h_1, h_2, \cdots, h_{n_R}\right)^T$. Nelle condizioni di fading di Rayleigh *piatto* se la separazione tra le antenne è sufficiente[10] il valo-



[8] Di qui in poi sarà possibile riferirsi agli istanti di simbolo come all'"uso del canale", ovvero pensando alla trasmissione di un simbolo $\boldsymbol{s}$ n_T-dimensionale come ad *un singolo uso* del canale.

[9] σ_h^2 dipende dalla somma delle intensità con cui le repliche del segnale trasmesso dall'antenna j giungono all'antenna i.

[10] Nel caso di un telefono *cellulare* sono presenti numerosi riflettori nelle vicinanze del ricevitore, producendo *nei downlink* fading incorrelati per distanze tra le antenne *dei mobili* di circa mezza lunghezza d'onda. Viceversa nel caso della *base station* fissa con cui il cellulare comunica, i cammini multipli *dell'uplink* hanno quasi tutti origine nei pressi del mobile, riducendo la gamma di angoli di incidenza dei raggi ricevuti, che iniziano ad essere indipendenti per distanze di decine di lunghezze d'onda: pertanto alla *base station* sono necessarie antenne molto più lontane tra loro.

re h_i relativo a ciascuna antenna è *incorrelato* da quello associato alle altre ovvero $E\{h_i h_j\} = 0$, e dunque *statisticamente indipendente* in virtù della gaussianità delle h_i. Pertanto se per una antenna si manifesta una forte attenuazione, ciò probabilmente non accade per le altre.

Sussistono tre possibili modi di sfruttare le n_R antenne di ricezione: preferire quella da cui si riceve più potenza, mediare equamente tra tutti i segnali ricevuti, oppure effettuare una somma *pesata*.

21.3.1.1 Selezione di diversità

Mostriamo innanzitutto come anche solo limitandosi a scegliere quale antenna utilizzare per la ricezione si riescano a conseguire risultati molto soddisfacenti. Consideriamo quindi un ricevitore per cui siano disponibili n_R *rami di diversità* indipendenti ed affetti da fading di Rayleigh, su ognuno dei quali (per ogni utilizzo del canale) si riceve un segnale con inviluppo complesso (vedi (21.1))

$$r_i = h_i s + n_i \tag{21.3}$$

con $i = 1, 2, \cdots, n_R$, in cui s è il valore complesso che individua un punto della costellazione adottata, trasmesso con potenza[11] $\mathcal{P}_T = \mathcal{E}_s f_s$. I valori $\rho_i = |h_i|$ sono v.a. di Rayleigh con d.d.p. $p(\rho) = \frac{\rho}{\sigma_h^2} e^{-\rho^2/2\sigma_h^2}$, uguale per tutti i rami i ma con valori incorrelati ovvero $E\{\rho_i \rho_j\} = 0$ con $i \neq j$, mentre $n_i \in \mathcal{CN}\{0, \sigma_n^2\}$ è un campione dell'inviluppo complesso di un processo gaussiano bianco passabanda.

A differenza di un canale AWGN, il rapporto SNR che compete alla (21.3) non è un valore deterministico, ma una v.a. che indichiamo come SNR *istantaneo, per simbolo*[12]

$$\gamma_i = \frac{|h_i|^2 E\{s^2\}}{\sigma_n^2} = |h_i|^2 \frac{\mathcal{E}_s}{\sigma_n^2} \tag{21.4}$$

e che si distribuisce con la medesima d.d.p. di $|h_i|^2$, che è (vedi eq. (20.13)) una *esponenziale*; pertanto

$$p(\gamma_i) = \frac{1}{\Gamma} e^{-\gamma_i/\Gamma}$$

(con $\gamma_i \geq 0$), dove Γ è l'SNR *medio* valutabile come

$$\Gamma = E\{\gamma_i\} = E\{|h_i|^2\} \frac{\mathcal{E}_s}{\sigma_n^2} = 2\sigma_h^2 \frac{\mathcal{E}_s}{\sigma_n^2} \tag{21.5}$$

(vedi § 20.5.1) uguale per tutti i rami i nelle ipotesi poste. A questo punto abbiamo tutte le relazioni necessarie a valutare la probabilità che *un singolo* ramo abbia un γ_i inferiore ad un valore δ, pari a (vedi eq. (22.3) a pag. 771)

$$Pr\{\gamma_i \leq \delta\} = \int_0^{\delta} p(\gamma_i)\, d\gamma_i = 1 - \int_{\delta}^{\infty} \frac{1}{\Gamma} e^{-\gamma_i/\Gamma} d\gamma_i = 1 - e^{-\delta/\Gamma}$$

mentre la probabilità che *tutti* gli n_R rami indipendenti presentino *contemporaneamente*

[11] Il valore $\mathcal{E}_s$ individua l'*energia per simbolo*, e misura il valore di $E\{s^2\}$.

[12] *Istantaneo* perché dipende da h_i che in linea di principio può variare da istante ad istante; *per simbolo* perché $\frac{\mathcal{E}_s}{\sigma_n^2} = \frac{\mathcal{P}_s}{\mathcal{P}_n} \frac{1}{f_s}$.

$\gamma_i < \delta$ vale

$$Pr\{\gamma_1, \gamma_2, \cdots, \gamma_M \leq \delta\} = (1 - e^{-\delta/\Gamma})^{n_R}$$

che indichiamo come $P_{n_R}(\delta)$, da cui otteniamo la probabilità che *almeno uno* dei rami consegua $\gamma_i \geq \delta$ come

$$Pr\{\gamma_i \geq \delta, \forall i\} = 1 - P_{n_R}(\delta) = 1 - (1 - e^{-\delta/\Gamma})^{n_R}$$

Esempio Consideriamo un ricevitore con quattro rami di diversità, ognuno affetto da fading di Rayleigh, e con un medesimo SNR medio Γ. Determinare la probabilità che l'SNR istantaneo γ_i di ciascun ramo si riduca contemporaneamente di 10 dB sotto il valor medio Γ, ossia che tutti i γ_i siamo inferiori ad un valore δ tale che $\delta/\Gamma = 0.1$, e confrontare il risultato con il caso di un ricevitore senza diversità.

Risulta che $Pr\{\gamma_1, \gamma_2, \gamma_3, \gamma_4 \leq \delta\} = P_4(\delta) = \left(1 - e^{-0.1}\right)^4 = 8.2 \cdot 10^{-5}$, mentre per un sistema SISO avremmo avuto $Pr\{\gamma_i \leq \delta\} = P_1(\delta) = 1 - e^{-0.1} = 9.5 \cdot 10^{-2}$. Considerando i 10 dB di differenza come il valore di un margine oltre il quale il collegamento diviene troppo rumoroso, l'uso di quattro rami di diversità corrisponde ad un miglioramento della probabilità di fuori servizio di più di mille volte!

L'approccio della *selezione di diversità* è facilmente realizzabile in quanto coinvolge solamente il ricevitore, dove viene comparata la potenza del segnale in arrivo sulle diverse antenne, e quindi il segnale più forte è inviato al ricevitore, ne più ne meno come nello schema anticipato al § 20.3.3.1. Dato che la stima del livello di potenza di ricezione per ciascun ramo si basa su di una media temporale, la decisione non avviene in modo prettamente istantaneo; d'altra parte, è sufficiente che avvenga con tempi inferiori al *tempo di coerenza* eq. (20.29).

E' possibile mostrare[13] che l'SNR medio Γ_{SD} per il ramo di volta in volta selezionato *migliora* all'aumentare del numero di antenne, anche se per ogni antenna aggiuntiva il miglioramento è sempre minore, risultando

$$\Gamma_{SD} = \Gamma \sum_{k=1}^{n_R} \frac{1}{k} \tag{21.6}$$

ovvero con un numero di 2, 3, 4 antenne si ottiene un fattore di miglioramento di 1.5, 1.83, 2.03... ma si può fare di meglio se vengono utilizzati *tutti* i rami in contemporanea, anziché uno solamente.

21.3.1.2 Combinazione di massimo rapporto - MRC

Riscriviamo la (21.3) espandendo il coefficiente h_i come $h_i = |h_i| e^{j\varphi_i}$, ovvero

$$r_i = |h_i| e^{j\varphi_i} s + n_i \tag{21.7}$$

con $|h_i|$ v.a. di Rayleigh e φ_i uniforme. La tecnica che stiamo per affrontare si basa sul calcolo di una combinazione lineare

$$\hat{r} = \sum_{i=1}^{n_R} |w_i| e^{j\phi_i} r_i \tag{21.8}$$

[13] Vedi D.G. Brennan, *Linear Diversity Combining Techniques*. Proc. IEEE, Vol. 91, N. 2, Feb 2003.

dei segnali ricevuti r_i mediante coefficienti complessi $w_i = |w_i|e^{-j\phi_i}$ ottenuti a partire da una stima[14] dei valori h_i. Per quanto riguarda la fase ϕ_i questa viene scelta pari a $-\varphi_i$, in modo da *annullare* il termine di fase in r_i e porsi nelle condizioni di *somma coerente*, dato che inserendo la (21.7) in (21.8) si ottiene

$$\begin{aligned}\hat{r} &= \sum_i |w_i| e^{-j\varphi_i} r_i = \sum_i |w_i| e^{-j\varphi_i} \left(|h_i| e^{j\varphi_i} s + n_i\right) = \\ &= \sum_i |w_i||h_i| s + \sum_i |w_i| e^{-j\varphi_i} n_i = \sum_i |w_i||h_i| s + \sum_i |w_i| n_i \end{aligned} \tag{21.9}$$

visto che la rotazione φ_i della v.a. complessa n_i non ne cambia la natura[15].

La scelta dei valori $|w_i|$ avviene secondo il criterio di rendere massimo l'SNR *momentaneo*[16] γ_{MR} di (21.9), da cui il nome di *maximal ratio combining* (MRC) della tecnica, in cui per *ratio* si intende l'SNR. A tale scopo osserviamo come alla componente di segnale di (21.9) competa una energia media (rispetto alla variabilità di s) $E\left\{\left(\sum_i |w_i||h_i|s\right)^2\right\} = \mathcal{E}_s \left(\sum_i |w_i||h_i|\right)^2$, mentre per la varianza del rumore si ottiene $E\left\{\left(\sum_i |w_i|n_i\right)^2\right\} = \sigma_n^2 \sum_i w_i^2$, essendo i campioni di rumore n_i statisticamente indipendenti: pertanto l'SNR di $\hat{r}$ (21.9) risulta pari a

$$\gamma_{MR} = \frac{\mathcal{E}_s}{\sigma_n^2} \frac{\left(\sum_{i=1}^{n_R} |w_i||h_i|\right)^2}{\sum_{i=1}^{n_R} w_i^2} = \frac{\mathcal{E}_s}{\sigma_n^2} \frac{\left|\mathbf{w}_m^T \cdot \mathbf{h}_m\right|^2}{\mathbf{w}_m^T \cdot \mathbf{w}_m} \tag{21.10}$$

dove con $\mathbf{w}_m = \left(|w_1|, \cdots, |w_{n_R}|\right)^T$ si è indicato il vettore del *modulo* dei coefficienti $\boldsymbol{w}$ e con $\mathbf{h}_m = \left(|h_1|, \cdots, |h_{n_R}|\right)^T$ quello del guadagno di ampiezza dei rami, in modo che la (21.10) sia espressa nei termini di *prodotto scalare* tra vettori. Con tale formalismo il massimo valore per γ_{MR} si ottiene applicando la disuguaglianza di Schwartz (§ 2.4.3) che asserisce

$$\left|\mathbf{w}_m^T \cdot \mathbf{h}_m\right|^2 \leq \left(\mathbf{h}_m^T \cdot \mathbf{h}_m\right) \cdot \left(\mathbf{w}_m^T \cdot \mathbf{w}_m\right)$$

con il segno di uguale solo quando $\mathbf{w}_m = \alpha\, \mathbf{h}_m$, ovvero i vettori sono *paralleli*. Con tale scelta la (21.10) diviene

$$\gamma_{MR} = \frac{\mathcal{E}_s}{\sigma_n^2} \frac{\left|\alpha\, \mathbf{h}_m^T \cdot \mathbf{h}_m\right|^2}{\alpha^2\, \mathbf{h}_m^T \cdot \mathbf{h}_m} = \frac{\mathcal{E}_s}{\sigma_n^2} \frac{\left(\sum_{i=1}^{n_R} |h_i|^2\right)^2}{\sum_{i=1}^{n_R} |h_i|^2} = \frac{\mathcal{E}_s}{\sigma_n^2} \sum_{i=1}^{n_R} |h_i|^2 = \sum_{i=1}^{n_R} \gamma_i \tag{21.11}$$

in cui l'ultima eguaglianza si basa sulla (21.4). L'SNR complessivo è dunque pari alla *somma* degli SNR dei singoli rami, e può così conseguire valori *accettabili* anche se nessuno dei rami lo ottiene individualmente.

In virtù dell'ipotesi che i valori h_i siano statisticamente indipendenti, dalla (21.11) otteniamo l'SNR *medio* Γ_{MR} a partire dai Γ_i dei singoli rami (eq. (21.5)) come

$$\Gamma_{MR} = \sum_{i=1}^{n_R} \Gamma_i \tag{21.12}$$

[14] Valida per un intervallo temporale minore del tempo di coerenza del canale.

[15] E' proprio in base a questa considerazione che il vettore gaussiano complesso costituito da campioni dell'inviluppo complesso di un processo di rumore passa banda prende il nome di *processo circolare*.

[16] Nel senso che dato che (come stiamo per vedere) γ_{MR} dipende dagli h_i che sono v.a., è una v.a. anch'esso. Ma allo stesso tempo gli h_i sono considerati costanti per tutto il tempo di coerenza, ed altrettanto accade a γ_{MR}.

Se ogni ramo presenta il medesimo SNR medio Γ_i si può confrontare (21.12) con (21.6), ed osservare che ora Γ_{MR} aumenta *linearmente* con il numero di antenne, ovvero migliora di 3 dB ad ogni raddoppio di n_R. Se invece (essendo gli $|h_i|$ v.a. indipendenti) i Γ_i sono differenti tra loro il miglioramento è inferiore, ma Γ_{MR} risulta comunque sempre *migliore del miglior* Γ_i.

Per poter essere operativi non resta che effettuare una scelta ragionata per il coefficiente di proporzionalità α introdotto dalla disuguaglianza di Schwartz, e dal punto di vista teorico di nessun impatto. Includendo anche il termine di rifasamento, nella pratica conviene scegliere dei pesi

$$w_i = \frac{h_i^*}{\sum_{i=1}^{n_R} |h_i|^2} \tag{21.13}$$

in modo da rendere $\hat{r}$ uno stimatore non polarizzato (§ 6.6.3): essendo infatti i termini n_i a media nulla, il valore atteso della (21.9) assume la forma

$$E\{\hat{r}\} = \sum_i w_i h_i s = \sum_i \frac{h_i^*}{\sum_i |h_i|^2} h_i s = \frac{\sum_i |h_i|^2}{\sum_i |h_i|^2} s = s \tag{21.14}$$

e dunque la tecnica MRC non altera *la dimensione* della costellazione $\mathcal{A}$ a cui s appartiene, e la decisione sul simbolo trasmesso può basarsi sul criterio di massima verosimiglianza

$$\hat{s} = \arg\max_{s \in \mathcal{A}} Prob\{\hat{r}/s\} \quad \text{ovvero} \quad \hat{s} = \arg\min_{s \in \mathcal{A}} (s - \hat{r}_i)^2$$

dove la seconda espressione individua una minima distanza euclidea[17] conseguenza del passaggio ai logaritmi e della gaussianità di n.

Probabilità di errore Con una procedura di media statistica analoga[18] a quella illustrata al § 20.5.1 si può arrivare a valutare la P_e^{bit} nel caso di costellazione BPSK ed in funzione del numero di ricevitori n_R e dell'SNR *medio ricevuto* Γ (lo stesso per tutte le antenne), che per valori di $E_b/N_0 \geq 10$ è bene approssimata come

$$P_{e,\,MRC,\,BPSK}^{bit} \simeq \binom{2n_R - 1}{n_R} \left[\frac{1}{2}\left(1 - \sqrt{\frac{\Gamma}{1+\Gamma}}\right)\right]^{n_R} \tag{21.15}$$

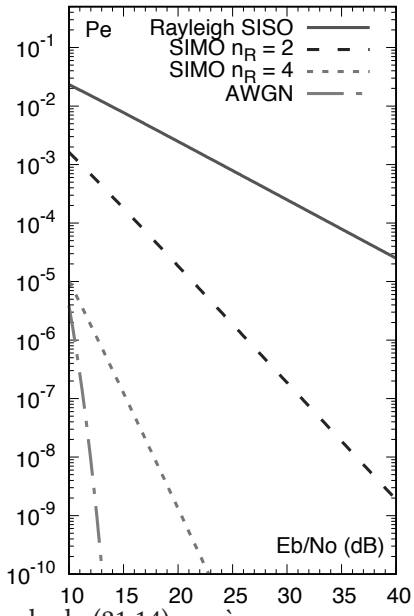
da confrontare con l'espressione (20.35) ottenuta per un sistema di trasmissione SISO, confronto possibile anche mediante la figura a pagina seguente, in cui oltre al caso del fading SISO compare anche la curva di P_e per il caso AWGN in cui non si verifica fading. Evidentemente il miglioramento è notevole!

Svolgiamo ora due ultime considerazioni a partire dalla (21.13). La prima è che nel caso di una potenza di rumore σ_n^2 uguale su tutti i rami, si può dire che la combinazione lineare attuata dalla (21.8) da maggior peso ai rami con l'SNR più elevato, dato che questo dipende da $|h_i|^2$. La seconda considerazione è che la scelta (21.13) rivela una analogia

[17] Con l'accortezza che per grandezze complesse il quadrato si valuta come prodotto per il coniugato, ossia $z^2 = z \cdot z^* = (\Re\{z\})^2 + (\Im\{z\})^2$.

[18] Nel senso di pervenire ad una espressione di $P_e(\gamma_{MR})$ e poi calcolarne il valore atteso rispetto alla v.a. γ_{MR}, che ora non è più esponenziale, bensì chi quadro (§ 6.6.5) con $2n_R$ gradi di libertà, essendo le h_i v.a. gaussiane complesse.

tra questa tecnica e quella del filtro adattato (§ 7.6) che pure trova motivazione nella massimizzazione dell'SNR, ottenuta in quel caso grazie ad una risposta in frequenza che è *il coniugato* del segnale da rivelare. Il paragone non è casuale dato che come per il filtro adattato, a supporto della scelta $\boldsymbol{w} \propto \boldsymbol{h}^*$ sussiste una argomentazione basata sulla disuguaglianza di Schwartz (§ 2.4.3)[19]; in altre parole, ciò che il filtro adattato compie nel dominio della frequenza, viene attuato dal ricevitore MRC nel dominio spaziale.



21.3.1.3 Combinazione equal gain

Torniamo all'ipotesi che rende massima la (21.12) ovvero quando i diversi rami subiscono un fading di intensità simile, cioè $|h_i| \simeq |h| \ \forall i$. In tal caso i pesi w_i ottimi per ottenere $\hat{r}$ tramite la (21.8) hanno tutti lo stesso modulo[20] pari a $|w_i| = \frac{1}{n_R |h_i|}$, in modo che la (21.14) ora è espressa come

$$E\{\hat{r}\} = \sum_i |w_i||h_i| s = \sum_i \frac{1}{n_R |h_i|} |h_i| s = \frac{n_R}{n_R} s = s \tag{21.16}$$

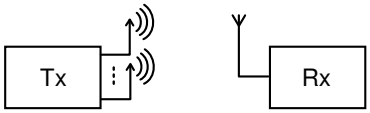
e quindi ogni antenna riceve lo stesso SNR medio Γ, conseguendo un SNR complessivo

$$\Gamma_{EG} = n_R \Gamma \tag{21.17}$$

ovvero n_R volte quello di ogni singolo ramo. In tal modo il ricevitore risulta molto semplificato, al punto da preferire a volte l'uso di pesi tutti uguali[21] anche in presenza di coefficienti h_i diversi tra loro, nel qual caso il valore (21.17) non viene raggiunto. Tale scelta viene indicata come metodo *equal gain*, ed offre risultati solo di poco inferiori a quelli di *massimo rapporto*, e comunque migliori del metodo *a selezione*.

21.3.2 Trasmettitore multiantenna

Ovviamente la presenza di una unica antenna ricevente impedisce di attuare strategie di scelta o di combinazione delle fonti di *diversità spaziale* disponibili. Con un numero $n_T > 1$ di trasmittenti i



[19] Vedi ad es. il riferimento della nota 13; per lo stesso risultato sussiste inoltre anche un'argomentazione basata sugli autovettori della matrice $\boldsymbol{h} \cdot \boldsymbol{h}^\dagger$ (l'apice † indica il trasposto coniugato), vedi B. Holter, G.E. Oien, *The Optimal Weights of a Maximum Ratio Combiner using an Eigenfilter Approach*, Proc. 5th IEEE Nordic Signal Processing Symposium, Hurtigruten, 4-6 October 2002, reperibile presso `https://www.ux.uis.no/norsig/norsig2002/Proceedings/papers/cr1099.pdf`, e che esamina anche il caso in cui il rumore si presenti con potenze differenti sui diversi rami.

[20] Rimane infatti la necessità di *rifasare* i rami per ottenere una somma *coerente*.

[21] Dove per uguali si intende $|w_i| = 1/n_R$ senza riguardo per $|h_i|$, causando nella (21.16) l'insorgenza di un fattore moltiplicativo, che può essere ignorato nel caso di una modulazione PSK ovvero con costellazione circolare.

simboli che giungono all'unico ricevente

$$r = \sum_{i=1}^{n_T} h_i s + n_i \tag{21.18}$$

risultano infatti *già combinati*. Ma come vedremo la conoscenza da parte *del ricevente* dei coefficienti complessi[22] h_i (che costituiscono l'unica riga della matrice $\boldsymbol{H}$ definita al § 21.2) permette comunque di sfruttare la diversità spaziale. Il modo con cui ciò avviene si basa su di una particolare tecnica di codifica di canale, nota come codifica *spazio-tempo* (STC) in quanto l'informazione da trasmettere viene *diluita* (ridondata) oltre che *nel tempo* (la sola dimensione possibile per un caso SISO[23]) anche *sulle antenne* (ovvero nello spazio), riuscendo a ottenere un *guadagno di diversità*.

Una seconda cosa degna di nota in questo passaggio è che in un canale bidirezionale[24] trasmettitore e ricevitore si scambiano continuamente di ruolo, dunque se anche in modalità MISO si riesce ad ottenere un miglioramento comparabile al caso SIMO, è sufficiente che solo una delle parti in comunicazione - tipicamente la stazione radio base o l'access point WiFi[25] - sia dotata di più antenne.

21.3.2.1 Codice a traliccio spazio - tempo

Un primo approccio alla realizzazione di un STC si è basato sulla *modulazione a traliccio* (pag. 541) e per questo indicato come *space time trellis code* o STTC. In questa tecnica i simboli codificati sono ripartiti tra le antenne, mente dal lato ricevente la decodifica avviene eseguendo un algoritmo di Viterbi *vettoriale*. Oltre ad un guadagno *di diversità* proporzionale al numero delle antenne, il metodo offre anche un guadagno *di codifica* legato al numero di stati del codice a traliccio soggiacente (a spese della complessità di decodifica), cosa che non avviene con la tecnica STBC descritta appresso, ma che è attualmente favorita nonostante questa carenza, più che compensata dal vantaggio di poter svolgere la decodifica mediante operazioni assai più semplici. Scegliamo pertanto di non addentraci nello studio degli STTC.

21.3.2.2 Codice a blocco spazio - tempo

A differenza dei codici di canale esaminati al § 17.4, in cui le quantità in ingresso ed in uscita erano semplici bit, uno STBC opera *a valle* della codifica di simbolo (indicata anche come *symbol mapper*), che ad un blocco di M bit fa corrispondere un valore complesso s_i che individua un punto di costellazione $\mathcal{A}$ tra $L = 2^M$. Lo STBC osserva un numero k di tali simboli, e fa corrispondere ad essi una codeword che consiste in una matrice $\boldsymbol{C}$ i cui elementi $c_{ij} \in \mathcal{A}$ identificano i simboli trasmessi dall'antenna

[22] Aleatori a componenti gaussiane a media nulla etc etc...

[23] In realtà ad es. nella modulazione COFDM (§ 16.8.10) la ridondanza viene distribuita anche sulle diverse sottoportanti, dunque *in frequenza*. Ma ci torniamo al § 21.7.2.

[24] Qualora la trasmissione nelle due direzioni avvenga sulla stessa portante è necessario che le parti si alternino nei ruoli (*time-duplex*), mentre invece possono trasmettere e ricevere allo stesso tempo se si adottano frequenze differenti (*frequency-duplex*).

[25] Anche perché le ridotte dimensioni dei device mobili rendono problematico realizzare antenne sufficientemente distanziate.

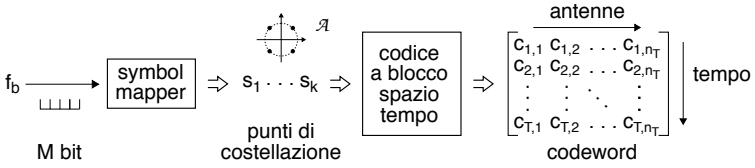


Figura 21.2: Generazione di un codice a blocco spazio-tempo

$j = 1, \cdots, n_T$ all'istante $i = 1, 2, \cdots, T$. Il tasso di codifica in questo caso è calcolato come il rapporto tra il numero k di simboli s_i ed il numero di istanti T impiegati per trasmetterli, ovvero

$$R_c = k/T \tag{21.19}$$

21.3.2.3 Codice di Alamouti

Esponiamo il metodo partendo dalla sua definizione iniziale dovuta ad *Alamouti*[26], relativa al caso di $n_T = 2$ antenne di trasmissione, che in due istanti di simbolo consecutivi trasmettono un totale di quattro simboli di canale c_{ij} legati ai due *di sorgente* s_1 ed s_2 secondo lo schema[27] illustrato in tabella 21.1. Le righe corrispondono ai due istanti di simbolo: durante il primo istante le antenne trasmettono rispettivamente s_1 la prima ed s_1 la seconda, mentre al secondo istante la prima trasmette $-s_2^*$ e la seconda s_1^*. Pertanto il codice agisce attraverso le due dimensioni della tabella, in cui le righe individuano la diversità *temporale*, mentre le colonne quella *spaziale*. Il calcolo del tasso di codifica fornisce $R_c = k/T = 2/2 = 1$, e dunque a differenza dei codici a blocco *tradizionali* non si verifica nessun aumento di banda. Vedremo che il codice di Alamouti è l'unico STBC a godere di tale proprietà, risultando in generale $R_c < 1$.

	antenna 1	antenna 2
tempo t	s_1	s_2
tempo $t + T_s$	$-s_2^*$	s_1^*

Tabella 21.1: Codice di Alamouti

Ricezione Applicando la (21.18) al segnale ricevuto nei due istanti, otteniamo

$$\begin{aligned} r_1 &= h_1 s_1 + h_2 s_2 + n_1 \\ r_2 &= -h_1 s_2^* + h_2 s_1^* + n_2 \end{aligned} \tag{21.20}$$

dove n_i è il solito rumore gaussiano complesso. Dalla seconda relazione si calcola

$$r_2^* = -h_1^* s_2 + h_2^* s_1 + n_2^* \tag{21.21}$$

e quindi, purché il ricevitore conosca i coefficienti h_1 ed h_2, i due valori r_1 ed r_2^* sono *combinati* in modo *simile* a quanto visto per l'MRC, questa volta allo scopo di ottenere

[26] S.M. Alamouti, *A simple transmit diversity technique for wireless communications*, IEEE Journal on Selected Areas in Communications, Oct 1998. Reperibile presso
`https://mast.queensu.ca/~fady/Math800/papers/Alamouti_JSAC98.pdf`

[27] Dunque le codeword $\boldsymbol{C}$ sono costituite dai simboli di sorgente stessi, una sorta di codice a ripetizione, se non fosse che ora ci sono anche i coniugati, e quel segno cambiato... in effetti la scelta di tab. 21.1 è un caso particolare di una regola più generale, ovvero costruire gli elementi di $\boldsymbol{C}$ come combinazione lineare dei simboli s_i, e dei loro coniugati.

una stima $\hat{s}_1$ e $\hat{s}_2$ per *entrambi* i simboli di sorgente, calcolando

$$\begin{aligned} \hat{s}_1 &= h_1^* r_1 + h_2 r_2^* \\ \hat{s}_2 &= h_2^* r_1 - h_1 r_2^* \end{aligned} \tag{21.22}$$

Sostituendo infatti le (21.20) e (21.21) in (21.22) si ottiene

$$\begin{aligned} \hat{s}_1 &= h_1^* \left(h_1 s_1 + h_2 s_2 + n_1\right) + h_2 \left(-h_1^* s_2 + h_2^* s_1 + n_2^*\right) \\ \hat{s}_2 &= h_2^* \left(h_1 s_1 + h_2 s_2 + n_1\right) - h_1 \left(-h_1^* s_2 + h_2^* s_1 + n_2^*\right) \end{aligned}$$

da cui

$$\begin{aligned} \hat{s}_1 &= \left(|h_1|^2 + |h_2|^2\right) s_1 + h_1^* n_1 + h_2 n_2^* \\ \hat{s}_2 &= \left(|h_1|^2 + |h_2|^2\right) s_2 + h_2^* n_1 - h_1 n_2^* \end{aligned} \tag{21.23}$$

dove i termini di interferenza tra simboli sono *scomparsi*, e si manifesta un *rinforzo* delle ampiezze dei simboli di sorgente pari alla somma dei quadrati dei guadagni di Rayleigh, come anche avveniva per l'MRC, e che indichiamo con $\alpha = |h_1|^2 + |h_2|^2$.

Prestazioni Se dividiamo le (21.23) per α otteniamo anche in questo caso delle stime non polarizzate dei simboli di sorgente

$$\begin{aligned} \tilde{s}_1 &= \tfrac{1}{\alpha}\hat{s}_1 = s_1 + \tilde{n}_1 \\ \tilde{s}_2 &= \tfrac{1}{\alpha}\hat{s}_2 = s_2 + \tilde{n}_2 \end{aligned} \tag{21.24}$$

in cui $\tilde{n}_1$ e $\tilde{n}_2$ sono v.a. gaussiane complesse con parti reale ed immaginaria incorrelate, a media nulla e varianza[28] $\sigma_{\tilde{n}}^2 = E\left\{\tilde{n}_i \tilde{n}_i^*\right\} = \sigma_n^2/\left(|h_1|^2+|h_2|^2\right)$ in cui $\sigma_n^2 = E\left\{n_i n_i^*\right\}$ è la potenza del rumore in ingresso al ricevitore. Il criterio di decodifica è come di consueto quello di massima verosimiglianza, che si riduce alla regola di decisione

$$s_i^\diamond = \arg\min_{s \in \mathcal{A}} \ (s - \tilde{s}_i)^2 \quad \text{con} \qquad i = 1, 2 \tag{21.25}$$

ovvero ad ognuno dei due istanti i si decide per la ricezione del simbolo $s_i^\diamond$ (che figura nella costellazione $\mathcal{A}$) *più vicino*[29] al simbolo (valore complesso) normalizzato $\tilde{s}_i$ espresso dalla (21.24), che possiamo riguardare come un valore di decodifica *soft*. In alternativa, se i simboli trasmessi (di sorgente) provengono da uno stadio di codifica di canale *esterno* per il quale è prevista una decodifica soft, i valori $\tilde{s}_i$ possono essere passati a quest'ultima *così come sono*.

La probabilità con cui la regola (21.25) dà esito errato è naturalmente legata in modo inverso al rapporto *SNR istantaneo* γ per le variabili $\tilde{s}_i$ (21.24) su cui si effettua

[28] Svolgendo infatti i calcoli per una di esse, ad esempio $\tilde{n_1} = \frac{1}{\alpha}\left(h_1^* n_1 + h_2 n_2^*\right)$, in virtù dell'incorrelazione tra n_1 ed n_2 si ottiene

$$\begin{aligned} \sigma_{\tilde{n}_1}^2 &= E\left\{\tilde{n}_i \tilde{n}_i^*\right\} = \frac{1}{\alpha^2} E\left\{\left(h_1^* n_1 + h_2 n_2^*\right)\left(h_1 n_1^* + h_2^* n_2\right)\right\} = \\ &= \frac{1}{\alpha^2}\left[|h_1|^2 E\left\{n_1 n_1^*\right\} + |h_2|^2 E\left\{n_2^* n_2\right\} + h_1^* h_2^* E\left\{n_1 n_2\right\} + h_1 h_2 E\left\{n_2^* n_1^*\right\}\right] = \\ &= \frac{1}{\alpha^2}\left[|h_1|^2 \sigma_{n_1}^2 + |h_2|^2 \sigma_{n_2}^2\right] = \frac{1}{\alpha^2}\left[\alpha \sigma_n^2\right] = \sigma_n^2 \frac{1}{|h_1|^2 + |h_2|^2} \end{aligned}$$

[29] Si ribadisce che per grandezze complesse il modulo quadro si calcola come prodotto per il coniugato, dunque la (21.25) diviene $(s - \tilde{s}_i)^2 = (s - \tilde{s}_i)\left(s^* - \tilde{s}_i^*\right) = s s^* + \tilde{s}_i \tilde{s}_i^* - s \tilde{s}_i^* - \tilde{s}_i s^*$, da valutare per ogni $s \in \mathcal{A}$.

la decisione, che risulta pari a

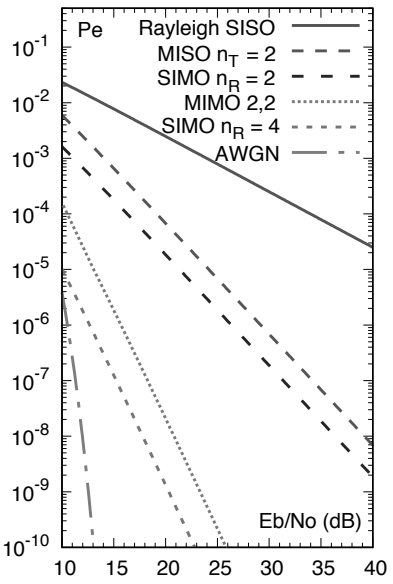
$$\gamma_{Ala} = \frac{E\left\{s_i s_i^*\right\}}{\sigma_{\tilde{n}}^2} = \frac{\mathcal{E}_s}{\sigma_n^2}\left(|h_1|^2 + |h_2|^2\right)$$

fornendo un *guadagno di diversità* identico a quello di un ricevitore SIMO-MRC (eq. (21.11)), tranne che... molto spesso le specifiche relative alla massima potenza irradiata impongono di suddividere la stessa in parti uguali tra le antenne di trasmissione. Ciò significa che anche l'energia per simbolo $\mathcal{E}_s$ della costellazione adottata risulta (con $n_T = 2$) *dimezzata*, cosicché il sistema MISO-STBC con $n_T = 2$ subisce una penalizzazione di 3 dB[30] rispetto alle prestazioni di SIMO-MRC con $n_R = 2$.

Il legame esatto tra la P_e legata all'uso della (21.25) e l'SNR medio

$$\Gamma = E\left\{\gamma_{Ala}\right\} = \frac{\mathcal{E}_s}{\sigma_n^2} E\left\{|h_1|^2 + |h_2|^2\right\}$$

dipende dal tipo di modulazione, ma per il caso BPSK la figura a lato riporta il risultato di simulazioni messe a confronto con le curve (pag. 711) dei casi SISO, SIMO ed AWGN, in modo da poter apprezzare l'effetto della perdita di 3 dB prima indicata sul risultato finale, poca cosa rispetto al miglioramento comunque conseguito. La curva indicata in figura come MIMO 2,2 si riferisce al caso che affrontiamo appresso.



21.3.2.4 Ricezione multiantenna di un codice di Alamouti

Eccoci dunque arrivati alla prima configurazione propriamente MIMO: vediamo come ricevere una trasmissione MISO-STBC mediante una schiera di n_R antenne di ricezione, conseguendo un ordine di diversità *massimo* pari a $2n_R$, ossia quello ottenuto con due antenne di trasmissione moltiplicato per n_R.

Il trasmettitore non modifica il suo operato, ed invia la codifica espressa in tabella 21.1 mediante le sue due antenne. Facciamo dapprima il caso di adottare $n_R = 2$, ricevendo così un totale di *quattro* diversi segnali

$$\begin{aligned} r_{11} &= h_{11}s_1 + h_{12}s_2 + n_1 \\ r_{21} &= h_{21}s_1 + h_{22}s_2 + n_2 \\ r_{12} &= -h_{11}s_2^* + h_{12}s_1^* + n_3 \\ r_{22} &= -h_{21}s_2^* + h_{22}s_1^* + n_4 \end{aligned} \tag{21.26}$$

in cui r_{ik} è il segnale ricevuto dall'antenna $i = 1, 2$ all'istante $k = 1, 2$, ed h_{ij} è il

[30] Questi 3 dB di differenza sono da interpretare come un guadagno *di array* legato al disporre di due antenne di ricezione, per cui in pratica viene ricevuta *il doppio* della potenza che si riceverebbe con una sola antenna, mentre in trasmissione ciò non si verifica, per la limitazione sulla potenza massima trasmessa.

coefficiente complesso del canale radio tra le antenne j e i. I valori r_{ik} delle (21.26) vengono quindi combinati per formare le grandezze di decisione

$$\begin{aligned}\hat{s}_1 &= h_{11}^* r_{11} + h_{12} r_{12}^* + h_{21}^* r_{21} + h_{22} r_{22}^* \\ \hat{s}_2 &= h_{12}^* r_{11} - h_{11} r_{12}^* + h_{22}^* r_{21} - h_{21} r_{22}^*\end{aligned} \tag{21.27}$$

e sostituendo le (21.26) nelle (21.27) dopo alcuni passaggi si ottiene

$$\begin{aligned}\hat{s}_1 &= \left(|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2\right) s_1 + h_{11}^* n_1 + h_{12} n_3^* + h_{21}^* n_2 + h_{22} n_4^* \\ \hat{s}_2 &= \left(|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2\right) s_2 + h_{12}^* n_1 - h_{11} n_3^* + h_{22}^* n_2 - h_{21} n_4^*\end{aligned} \tag{21.28}$$

riproducendo quindi la situazione già osservata alle (21.23), con la differenza che ora il *rinforzo* per le ampiezze dei simboli di sorgente vale $\alpha = |h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2$, ottenendo un'ordine di diversità complessivo che è *il prodotto* degli ordini ai due estremi multiantenna.

Gli stessi risultati sono facilmente estensibili ad un numero $n_R > 2$ di antenne riceventi, dato che ponendo le (21.26) nella forma

$$\begin{aligned}r_{i1} &= h_{i1} s_1 + h_{i2} s_2 + n_{i1} \\ r_{i2} &= -h_{i1} s_2^* + h_{i2} s_1^* + n_{i2}\end{aligned} \tag{21.29}$$

con $i = 1, 2, \cdots n_R$, le (21.27) si estendono come

$$\begin{aligned}\hat{s}_1 &= \textstyle\sum_{i=1}^{n_R} \left(h_{i1}^* r_{i1} + h_{i2} r_{i2}^*\right) \\ \hat{s}_2 &= \textstyle\sum_{i=1}^{n_R} \left(h_{i2}^* r_{i1} - h_{i1} r_{i2}^*\right)\end{aligned} \tag{21.30}$$

ed al posto delle (21.28) otteniamo

$$\begin{aligned}\hat{s}_1 &= \textstyle\sum_{i=1}^{n_R} \left[\left(|h_{i1}|^2 + |h_{i2}|^2\right) s_1 + h_{i1}^* n_{i1} + h_{i2} n_{i2}^*\right] \\ \hat{s}_2 &= \textstyle\sum_{i=1}^{n_R} \left[\left(|h_{i1}|^2 + |h_{i2}|^2\right) s_2 + h_{i2}^* n_{i1} - h_{i1} n_{i2}^*\right]\end{aligned} \tag{21.31}$$

Ortogonalità Mostriamo come i risultati fin qui ottenuti siano una diretta conseguenza del fatto che le righe della matrice delle codeword $\boldsymbol{C} = \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}$ di Alamouti sono *ortogonali* tra loro, ossia il prodotto scalare tra due diverse righe è nullo, e dunque

$$\begin{aligned}\boldsymbol{C}\boldsymbol{C}^\dagger &= \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix} \begin{bmatrix} s_1^* & -s_2 \\ s_2^* & s_1 \end{bmatrix} = \begin{bmatrix} s_1 s_1^* + s_1 s_2^* & -s_1 s_2 + s_1 s_2 \\ -s_1^* s_2^* + s_1^* s_2^* & s_2 s_2^* + s_1 s_1^* \end{bmatrix} = \\ &= \begin{bmatrix} |s_1|^2 + |s_2|^2 & 0 \\ 0 & |s_1|^2 + |s_2|^2 \end{bmatrix}\end{aligned}$$

in cui l'operatore † è noto come *Hermitiano* ed esegue il coniugato della trasposta della matrice[31], equivalente per le matrici complesse dell'operazione di coniugazione valida per gli scalari. E' stato dimostrato che il codice di Alamouti è l'unico STBC a simboli complessi con codeword ortogonali.

Esprimendo ora la (21.29) in forma matriciale come

$$\begin{bmatrix} r_{i1} \\ r_{i2}^* \end{bmatrix} = \begin{bmatrix} h_{i1} & h_{i2} \\ h_{i2}^* & -h_{i1}^* \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} n_{i1} \\ n_{i2}^* \end{bmatrix} \quad \text{ovvero} \quad \boldsymbol{r}_i = \boldsymbol{G}_i \boldsymbol{s} + \boldsymbol{n}_i$$

[31] Indicato anche come operatore *aggiunto*, mentre $\boldsymbol{A}^\dagger$ è detta *matrice aggiunta* di $\boldsymbol{A}$.

osserviamo come anche la matrice $\boldsymbol{G}_i = \begin{bmatrix} h_{i1} & h_{i2} \\ h_{i2}^* & -h_{i1}^* \end{bmatrix}$ sia *ortogonale*, per qualunque valore dei coefficienti h, e dunque $\boldsymbol{G}_i \boldsymbol{G}_i^\dagger = \alpha_i \boldsymbol{I}_2$ in cui $\boldsymbol{I}_2$ è la matrice identità 2×2 ed $\alpha_i = |h_{i1}|^2 + |h_{i2}|^2$. Pre-moltiplicando il vettore ricevuto dalla $i-$esima antenna $\boldsymbol{r}_i$ per $\boldsymbol{G}_i^\dagger = \begin{bmatrix} h_{i1}^* & h_{i2} \\ h_{i2}^* & -h_{i1} \end{bmatrix}$, come avviene per le (21.22), otteniamo quindi

$$\tilde{\boldsymbol{r}}_i = \boldsymbol{G}_i^\dagger \boldsymbol{r}_i = \boldsymbol{G}_i^\dagger \boldsymbol{G}_i \, \boldsymbol{s} + \boldsymbol{G}_i^\dagger \boldsymbol{n}_i = \alpha_i \boldsymbol{s} + \tilde{\boldsymbol{n}}_i$$

dove $\tilde{\boldsymbol{n}}_i$ è ancora gaussiano complesso a media nulla e componenti incorrelate. A questo punto la decisione di massima verosimiglianza per l'antenna i assume la forma

$$\boldsymbol{s}^\diamond = \arg \min_{\boldsymbol{s} \in \mathcal{A}^2} (\tilde{\boldsymbol{r}}_i - \alpha_i \boldsymbol{s}) \, (\tilde{\boldsymbol{r}}_i - \alpha_i \boldsymbol{s})^* \tag{21.32}$$

ma essendo come osservato $\tilde{\boldsymbol{n}}_i = \tilde{\boldsymbol{r}}_i - \alpha_i \, \boldsymbol{s}$ a componenti incorrelate, la (21.32) si scompone in due minimizzazioni indipendenti, come espresso dalle (21.25) e (21.31).

21.3.3 Prestazioni limite

Abbandoniamo gli sviluppi ottenuti per il codice di Alamouti per tornare al caso più generale, e definire uno STBC come una procedura per ottenere una codeword definita da una matrice ad elementi complessi

$$\boldsymbol{C}^j = \begin{bmatrix} c_{1,1}^j & c_{1,2}^j & \cdots & c_{1,n_T}^j \\ c_{2,1}^j & c_{2,2}^j & \cdots & c_{2,n_T}^j \\ \vdots & \vdots & \ddots & \vdots \\ c_{T,1}^j & c_{T,2}^j & \cdots & c_{T,n_T}^j \end{bmatrix}$$

con $j = 1, 2, \cdots, 2^{kM}$ in corrispondenza di ciascuna delle altrettante possibili sequenze di k simboli complessi a 2^M valori, come descritto al § 21.3.2.2, da trasmettere da parte di n_T antenne in $T \geq n_T$ istanti e ricevere mediante n_R antenne, cercando di individuare quali siano i fattori principali che concorrono alle prestazioni del codice, in termini di probabilità di errore.

A fronte di sviluppi analitici che tralasciamo[32], si dimostra che la probabilità di decidere erroneamente per la codeword $\boldsymbol{C}^i$ quando viene trasmessa $\boldsymbol{C}^j$ per valori elevati di SNR medio Γ non supera un valore

$$Prob\left\{\boldsymbol{C}^j \rightarrow \boldsymbol{C}^i\right\} \leq \frac{1}{\left(G_c \, \Gamma \cdot 1/4\right)^{G_d}} \tag{21.33}$$

in cui $G_d = r \cdot n_R$ esprime il *guadagno di diversità* che dipende oltre che da n_R anche dal *rango*[33] r della matrice $\boldsymbol{A} = \left(\boldsymbol{C}^i - \boldsymbol{C}^j\right)^\dagger \left(\boldsymbol{C}^i - \boldsymbol{C}^j\right)$, mentre $G_c = \sqrt[r]{\prod_{n=1}^{r} \lambda_n}$ misura il *guadagno del codice* e dipende dagli autovalori λ_n diversi da zero della matrice $\boldsymbol{A}$.

[32] Che ho trovato accennati, con i dovuti rimandi, su E. KROUK, S. SEMENOV, *Modulation and coding techniques in wireless communications*, 2011 John Wiley & Sons Ltd.

[33] Il rango di una matrice quadrata $\boldsymbol{A}$ $n \times n$ è definito come il numero delle sue righe (o colonne) linearmente indipendenti, ma è anche uguale al numero di autovalori diversi da zero, dove gli autovalori sono gli zeri del polinomio caratteristico definito come $\det(A - \lambda I_n)$

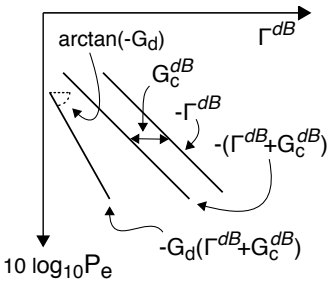
Pertanto il limite superiore (21.33) per la P_e può essere ridotto aumentando sia G_c che G_d, e se per tutte le coppie di codeword i, j la matrice $\boldsymbol{A}$ è a rango pieno (pari ad n_T), allora *l'ordine* di diversità conseguito dal codice è pari a $n_T \cdot n_R$.

Esempio Il simbolo vettoriale $\boldsymbol{s} = [s_1, s_2]^T$ è codificato secondo Alamouti come $\boldsymbol{C} = \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}$, mentre un diverso input $\boldsymbol{s'} = [s'_1, s'_2]^T$ come $\boldsymbol{C'} = \begin{bmatrix} s'_1 & s'_2 \\ -s'^*_2 & s'^*_1 \end{bmatrix}$, dunque $\boldsymbol{C} - \boldsymbol{C'} = \begin{bmatrix} s_1 - s'_1 & s_2 - s'_2 \\ -s_2^* + s'^*_2 & s_1^* - s'^*_1 \end{bmatrix}$ il cui determinante vale $|s_1 - s'_1|^2 + |s_2 - s'_2|^2 \neq 0$ per qualsiasi $\boldsymbol{s} \neq \boldsymbol{s'}$. Pertanto la matrice $\boldsymbol{A}$ ha sempre rango pieno[34], ed il codice offre un ordine di diversità pari a $n_T \cdot n_R$.

Prestazioni asintotiche Prima di procedere svolgiamo una interessante considerazione sulla legge di dipendenza della P_e a simbolo da G_c e G_d, come espressa dalla (21.33), che per SNR elevato possiamo generalizzare come $P_e \propto (G_c \Gamma)^{-G_d}$ o, esprimendo le grandezze in dB

$$10 \log_{10} P_e \propto 10 \log_{10} (G_c \Gamma)^{-G_d} = -G_d \left(G_c^{dB} + \Gamma^{dB} \right)$$

che rappresentiamo alla figura a lato[35] in modo da meglio apprezzare come, mentre il guadagno di codifica G_c^{dB} determina una semplice *traslazione a sinistra* della curva di prestazione, il guadagno di diversità G_d *ne modifica l'inclinazione*, e dunque ha un effetto sempre più pronunciato all'aumentare dell'SNR.



21.3.4 Codici sub ottimi

Concludiamo la sezione sulla diversità spaziale citando le alternative che, utilizzando più di due antenne in trasmissione, conseguono un ordine di diversità pari a $n_T \cdot n_R$ a spese di un tasso di codifica $R_c = k/T < 1$, oppure conseguono $R_c = 1$ ma con un ordine di diversità ridotto.

Codice ortogonale reale quadrato Un primo risultato da citare è che limitando il campo di applicazione a simboli *reali* (ovvero ad un modulazione L-ASK) possono essere definiti STBC *ortogonali*[36] con codeword *quadrate* ovvero $n_T = T$ pari a 2, 4 ed 8, ed in grado di offrire $R_c = 1$ e diversità $n_T \cdot n_R$, prestazioni che per simboli complessi sono invece conseguite esclusivamente dal codice di Alamouti con $n_T = 2$. Un esempio di STBC che usa $n_T = 4$ antenne per trasmettere $k = 4$ simboli *reali* in $T = 4$ istanti è indicato come $\boldsymbol{C}_4$ nello schema che segue:

[34] Infatti il determinante è anche pari al prodotto degli autovalori, e dunque il suo essere $\neq 0$ implica che non vi siano autovalori nulli.

[35] Anziché mostrare la P_e in scala logaritmica come si fa di solito, per coerenza con l'espressione $10 \log_{10} P_e \propto= -G_d \left(G_c^{dB} + \Gamma^{dB} \right)$ sulle ordinate è mostrato il log di P_e, ed essendo $P_e < 1$, il suo log è negativo.

[36] L'ortogonalità delle codeword è la proprietà che consente di conseguire la piena diversità spaziale e che permette la decodifica semplificata discussa precedentemente.

$$\boldsymbol{C}_4 = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ -s_1 & s_2 & -s_3 & s_4 \\ -s_1 & s_2 & s_3 & -s_4 \\ -s_1 & -s_2 & s_3 & s_4 \end{bmatrix} \quad \boldsymbol{C}_{8\times4}^T = \begin{bmatrix} s_1 & -s_1 & -s_1 & -s_1 & s_1^* & -s_1^* & -s_1^* & -s_1^* \\ s_2 & s_2 & s_2 & -s_2 & s_2^* & s_2^* & s_2^* & -s_2^* \\ s_3 & -s_3 & s_3 & s_3 & s_3^* & -s_3^* & s_3^* & s_3^* \\ s_4 & s_4 & -s_4 & s_4 & s_4^* & s_4^* & -s_4^* & s_4^* \end{bmatrix}$$

Codice complesso associato Un secondo risultato è che è possibile ottenere un codice ortogonale *complesso* a partire da uno reale trasmettendo la stessa struttura di codeword ma ad elementi complessi, fatta seguire da altrettanti T istanti in cui viene trasmessa la codeword coniugata, come esemplificato dalla matrice $\boldsymbol{C}_{8\times4} = \begin{bmatrix} \boldsymbol{C}_4 \\ \boldsymbol{C}_4^* \end{bmatrix}$, sopra rappresentata *trasposta* per ragioni di spazio. Come evidente l'adozione di $\boldsymbol{C}_{8\times4}$ comporta un raddoppio del numero di istanti necessari alla trasmissione, e dunque un dimezzamento di R_c; d'altra parte lo stesso raddoppio di T determina anche un miglioramento di 3 dB per l'SNR di decisione, in quanto uno stesso simbolo s_i viene ricevuto il doppio delle volte. D'altra parte le esigenze di *tempo reale* della trasmissione comportano che il raddoppio del numero di istanti T richieda il dimezzamento della loro durata, e quindi la ricezione di metà dell'energia per simbolo: tale peggioramento è compensato dal guadagno di 3 dB legato alla doppia trasmissione.

Codice reale non quadrato E' anche possibile realizzare STBC *reali* con codeword *non quadrate* ovvero con $n_T \neq T$, purché il numero di elementi $T \cdot n_T$ nella matrice del codice sia un multiplo intero del numero k di simboli da codificare; tali soluzioni esistono per qualunque n_T, conseguono massima diversità e decodifica lineare, e quando utilizzano un numero di istanti T pari a quello k dei simboli da rappresentare esibiscono anche un tasso di codifica $R_c = 1$

Esempio Citiamo il caso di un SBTC reale con $n_T = 3$ e $k = T = 4$ mostrato a lato, che ripete ciascun simbolo tre volte e consegue $R_c = 1$; quando da questo se ne ricava un SBTC complesso concatenando la matrice con la coniugata, il tasso si riduce ad $1/2$.

$$\boldsymbol{C}_{4\times3} = \begin{bmatrix} s_1 & s_2 & s_3 \\ -s_2 & s_1 & -s_4 \\ -s_3 & s_4 & s_1 \\ -s_4 & -s_3 & s_2 \end{bmatrix}$$

Ritardo minimo Finché ci si limita a raddoppiare un codice reale per farne uno complesso come nell'esempio precedente, il tasso non può essere maggiore di $1/2$; sono però stati scoperti codici complessi in grado di conseguire $R_c = 3/4$ che, a parità del numero n_T di antenne, riducono il numero T di istanti necessari, conseguendo anche un minor *ritardo* di codifica.

Codici quasi-ortogonali Anziché sacrificare il tasso R_c per l'ortogonalità, è possibile fare l'opposto e mantenere $R_c = 1$ assieme alla piena diversità (con $n_T > 2$) mediante STBC quasi-ortogonali[37], penalizzati da una modesta perdita di prestazioni (P_e) ed una maggior complessità di decodifica.

[37] Ovvero per i quali il prodotto $\boldsymbol{C} \cdot \boldsymbol{C}^\dagger$ non assume la forma $\alpha \boldsymbol{I}_T$.

21.4 Capacità di canale con fading di Rayleigh

Le tecniche di *valorizzazione* della diversità spaziale fin qui discusse apportano un beneficio alla comunicazione che è stato quantificato come un aumento dell'SNR per la variabile di decisione. Tale aumento può essere *capitalizzato* aumentando il numero M di bit (e di livelli $L = 2^M$) rappresentati da ciascun simbolo s_i del messaggio da trasmettere, aumentando così la velocità di trasmissione, a parità di periodo di simbolo e di banda occupata. Il legame tra SNR di un canale AWGN e la conseguente massima velocità di trasmissione è stato analizzato al § 17.3 ed espresso nei termini della *capacità* C del canale; affrontiamo ora l'adattamento di tale formulazione al caso di canale affetto da fading di Rayleigh ed in presenza di più antenne trasmittenti e/o riceventi, descritto dalla relazione

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n} \tag{21.34}$$

come illustrato al § 21.2, con il vincolo di mantenere la potenza $\mathcal{P}_s = E\left\{\boldsymbol{s}^2\right\} \cdot f_s$ emessa *collettivamente* da tutte le n_T antenne trasmittenti limitata a $\mathcal{P}_T$.

Canale SISO Per avere un termine di paragone rispetto al quale confrontare i risultati valutiamo per prima l'espressione di C con $n_T = n_R = 1$, nel cui caso la (21.34) è una semplice relazione scalare (a valori complessi) $r = hs + n$. L'espressione (eq. (17.18)) della capacità AWGN $C = B \cdot \log_2\left(1 + \frac{\mathcal{P}_T}{N_0 B}\right)$ [bit/sec] viene riscritta come[38]

$$C_{SISO} = \log_2\left(1 + \rho\,|h|^2\right) \quad \text{bit/secondo/Hertz} \tag{21.35}$$

in cui ρ è l'SNR che si sarebbe ricevuto in assenza di fading, $|h|$ ha una d.d.p. di Rayleigh dovuta al fading, ed il risultato viene espresso *per Hertz* di banda occupata anziché in forma cumulativa, e dunque la (21.35) ha il significato di massima *efficienza spettrale* (pag. 498) conseguibile. La stessa (21.35) può essere equivalentemente misurata in bit/simbolo o bit/uso del canale.

[38] Ripercorriamo lo sviluppo svolto al § 17.3 per un canale a valori *reali*, che definisce $C = \max_{p(s)}\{I(S;R)\}$ in cui $I(S;R) = h(R) - h(R/S)$ è l'informazione mutua media tra simbolo s trasmesso e valore ricevuto r. Il termine $h(R/S)$ è dovuto al solo rumore, e per esso rimane valido il risultato (17.16) ovvero $h(R/S) = \frac{1}{2}\log_2\left(2\pi\mathrm{e}\sigma_n^2\right)$; il massimo di C si ottiene quindi massimizzando $h(R)$, che come noto (§ 9.7.2) avviene con r gaussiano, fornendo $h(R) = \frac{1}{2}\log_2\left(2\pi\mathrm{e}\sigma_r^2\right)$ in cui per σ_r^2 si ottiene

$$\sigma_r^2 = E\left\{r^2\right\} = E\left\{(hs+n)(hs+n)\right\} = |h|^2 E\left\{s^2\right\} + E\left\{n^2\right\} = |h|^2\sigma_s^2 + \sigma_n^2$$

in virtù dell'incorrelazione tra s ed n, entrambi a media nulla. Procedendo come al § 17.3 si ha

$$C = \max_{p(s)}\{I(S;R)\} = \max_{p(s)}\{h(R) - h(R/S)\} = \frac{1}{2}\log_2\left(2\pi\mathrm{e}\left(\sigma_n^2 + \sigma_s^2|h|^2\right)\right) - \frac{1}{2}\log_2\left(2\pi\mathrm{e}\sigma_n^2\right) =$$
$$= \frac{1}{2}\log_2\frac{\sigma_n^2 + \sigma_s^2|h|^2}{\sigma_n^2} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_s^2}{\sigma_n^2}|h|^2\right)$$

in cui definiamo $\frac{\sigma_s^2}{\sigma_n^2} = \rho$ pari all'SNR in assenza di fading. Per arrivare alla (21.35) osserviamo come un canale a valori *complessi* equivalga a due canali reali indipendenti (basti pensare al mo-demodulatore in fase e quadratura), essenzialmente in virtù della indipendenza delle parti reale ed immaginaria del coefficiente complesso h e del rumore n. Pertanto la capacità del canale complesso è *il doppio* di quanto ora calcolato, come espresso dalla (21.35). Qualcuno può giustamente chiedersi ora se l'SNR ρ sia da riferirsi al singolo ramo (I o Q), oppure al segnale complesso. In realtà i due valori sono *equivalenti*, perché per il canale complesso sia σ_s^2 che σ_n^2 *raddoppiano*.

Capacità ergodica e garantita Osserviamo che il valore (21.35), come quelli che seguiranno, è in realtà una v.a., in quanto h lo è. A partire dall'espressione (21.35) si possono ricavare due *numeri*: il primo si ottiene eseguendo il valore atteso di C rispetto alla variabilità di h, ottenendo un valore *medio* indicato anche come capacità *ergodica*. Il secondo è invece riferito ad un grado di servizio g ed esprime il valore di capacità *superato* nel $g\%$ del tempo, ed in questo senso detta capacità *garantita* (per il $g\%$ del tempo). Ma ai fini di quel che segue sono presi in considerazione casi in cui h non varia per tutta la durata del collegamento, in modo da permettere il confronto tra risultati.

Canale SIMO In presenza di n_R antenne riceventi al § 21.3.1.2 si è analizzato come un ricevitore MRC determini un aumento dell'SNR di una quantità pari a $\sum_{i=1}^{n_R} |h_i|^2$, e difatti anche il valore teorico di capacità risulta pari a

$$C_{SIMO} = \log_2\left(1 + \rho \sum_{i=1}^{n_R} |h_i|^2\right) \quad \text{bit/secondo/Hertz} \tag{21.36}$$

in cui h_i è il guadagno complesso per l'$i-$esima antenna. Notiamo che la dipendenza tra n_R e C_{SIMO} è di tipo *logaritmico*, e dunque l'incremento di C_{SIMO} è via via minore all'aumentare di n_R.

Canale MISO Nel caso di n_T antenne trasmittenti, ma in assenza di conoscenza da parte del trasmettitore dei valori che compaiono nella matrice $\boldsymbol{H}$, ogni antenna utilizza un n_T-esimo della potenza totale di trasmissione, e la capacità risulta pari a

$$C_{MISO} = \log_2\left(1 + \frac{\rho}{n_T} \sum_{i=1}^{n_T} |h_i|^2\right) \quad \text{bit/secondo/Hertz} \tag{21.37}$$

ovvero oltre all'incremento solamente logaritmico con n_T, si assiste anche alla progressiva riduzione dell'SNR di ricezione, in virtù del vincolo sulla potenza totale di trasmissione che va suddivisa su tutte le n_T antenne.

21.4.1 Capacità del canale MIMO

Il calcolo della capacità in presenza di più antenne sia in trasmissione (n_T) che in ricezione (n_R) segue le linee guida indicate alla nota 38, con l'obiettivo di rendere massima l'informazione mutua media

$$I(\boldsymbol{s};\boldsymbol{r}) = h(\boldsymbol{r}) - h(\boldsymbol{r}/\boldsymbol{s})$$

tra i vettori dei simboli trasmessi $\boldsymbol{s}$ e di quelli ricevuti $\boldsymbol{r}$ in modo da ottenere $C = \max_{p(\boldsymbol{s})} \{I(\boldsymbol{s};\boldsymbol{r})\}$. Dato che si suppone la matrice di canale $\boldsymbol{H}$ perfettamente nota in ricezione la sua incertezza condizionata ad $\boldsymbol{s}$ è nulla, dunque

$$h(\boldsymbol{r}/\boldsymbol{s}) = h(\boldsymbol{H}\boldsymbol{s} + \boldsymbol{n}/\boldsymbol{s}) = h(\boldsymbol{n}/\boldsymbol{s}) = h(\boldsymbol{n})$$

in quanto $\boldsymbol{n}$ ed $\boldsymbol{s}$ sono statisticamente indipendenti, e quindi

$$I(\boldsymbol{s};\boldsymbol{r}) = h(\boldsymbol{r}) - h(\boldsymbol{n}) \tag{21.38}$$

L'entropia differenziale $h(\boldsymbol{n})$ del vettore gaussiano complesso $\boldsymbol{n}$ è calcolata al § 21.9.1 come $h(\boldsymbol{n}) = \log(\det(\pi e \boldsymbol{\Sigma}_n))$, dove $\boldsymbol{\Sigma}_{\boldsymbol{n}} = E\left\{\boldsymbol{n}\boldsymbol{n}^\dagger\right\}$ è la matrice di covarianza del rumore $\boldsymbol{n}$. Per massimizzare la (21.38) occorre quindi massimizzare $h(\boldsymbol{r})$, il che avviene

quando anche $\boldsymbol{r}$ è un vettore gaussiano complesso, per il quale nuovamente $h(\boldsymbol{r}) = \log(\det(\pi e \Sigma_r))$, in cui la matrice di covarianza Σ_r del vettore (a media nulla) ricevuto risulta pari a

$$\begin{aligned}\Sigma_r &= E\left\{\boldsymbol{r}\boldsymbol{r}^\dagger\right\} = E\left\{(\boldsymbol{H}\boldsymbol{s}+\boldsymbol{n})(\boldsymbol{H}\boldsymbol{s}+\boldsymbol{n})^\dagger\right\} = E\left\{\boldsymbol{H}\boldsymbol{s}(\boldsymbol{H}\boldsymbol{s})^\dagger\right\} + E\left\{\boldsymbol{n}\boldsymbol{n}^\dagger\right\} = \\ &= \boldsymbol{H}E\left\{\boldsymbol{s}\boldsymbol{s}^\dagger\right\}\boldsymbol{H}^\dagger + \Sigma_n = \boldsymbol{H}\Sigma_s\boldsymbol{H}^\dagger + \Sigma_n\end{aligned}$$

data l'incorrelazione tra $\boldsymbol{s}$ ed $\boldsymbol{n}$, ed avendo sostituito $(\boldsymbol{H}\boldsymbol{s})^\dagger = \boldsymbol{s}^\dagger\boldsymbol{H}^\dagger$. Mettendo tutto assieme otteniamo dunque

$$\begin{aligned}C_{MIMO} &= \max_{p(\boldsymbol{s})}\{I(\boldsymbol{s};\boldsymbol{r})\} = \max_{p(\boldsymbol{s})}\{h(\boldsymbol{r}) - h(\boldsymbol{n})\} = \\ &= \log_2\left(\det\left[\pi e\Sigma_r\right]\right) - \log_2\left(\det\left[\pi e\Sigma_n\right]\right) = \log_2\left(\frac{\det\left[\pi e\Sigma_r\right]}{\det\left[\pi e\Sigma_n\right]}\right) = \\ &= \log_2\left(\det\left[\Sigma_r\Sigma_n^{-1}\right]\right) = \log_2\left(\det\left[\left(\boldsymbol{H}\Sigma_s\boldsymbol{H}^\dagger + \Sigma_n\right)\Sigma_n^{-1}\right]\right) = \\ &= \log_2\left(\det\left[\boldsymbol{H}\Sigma_s\boldsymbol{H}^\dagger\Sigma_n^{-1} + \Sigma_n\Sigma_n^{-1}\right]\right) = \\ &= \log_2\left(\det\left[\boldsymbol{I}_{n_R} + \Sigma_n^{-1}\boldsymbol{H}\Sigma_s\boldsymbol{H}^\dagger\right]\right) \quad \text{bit/sec/Hz}\end{aligned} \tag{21.39}$$

in cui la capacità MIMO viene a dipendere dalla covarianza di sorgente Σ_s, da quella di rumore Σ_n, e dalla matrice di canale $\boldsymbol{H}$. Vediamo ora come diverse ipotesi su tali grandezze permettano di arrivare a risultati adatti ai singoli casi.

Rumore incorrelato tra antenne La circostanza di avere un rumore *spazialmente bianco*, che si verifica quando per le componenti n_i del vettore $\boldsymbol{n}$ risulta $E\left\{n_i n_j^*\right\} = 0$ con $i \neq j$ oppure σ_n^2 se $i = j$, può essere interpretata come una assenza di interferenti *in comune* tra le antenne di ricezione. A ciò consegue una covarianza *del rumore* nella forma $\Sigma_n = \sigma_n^2\boldsymbol{I}_{n_R}$, ossia una matrice nulla tranne che sulla diagonale, dove vale σ_n^2. In tal caso la (21.39) diviene

$$C_{MIMO-WN} = \log_2\left(\det\left[\boldsymbol{I}_{n_R} + \frac{1}{\sigma_n^2}\boldsymbol{H}\Sigma_s\boldsymbol{H}^\dagger\right]\right) \quad \text{bit/sec/Hz} \tag{21.40}$$

Potenza trasmessa uniforme e simboli incorrelati In generale il trasmettitore non è a conoscenza dei valori di $\boldsymbol{H}$, e l'unica cosa che può fare è trasmettere con la stessa potenza su tutte le n_T antenne. Allo stesso tempo è lecito considerare i simboli trasmessi incorrelati, ossia per due elementi s_i, s_j di $\boldsymbol{s}$ si ha $E\left\{s_i s_j\right\} = 0$ se $i \neq j$. Sotto tali condizioni la covarianza del segnale è pari a $\Sigma_s = \frac{\mathcal{E}_s}{n_T}\boldsymbol{I}_{n_T}$ ossia è tutta nulla tranne che sulla diagonale, dove vale un n_T-esimo dell'energia per simbolo $\mathcal{E}_s = E\left\{\boldsymbol{s}^\dagger\boldsymbol{s}\right\}$. In queste circostanze la (21.40) diviene

$$C_{MIMO-EP} = \log_2\left(\det\left[\boldsymbol{I}_{n_R} + \frac{\rho}{n_T}\boldsymbol{H}\boldsymbol{H}^\dagger\right]\right) \quad \text{bit/sec/Hz} \tag{21.41}$$

in cui $\rho = \frac{\mathcal{E}_s}{\sigma_n^2}$ è l'SNR *per simbolo* del segnale ricevuto da ciascuna delle n_R antenne a partire da *tutte* le n_T trasmittenti.

Derivazione della capacità MISO e SIMO Notiamo che la condizione di equipartizione di $\mathcal{P}_T$ è in comune alla configurazione MISO che conduce alla (21.37), ma mentre

in quel caso la capacità aumenta con legge logaritmica rispetto ad n_T, si dimostra che la (21.41) aumenta *linearmente* con $m = \min(n_T, n_R)$. Osserviamo inoltre che (21.41) diviene esattamente pari a C_{MISO} (eq. (21.37)) quando $n_R = 1$, dato che in questa circostanza $\boldsymbol{H}$ è un vettore riga, e dunque $\boldsymbol{H}\boldsymbol{H}^\dagger = \sum_{i=1}^{n_T} |h_i|^2$. Per lo stesso motivo (21.41) diviene pari a C_{SIMO} (eq. (21.36)) quando $n_T = 1$.

Il contributo di $\boldsymbol{H}$ Quando entrambi i lati del collegamento sono equipaggiati con più antenne la (21.41) può essere riscritta in una forma che permette di individuare quali caratteristiche della matrice di canale MIMO $\boldsymbol{H}$ di dimensioni $n_R \times n_T$ rendono il valore $C_{MIMO-EP}$ più o meno grande. Andiamo a mostrare che ciò dipende dagli *autovalori non nulli* $\lambda_1, \lambda_2, \cdots, \lambda_m$ della matrice $n_R \times n_R$

$$\boldsymbol{W} = \boldsymbol{H}\boldsymbol{H}^\dagger$$

simmetrica, semidefinita positiva e di rango $m \leq \min(n_T, n_R)$[39], che può quindi essere espressa (§ 6.7.3) come $\boldsymbol{W} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\dagger$ in cui $\boldsymbol{\Lambda}$ è una matrice quadrata tutta nulla tranne per gli autovalori λ_i sulla diagonale, e $\boldsymbol{\Gamma}$ è una matrice unitaria[40] con colonne pari agli autovettori di $\boldsymbol{W}$, posti nello stesso ordine con cui gli autovalori compaiono in $\boldsymbol{\Lambda}$. Partendo dalla (21.41) scriviamo pertanto

$$\begin{aligned} C_{MIMO-EP} &= \log_2\left(\det\left[\boldsymbol{I}_{n_R} + \frac{\rho}{n_T}\boldsymbol{H}\boldsymbol{H}^\dagger\right]\right) = \log_2\left(\det\left[\boldsymbol{I}_{n_R} + \frac{\rho}{n_T}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\dagger\right]\right) = \\ &= \log_2\left(\det\left[\boldsymbol{I}_m + \frac{\rho}{n_T}\boldsymbol{\Gamma}^\dagger\boldsymbol{\Gamma}\boldsymbol{\Lambda}\right]\right) = \log_2\left(\det\left[\boldsymbol{I}_m + \frac{\rho}{n_T}\boldsymbol{\Lambda}\right]\right) = \\ &= \log_2\left(\left(1 + \frac{\rho}{n_T}\lambda_1\right)\left(1 + \frac{\rho}{n_T}\lambda_2\right)\cdots\left(1 + \frac{\rho}{n_T}\lambda_m\right)\right) = \\ &= \sum_{i=1}^{m} \log_2\left(1 + \frac{\rho}{n_T}\lambda_i\right) \quad \text{bit/sec/Hz} \end{aligned} \tag{21.42}$$

in cui la terza eguaglianza deriva dall'identità riportata alla nota 39, la quarta dall'essere $\boldsymbol{\Gamma}$ unitaria, mentre alla quinta si è sviluppato il determinante come prodotto dei termini sulla diagonale. Discutiamo del risultato non appena ottenuto:

- la (21.42) mostra come $C_{MIMO-EP}$ sia pari alla *somma* delle capacità di m canali *virtuali* di tipo SISO (eq. (21.35)) *indipendenti*, ognuno con SNR per simbolo pari a $\frac{\rho}{n_T} = \frac{1}{n_T}\frac{\mathcal{E}_s}{\sigma_n^2}$ e guadagno di Rayleigh $|h|^2 = \lambda_i$;
- essendo il determinante pari al prodotto degli autovalori, la matrice $\boldsymbol{I}_m + \frac{\rho}{n_T}\boldsymbol{\Lambda} = \begin{bmatrix} 1 + \frac{\rho}{n_T}\lambda_1 & 0 & \cdots & 0 \\ 0 & 1 + \frac{\rho}{n_T}\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 + \frac{\rho}{n_T}\lambda_m \end{bmatrix}$ di dimensioni $m \times m$ ha auto-

[39] La definizione $\boldsymbol{W} = \boldsymbol{H}\boldsymbol{H}^\dagger$ è valida quando $n_R \leq n_T$, mentre se $n_T < n_R$ conviene scrivere $\boldsymbol{W} = \boldsymbol{H}^\dagger\boldsymbol{H}$, con dimensioni $n_R \times n_R$ e $n_T \times n_T$ rispettivamente. Da questo punto di vista è da notare che in base al teorema di *Weinstein–Aronszajn* sussiste l'identità

$$\det\left[\boldsymbol{I}_{n_R} + \tfrac{\rho}{n_T}\boldsymbol{H}\boldsymbol{H}^\dagger\right] = \det\left[\boldsymbol{I}_{n_T} + \tfrac{\rho}{n_T}\boldsymbol{H}^\dagger\boldsymbol{H}\right]$$

vedi ad es. `https://en.wikipedia.org/wiki/Weinstein-Aronszajn_identity`: date due matrici A e B di dimensioni $m \times n$ ed $n \times m$, risulta $\det(I_m + AB) = \det(I_n + BA)$.

[40] Ovvero per la quale $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\dagger = \boldsymbol{\Gamma}^\dagger\boldsymbol{\Gamma} = \boldsymbol{I}$, ossia la matrice identità.

valori $1 + \frac{\rho}{n_T}\lambda_i$, così come anche $\boldsymbol{I}_{n_R} + \frac{\rho}{n_T}\boldsymbol{H}\boldsymbol{H}^\dagger$;

- essendo $\boldsymbol{W}$ semidefinita positiva[41] i suoi autovalori λ_i sono tutti reali e non negativi, risultando inoltre[42] $\det \boldsymbol{W} = \det \Lambda$. Dunque $\boldsymbol{I}_m + \frac{\rho}{n_T}\Lambda$ ha autovalori reali e positivi, così come $\boldsymbol{I}_{n_R} + \frac{\rho}{n_T}\boldsymbol{H}\boldsymbol{H}^\dagger$;
- la (21.42) è tanto più grande quanto maggiori sono gli autovalori λ_i, o meglio (dopo averli ordinati in ordine decrescente) quanto più il loro valore *non decade* troppo rapidamente. Autovalori piccoli o peggio ancora nulli (qualora il rango di $\boldsymbol{W}$ sia minore di n_R) sono indicativi di antenne poco distanziate, ovvero di una matrice $\boldsymbol{H}$ con righe (o colonne) circa pari alla combinazione lineare di altre righe (o colonne), ossia con valori h_{ij} correlati, da attribuirsi ad una ridotta variabilità del multipath, eventualmente anche dovuta ad un elevato *fattore di Rice* ossia una eccessiva visibilità tra le antenne (§ 20.3.1 e pag. 686);
- gli elementi della matrice *aleatoria* $\boldsymbol{W} = \boldsymbol{H}\boldsymbol{H}^\dagger$ sono pari a $w_{ij} = \sum_{k=1}^{n_T} h_{ik}h_{jk}^*$ ossia al *prodotto scalare* tra le righe i e j di $\boldsymbol{H}$, di valore tanto maggiore quanto più i guadagni complessi tra ciascuna delle due antenne riceventi i e j e tutte le n_T trasmittenti sono *paralleli*. Da questo punto di vista gli elementi w_{ij} sono proporzionali ad una *stima campionaria* della correlazione tra i valori ricevuti dalle antenne i e j.

21.4.1.1 Trasmissione a potenza differenziata

Rimuoviamo ora l'ipotesi $\boldsymbol{\Sigma}_s = \frac{\mathcal{E}_s}{n_T}\boldsymbol{I}_{n_T}$ adottata per giungere alle (21.41) e (21.42) ed investighiamo se esista una diversa matrice $\boldsymbol{\Sigma}_s$ tale da rendere il termine $\det\left[\boldsymbol{I}_{n_R} + \frac{1}{\sigma_n^2}\boldsymbol{H}\boldsymbol{\Sigma}_s\boldsymbol{H}^\dagger\right]$ che compare nella (21.40) maggiore di quanto ottenuto per la (21.42), mantenendo il vincolo[43]

$$\operatorname{tr}(\boldsymbol{\Sigma}_s) = \sum_{i=1}^{n_T} E\left\{s_i s_i^*\right\} = \mathcal{E}_s \leq \frac{\mathcal{P}_T}{f_s}$$

A questo scopo consideriamo per $\boldsymbol{H}$ la sua *scomposizione ai valori singolari* (SVD[44])

$$\boldsymbol{H} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\dagger$$

dove $\boldsymbol{D}$ è una matrice $n_R \times n_T$ tutta nulla tranne che per gli elementi sulla diagonale, noti come *valori singolari* di $\boldsymbol{H}$, e pari alla *radice quadrata* degli autovalori non nulli λ_i ($i = 1, \cdots, m$) di $\boldsymbol{H}\boldsymbol{H}^\dagger$ (ma anche di $\boldsymbol{H}^\dagger\boldsymbol{H}$), ovvero $(\boldsymbol{D})_{ii} = \lambda_i^{1/2}$, mentre $\boldsymbol{U}$ e $\boldsymbol{V}$ sono

[41] Ossia tale che $\Re\left\{\mathbf{c}^\dagger\boldsymbol{W}\mathbf{c}\right\} \geq 0$ per $\forall \boldsymbol{c}$, vedi anche § 6.7.3. Per la precisione $\boldsymbol{W}$ è una matrice *Hermitiana*, ossia tale che $\boldsymbol{W} = \boldsymbol{W}^\dagger$, e la condizione precedente diviene $\mathbf{c}^\dagger\boldsymbol{W}\mathbf{c} \geq 0$, vedi ad es. `https://it.wikipedia.org/wiki/Matrice_definita_positiva`. Anzi, ad essere ancora più precisi, per la sua natura aleatoria $\boldsymbol{W}$ è nota come *matrice di Wishart*, e la legge di distribuzione probabilistica dei suoi valori è descritta dalla omonima distribuzione, vedi ad es. `https://it.wikipedia.org/wiki/Distribuzione_di_Wishart`.

[42] In quanto $\boldsymbol{W}$ e Λ sono matrici *simili*, vedi ad es. `https://it.wikipedia.org/wiki/Similitudine_tra_matrici`.

[43] Ricordiamo che la *traccia* di una matrice quadrata è *la somma* degli elementi sulla diagonale.

[44] Vedi ad es. `https://it.wikipedia.org/wiki/Decomposizione_ai_valori_singolari`

matrici *unitarie* di dimensione $n_R \times n_R$ e $n_T \times n_T$ rispettivamente, le cui colonne sono (per $\boldsymbol{U}$) gli autovettori di $\boldsymbol{H}\boldsymbol{H}^\dagger$, e per $\boldsymbol{V}$ gli autovettori di $\boldsymbol{H}^\dagger\boldsymbol{H}$.

Con tali posizioni per l'argomento di (21.41) otteniamo

$$\begin{aligned}&\det\left[\boldsymbol{I}_{n_R} + \tfrac{1}{\sigma_n^2}\boldsymbol{H}\boldsymbol{\Sigma}_s\boldsymbol{H}^\dagger\right] = \\ &= \det\left[\boldsymbol{I}_{n_R} + \tfrac{1}{\sigma_n^2}\boldsymbol{U}\boldsymbol{D}\,\boldsymbol{V}^\dagger\boldsymbol{\Sigma}_s\boldsymbol{V}\boldsymbol{D}^\dagger\,\boldsymbol{U}^\dagger\right] = \\ &= \det\left[\boldsymbol{I}_{n_R} + \tfrac{1}{\sigma_n^2}\boldsymbol{D}\,\boldsymbol{V}^\dagger\boldsymbol{\Sigma}_s\boldsymbol{V}\boldsymbol{D}^\dagger\right]\end{aligned} \tag{21.43}$$

dove all'ultima eguaglianza si è invocata l'identità di cui alla nota 39 e ricordato che $\boldsymbol{U}^\dagger\boldsymbol{U} = \boldsymbol{I}_{n_R}$. Evocando ora la *disuguaglianza di Hadamard*[45], che recita

> *per una matrice $\boldsymbol{A}$ semidefinita positiva risulta $\det(\boldsymbol{A}) \leq \prod_i a_{ii}$, con il segno di uguale solo se $\boldsymbol{A}$ è diagonale*

osserviamo come per massimizzare la (21.43) occorra che $\boldsymbol{I}_{n_R} + {}^1\!/_{\sigma_n^2}\boldsymbol{D}\,\boldsymbol{V}^\dagger\boldsymbol{\Sigma}_s\boldsymbol{V}\boldsymbol{D}^\dagger$ sia diagonale. A tale scopo eseguiamo il cambio di variabile[46] $\tilde{\boldsymbol{s}} = \boldsymbol{V}^\dagger\boldsymbol{s}$ in modo da poter scrivere

$$\boldsymbol{\Sigma}_{\tilde{s}} = E\left\{\tilde{\boldsymbol{s}}\tilde{\boldsymbol{s}}^\dagger\right\} = E\left\{\boldsymbol{V}^\dagger\boldsymbol{s}\boldsymbol{s}^\dagger\boldsymbol{V}\right\} = \boldsymbol{V}^\dagger\boldsymbol{\Sigma}_s\boldsymbol{V}$$

ovvero $\boldsymbol{\Sigma}_s = \boldsymbol{V}\boldsymbol{\Sigma}_{\tilde{s}}\boldsymbol{V}^\dagger$, e quindi riformulare la (21.43) come

$$\begin{aligned}\det\left[\boldsymbol{I}_{n_R} + \tfrac{1}{\sigma_n^2}\boldsymbol{D}\,\boldsymbol{V}^\dagger\boldsymbol{\Sigma}_s\boldsymbol{V}\boldsymbol{D}^\dagger\right] &= \det\left[\boldsymbol{I}_{n_R} + \tfrac{1}{\sigma_n^2}\boldsymbol{D}\,\boldsymbol{V}^\dagger\boldsymbol{V}\boldsymbol{\Sigma}_{\tilde{s}}\boldsymbol{V}^\dagger\boldsymbol{V}\boldsymbol{D}^\dagger\right] = \\ &= \det\left[\boldsymbol{I}_{n_R} + \tfrac{1}{\sigma_n^2}\boldsymbol{D}\,\boldsymbol{\Sigma}_{\tilde{s}}\boldsymbol{D}^\dagger\right]\end{aligned} \tag{21.44}$$

che è massimo quando $\boldsymbol{I}_{n_R} + \frac{1}{\sigma_n^2}\boldsymbol{D}\,\boldsymbol{\Sigma}_{\tilde{s}}\boldsymbol{D}^\dagger$ è diagonale, ovvero quando $\boldsymbol{\Sigma}_{\tilde{s}}$ lo è. Ma l'elemento $i-$esimo sulla diagonale di $\boldsymbol{\Sigma}_{\tilde{s}}$ è pari all'energia per simbolo $\mathcal{E}_{s_i}$ trasmesso dall'$i-$esima antenna[47]: scriviamo dunque $\boldsymbol{\Sigma}_{\tilde{s}} = \mathrm{diag}\left(\mathcal{E}_{s_1}, \mathcal{E}_{s_2}, \cdots, \mathcal{E}_{s_{n_T}}\right)$ in modo da ottenere dalle (21.40), (21.43) e (21.44)

$$\begin{aligned}C_{MIMO-NEP} &= \log_2\left(\det\left[\boldsymbol{I}_{n_R} + \tfrac{1}{\sigma_n^2}\boldsymbol{D}\,\boldsymbol{\Sigma}_{\tilde{s}}\boldsymbol{D}^\dagger\right]\right) = \\ &= \log_2\left(\textstyle\prod_{i=1}^{n_T}\left(1 + \frac{1}{\sigma_n^2}\lambda_i\mathcal{E}_{s_i}\right)\right) = \sum_{i=1}^{n_T}\log_2\left(1 + \frac{\mathcal{E}_{s_i}}{\sigma_n^2}\lambda_i\right) = \\ &= \textstyle\sum_{i=1}^{m}\log_2\left(1 + \rho_i\lambda_i\right) \quad \text{bit/sec/Hz}\end{aligned} \tag{21.45}$$

del tutto simile alla (21.42), tranne che ora ogni antenna trasmittente può avere un diverso SNR per simbolo, pari a $\rho_i = \frac{\mathcal{E}_{s_i}}{\sigma_n^2}$. Notiamo che all'ultimo passaggio la sommatoria arriva fino al numero m di autovalori non nulli, in quanto per quelli nulli il contributo sarebbe stato pari a $\log_2(1) = 0$.

---

[45] Vedi ad es. `https://en.wikipedia.org/wiki/Hadamard's_inequality`

[46] Che, essendo $\boldsymbol{V}$ unitaria ossia una rotazione, non modifica le caratteristiche informative ed energetiche delle quantità in gioco.

[47] In quanto

$$\left(\boldsymbol{\Sigma}_{\tilde{s}}\right)_{i,i} = E\left\{\tilde{s}_i\tilde{s}_i^*\right\} = E\left\{\left(\boldsymbol{V}^\dagger\right)_i s_i s_i^* \left(\boldsymbol{V}\right)_i\right\} = E\left\{s_i s_i^*\right\}\left(\boldsymbol{V}^\dagger\right)_i\left(\boldsymbol{V}\right)_i = \mathcal{E}_{s_i}$$

ovvero l'energia per il simbolo trasmesso dall'antenna i. Con $\left(\boldsymbol{V}^\dagger\right)_i$ si è indicata l'$i-$esima riga di $\boldsymbol{V}^\dagger$ e con $\left(\boldsymbol{V}\right)_i$ l'$i-$esima colonna di $\boldsymbol{V}$, il cui prodotto scalare è uno, essendo la matrice unitaria. La trasformazione $\tilde{\boldsymbol{s}} = \boldsymbol{V}^\dagger\boldsymbol{s}$ è infatti una *rotazione* che non altera la norma di $\boldsymbol{s}$, e dunque $\boldsymbol{\Sigma}_s$ e $\boldsymbol{\Sigma}_{\tilde{s}}$ hanno la stessa diagonale.

21.4.1.2 Codifica a riempimento d'acqua

A questo punto non rimane che individuare la distribuzione delle potenze di trasmissione $\mathcal{P}_i = \mathcal{E}_{s_i} f_s$ (e dunque dei corrispondenti SNR per simbolo $\rho_i = \mathcal{E}_{s_i}/\sigma_n^2$) tale da rendere massima (21.45), nel rispetto del vincolo che $\sum_{i=1}^{n_T} \rho_i - \frac{\mathcal{E}_s}{\sigma_n^2} = 0$ con $\mathcal{E}_s = E\left\{\boldsymbol{s}^\dagger \boldsymbol{s}\right\}$ e $\mathcal{P}_T = \mathcal{E}_s f_s$. Si tratta quindi di un *classico* problema di *ottimizzazione vincolata*, da affrontare con il metodo dei moltiplicatori di Lagrange (§ 9.7.1): a questo scopo definiamo la funzione *Lagrangiana* $\mathcal{L}$ con moltiplicatore μ come

$$\mathcal{L} = \sum_{i=1}^{m} \log_2 (1 + \rho_i \lambda_i) + \mu\left(\sum_{i=1}^{m} \rho_i - \rho_T\right)$$

avendo posto $\rho_T = \frac{\mathcal{E}_s}{\sigma_n^2}$, ed eguagliamone a zero le derivate rispetto ai ρ_i e rispetto a μ:

$$\begin{aligned} \frac{\partial}{\partial \rho_i}\mathcal{L} &= \frac{\lambda_i}{1 + \rho_i \lambda_i} + \mu = 0 \quad i = 1, 2, \cdots, m \\ \frac{\partial}{\partial \mu}\mathcal{L} &= \sum_{i=1}^{m} \rho_i - \rho_T = 0 \end{aligned}$$

da cui

$$\begin{aligned} &\frac{1}{\frac{1}{\lambda_i} + \rho_i} = -\mu \quad \text{ossia} \quad \frac{1}{\lambda_i} + \rho_i = -\frac{1}{\mu} \quad i = 1, 2, \cdots, m \\ &\sum_{i=1}^{m} \rho_i = \rho_T \end{aligned} \tag{21.46}$$

dove la seconda riga è ancora il vincolo sulla potenza. Il valore del moltiplicatore μ si ottiene sommando (21.46) su tutti gli i e tenendo conto del vincolo, ovvero

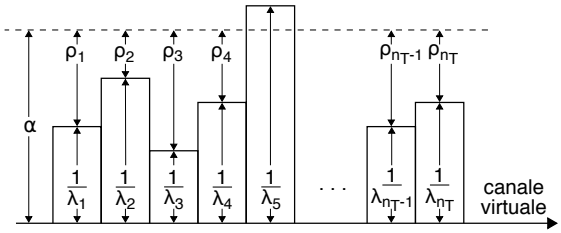
$$\sum_{i=1}^{m}\left(\frac{1}{\lambda_i} + \rho_i\right) = \sum_{i=1}^{m} \frac{1}{\lambda_i} + \rho_T = -\frac{m}{\mu} \quad \rightarrow \quad -\frac{1}{\mu} = \frac{1}{m}\left(\sum_{i=1}^{m} \frac{1}{\lambda_i} + \rho_T\right) = \alpha \tag{21.47}$$

in modo che dalla (21.46) si ottenga

$$\begin{cases} \rho_i = \alpha - \frac{1}{\lambda_i} & \text{se } \rho_i > 0 \\ \rho_i = 0 & \text{altrimenti} \end{cases} \tag{21.48}$$

che come è facile verificare[48] rispetta il vincolo, mentre la limitazione $\rho_i \geq 0$ ha un evidente significato fisico.

Ricordando la discussione svolta a riguardo della (21.42), i termini λ_i rappresentano i guadagni di Rayleigh dei *canali virtuali* di tipo SISO di cui risulta costituito il canale MIMO. La soluzione (21.48) è detta *a riempimento d'acqua* perché, potendo essere scritta anche come $\rho_i + \frac{1}{\lambda_i} = \alpha$, impone che la somma $\rho_i + \frac{1}{\lambda_i}$ sia *la stessa* per ogni canale, aumentando quindi la potenza (ossia l'SNR per simbolo ρ_i) laddove $\frac{1}{\lambda_i}$ è minore, ovvero λ_i è maggiore, ossia quando il canale virtuale i è più *affidabile*. Il senso di tale strategia è quello di affidarsi in misura maggiore ai canali migliori, e non sprecare potenza su quelli peggiori.



[48] Basta calcolare $\sum_{i=1}^{m} \rho_i = \sum_{i=1}^{m}\left(\alpha - \frac{1}{\lambda_i}\right) = \sum_{i=1}^{m} \frac{1}{\lambda_i} + \rho_T - \sum_{i=1}^{m} \frac{1}{\lambda_i} = \rho_T$

Sostituendo i valori forniti dalla (21.48) per ρ_i nella (21.45) otteniamo infine[49]

$$C_{MIMO-WF} = \sum_{\substack{i=1 \\ i \notin \mathcal{B}}}^{m} \log_2(\alpha \lambda_i) \quad \text{bit/sec/Hz} \tag{21.49}$$

dove $\mathcal{B}$ è l'insieme degli indici i per cui $\frac{1}{\lambda_i} \geq \alpha = \frac{1}{m}\left(\sum_{i=1}^{m} \frac{1}{\lambda_i} + \rho_T\right)$.

Architettura di trasmissione Dopo questa analisi così particolare, mostriamo come la stessa SVD della matrice $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\dagger$ possa essere *di guida* per la realizzazione di un sistema di trasmissione MIMO in grado di *approcciare*[50] la velocità per simbolo espressa dalla (21.49). Per riuscirvi è necessario che il lato trasmittente conosca gli autovalori non nulli λ_i di $\boldsymbol{H}\boldsymbol{H}^\dagger$, la matrice $\boldsymbol{V}$, e la potenza di rumore σ_n^2 in ricezione.

A partire dalla potenza $\mathcal{P}_T$ a disposizione, il trasmittente valuta l'SNR per simbolo

$$\rho_T = \frac{1}{\sigma_n^2}\frac{\mathcal{P}_T}{f_s} = \frac{\mathcal{E}_{\boldsymbol{s}}}{\sigma_n^2} = \frac{E\{\boldsymbol{s}^\dagger \boldsymbol{s}\}}{\sigma_n^2}$$

da utilizzare nella (21.47) per calcolare il valore α, e da questo ottiene i valori ρ_i come prescritto dalla (21.48). Le ampiezze dei simboli s_i da inviare alle antenne *non spente* sono quindi (più o meno) amplificati in modo da ottenere $\frac{E\{s_i s_i^*\}}{\sigma_n^2} = \rho_i$ come calcolato, ed il vettore di simboli $\boldsymbol{s}^{WF}$ così ottenuto viene moltiplicato per la matrice $\boldsymbol{V}$ fornendo il nuovo vettore $\tilde{\boldsymbol{s}} = \boldsymbol{V}\boldsymbol{s}^{WF}$, i cui elementi $\tilde{s}_i$ sono usati per generare il segnale dati che modula la portante del segnale trasmesso *realmente* dalle antenne.

Dall'altro lato del collegamento (vedi fig. 21.3-a)) viene ora ricevuto il vettore $\tilde{\boldsymbol{r}} = \boldsymbol{H}\tilde{\boldsymbol{s}} + \boldsymbol{n}$, ma al suo posto il ricevitore prende in considerazione il prodotto di $\tilde{\boldsymbol{r}}$ per la matrice $\boldsymbol{U}^\dagger$, ottenendo così un nuovo vettore di simboli

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{U}^\dagger \tilde{\boldsymbol{r}} = \boldsymbol{U}^\dagger \boldsymbol{H}\tilde{\boldsymbol{s}} + \boldsymbol{U}^\dagger \boldsymbol{n} = \boldsymbol{U}^\dagger \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\dagger \boldsymbol{V}\boldsymbol{s}^{WF} + \boldsymbol{U}^\dagger \boldsymbol{n} = \\ &= \boldsymbol{D}\,\boldsymbol{s}^{WF} + \tilde{\boldsymbol{n}} \end{aligned}$$

e dato che $\boldsymbol{D} = \text{diag}\left(\lambda_1^{1/2}, \lambda_2^{1/2}, \cdots, \lambda_m^{1/2}\right)$ ciò equivale a ricevere m segnali *indipendenti*

$$r_i = \lambda_i^{1/2} s_i^{WF} + \tilde{n}_i \qquad i = 1, 2, \cdots, m$$

in cui $\tilde{n}_i$ ha esattamente le stesse caratteristiche statistiche (media nulla e varianza σ_n^2) del valore n_i, in virtù dell'essere $\boldsymbol{U}^\dagger$ una matrice unitaria[51]. Pertanto in questo modo il ricevitore *vede* direttamente gli m canali virtuali, tutti di tipo SISO.

---

[49] Infatti

$$\sum_{i=1}^{m} \log_2(1 + \rho_i \lambda_i) = \sum_{\substack{i=1 \\ i \notin \mathcal{B}}}^{m} \log_2(1 + (\alpha - \tfrac{1}{\lambda_i})\lambda_i) = \sum_{\substack{i=1 \\ i \notin \mathcal{B}}}^{m} \log_2(1 + \alpha\lambda_i - 1)$$

dove i canali per i quali in base alla (21.48) si ottiene $\rho_i = 0$ (ovvero elementi di $\mathcal{B}$) avrebbero dato un contributo $\log_2(1 + \rho_i \lambda_i) = \log_2(1) = 0$.

[50] Ricordiamo che il valore di capacità è un *limite massimo* di velocità per la trasmissione senza errori, rispetto al quale confrontare le prestazioni della codifica di canale in uso.

[51] Considerando l'intero vettore $\tilde{\boldsymbol{n}} = \boldsymbol{U}^\dagger \boldsymbol{n}$, le sue componenti $\tilde{n}_i$ sono v.a. gaussiane complesse a media nulla in quanto combinazioni lineari di v.a. della stessa natura. Per la covarianza si ottiene

$$\Sigma_{\tilde{\boldsymbol{n}}} = E\{\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}^\dagger\} = E\{\boldsymbol{U}^\dagger \boldsymbol{n}(\boldsymbol{U}^\dagger \boldsymbol{n})^\dagger\} = E\{\boldsymbol{U}^\dagger \boldsymbol{n}\boldsymbol{n}^\dagger \boldsymbol{U}\} = \boldsymbol{U}^\dagger \Sigma_{\boldsymbol{n}} \boldsymbol{U} = \sigma_n^2 \boldsymbol{U}^\dagger \boldsymbol{I}\boldsymbol{U} = \sigma_n^2 \boldsymbol{I}$$

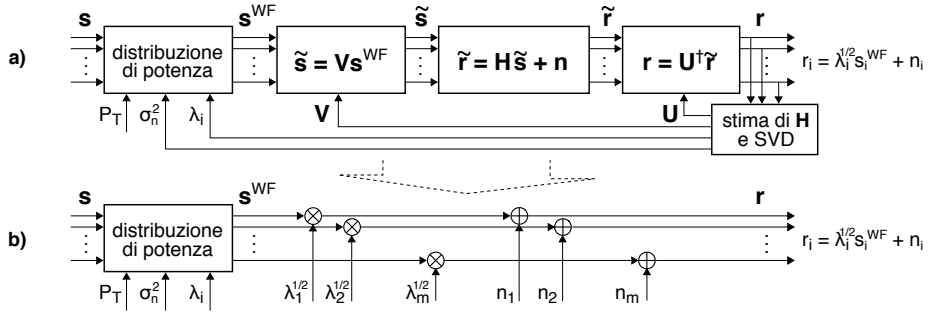


Figura 21.3: Trasmissione MIMO con CSI in trasmissione e *water-filling*: a) - schema di elaborazione, b) - canali virtuali equivalenti

Esempio Per un canale MIMO con $\boldsymbol{H} = \begin{bmatrix} .1 & .3 & .7 \\ .5 & .4 & .1 \\ .2 & .6 & .8 \end{bmatrix}$ si ottiene una SVD $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\dagger}$ pari a

$$\boldsymbol{H} = \begin{bmatrix} -.555 & .3764 & -.7418 \\ -.3338 & -.9176 & -.2158 \\ -.7619 & .1278 & .6349 \end{bmatrix} \begin{bmatrix} 1.3333 & 0 & 0 \\ 0 & 0.5129 & \\ 0 & 0 & 0.0965 \end{bmatrix} \begin{bmatrix} -.2811 & -.7713 & -.5710 \\ -.5679 & -.3459 & .7469 \\ -.7736 & .5342 & -.3408 \end{bmatrix}.$$

Notiamo che i canali virtuali presentano guadagni molto differenti, ed in particolare il terzo ($\lambda_3^{1/2} = 0.0965$, dunque $\lambda_3 \simeq 0.009$) produrrà un contributo alla capacità (21.49) trascurabile, ed in base alla (21.48) si vedrà allocata una potenza di molto ridotta.

Considerazioni applicative La diseguale distribuzione della potenza disponibile tra le antenne di trasmissione permette di raggiungere velocità più elevate di quanto possibile per una distribuzione uniforme, ma richiede la conoscenza da parte del trasmettitore di informazioni relative al canale MIMO, dette *channel status information* (CSI), che devono essere *stimate* dal lato ricevente[52]. Nel caso di una trasmissione *half-duplex* in cui le parti si scambiano alternativamente di ruolo utilizzando la medesima frequenza portante può essere sfruttata la reciprocità[53] della matrice $\boldsymbol{H}$, e dunque ognuna delle due parti effettua la stima per proprio conto. Ciò non è possibile qualora la trasmissione sia di tipo *full-duplex*[54], rendendo necessario comunicare la CSI da ricevente a trasmittente, con due inconvenienti: il primo è che per evitare l'impegno di una eccessiva banda a ritroso tale informazione deve necessariamente essere quantizzata, introducendo errori che possono inficiare i benefici legati alla conoscenza della CSI da parte del trasmittente; il secondo è che in caso di mobilità la velocità di variazione del canale MIMO può essere troppo elevata rispetto ai tempi necessari alla stima della CSI

[52] Le tecniche usate a questo scopo sono generalmente basate sulla trasmissione preventiva di *sequenze di training* o di *toni pilota* concordati, in base alla ricezione (alterata) dei quali il ricevitore è in grado di ricostruire le alterazioni subite dal segnale trasmesso. In linea generale questo processo richiede un tempo proporzionale ad n_T, vedi ad es. `https://en.wikipedia.org/wiki/Channel_state_information`

[53] Nel senso che la risposta in frequenza h_{ij} del canale radio tra l'antenna j di trasmissione e quella i di ricezione è la stessa di quella in senso inverso. Pertanto se un ricevitore con n_R antenne stima una $\boldsymbol{H}_{avanti}$ relativa ai segnali ricevuti, quando lo stesso si comporta da trasmettitore (con uguale numero di antenne) *traspone* la matrice stimata e la usa come CSI per il canale MIMO in direzione opposta, ovvero $\boldsymbol{H}_{indietro} = \boldsymbol{H}_{avanti}^{T}$.

[54] Infatti in tal caso occorre adottare due portanti differenti nelle due direzioni, per non incorrere in una forte auto-interferenza in ricezione.

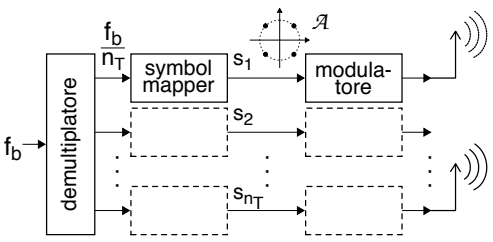
ed alla sua trasmissione, con il rischio di introdurre errori anche per questo motivo.

Rispetto al miglioramento di prestazioni, come già osservato esso dipende dalla particolare realizzazione della matrice $\boldsymbol{H}$, dal suo rango, e dalla distribuzione degli autovalori. Il miglioramento inoltre diminuisce all'aumentare dell'SNR di ricezione, sino a (di fatto) annullarsi per $SNR > 30$ dB[55]. Per bassi valori di SNR il miglioramento è invece sensibile, fino a più che raddoppiare il valore di capacità per $SNR < 0$ dB.

21.5 Multiplazione spaziale

Mentre nei collegamenti con un basso SNR il principale beneficio del disporre di più antenne è quello di poter sfruttare la diversità spaziale (§ 21.3) per conseguire una P_e altrimenti insufficiente, se al contrario l'SNR è abbastanza elevato da garantire una comunicazione affidabile si possono sfruttare le antenne per trasmettere con ognuna di esse *un diverso* segnale dati, moltiplicando la velocità di trasmissione di un fattore pari a $g_M = \min(n_T, n_R)$, detto *guadagno di multiplazione*.

Dato che come osservato nel caso di un buon SNR il vantaggio derivante dalla scomposizione del collegamento in canali virtuali associati alla SVD di $\boldsymbol{H}$ è ridotto, si preferisce che il flusso di simboli s_i convogliato dall'antenna $i-$esima sia frutto della suddivisione del flusso originario a velocità f_b operata *a monte* della trasmissione, da parte di un *demultiplatore* che assegna ad ogni antenna un flusso di f_b/n_T bit/sec. Ogni gruppo di M bit di ciascuno degli n_T flussi viene quindi *mappato* su di un simbolo a valori complessi $s_i,\ i = 1, 2, \cdots, n_T$ appartenente ad un alfabeto $\mathcal{A}$ definito dagli $L = 2^M$ punti di costellazione della modulazione adottata; l'insieme dei simboli trasmessi *contemporaneamente e sulla stessa portante* viene quindi descritto nei termini di un vettore $\boldsymbol{s} = \left(s_1, s_2, \cdots s_{n_T}\right)^T$ a valori complessi.



Osserviamo ora che per ricevere correttamente tutti gli n_T diversi flussi dati occorre disporre di un numero n_R di antenne in ricezione almeno pari ad n_T ovvero deve essere $n_R \geq n_T$; le antenne *in più* in ricezione possono essere usate per conseguire anche un guadagno *di diversità*. In questo contesto l'equazione (21.2) del canale MIMO

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n} \tag{21.50}$$

con $\boldsymbol{r} \in \mathbb{C}^{n_R}$, $\boldsymbol{s} \in \mathcal{A}^{n_T}$, $E\left\{\boldsymbol{s}\boldsymbol{s}^\dagger\right\} = \frac{\mathcal{E}_s}{n_T}\boldsymbol{I}_{n_T}$, $\boldsymbol{H} \in \mathbb{C}^{n_R \times n_T}$ e $\boldsymbol{n} \in CN\left(0, \sigma_n^2\right)$, individua una *trasformazione lineare* affetta da rumore, formalmente assimilabile alla distorsione lineare subita da un segnale in transito su di un canale selettivo in frequenza e rumoroso, con la differenza che l'alterazione agisce ora su di un dominio spaziale anziché frequenziale. In tal senso il tentativo di risalire al vettore $\boldsymbol{s}$ a partire da quello $\boldsymbol{r}$ noto al ricevitore è simile ad una operazione di *equalizzazione* (§ 18.4) al punto che, con un abuso di terminologia, è lecito riferirsi al primo come ad una *equalizzazione spaziale*.

[55] E' lecito pensare come per SNR elevati *l'acqua sia profonda*, e dunque l'allocazione di potenza sia circa la stessa per tutti i canali virtuali.

21.5.1 Ricevitore a massima verosimiglianza (ML)

Effettua la stima del vettore trasmesso $\boldsymbol{s}$ come

$$\hat{\boldsymbol{s}} = \arg \max_{\boldsymbol{s} \in \mathcal{A}^{n_T}} Pr\{\boldsymbol{r}/\boldsymbol{s}\}$$

ed essendo il vettore $\boldsymbol{n}$ nella (21.50) costituito da v.a. gaussiane complesse *circolari* ed incorrelate, equivale ad un criterio di *minima distanza*, ovvero

$$\hat{\boldsymbol{s}} = \arg \min_{\boldsymbol{s} \in \mathcal{A}^{n_T}} \left|\boldsymbol{r} - \boldsymbol{H}\boldsymbol{s}\right|^2 \tag{21.51}$$

dove l'argomento minimizzato dalla (21.51) è la norma quadratica del vettore $\boldsymbol{n}$. La ricerca del minimo si estende a tutte le *disposizioni con ripetizione* sulle n_T antenne degli L possibili valori per il generico simbolo s_i, ovvero L^{n_T} diverse disposizioni, causando una complessità *esponenziale* nel numero di antenne: ad es. adottando una semplice modulazione 16-QAM su $n_T = 4$ antenne (corrispondenti a 16 bit/simbolo $\boldsymbol{s}$), si ottengono... $16^4 = 65536$ combinazioni, decisamente una quantità *spropositata*.

Soluzioni sub-ottime Sebbene la decodifica ML sia quella ottimale, la sua complessità ha spinto la ricerca di soluzioni sub-ottime ma approcciabili. Una soluzione *algoritmica* è quella nota come *sphere decoding*[56], che non approfondiamo. Una *famiglia* di soluzioni alternative sono quelle *lineari*, che (in analogia al caso dell'equalizzazione) pervengono ad una matrice $\boldsymbol{G}$ di dimensioni $n_T \times n_R$ tale da poter scrivere

$$\tilde{\boldsymbol{s}} = \boldsymbol{G}\boldsymbol{r}$$

in cui si è rimosso il vincolo di voler direttamente ottenere un vettore appartenente ad $\mathcal{A}^{n_T}$. Il valore $\tilde{\boldsymbol{s}}$ viene quindi indicato come vettore *soffice*, il cui corrispondente vettore *hard* $\hat{\boldsymbol{s}} \in \mathcal{A}^{n_T}$ si ottiene decidendo per ogni elemento s_i il valore *più vicino* ad uno di quelli ammessi, ossia

$$\hat{s}_i = \arg \min_{s \in \mathcal{A}} (\tilde{s}_i - s)^2 \tag{21.52}$$

Nel seguito descriviamo due criteri per scegliere $\boldsymbol{G}$, oltre ad un ulteriore metodo di soluzione che adotta una proiezione lineare nel contesto di una procedura di *cancellazione ordinata,* coniugando efficienza con precisione, la cui implementazione più diffusa è nota come V-BLAST.

21.5.2 Ricevitore zero-forcing

Individua il vettore $\tilde{\boldsymbol{s}}$ che (senza rispettare il vincolo $\tilde{\boldsymbol{s}} \in \mathcal{A}^{n_T}$ della (21.51)) risolve il problema

$$\tilde{\boldsymbol{s}} = \arg \min_{\boldsymbol{s}} \left|\boldsymbol{r} - \boldsymbol{H}\boldsymbol{s}\right|^2 \tag{21.53}$$

calcolando una matrice $\boldsymbol{G}_{n_T \times n_R}$ tale che $\tilde{\boldsymbol{s}} = \boldsymbol{G}\boldsymbol{r}$.

[56] Vedi ad es. *B. Hassibi, H. Vikalo*, `On the Sphere-Decoding Algorithm, Expected Complexity`, ma anche *Mathworks*, `comm.SphereDecoder`. In estrema sintesi, questa tecnica riduce la complessità della ricerca del ricevitore ML limitandola ai vettori $\boldsymbol{s}$ che ricadono (in qualche modo) all'interno di una sfera di raggio fisso e centrata sul vettore ricevuto $\boldsymbol{r}$. All'aumentare del raggio aumenta la complessità ed al limite si ottiene la soluzione ML, ma anche per complessità ridotte la soluzione trovata non si discosta molto da quella ottima.

Nel caso generale in cui $n_R \geq n_T$ si ottiene

$$\boldsymbol{G}_{ZF} = \left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger = \boldsymbol{H}_{PI} \tag{21.54}$$

nota anche come *pseudo-inversa*[57] $\boldsymbol{H}_{PI}$ di $\boldsymbol{H}$, che qualora $\boldsymbol{H}$ sia quadrata ($n_R = n_T$) ed invertibile si riduce a $\boldsymbol{G} = \boldsymbol{H}^{-1}$. In entrambi i casi la (21.53) fornisce

$$\tilde{\boldsymbol{s}} = \boldsymbol{G}_{ZF}\boldsymbol{r} = \left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger \left(\boldsymbol{H}\boldsymbol{s} + \boldsymbol{n}\right) = \boldsymbol{s} + \left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger \boldsymbol{n} \tag{21.55}$$

ottenendo l'effetto di invertire $\boldsymbol{H}$, come dire *annullare*, da cui il nome *forzante a zero*. Il costo computazionale associato è circa cubico in n_T per matrici quadrate, ma ottenere $\hat{\boldsymbol{s}} \in \mathcal{A}^{n_T}$ da $\tilde{\boldsymbol{s}}$ applicando la (21.52) è lineare in n_T.

Notiamo ora che, in perfetta analogia con quanto osservato a pag. 625 per l'omonimo equalizzatore, il valore $\tilde{\boldsymbol{s}}$ ottenuto mediante la (21.55) è gravato da un termine di rumore $\tilde{\boldsymbol{n}} = \left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger \boldsymbol{n}$ le cui componenti sono *correlate* tra antenne riceventi, e la cui entità può risultare assai rilevante qualora $\boldsymbol{H}$ sia *mal condizionata*[58] ovvero $\boldsymbol{H}^\dagger \boldsymbol{H}$ presenti autovalori per cui $\frac{|\lambda_{max}|}{|\lambda_{min}|} \gg 1$, in quanto[59]

$$\begin{aligned}\boldsymbol{\Sigma}_{\tilde{\boldsymbol{n}}} &= E\left\{\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}^\dagger\right\} = E\left\{\left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger \boldsymbol{n}\boldsymbol{n}^\dagger \boldsymbol{H}\left(\boldsymbol{H}\boldsymbol{H}^\dagger\right)^{-1}\right\} = \\ &= \sigma_n^2 \left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger \boldsymbol{H}\left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} = \sigma_n^2 \left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1}\end{aligned}$$

e l'inversione di $\boldsymbol{H}\boldsymbol{H}^\dagger$ comporta la divisione per il suo determinante[60], che essendo pari al prodotto degli autovalori può nelle condizioni indicate risultare molto piccolo; in particolare, il rumore è particolarmente amplificato per le antenne trasmittenti che corrispondono alle colonne di $\boldsymbol{H}$ associate ai valori singolari $\lambda_i^{1/2}$ più piccoli.

21.5.3 Ricevitore lineare a minimo errore medio quadratico L-MMSE

Per ovviare al problema evidenziato si può adottare il diverso criterio di trovare la matrice $\boldsymbol{G}$ (con cui calcolare $\tilde{\boldsymbol{s}} = \boldsymbol{G}\boldsymbol{r}$) che rende minimo l'*errore quadratico medio* (MSE), ovvero

$$\boldsymbol{G}^\diamond = \arg\min_{\boldsymbol{G}} E\left\{|\boldsymbol{s} - \boldsymbol{G}\boldsymbol{r}|^2\right\} \tag{21.56}$$

Si dimostra[61] che la matrice $\boldsymbol{G}$ che verifica la (21.56) deve necessariamente soddisfare

[57] In questo testo la definizione di pseudo inversa è scaturita all'eq. (7.39) nel contesto della *regressione lineare multipla* (§ 7.7.1), con l'analogia che diviene evidente qualora si confronti la (21.53) con la (7.38) e si sostituisca $\boldsymbol{r}, \boldsymbol{H}, \boldsymbol{s}$ con $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}$. La discussione ivi svolta nel nostro caso significa che se fosse stato trasmesso $\tilde{\boldsymbol{s}}$ che soddisfa (21.53) avremmo ricevuto $\tilde{\boldsymbol{r}} = \boldsymbol{H}\tilde{\boldsymbol{s}} + \boldsymbol{n}$ vincolato (a parte per il rumore) a giacere nello spazio esplorato dalle colonne di $\boldsymbol{H}$, e dunque la differenza $\boldsymbol{\epsilon} = \tilde{\boldsymbol{r}} - \boldsymbol{r} = \boldsymbol{H}(\tilde{\boldsymbol{s}} - \boldsymbol{s})$ deve essere *ortogonale* alle stesse colonne, e quindi $\boldsymbol{H}^\dagger \cdot \boldsymbol{\epsilon} = \boldsymbol{0}$, dopodiché i passaggi sono gli stessi utilizzando l'hermitiano † anziché il trasposto $^\top$.

[58] Vedi ad es. `https://en.wikipedia.org/wiki/Condition_number`

[59] Alla seconda eguaglianza manca il passaggio

$$\tilde{\boldsymbol{n}}^\dagger = \left(\left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger \boldsymbol{n}\right)^\dagger = \boldsymbol{n}^\dagger \left(\left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1} \boldsymbol{H}^\dagger\right)^\dagger = \boldsymbol{n}^\dagger \boldsymbol{H}\left(\left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1}\right)^\dagger = \boldsymbol{n}^\dagger \boldsymbol{H}\left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1}$$

dato che $\left(\boldsymbol{A}^{-1}\right)^\dagger = \left(\boldsymbol{A}^\dagger\right)^{-1}$

[60] Vedi ad es. `https://it.wikipedia.org/wiki/Matrice_invertibile`

[61] Indicando l'argomento interno di (21.56) come l'errore $\boldsymbol{\epsilon}(\boldsymbol{G}) = \boldsymbol{s} - \boldsymbol{G}\boldsymbol{r}$ ed il suo valore atteso quadratico come $J(\boldsymbol{G}) = E\left\{\boldsymbol{\epsilon}^2\right\} = E\left\{\boldsymbol{\epsilon}^\dagger \boldsymbol{\epsilon}\right\}$, quando J è minimo tutte le componenti del suo gradiente devono annullarsi, ovvero $\nabla_{\boldsymbol{G}} J(\boldsymbol{G})\big|_{\boldsymbol{G}=\boldsymbol{G}^\diamond} = \boldsymbol{0}_{n_T \times n_R}$ (cioè si annullano le derivate di J rispetto a tutti gli

anche la relazione

$$E\left\{\left(\boldsymbol{s}-\boldsymbol{G}^{\diamond}\boldsymbol{r}\right)\boldsymbol{r}^{\dagger}\right\}=\boldsymbol{0}_{n_T\times n_R} \tag{21.57}$$

nota come principio *di ortogonalità*[62], sviluppando la quale (dopo un po' di conti) si perviene al risultato[63]

$$\boldsymbol{G}_{MMSE}=\boldsymbol{G}^{\diamond}=\left(\boldsymbol{H}\boldsymbol{H}^{\dagger}+\frac{1}{\rho}\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{\dagger} \tag{21.58}$$

in cui

$$\rho=\frac{\mathcal{E}_s}{\sigma_n^2}=\frac{\mathcal{P}_T}{f_s n_T}\frac{n_R}{E\{\boldsymbol{n}^{\dagger}\boldsymbol{n}\}}$$

è l'SNR per simbolo e per antenna ricevente, comprensivo dell'eventuale guadagno di diversità n_R/n_T. Notiamo quindi che per ρ elevato la (21.58) equivale alla (21.54), cancellando gli interferenti ma senza considerare il rumore; qualora invece ρ tenda a zero la (21.58) tende ad $\boldsymbol{H}^{\dagger}$, un risultato simile al MRC[64] (§ 21.3.1.2), e che non considera gli interferenti. Dopodiché il valore di $\hat{\boldsymbol{s}}\in\mathcal{A}^{n_T}$ si ottiene applicando anche per questo caso la (21.52).

21.5.4 Ricevitore a cancellazioni successive - VBLAST

Nei casi precedenti la decisione da parte del ricevitore in merito al simbolo s_i trasmesso dall'i–esima antenna viene presa *congiuntamente* a tutti gli altri, calcolando il vettore

---

elementi di $\boldsymbol{G}$). Possiamo quindi scrivere

$$\frac{\partial J}{\partial \boldsymbol{G}}=\frac{\partial}{\partial \boldsymbol{G}}E\{\boldsymbol{\epsilon}^2\}=2E\left\{\boldsymbol{\epsilon}\cdot\frac{\partial \boldsymbol{\epsilon}^{\dagger}}{\partial \boldsymbol{G}}\right\}=\boldsymbol{0}$$

e dato che $\frac{\partial \boldsymbol{\epsilon}^{\dagger}}{\partial \boldsymbol{G}}=\frac{\partial}{\partial \boldsymbol{G}}\left(\boldsymbol{s}-\boldsymbol{G}\boldsymbol{r}\right)^{\dagger}=-\boldsymbol{r}^{\dagger}$ si ottiene J minimo quando

$$E\left\{\boldsymbol{\epsilon}_{MMSE}\cdot\frac{\partial \boldsymbol{\epsilon}^{\dagger}}{\partial \boldsymbol{G}}\right\}=E\left\{\left(\boldsymbol{s}-\boldsymbol{G}^{\diamond}\boldsymbol{r}\right)\boldsymbol{r}^{\dagger}\right\}=\boldsymbol{0}_{n_T\times n_R}$$

[62] In effetti ciò che la (21.57) afferma è l'incorrelazione tra ogni componente del vettore di errore minimo rispetto ad ogni componente di quello ricevuto, ma vedi anche `https://en.wikipedia.org/wiki/Orthogonality_principle`.

[63] A partire dalla (21.57) otteniamo $E\left\{\left(\boldsymbol{s}-\boldsymbol{G}^{\diamond}\boldsymbol{r}\right)\boldsymbol{r}^{\dagger}\right\}=E\left\{\boldsymbol{s}\boldsymbol{r}^{\dagger}-\boldsymbol{G}^{\diamond}\boldsymbol{r}\boldsymbol{r}^{\dagger}\right\}=\Sigma_{SR}-\boldsymbol{G}^{\diamond}\Sigma_{RR}=0$ per cui deve risultare $\boldsymbol{G}^{\diamond}=\Sigma_{SR}/\Sigma_{RR}$, dove le matrici di covarianza $\boldsymbol{\Sigma}$ coincidono con quelle di correlazione, dato il valore atteso nullo di $\boldsymbol{s}$ e di $\boldsymbol{r}$. Calcoliamone il valore:

$$\Sigma_{SR}=E\{\boldsymbol{s}\boldsymbol{r}^{\dagger}\}=E\{\boldsymbol{s}\left(\boldsymbol{H}\boldsymbol{s}+\boldsymbol{n}\right)^{\dagger}\}=E\{\boldsymbol{s}\boldsymbol{s}^{\dagger}\boldsymbol{H}^{\dagger}\}+E\{\boldsymbol{s}\boldsymbol{n}^{\dagger}\}=\Sigma_S\boldsymbol{H}^{\dagger}=\mathcal{E}_s\boldsymbol{H}^{\dagger}$$
$$\Sigma_{RR}=E\{\left(\boldsymbol{H}\boldsymbol{s}+\boldsymbol{n}\right)\left(\boldsymbol{H}\boldsymbol{s}+\boldsymbol{n}\right)^{\dagger}\}=E\{\boldsymbol{H}\boldsymbol{s}\boldsymbol{s}^{\dagger}\boldsymbol{H}^{\dagger}+\boldsymbol{H}\boldsymbol{s}\boldsymbol{n}^{\dagger}+\boldsymbol{n}\boldsymbol{s}^{\dagger}\boldsymbol{H}^{\dagger}+\boldsymbol{n}\boldsymbol{n}^{\dagger}\}=\boldsymbol{H}\Sigma_S\boldsymbol{H}^{\dagger}+\Sigma_N=\mathcal{E}_s\boldsymbol{H}\boldsymbol{H}^{\dagger}+\sigma_n^2\boldsymbol{I}$$

dato che $E\{\boldsymbol{s}\boldsymbol{n}^{\dagger}\}$, $E\{\boldsymbol{H}\boldsymbol{s}\boldsymbol{n}^{\dagger}\}$ e $E\{\boldsymbol{n}\boldsymbol{s}^{\dagger}\boldsymbol{H}^{\dagger}\}$ sono nulli in virtù dell'incorrelazione tra $\boldsymbol{s}$ ed $\boldsymbol{n}$, e si è posto $\Sigma_S=\mathcal{E}_s\boldsymbol{I}$ e $\Sigma_N=\sigma_n^2\boldsymbol{I}$ in virtù dell'incorrelazione tra simboli delle diverse antenne, nonché tra campioni di rumore, in cui $\mathcal{E}_s=E\{s_i^*s_i\}=\mathcal{P}_T/f_sn_T$ è pari all'energia per simbolo e per antenna, uguale per tutte, mentre $\sigma_n^2=E\{n_i^*n_i\}=1/n_RE\{\boldsymbol{n}^{\dagger}\boldsymbol{n}\}$ è la potenza del campione di rumore in ingresso all'i–esima antenna ricevente. Sostituendo ora Σ_{SR} e Σ_{RR} nella relazione $\boldsymbol{G}^{\diamond}=\Sigma_{SR}/\Sigma_{RR}$ otteniamo il risultato cercato

$$\boldsymbol{G}^{\diamond}=\frac{\Sigma_{SR}}{\Sigma_{RR}}=\frac{\mathcal{E}_s\boldsymbol{H}^{\dagger}}{\mathcal{E}_s\boldsymbol{H}\boldsymbol{H}^{\dagger}+\sigma_n^2\boldsymbol{I}}=\frac{\boldsymbol{H}^{\dagger}}{\boldsymbol{H}\boldsymbol{H}^{\dagger}+\frac{\sigma_n^2}{\mathcal{E}_s}\boldsymbol{I}}$$

[64] Nel senso che il vettore di pesi $\boldsymbol{w}$ con coefficienti (21.13) è il coniugato dell'unica colonna di $\boldsymbol{H}$ presente nel caso SIMO, come ora la riga i–esima di $\boldsymbol{H}^{\dagger}$ è coniugata dell'i–esima colonna di $\boldsymbol{H}$ - ma vedi anche la discussione seguente.

$\tilde{\boldsymbol{s}} = \boldsymbol{G}\boldsymbol{r}$ mediante una unica operazione. Al contrario, l'approccio che stiamo per illustrare[65] decodifica i simboli s_i antenna (trasmittente) per antenna, *cancellando* ogni volta il contributo ad $\boldsymbol{r}$ dovuto ai simboli già decodificati.

Prima di procedere torniamo all'espressione $\tilde{\boldsymbol{s}} = \boldsymbol{G}\boldsymbol{r}$ in cui $\boldsymbol{G}$ può essere ottenuta con il metodo ZF oppure MMSE, ma specializzando il ragionamento al caso ZF osserviamo che il valore per l'elemento $\tilde{s}_i$ si ottiene da $\boldsymbol{r}$ come

$$\tilde{s}_i = \boldsymbol{w}_i^T \boldsymbol{r} \tag{21.59}$$

in cui $\boldsymbol{w}_i$ è il vettore con elementi pari alla $i-$esima riga di $\boldsymbol{G}$, che indichiamo come $(\boldsymbol{G})_i$, ovvero $\boldsymbol{w}_i = (\boldsymbol{G})_i^T$. Ebbene, accade che $\boldsymbol{w}_i$ è *ortogonale* a tutte le colonne di $\boldsymbol{H}$ tranne l'$i-$esima: ciò è evidente nel caso di matrice $\boldsymbol{H}$ quadrata, per la quale si ottiene $\boldsymbol{G}_{ZF} = \boldsymbol{H}^{-1}$ e dunque ogni sua riga i è ortogonale alle colonne $j \neq i$ di $\boldsymbol{H}$ in quanto $\boldsymbol{H}^{-1}\boldsymbol{H} = \boldsymbol{I}$. Ma anche se $\boldsymbol{H}$ non è quadrata, la sua pseudo-inversa $\boldsymbol{G}_{ZF} = \boldsymbol{H}_{PI}$ data dalla (21.54) ha righe ortogonali alle colonne di $\boldsymbol{H}$ non omologhe, risultando[66] $\boldsymbol{H}_{PI}\boldsymbol{H} = \boldsymbol{I}$. Tale proprietà di ortogonalità è alla base dell'effetto di inversione del sistema ottenuto con la (21.55).

Consideriamo ora che dopo la decodifica del primo simbolo ($i = 1$)

$$\tilde{s}_i = \boldsymbol{w}_i^T \boldsymbol{r}$$

ed alla sua quantizzazione in $\hat{s}_i \in \mathcal{A}$, nel caso in cui l'$SNR$ sia sufficiente ad avere una P_e abbastanza piccola si può procedere alla *cancellazione* dell'effetto di $\hat{s}_i$ su $\boldsymbol{r}$, semplicemente valutando[67] un nuovo vettore

$$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \hat{s}_i \, (\boldsymbol{H})_i$$

in cui $(\boldsymbol{H})_i$ è l'$i-$esima colonna di $\boldsymbol{H}$: per il nuovo $\boldsymbol{r}_{i+1}$ è quindi come se l'antenna $i-$esima fosse *spenta*, ovvero come se la colonna $(\boldsymbol{H})_i$ fosse tutta nulla: indichiamo quindi con $\boldsymbol{H}_0^i$ la matrice $\boldsymbol{H}$ a cui sono state *azzerate* le colonne $j = 1, 2, \cdots, i$. Il vettore $\boldsymbol{w}_{i+1}$ con cui decodificare il successivo simbolo

$$\tilde{s}_{i+1} = \boldsymbol{w}_{i+1}^T \boldsymbol{r}_{i+1}$$

deve quindi essere ortogonale *non a tutte* le colonne di $\boldsymbol{H}$, ma solo a quelle non nulle di $\boldsymbol{H}_0^i$, coerentemente con il ridotto numero di interferenti ancora presenti nel vettore $\boldsymbol{r}_{i+1}$. Il vettore $\boldsymbol{w}_{i+1}^T$ viene quindi posto pari alla $(i+1)$ $-$esima *riga della pseudo-inversa* di $\boldsymbol{H}_i^0$.

Procedendo in questo modo i simboli decodificati per ultimi godono di un ordine di diversità maggiore di quelli decodificati per primi, in quanto il numero di antenne trasmittenti *si riduce* via via. D'altra parte nel caso in cui la P_e non sia trascurabile si

[65] Noto come V-BLAST, vedi P.W. WOLNIANSKY ET AL, *V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel*, 1998 URSI Int. Symposium Conference Proceedings, reperibile presso `https://www.ee.columbia.edu/~jiantan/E6909/wolnianskyandfoschini.pdf`

[66] Vedi ad es. `https://it.wikipedia.org/wiki/Pseudo-inversa`

[67] Infatti (a parte il rumore) è possibile scrivere $\boldsymbol{r} = \sum_{k=1}^{n_T} s_k \, (\boldsymbol{H})_k$, dunque

$$\boldsymbol{r}' = \boldsymbol{r} - \hat{s}_i \, (\boldsymbol{H})_i = \sum_{\substack{k=1 \\ k \neq i}}^{n_T} s_k \, (\boldsymbol{H})_k$$

può verificare un effetto di propagazione degli errori, specialmente se questi avvengono nei simboli decodificati per primi. Occorre quindi stabilire un criterio di ordinamento in grado di garantire le migliori prestazioni: si dimostra che l'ordinamento ottimale si ottiene scegliendo ogni volta l'indice i_M dell'antenna che trasmette il simbolo s_{i_M} per il quale l'SNR dopo decodifica è il più grande, ovvero $i_M = \arg\max_i SNR\,(i)$. Si può altresì dimostrare[68] che

$$SNR\,(i) = \frac{E\left\{|s_i|^2\right\}}{\sigma_n^2 |\boldsymbol{w}_i|^2} \tag{21.60}$$

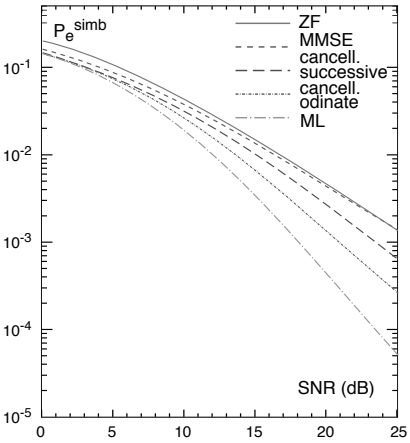


Figura 21.4: Prestazioni della multiplazione spaziale

e quindi se ogni antenna adotta la stessa costellazione $\mathcal{A}$ la massimizzazione di (21.60) si ottiene scegliendo $i_M = \arg\min_i |\boldsymbol{w}_i|^2$; dato che $\boldsymbol{w}_i$ è pari alla $i-$esima riga della pseudo-inversa di $\boldsymbol{H}_i^0$, quest'ultima dopo essere stata ricalcolata ad ogni iterazione, viene esplorata per individuare la riga con norma minore.

Qualora anziché la pseudo inversa dello ZF si adotti la soluzione MMSE data dalla (21.58), il metodo è ancora applicabile, con un lieve vantaggio per SNR basso, come mostrato nella figura a fianco[69], che confronta le prestazioni per diverse opportunità di decodifica di un sistema MIMO 2×2 con modulazione QPSK per una tipica matrice $\boldsymbol{H}$.

21.5.5 Compromesso diversità - multiplazione

Quando in ricezione si dispone di un numero di antenne maggiore del minimo (ossia $n_R > n_T$) le antenne *in più* non contribuiscono al guadagno di multiplazione (al massimo pari a $\min(n_T, n_R)$), ma consentono di conseguire un ordine di diversità *almeno* pari a $n_R - n_T + 1$ (con il metodo di cancellazione ordinata anche di più, dato che in pratica n_T *diminuisce* man mano). Ma anche se $n_R = n_T$ è comunque possibile (e necessario se $n_R < n_T$) sacrificare parte della velocità di trasmissione per migliorare le prestazioni nei confronti del rumore, sfruttando la diversità spaziale frutto del fading indipendente. Si può infatti ad esempio scegliere di dimezzare il numero di antenne usate per la multiplazione, ed impiegare la restante metà per trasmettere la ridondanza associata ad un codice spazio-tempo (§ 21.3.2.2). Ma anziché provare le

[68] Sostituendo in $\tilde{s}_i = \boldsymbol{w}_i^T \boldsymbol{r}$ l'espressione del vettore ricevuto $\boldsymbol{r} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n}$ otteniamo

$$\tilde{s}_i = \boldsymbol{w}_i^T \sum_{k=1}^{n_T} s_k\,(\boldsymbol{H})_k + \boldsymbol{w}_i^T \boldsymbol{n} = s_i \boldsymbol{w}_i^T\,(\boldsymbol{H})_i + \boldsymbol{w}_i^T \boldsymbol{n} = s_i + \boldsymbol{w}_i^T \boldsymbol{n}$$

in cui si tiene conto dell'ortogonalità tra $\boldsymbol{w}_i$ e le colonne $(\boldsymbol{H})_j$ con $j \neq i$, e del fatto che $\boldsymbol{w}_i^T\,(\boldsymbol{H})_i = 1$ essendo $\boldsymbol{w}_i^T$ pari all'$i-$esima riga della pseudoinversa di $\boldsymbol{H}$. Pertanto l'SNR risulta pari a

$$\rho = \frac{E\left\{|s_i|^2\right\}}{E\left\{|\boldsymbol{w}_i^T \boldsymbol{n}|^2\right\}} = \frac{E\left\{|s_i|^2\right\}}{E\left\{\boldsymbol{w}_i^T \boldsymbol{n}\boldsymbol{n}^T \boldsymbol{w}_i\right\}} = \frac{E\left\{|s_i|^2\right\}}{\sigma_n^2\, \boldsymbol{w}_i^T \boldsymbol{w}_i}$$

[69] In realtà la figura è frutto di un *missaggio* di due prelevate da lavori differenti, ed i valori mostrati sono da ritenersi indicativi e non esatti, oltre che frutto della $\boldsymbol{H}$ adottata.

diverse combinazioni possibili, riferiamo di un risultato generale[70]. La relazione tra guadagno di multiplazione spaziale r e di diversità d può essere ottenuta dopo avere definito queste due grandezze nei termini della rispettiva legge di dipendenza *asintotica* da SNR, ovvero

$$r = \lim_{SNR\to\infty} \frac{R\,(SNR)}{\log_2 SNR}; \qquad d = -\lim_{SNR\to\infty} \frac{\log P_e\,(SNR)}{\log SNR} \tag{21.61}$$

rappresentando così il risultato che per SNR elevato la velocità di trasmissione R aumenta come $r \log_2 SNR$([71]), mentre la probabilità di errore decade come $1/SNR^d$([72]).

In questo modo può essere tracciata la curva $d\,(r)$ mostrata in fig. 21.5 e che *raccorda* i valori estremi di massima diversità $d_{\max} = n_T \cdot n_R$ e massima multiplazione $r_{\max} = \min\left(n_T, n_R\right)$ mediante la relazione

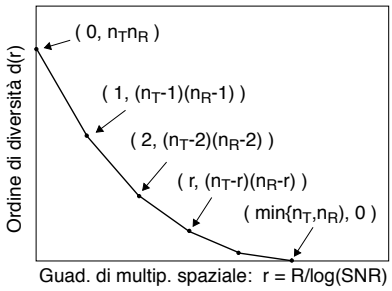


Figura 21.5: Compromesso diversità - multiplazione

$$d_{opt}\,(r) = (n_T - r)\,(n_R - r) \quad \text{con} \quad r = 0, 1, \cdots, \min\left(n_T, n_R\right) \tag{21.62}$$

che individua il *massimo* ordine di diversità conseguibile per un guadagno di multiplazione assegnato.

Esempio Desiderando ottenere una velocità R di 15 bps/Hz da un sistema MIMO con $n_T = n_R = 8$ ed $SNR = 15$ dB, qual'è il massimo guadagno di diversità d_{opt} che il sistema può offrire? **R:** dalla relazione $R = r \log_2 SNR$ con $SNR = 15$ dB otteniamo $15 = r \log_2 10^{1.5}$ e dunque $r = 3.01 \simeq 3$. Pertanto in linea teorica 3 delle 8 antenne sono usate per la multiplazione, e 5 per la diversità, con un guadagno di diversità che *al massimo* può arrivare a $d_{opt}\,(r) = (n_T - r)\,(n_R - r) = (8-3)\,(8-3) = 25$. Nella pratica *naif*, se si adotta un codice di Alamouti che utilizza 6 antenne per trasmettere 3 flussi codificati, ognuno dei quali è ricevuto da due antenne, l'ordine di diversità per ogni flusso risulta pari a $2n_R = 2 \cdot 2 = 4$ (vedi § 21.3.2.4), e dunque complessivamente pari solamente a $3 \cdot 4 = 12$. In compenso, la potenza non usata dalle 2 antenne spente in trasmissione può essere ridistribuita sulle 6 attive, così come le due antenne *in più* di ricezione possono essere impiegate nella ricezione di 2 dei 3 flussi, portando l'ordine di diversità complessivo a $6 + 6 + 4 = 16$. Alla nota[73] alcuni approcci in grado di avvicinarsi di più al risultato (21.62).

[70] Vedi L.Zheng, D.N.C.Tse, *Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels*, IEEE Trans. on Inf. Theory, May 2003, dove sono riportate le considerazioni e le ipotesi che determinano il risultato; trovo una copia *libera* presso l'Univ. di Stanford.

[71] Dato che la capacità per simbolo $C = \log_2(1 + SNR)$ di un canale SISO tende a $\log_2 SNR$ con $SNR \to \infty$, il valore r nella (21.61) esprime il numero di canali SISO equivalenti.

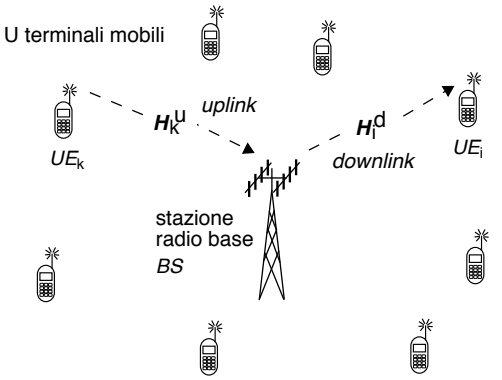
[72] Probabilmente ci vorrebbe qualche passaggio in più, ma la relazione $P_e \propto 1/SNR^d$ deriva dal confronto tra la (21.15) del ricevitore MRC che sfrutta la diversità, e la (20.35) valida per un canale SISO.

[73] S.Sfar, L.Dai, K.B.Letaief, *Optimal Diversity-Multiplexing Tradeoff With Group Detection for MIMO Systems*, IEEE Tr. on Comm. 2005; P. Elia ed al, *Explicit Space-Time Codes Achieving The Diversity-Multiplexing Gain Tradeoff*, IEEE Tr. on Inf. Th. 2006

21.6 Trasmissione multiutente o MU - MIMO

Fino ad ora abbiamo discusso di collegamenti MIMO punto-punto, ma in generale esiste una forte asimmetria nel numero antenne presenti ai due lati di un collegamento *radiomobile*, che sono molte di più presso una *stazione radio base*[74] o BS (il cui numero, decine o anche centinaia, indichiamo ora come n_{BS}) rispetto a quelle (che ora chiamiamo n_k) di cui è equipaggiato il $k-$esimo *User Equipment* UE_k (o terminale, o dispositivo) ad essa *associato* - che in gran parte dei casi, dispone di una sola antenna.

Come poter sfruttare in questo caso l'elevata velocità di trasmissione della BS, resa possibile dalla multiplazione spaziale di cui è capace grazie alle sue n_{BS} antenne? Molto semplice: dedicando ad ogni diverso terminale mobile UE_k una (o più) differente antenna di trasmissione della BS, realizzando una sorta di MIMO *distribuito*, anche se il passaggio non è così *indolore*, come andiamo ad approfondire dopo aver distinto tra le due direzioni di trasmissione, indicando come *uplink* il verso tra i terminali e la BS, e come *downlink* la direzione opposta, il cui insieme dei singoli collegamenti è indicato anche come *canale broadcast*.



Uplink In questa direzione le cose non sono molto diverse da quanto già approfondito. La BS riceve un segnale[75]

$$\boldsymbol{r}_{BS} = \sum_{k=1}^{U} \boldsymbol{H}_k^u \boldsymbol{s}_k + \boldsymbol{n} = \boldsymbol{H}^u \boldsymbol{s} + \boldsymbol{n}$$

in cui $\boldsymbol{H}_k^u$ è la matrice $(N \times n_k)$ di canale MIMO di *uplink* tra UE_k e BS, $\boldsymbol{s}_k$ è il vettore $(n_k \times 1)$ dei simboli inviati da UE_k, $\boldsymbol{H}^u = \left[\boldsymbol{H}_1^u \; \boldsymbol{H}_2^u \; \cdots \; \boldsymbol{H}_U^u\right]$ è la matrice $(n_{BS} \times \sum_{k=1}^{U} n_k)$ di uplink da *tutti* gli U terminali UE_k, ed $\boldsymbol{s} = \left[\boldsymbol{s}_1^T \; \boldsymbol{s}_2^T \; \cdots \; \boldsymbol{s}_U^T\right]^T$ è il vettore dei simboli complessivamente trasmessi.

Ogni UE_k con più di una antenna può trasmettere più flussi dati ricorrendo alla multiplazione spaziale, oppure se $n_k = 1$ il singolo collegamento è di tipo SIMO. In entrambi i casi la BS deve *separare* tra loro i contributi ricevuti dagli U utenti adottando le tecniche illustrate al § 21.5 e basate sulla conoscenza della $\boldsymbol{H}^u$ *di uplink*, che la BS può stimare nei confronti degli UE_k attivi. Fortunatamente la dispersione spaziale degli utenti fa si che $\boldsymbol{H}^u$ sia ben condizionata[76], senza dunque porre problemi per la sua

[74] Il punto accesso di una rete di telefonia mobile, o di una rete WiMax, o l'access point di una rete WiFi casalinga, vedi ad es. `https://it.wikipedia.org/wiki/Stazione_radio_base`, `https://it.wikipedia.org/wiki/WiMAX`, `https://it.wikipedia.org/wiki/Access_point`

[75] Supponiamo qui che i terminali mobili siano sincronizzati temporalmente, e che il sistema sia in grado di compensare i diversi ritardi di propagazione per la corretta determinazione degli istanti di campionamento dei simboli.

[76] Infatti ogni canale utente-BS sperimenta cammini multipli completamente differenti ed incorrelati, rendendo le colonne di $\boldsymbol{H}$ indipendenti, al punto da definirle come una *firma spaziale* di ciascun utente, dando al sistema di trasmissione la denominazione di *space division multiple access* o SDMA.

inversione.

Downlink In questa direzione il generico dispositivo UE$_k$ con n_k antenne riceve un vettore $\boldsymbol{r}_k$ di dimensione $n_k \times 1$ pari a

$$\boldsymbol{r}_k = \boldsymbol{H}_k^d \boldsymbol{s} + \boldsymbol{n} \tag{21.63}$$

in cui $\boldsymbol{H}_k^d$ $(n_k \times n_{BS})$ individua il canale MIMO *di broadcast* tra *tutte* le antenne di BS e le n_k di UE$_k$, e $\boldsymbol{s}$ $(n_{BS} \times 1)$ è il vettore di simboli destinati a *tutti* gli U utenti. Purtroppo essendo $n_k < n_{BS}$, il dispositivo mobile non può adottare le soluzioni previste al § 21.5 per cancellare i termini di interferenza, più evidenti particolarizzando la (21.63) al caso di un terminale ad antenna singola, che riceve *un valore*

$$r_k = \boldsymbol{h}_k^T \boldsymbol{s} + n = h_k s_k + \sum_{\substack{h=1 \\ h \neq k}}^{U} h_h s_h + n \tag{21.64}$$

in cui $\boldsymbol{h}_k$ è il vettore[77] $(n_{BS}, 1)$ dei guadagni complessi tra le n_{BS} antenne di trasmissione e l'unica di ricezione, h_k è *l'elemento* di $\boldsymbol{h}_k$ corrispondente all'antenna della BS che trasmette il simbolo s_k diretto all'utente k, ed il termine $\sum_{h=1,h\neq k}^{U} h_h s_h$ individua *l'interferenza* da parte degli altri e che si aggiunge al contributo del rumore[78], in modo che si possa parlare di rapporto *segnale / interferenti più rumore* o $SINR$.

21.6.1 Precodifica

Nell'impossibilità per i terminali mobili con antenna singola di sopprimere gli interferenti, occorre trovare una soluzione semplice[79] che possa essere attuata presso la BS: fortunatamente questa esiste ed è *la duale* di quella esposta ai § 21.5.2-21.5.3, necessitando come precondizione la conoscenza da parte della BS di tutti i vettori $\boldsymbol{h}_k$ *visti* dagli utenti, in modo da poter ricostruire la matrice

$$\boldsymbol{H}^d = \left[\boldsymbol{h}_1 \; \boldsymbol{h}_2 \; \cdots \; \boldsymbol{h}_U\right]^T \tag{21.65}$$

di dimensioni $U \times n_{BS}$ e che caratterizza il *canale broadcast* del downlink tra le antenne della BS e quelle (singole) degli U utenti.

Zero forcing precoding Una volta nota $\boldsymbol{H}^d$, la BS ne può calcolare la *pseudo inversa* (eq. (21.54))

$$\boldsymbol{P} = \left(\boldsymbol{H}^{d\,\dagger} \boldsymbol{H}^d\right)^{-1} \boldsymbol{H}^{d\,\dagger} \tag{21.66}$$

e sostituire al vettore di simboli $\boldsymbol{s}$ da inviare agli U terminali un nuovo vettore *precodificato*

$$\boldsymbol{s}_P = \boldsymbol{P}\boldsymbol{s} = \sum_{h=1}^{U} \boldsymbol{p}_h s_h$$

dove $\boldsymbol{p}_h$ è pari alla $h-$esima colonna di $\boldsymbol{P}$, di fatto *spargendo* ogni simbolo s_h destinato a ciascun utente h su *tutte* le n_{BS} antenne. Osserviamo ora che in virtù della definizione

[77] Corrispondente all'unica riga di $\boldsymbol{H}_k^d$ pertinente all'utente k qualora $n_k = 1$. Il vettore $\boldsymbol{h}_k$ può essere pensato nel senso *letterale* di indicare *una direzione* (a coordinate complesse) nello spazio generato dalle antenne della BS, dunque non la direzione *fisica* dell'utente, ma che lo distingue comunque dagli altri.

[78] A ben vedere, per un numero U elevato di utenti si applica il teorema centrale del limite, ed il termine $\sum_{h=1,h\neq k}^{U} h_h s_h$ tende ad una v.a. gaussiana.

[79] La soluzione *ottima* è nota come *dirty paper coding* (DPC), di difficile implementazione, vedi ad es. `https://it.wikipedia.org/wiki/Dirty_paper_coding`

di pseudoinversa si ha $\boldsymbol{H}^d\boldsymbol{P} = \boldsymbol{I}$, e dunque ogni colonna $\boldsymbol{p}_k$ di $\boldsymbol{P}$ è ortogonale a tutte le righe di $\boldsymbol{H}^d$ tranne che alla $k-$esima, ovvero $\boldsymbol{h}_k^T\boldsymbol{p}_h = \delta_{hk}$, in modo che la (21.64) diviene

$$r_k = \boldsymbol{h}_k^T\boldsymbol{s}_P + n = \boldsymbol{h}_k^T\sum_{h=1}^{U}\boldsymbol{p}_h s_h + n = \boldsymbol{h}_k^T\boldsymbol{p}_k s_k + n = s_k + n \tag{21.67}$$

da cui *scompare* il termine di interferenza da parte dei simboli s_h destinati agli altri utenti, ed il simbolo ricevuto può essere stimato come $\hat{s}_k = \arg\min_{s\in\mathcal{A}} |s - s_k|^2$.

MMSE precoding Tornando all'espressione (21.64), occorre distinguere il caso in cui le prestazioni del sistema sono *limitate dal rumore* e cioè quando $E\left\{|n|^2\right\} > E\left\{|\sum_{h=1,h\neq k}^{U} h_h s_h|^2\right\}$, dal caso in cui sono invece limitate *dagli interferenti*, dove la diseguaglianza cambia verso. La precodifica (21.66) si comporta bene solo nel secondo caso, mentre ponendo come obiettivo della matrice $\boldsymbol{P}$ quello di rendere minimo l'errore medio quadratico $E\left\{|s - s_k|^2\right\}$ si ottiene una soluzione che pur non cancellando completamente gli interferenti permette di bilanciare la riduzione della loro potenza con la riduzione della potenza di rumore; tale soluzione è indicata come versione *regolarizzata* di (21.66) e fornisce

$$\boldsymbol{P}_{MMSE} = \left(\boldsymbol{H}^{d\,\dagger}\boldsymbol{H}^d + \frac{1}{\rho}\boldsymbol{I}\right)^{-1}\boldsymbol{H}^{d\,\dagger} \tag{21.68}$$

in cui $\rho = {}^{E\left\{|\boldsymbol{P}\boldsymbol{s}|^2\right\}}/_{E\left\{|\boldsymbol{n}|^2\right\}}$ è una misura dell'SNR.

21.6.2 Controllo di potenza

I valori ottenuti per le colonne della matrice di precodifica $\boldsymbol{P}$ devono essere alterati, essenzialmente per due motivi:

- rispettare il vincolo sulla potenza trasmessa $P_T = E\left\{|\boldsymbol{P}\boldsymbol{s}|^2\right\}\cdot f_s$;
- rendere massimo un criterio di qualità complessiva come ad es. la velocità aggregata, cioè la somma di quella per ciascun utente.

Il secondo obiettivo è quello più *intrigante*, dato che aumentare la potenza per un utente significa aumentare la capacità del suo canale, a discapito di quella degli altri per i quali si verifica invece un aumento della potenza interferente, e (per il vincolo sulla potenza complessiva) la riduzione della propria. Inoltre la potenza allocata ad ogni utente (ossia al simbolo a lui diretto) deve essere determinata in base all'SNR di ricezione, altra informazione che deve essere comunicata alla BS, in modo che quest'ultima possa attuare tecniche di *water filling*, ed assegnare più potenza a chi ha un SNR migliore. Non ci addentriamo nelle strategie di ottimizzazione congiunta, alcune delle quali iterative: questo è solamente un testo introduttivo.

21.6.3 Prioritizzazione degli utenti

Fino ad ora si è implicitamente supposto che $U < n_{BS}$, ma gli utenti che intendono comunicare (tra quelli registrati presso una BS) possono essere ovviamente più delle antenne a disposizione. Tutti i terminali che richiedono il servizio comunicano alla BS le informazioni sullo stato del proprio canale o CSI ($\boldsymbol{h}_k^d$, ρ_k) in modo da permetterle di calcolare la $\boldsymbol{P}$ di precodifica. La BS deve quindi selezionare un sottoinsieme di

$U_M \leq n_{BS}$ utenti *scegliendoli* in modo da massimizzare la velocità somma; ma la velocità di ciascun utente ha come limite superiore la capacità del suo canale, che a sua volta dipende da relativo ρ_k. La ricerca esaustiva del migliore gruppo di $U_M < U$ utenti verso cui trasmettere ha complessità esponenziale, per cui è necessario ricorre a strategie euristiche. Una tecnica (cosiddetta *greedy* o *vorace*) consiste nello scegliere per primo l'utente che può dare il maggior contributo di velocità, poi il secondo migliore, e così via. Ovviamente l'obiettivo da massimizzare, anziché tenere conto solamente della velocità-somma, può prendere in considerazione anche altri fattori, come la lunghezza di coda (da minimizzare) od altro ancora, vedi appresso.

21.6.4 Precodifica con feedback limitato

Il calcolo della matrice $\boldsymbol{P}$ (eq. (21.66) o (21.68)) da parte della BS necessita della conoscenza da parte di quest'ultima della matrice $\boldsymbol{H}^d$ (21.65), la cui riga $\boldsymbol{h}_k^d$ descrive il canale tra BS e UE_k, stimato presso quest'ultimo. Nel caso di un sistema *full duplex* basato su portanti differenti per le due direzioni del collegamento il canale *non è reciproco* (nota 53), e quindi occorre un canale di uplink *di ritorno* attraverso il quale i terminali mobili possano trasmettere alla BS la loro stima di $\boldsymbol{h}_k^d$, codificata mediante B bit frutto della quantizzazione dei valori di $\boldsymbol{h}_k^d$ e di ρ_k. Come noto dalla teoria dell'informazione (§ 9.6.2), una distorsione nulla si ottiene solamente per $B \to \infty$, dunque per B finito è inevitabile la presenza di errori negli $\boldsymbol{h}_k^d$ quantizzati, che a loro volta causano errori nella determinazione delle colonne $\boldsymbol{p}_k$ della matrice di precodifica. Scrivendo quindi $\tilde{\boldsymbol{p}}_k = \boldsymbol{p}_k + \boldsymbol{e}_k$ in cui $\boldsymbol{p}_k$ e $\tilde{\boldsymbol{p}}_k$ sono i vettori di precodifica calcolati per una CSI esatta oppure quantizzata, la (21.67) diviene

$$\begin{aligned} r_k &= \boldsymbol{h}_k^T \textstyle\sum_{h=1}^{U} \tilde{\boldsymbol{p}}_h s_h + n = \boldsymbol{h}_k^T \sum_{h=1}^{U} \boldsymbol{p}_h s_h + \boldsymbol{h}_k^T \sum_{h=1}^{U} \boldsymbol{e}_h s_h + n = \\ &= \boldsymbol{h}_k^T \boldsymbol{p}_k s_k + \boldsymbol{h}_k^T \textstyle\sum_{h=1}^{U} \boldsymbol{e}_h s_h + n = s_k + n_I + n \end{aligned} \tag{21.69}$$

in cui $n_I = \boldsymbol{h}_k^T \sum_{h=1}^{U} \boldsymbol{e}_h s_h$ costituisce un termine di rumore *interferente*.

Legame feedback - SNR Il termine n_I nella (21.69) è tanto maggiore quanto peggiore è la risoluzione della quantizzazione di $\boldsymbol{h}_k^d$, ovvero quanto minore è il numero B di bit di feedback inviati da ciascun terminale. Mantenendo B fisso si produce un effetto *piattaforma* (da *error floor*) nelle prestazioni (sia di velocità che di errore) in funzione dell'SNR, dato che n_I non cambia anche se il rumore n si riduce. Il rimedio è quello di richiedere una maggiore accuratezza (ossia un maggior numero di bit B_k) nei valori trasmessi $\boldsymbol{h}_k^d$ in forma quantizzata per quei terminali UE_k che sperimentano un rumore termico n minore[80], in modo da mantenere i due tipi di rumore (n e n_I) alla stessa potenza. Inoltre essendo i vettori $\boldsymbol{h}_k$ e $\boldsymbol{p}_k$ di dimensione n_{BS}, l'espressione $n_I = \boldsymbol{h}_k^T \sum_{h=1}^{U} \boldsymbol{e}_h s_h$ implica una somma di n_{BS} termini, cosicché anche il numero di antenne di trasmissione contribuisce in modo diretto ad n_I. Qualora l'SNR sia uguale per tutti, è stato trovato[81] che la BS *può* conseguire un guadagno di multiplazione massimo

[80] Ovvero un SNR ρ_k maggiore.

[81] N. JINDAL, *MIMO Broadcast Channels With Finite-Rate Feedback*, IEEE Trans. on Inf. Th. Nov. 2006, ne trovo una copia *libera* presso Univ. of Minnesota

(n_{BS}) quando ogni terminale invia

$$B = (n_{BS} - 1)\log_2 SNR \approx \frac{n_{BS} - 1}{3} SNR_{dB} \quad \text{bit} \tag{21.70}$$

soffrendo una penalizzazione di soli 3 dB di prestazione rispetto alla conoscenza perfetta della CSI.

Quantizzazione vettoriale del canale di downlink Approfondiamo un minimo come vengono prodotti i B bit di feedback. La QV, introdotta al § 10.1.2.4, è la tecnica per eseguire la codifica di sorgente $\tilde{\boldsymbol{h}}_k = Q(\boldsymbol{h}_k)$ della grandezza vettoriale $\boldsymbol{h}_k$. Consiste in un codebook[82] $\mathcal{H} = \left\{\boldsymbol{h}_{q1}, \boldsymbol{h}_{q2}, \cdots, \boldsymbol{h}_{q2^B}\right\}$ di 2^B vettori di canale *a norma unitaria*, tra cui si sceglie il vettore $\boldsymbol{h}_{opt}$ *più vicino* a quello misurato $\boldsymbol{h}_k$ (reso a norma unitaria anch'esso) usando come criterio di distorsione il *modulo quadro* del *prodotto scalare*[83]. Si sceglie pertanto l'indice[84]

$$n_{opt} = \arg\max_{\boldsymbol{h}_n \in \mathcal{H}} \left|\boldsymbol{h}_n^\dagger \boldsymbol{h}_k\right|^2 = \arg\max_{\boldsymbol{h}_n \in \mathcal{H}} \cos^2\left[\angle\left(\boldsymbol{h}_n^\dagger, \boldsymbol{h}_k\right)\right]$$

decidendo quindi per $\tilde{\boldsymbol{h}}_k = Q(\boldsymbol{h}_k) = \boldsymbol{h}_{opt}$, e si trasmettono i B bit della codifica binaria di n_{opt}. Anche la BS è a conoscenza dello stesso CB $\mathcal{H}$, da cui preleva $\boldsymbol{h}_{n_{opt}}$ e lo usa per costruire $\boldsymbol{H}^d$ da cui ricavare $\boldsymbol{P}$. Anche se i vettori del codebook sono a norma unitaria, si ottengono vettori di precoding (ad es., mediante *zero forcing*) comunque in grado di cancellare gli interferenti.

Qualora il terminale rilevi una variazione di SNR per cui la (21.70) indica la necessità di variare B, l'adozione di un CB organizzato *gerarchicamente* semplifica l'aumento (o la riduzione) della sua cardinalità, che raddoppia per ogni bit aggiunto a B. Infine, qualora il numero di utenti attivi sia il risultato di una strategia di prioritizzazione tra un insieme più ampio, si possono ottenere risultati migliori scegliendo quei terminali che abbiano comunicato vettori $\tilde{\boldsymbol{h}}_k$ il più possibile ortogonali tra loro.

Quantizzazione del precoder Osserviamo ora che l'errore $\boldsymbol{e}_h = \boldsymbol{p}_h - \tilde{\boldsymbol{p}}_h$ che compare nell'espressione (21.69) del segnale ricevuto r_k come un termine di *rumore interferente* $n_I = \boldsymbol{h}_k^T \sum_{h=1}^{U} \boldsymbol{e}_h s_h$, essendo causato dal calcolo delle colonne della matrice $\boldsymbol{P}$ di pre-

[82] La scelta degli elementi del codebook può avvenire in accordo ad un modello di d.d.p. $p(\boldsymbol{h}^d)$ nel contesto dell'algoritmo di *Lloyd-Max* (nota 26 a pag. 100), oppure a partire da una base dati di vettori osservati su cui eseguire un algoritmo di clusterizzazione, vedi ad es. `https://en.wikipedia.org/wiki/K-means_clustering`. Infine è da citare la possibilità di popolare il CB in modo *casuale*, a partire da un seme iniziale, trasmettendo il quale il CB stesso può essere autonomamente *ri-generato* alla BS.

[83] Vedi § 2.4.3. In questo modo due vettori sono tanto più simili quanto più condividono lo stesso orientamento; la normalizzazione ne ha portato *la punta* sulla superficie di una sfera di raggio unitario, e qui si va a finire negli *spazi Grassmaniani*. Qualcuno può chiedersi: come mai la distorsione è nulla anche per vettori *opposti*? Una possibile risposta è che i vettori quantizzati $\tilde{\boldsymbol{h}}_h$ servono alla BS per calcolare vettori $\boldsymbol{p}_k$ di precoding *ortogonali* ad $\tilde{\boldsymbol{h}}_{h \neq k}$. L'ortogonalità dunque permane anche nei confronti di un vettore con il segno cambiato.

[84] Per evitare che più terminali possano scegliere lo stesso vettore del CB, ognuno di essi può utilizzare un diverso CB. Dato che questi devono essere tutti noti presso la BS, la questione può essere semplificata nel caso di CB generati casualmente a partire da un seme.

codifica a partire dalla $\boldsymbol{H}^d$ *quantizzata*, è un errore *causato da un altro errore*. Tanto vale allora cercare di quantizzare direttamente le colonne di $\boldsymbol{P}$, e lasciare che siano i terminali stessi a decidere quale vettore di precodifica $\tilde{\boldsymbol{p}}_k$ è il più idoneo moltiplicatore dei simboli destinati a ciascuno di essi, senza necessità di comunicare la CSI.

A questo fine si adotta un codebook $\mathcal{P} = \left\{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{2^B}\right\}$ *casuale* con vettori complessi $\boldsymbol{p}_h$ ($n_{BS} \times 1$) a norma unitaria ed uniformemente distribuiti in *tutte le direzioni*[85], noto sia alla BS che a tutti i terminali UE_k. Ogni dispositivo, dopo aver stimato il proprio vettore di canale $\boldsymbol{h}_k^d$, può individuare il vettore $\boldsymbol{p}_{opt} \in \mathcal{P}$ *più parallelo* ad $\boldsymbol{h}_k^d$ e dunque in grado di massimizzare il proprio $SINR$[86], ovvero

$$SINR_{k,m} = \frac{|\boldsymbol{h}_k^T \boldsymbol{p}_m|^2}{1/\rho + \sum_{h=1, h\neq m}^{U} |\boldsymbol{h}_k^T \boldsymbol{p}_k|^2} \qquad \boldsymbol{p}_{opt} = \arg\max_{\boldsymbol{p}_m \in \mathcal{P}} SNR_{k,m} \tag{21.71}$$

dove $SINR_{k,m}$ è l'$SINR$ per UE_k qualora i simboli ad esso destinati siano precodificati con il vettore $\boldsymbol{p}_m$. Il terminale quindi confronta $SINR_{k,opt}$ con un valore soglia γ che gli è stato comunicato dalla BS, e *se lo supera* può inviare sia $SINR_{k,opt}$ che i B bit della codifica binaria dell'indice *opt*, indicando così il desiderio che i simboli s_k a lui diretti siano precodificati mediante $\boldsymbol{p}_{opt}$. Dal canto suo dopo aver ricevuto il feedback di ciascuno, la BS individua per ogni vettore $\boldsymbol{p}_h$ il terminale k (ai cui dati s_k verrà applicato $\boldsymbol{p}_h$) che ha comunicato il valore $SNR_{k,h}$ più grande, ovvero per il quale $\boldsymbol{p}_h$ porta più beneficio. Con tali vettori $\boldsymbol{p}_h$ costruisce la matrice $\boldsymbol{P}$, ed effettua la trasmissione dedicata ai terminali selezionati.

Notiamo infine che adottando un codebook $\mathcal{P}$ costituito dalle n_{BS} colonne di una matrice di precoding *identità*, ovvero $\boldsymbol{P} = \boldsymbol{I}_{n_{BS}}$, si ottengono vettori $\boldsymbol{p}_h$ tutti nulli tranne che per l'$h-$esimo elemento pari ad uno. Con la scelta (21.71) ciascun terminale indica in tal caso l'antenna della BS rispetto alla quale sperimenta le migliori condizioni di propagazione, come avviene per la selezione di diversità (§ 21.3.1.1). Specialmente nel caso di un ridotto numero di utenti, ciò permette di organizzare le trasmissioni di downlink in modalità *a divisione di tempo*, trasmettendo di volta in volta con *tutta* la

[85] Ovvero tali da rendere *massima* la separazione angolare *minima* $d(\mathcal{P})$ tra due elementi di $\mathcal{P}$, avendo definito $d(\mathcal{P}) = \min_{1\leq i\leq j\leq 2^B} \sin\left(\theta_{i,j}\right)$ e $\theta_{i,j} = \arccos\left(\boldsymbol{p}_i^{\dagger} \boldsymbol{p}_j\right)$, e quindi $\mathcal{P} : d(\mathcal{P}) = \max$. Di nuovo, è un problema noto, vedi J.H. Conway, R.H. Hardin, N.J.A. Sloane, *Packings in Grassmannian Spaces*, Experimental Mathematics 5:2, 1996, vedi `http://www2.stat.duke.edu/~sayan/SAMSI/lec/conway.pdf`. Notiamo che qualora $2^B > n_{BS}$ non è possibile trovare 2^B vettori mutuamente ortogonali.

[86] Consideriamo la (ipotetica) trasmissione da parte della BS del segnale $\boldsymbol{s}_P = \sum_{h=1}^{2^B} \boldsymbol{p}_h s_h$, che determina presso UE_k la ricezione (vedi eq. (21.67)) del segnale $r_k = \boldsymbol{h}_k^T \boldsymbol{s}_P + n = \sum_{h=1}^{2^B} \boldsymbol{h}_k^T \boldsymbol{p}_h s_h + n$. Si intende trovare l'indice h associato al vettore $\boldsymbol{p}_h$ che determina le migliori condizioni di ricezione per UE_k. Se consideriamo come componente di segnale quella relativa al generico indice m ovvero $\boldsymbol{h}_k^T \boldsymbol{p}_m s_m$, il termine di rumore è $\sum_{h=1,h\neq m}^{2^B} \boldsymbol{h}_k^T \boldsymbol{p}_h s_h + n$, e dunque il valore di $SINR_{k,m}$ si calcola come

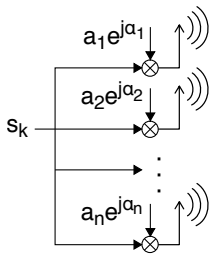
$$SINR_{k,m} = \frac{E\left\{|\boldsymbol{h}_k^T \boldsymbol{p}_m s_m|^2\right\}}{\sigma_n^2 + E\left\{|\sum_{h=1,h\neq m}^{2^B} \boldsymbol{h}_k^T \boldsymbol{p}_h s_h|^2\right\}} = \frac{\mathcal{E}_s \cdot |\boldsymbol{h}_k^T \boldsymbol{p}_m|^2}{\sigma_n^2 + \mathcal{E}_s \cdot \sum_{h=1,h\neq m}^{2^B} |\boldsymbol{h}_k^T \boldsymbol{p}_h|^2} = \frac{|\boldsymbol{h}_k^T \boldsymbol{p}_m|^2}{1/\rho + \sum_{h=1,h\neq m}^{2^B} |\boldsymbol{h}_k^T \boldsymbol{p}_h|^2}$$

in cui $\rho = \mathcal{E}_s/\sigma_n^2$ e $\mathcal{E}_s = E\left\{|s_h|^2\right\}$, ed il passaggio a denominatore è possibile considerando i $\boldsymbol{p}_h$ *ortonormali*. $SINR_{k,m}$ è dunque massimo quando lo è il numeratore, ossia quando lo è il prodotto scalare tra $\boldsymbol{h}_k$ e $\boldsymbol{p}_m$.

potenza disponibile ed alla *massima* velocità, mediante una unica antenna m a beneficio del terminale k che ha dichiarato un miglior $SINR_{k,m}$.

21.6.5 Beamforming

L'effetto di uno specifico vettore di precodifica $\boldsymbol{p}_k$ a norma unitaria è quello di moltiplicare il simbolo s_k per un valore complesso $a_{k,i}e^{j\alpha_{k,i}}$ diverso per ciascuna antenna $i = 1, 2, \cdots, n_{BS}$ della BS. Di fatto questo è il modo di operare tipico delle *antenne intelligenti*[87] che realizzano un diagramma di radiazione *direttivo ed orientabile* senza dove *muovere* nulla, ma adottando una schiera di antenne e *sfasando* con valori α_i crescenti il segnale diretto all'$i-$esima antenna.



Al diagramma di radiazione ottenuto è stato dato il nome di *beam* ed alla tecnica quello di *beamforming*, in cui beam può essere tradotto come *trave, raggio, fascio*; data la somiglianza formale con l'operazione di precodifica l'uso del termine di *beamforming* si è esteso ad indicare non solo le operazioni di precoding, ma anche quelle (§ 21.4.1.1) basate sulla SVD ed il water-filling per ottenere un numero di canali *indipendenti* pari al rango di $\boldsymbol{H}$, modalità operativa indicata anche come *eigen-beamforming*.

Il vantaggio più evidente del beamforming *propriamente detto* è che, trasmettendo in direzione (geometrica) del destinatario, si riduce la potenza interferente per gli altri utenti dislocati altrove e che *non rientrano* nel fascio. Per ottenere tale risultato è però necessario che la BS conosca la posizione geografica dei terminali mobili. In alternativa si possono generare *più beam* in direzioni casuali diverse[88] e lasciare che siano i terminali ad indicare per quale beam si verifica una ricezione migliore, un po' come illustrato per il caso della quantizzazione del precoder, che viene infatti descritto anche con il termine di beamforming *opportunistico*.

Nel caso di propagazione affetta da fading di Rayleigh (e dunque priva di un cammino diretto prevalente) il concetto di direttività delle antenne perde un po' di significato, mentre continua a rivestire importanza per celle di raggio più ampio, e in cui antenne (della BS) siano situate su di una torre sufficientemente elevata rispetto agli ostacoli presenti. In tal caso si determina una forte componente di onda diretta ed un fattore di Rice non trascurabile, riducendo di fatto il grado di diversità spaziale: in questo caso l'adozione del beamforming *geografico* non è solo vantaggiosa, ma necessaria. Al contrario, per un ambiente privo di percorsi diretti e ricco di scattering è preferibile attuare il precoding atto a massimizzare l'$SINR$ (eq. (21.71)) e capace di far combinare in modo coerente le repliche che viaggiano su cammini multipli differenti, come esemplificato nella animazione presso `https://youtu.be/XBb481RNqGw`.

[87] O *phased array*, vedi ad es. `https://en.wikipedia.org/wiki/Phased_array` ma anche `https://www.vialattea.net/content/1875/`.

[88] Cioè pilotare i modulatori collegati alle antenne con un vettore $\boldsymbol{s}_B = \sum_{h=1}^{n_B} \boldsymbol{b}_h s_h$ in cui il simbolo s_h destinato all'$h-$esimo utente viene *affasciato* nella direzione stabilita dal beam $\boldsymbol{b}_h$, ne più ne meno come avviene nel precoding.

21.7 Trasmissione MIMO - OFDM

La necessità di subire un fading *piatto*, ovvero che il segnale modulato occupi una banda inferiore alla *banda di coerenza*, determina *un limite* alla massima velocità binaria della trasmissione MIMO *a portante singola*. Infatti per contenere la banda si dovrebbe ricorrere ad una costellazione con un numero molto elevato di livelli, incappando in un rapido degrado delle prestazioni; lasciare invece che le banda aumenti in proporzione alla velocità di simbolo significa dover equipaggiare ogni antenna di ricezione con uno stadio di equalizzazione adattativa, con evidente aumento di complessità, tanto maggiore con l'aumento della velocità. D'altra parte l'adozione di una modulazione a spettro espanso che affronta l'equalizzazione mediante una architettura di ricezione *Rake* (§ 20.5.2) determina una complessità che cresce in modo quadratico con la velocità di trasmissione[89].

Al contrario, se il segnale trasmesso da ciascuna antenna adotta una modulazione OFDM (§ 16.8) i problemi appena esposti semplicemente *svaniscono*, ed al tempo stesso si aprono possibilità di ulteriore miglioramento della qualità e delle prestazioni del collegamento. Tutto nasce dal fatto che, ripartendo il flusso binario da trasmettere su P canali ortogonali (in frequenza) con una occupazione di banda ridotta, per ciascuno di essi risulta verificata l'ipotesi di fading piatto. Inoltre l'adozione del *prefisso ciclico* consente di annullare gli effetti dell'ISI senza dovere ricorrere ad uno stadio di equalizzazione, se non per la componente di stima di canale alla frequenza delle sottoportanti, che è comunque necessaria per la ricezione di trasmissioni MIMO. Infine l'utilizzo di una FFT per realizzare i processi di modulazione e demodulazione fa si che la complessità dei dispositivi aumenti *solo* con legge $P \log_2 P$ al crescere del numero di sottoportanti P, e dunque della velocità di trasmissione.

Vantaggio di diversità La trasmissione MIMO-OFDM consente inoltre di sfruttare, oltre a quella *spaziale* dovuta alla molteplicità delle antenne, anche la diversità *in frequenza*[90]. Pur se il canale associato a ciascuna sottoportante è considerato sede di fading *piatto*, il valore h del suo guadagno complesso è una v.a. che può assumere una realizzazione *diversa* per sottoportanti differenti[91], aspetto che nelle trasmissioni SISO viene affrontato con la tecnica COFDM (§ 16.8.10). In particolare si dimostra[92] che un canale per il quale sono individuabili L cammini multipli (a cui si deve la selettività in frequenza, § 20.4.5) ed equipaggiato da n_T, n_R antenne ai lati del collegamento, offre

[89] La fonte di questa affermazione (Wikipedia) cita la tesi di dottorato G.RALEIGH, *On Multivariate Communication Theory and Data Rate Multiplying Techniques for Multipath Channels*, 1998, ma il link a cui puntava non risulta più attivo.

[90] La diversità in frequenza è esattamente quella discussa al § 20.3.3.1 e dovuta ai cammini multipli che rendono il canale selettivo in frequenza (§ 20.4.5).

[91] Possiamo notare che come le antenne riescono ad offrire diversità spaziale solo qualora le stesse siano sufficientemente distanziate rispetto alla lunghezza d'onda λ della trasmissione, così le sottoportanti offrono diversità in frequenza solo su canali separati da un intervallo di frequenza maggiore della banda di coerenza.

[92] H. BOLCSKEI, A.J. PAULRAJ, *Space-frequency coded broadband OFDM systems*, 2000 IEEE Wireless Comm. and Networking Conference

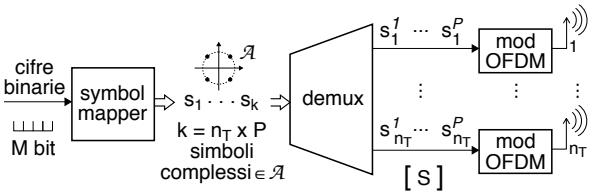
un ordine di diversità *massimo*[93] pari al prodotto di queste quantità, ovvero pari a

$$L \cdot n_T \cdot n_R \tag{21.72}$$

Ma procediamo con ordine.

21.7.1 Modello di canale MIMO-OFDM

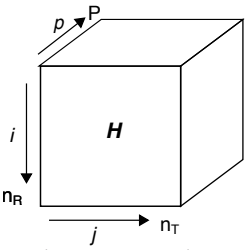
Partiamo dal caso più generale in figura, in cui $n_T \times P$ simboli complessi vengono sistematicamente[94] ripartiti su n_T antenne mediante ciascuna delle quali viene trasmesso un simbolo OFDM costituito da P sottoportanti. Anziché un semplice vettore $\boldsymbol{s}$ di n_T simboli, per ogni utilizzo del canale viene ora trasmessa una matrice $\boldsymbol{S}$ di $n_T \times P$ elementi complessi s_{jp}; similmente, in ricezione si ottiene la matrice $\boldsymbol{R}$ di $n_R \times P$ elementi r_{ip}, secondo la relazione



$$r_{ip} = \sum_{j=1}^{n_T} h_{ijp} \cdot s_{jp} + n_{ip} \tag{21.73}$$

con $i = 1 \cdots n_R$ e $p = 1 \cdots P$, in cui h_{ijp} è la risposta in frequenza tra le antenne j e i alla frequenza p (supposta costante almeno per la durata del simbolo OFDM), la sommatoria prende in considerazione tutte le antenne di trasmissione, e n_{ip} è il corrispondente campione complesso di rumore gaussiano circolare.

La classica matrice $\boldsymbol{H}$ dei guadagni complessi tra le n_R antenne di ricezione (le righe i) e la n_T di trasmissione (colonne j) che compare nella (21.2) si è dunque *arricchita* di un ulteriore indice, quello che individua il *piano p* corrispondente ad ognuna delle sottoportanti usate dall'OFDM, in modo che tutti i $n_R \cdot n_T \cdot P$ valori possano essere individuati come $(\boldsymbol{H})_{i,j,p} = h_{ijp}$. La relazione (21.73) può quindi essere riscritta in forma matriciale considerando le matrici $\boldsymbol{S}$ ed $\boldsymbol{R}$ dei simboli trasmessi e ricevuti lette per colonne, ottenendo rispettivamente i vettori $\boldsymbol{s}_p$ degli n_T valori trasmessi alla frequenza p, e quelli $\boldsymbol{r}_p$ degli n_R valori ricevuti alla medesima frequenza, in modo da poter scrivere



$$\boldsymbol{r}_p = \boldsymbol{H}_p \cdot \boldsymbol{s}_p + \boldsymbol{n}_p \quad \text{con} \quad p = 1, 2, \cdots, P$$

in cui $\boldsymbol{H}_p$ di dimensione $n_R \times n_T$ è *il piano p* della matrice tridimensionale $\boldsymbol{H}$ e corrisponde alla tradizionale relazione (21.2) tra le antenne, specializzata per la frequenza p, ed $\boldsymbol{n}_p$ di dimensione n_R rappresenta il rumore alla stessa frequenza. La costruzione della matrice $\boldsymbol{R}$ *completa* si realizza quindi concatenando le sue colonne $\boldsymbol{r}_p$, ovvero

$$\boldsymbol{R} = [\boldsymbol{r}_1\, \boldsymbol{r}_2\, \cdots\, \boldsymbol{r}_P] = [\boldsymbol{H}_1\boldsymbol{s}_1\, \boldsymbol{H}_2\boldsymbol{s}_2\, \cdots\, \boldsymbol{H}_P\boldsymbol{s}_P]$$

[93] Con il massimo conseguito solo per il caso di indipendenza di tutti i percorsi alternativi.

[94] In alternativa la ripartizione può contemplare elaborazioni più complesse come quelle descritte al § 21.7.2, o prevedere anche uno stadio di precoding, od anche di codifica di canale e/o interleaving..

La variazione di $\boldsymbol{H}_p$ con l'indice di portante p è dunque l'elemento che aggiunge al MIMO-OFDM anche la componente di diversità *in frequenza* oltre che spaziale.

21.7.2 Codice spazio-tempo-frequenza

Per beneficiare del guadagno di diversità spaziale la trasmissione MIMO-OFDM si deve *affidare* ai codici spazio-tempo affrontati al § 21.3.2.2. A tale scopo lo schema di pag. 713 si modifica come indicato in fig. 21.6, che mostra come ad una sequenza di k simboli s_i ad $L = 2^M$ valori venga fatta corrispondere una codeword

$$\boldsymbol{C} = \begin{array}{c} \rightarrow \text{ antenne } \rightarrow \\ \begin{bmatrix} \boldsymbol{c}_1^1 & \boldsymbol{c}_2^1 & \cdots & \boldsymbol{c}_{n_T}^1 \\ \boldsymbol{c}_1^2 & \boldsymbol{c}_2^2 & \cdots & \boldsymbol{c}_{n_T}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{c}_1^T & \boldsymbol{c}_2^T & \cdots & \boldsymbol{c}_{n_T}^T \end{bmatrix} \\ \end{array} \quad \begin{array}{c} \downarrow \\ \text{istanti} \\ \downarrow \end{array}$$

i cui elementi $\boldsymbol{c}_t^j$ sono ora vettori di P elementi che individuano i punti di costellazione $L-$aria con cui modulare le sottoportanti del simbolo OFDM trasmesso all'istante t dall'antenna j. In questo contesto il tasso di codifica (eq. (21.19)) prende in considerazione anche il numero di portanti, ed è definito come $R_c = \frac{k}{PT}$.

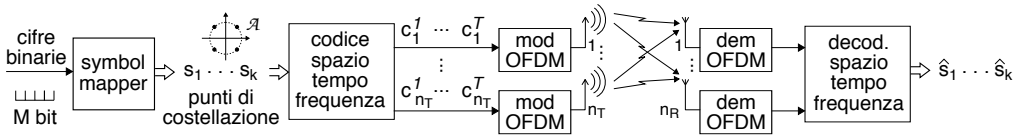


Figura 21.6: Codifica spazio-tempo-frequenza per un collegamento MIMO-OFDM

Si aprono ora diverse possibilità: in fig.21.7-a viene mostrata l'applicazione di un codice STBC di Alamouti su di una unica sottoportante, trasmesso da due antenne in due istanti temporali: si ottiene $R_c = 1$ in quanto si trasmettono due simboli in due istanti, e lo stesso può essere fatto su tutte le P portanti. La fig. 21.7-b mostra invece lo stesso codice trasmesso sempre da due antenne, ma applicato su due portanti *dello stesso simbolo*, conseguendo nuovamente $R_c = 1$: questo caso viene indicato come *space-frequency block code* (SFBC). E' evidente che mentre nel primo caso possiamo trarre vantaggio solo sul fronte della diversità spaziale, nel secondo vorremmo sfruttare anche la diversità frequenziale, sempre nei limiti di quanto evidenziato alla nota 91; in entrambi i casi non si riesce però a conseguire[95] tutta la diversità offerta sia nello spazio che in frequenza, vedi eq. (21.72). Qualcosa di meglio si riesce a fare secondo gli approcci di fig. 21.7-c e -d, che raddoppiano il numero di portanti su cui si sviluppa il codice, con un dimezzamento di R_c; per entrambi se il numero di repliche del codice eguaglia quello (L) dei cammini multipli, lo schema può offrire il massimo guadagno (21.72) con un rate $R_c = 1/L$. Tuttavia, l'implementazione di fig. 21.7-d presenta una maggiore complessità di decodifica.

[95] Una buona sintesi storica di questo filone di studio si trova in W Zhang, XG Xia, KB Letaief, *Space-time/frequency coding for MIMO-OFDM in next generation broadband wireless systems*, IEEE Wireless Comm. June 2007, di cui trovo una copia presso `https://www.eecis.udel.edu/~xxia/WeiZhang1.pdf`

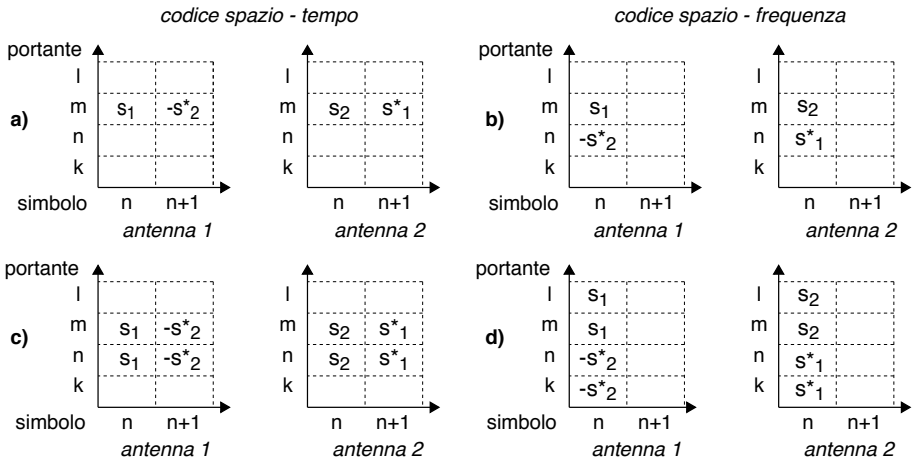


Figura 21.7: Schemi di codifica spazio-tempo e spazio-frequenza

Rimandando alla nota (95) per gli approfondimenti, qui citiamo solamente che sono stati individuati codici SFBC in grado di sfruttare il massimo guadagno di diversità (21.72) con $R_c = 1$, ed anche (impiegando codici *algebrici*[96]) con $R_c = n_T$, ovvero capaci di trasmettere un diverso flusso per antenna. Ovviamente, tutto ciò al prezzo di una complessità di decodifica ancora maggiore.

Ma non è finita! Nel caso in cui il canale presenti una accentuata variabilità temporale, con valori di $\boldsymbol{H}$ da considerare costanti nell'ambito di un blocco (in cui entrano uno o pochi simboli OFDM) ma variabili da un blocco all'altro, il canale viene definito come soggetto a *block-fading*, ed una ulteriore categoria di codici STFBC[97] che si estende su di n_B blocchi con fading *indipendente* può conseguire un ordine di diversità ancora maggiore, e pari a $L \cdot n_T \cdot n_R \cdot n_B$, con un tasso $R_C = n_T$.

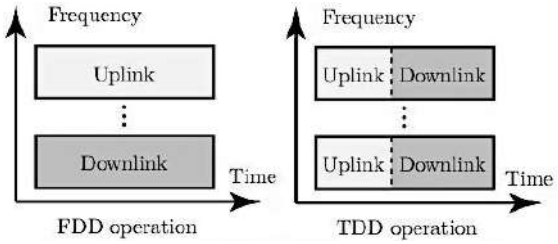
21.7.3 Sistema multiutente MU-MIMO-OFDM

La trasmissione MIMO-OFDM si è rivelata particolarmente idonea a realizzare un sistema ad accesso multiplo come introdotto al § 21.6, in cui una stazione radio base BS con n_{BS} antenne comunica con U terminali UE_k ad antenna singola, sia in direzione downlink (DL o canale broadcast) che in uplink (UL o accesso multiplo). L'unica antenna di ricezione dei terminali impedisce l'uso di codici spazio-tempo, ma non di quelli spazio-frequenza.

Alternanza temporale Nel caso OFDM si rendono possibili considerazioni che portano a preferire uno schema di condivisione delle risorse radio in cui la trasmissione nei due versi *si alterna* nel tempo secondo una modalità nota come *time division duplex* o TDD, ed occupa la medesima banda nelle due direzioni, banda che per l'OFDM è di

[96] Vedi H. El Gamal, M. O. Damen, *Universal Space–Time Coding*, IEEE Trans. on Inf. Th., May 2003, reperibile presso `http://www2.ece.ohio-state.edu/~elgamal/print12.pdf`

[97] Vedi W. Zhang, X. Xia, P. C. Ching, *High-Rate Full-Diversity Space–Time–Frequency Codes for Broadband MIMO Block-Fading Channels*, IEEE Trans. on Comm., January 2007, di cui trovo una copia su `CiteseerX`



estensione ben maggiore rispetto al caso a portante singola, coprendo un intervallo pari a diversi multipli della banda di coerenza B_c. Lo schema TDD consente una serie di semplificazioni rispetto al *frequency division duplex* (FDD), prima tra tutte la possibilità per gli utenti mobili di evitare di dover comunicare alla BS la stima di canale da essi effettuata (§ 21.7.3.2), in virtù della reciprocità del canale (vedi nota 53).

Uplink La BS alla portante p riceve un segnale

$$\boldsymbol{r}_{BS}^{p} = \sum_{k=1}^{U} \boldsymbol{h}_{k}^{p} s_{k}^{p} + \boldsymbol{n} = \boldsymbol{H}_{p}^{u} \boldsymbol{s}_{p} + \boldsymbol{n} \tag{21.74}$$

in cui $\boldsymbol{h}_k^p$ è il vettore $(n_{BS} \times 1)$ del canale MIMO di *uplink* per la portante p tra UE_k e le antenne della BS, s_k^p è il simbolo inviato dall'unica antenna di UE_k sulla portante p con energia $\mathcal{E}_k = E\left\{|s_k^p|^2\right\}$, $\boldsymbol{H}_p^u = \left[\boldsymbol{h}_1^p\ \boldsymbol{h}_2^p\ \cdots\ \boldsymbol{h}_U^p\right]$ è la matrice $(n_{BS} \times U)$ di uplink da *tutti* gli U dispositivi UE_k, ed $\boldsymbol{s}_p$ è il vettore $(U \times 1)$ ad elementi s_k^p; qualora uno UE non trasmetta sulla portante p, pone il valore s_k^p a zero. Infine, $\boldsymbol{n}$ è un vettore aleatorio gaussiano complesso a media nulla e covarianza $\sigma_{UL}^2 \cdot I_{n_{BS} \times n_{BS}}$.

Per effettuare la detezione del simbolo s_j^p la BS calcola la matrice $\boldsymbol{G}_p$ a partire da $\boldsymbol{H}_p$ con uno dei metodi discussi al § 21.5 e valuta il prodotto tra la $j-$esima *riga* $\boldsymbol{g}_j^p$ di $\boldsymbol{G}_p$ e la (21.74), ottenendo

$$\tilde{s}_j^p = \boldsymbol{g}_j^p \boldsymbol{r}_{BS}^p = \boldsymbol{g}_j^p \boldsymbol{h}_j^p s_j^p + \sum_{\substack{k=1 \\ k \neq j}}^{U} \boldsymbol{g}_j^p \boldsymbol{h}_k^p s_k^p + \boldsymbol{g}_j^p \boldsymbol{n} \tag{21.75}$$

in cui il primo termine è quello desiderato, il secondo sono gli interferenti (che se in numero elevato possono essere ritenuti a somma gaussiana) e l'ultimo è il nuovo termine di rumore; dopodiché si valuta $\hat{s}_j = \arg\min_{s \in \mathcal{A}} \left(\tilde{s}_j - s\right)^2$.

Notiamo che qualora $\boldsymbol{G}_p$ sia ottenuta con l'approccio *zero forcing* (§ 21.5.2), il termine di interferenza si annulla; notiamo inoltre che il calcolo (21.75) può essere svolto in contemporanea per tutti gli utenti, come $\tilde{\boldsymbol{s}} = \boldsymbol{G}_p \boldsymbol{r}_{BS}^p = \boldsymbol{G}_p \boldsymbol{H}_p^u \boldsymbol{s}_p + \boldsymbol{G}_p \boldsymbol{n}$.

Downlink In questa direzione le n_{BS} antenne inviano (sulla portante p) il segnale

$$\boldsymbol{x}_p = \sum_{k=1}^{U} \boldsymbol{p}_k^p s_k^p = \boldsymbol{P}_p \boldsymbol{s}_p \tag{21.76}$$

dove $\boldsymbol{p}_k^p$ è il vettore di *precodifica* (§ 21.6.1) a norma unitaria relativo al terminale k, ovvero la $k-$esima colonna dell'omonima matrice $\boldsymbol{P}_p$ $(n_{BS} \times U)$ ottenuta a partire dalla conoscenza della matrice $(U \times n_{BS})$ di DL $\boldsymbol{H}_p^d = \left(\boldsymbol{H}_p^u\right)^T$, trasposta di quella di UL per la reciprocità del canale MIMO alla stessa frequenza p, e che determina la direttività *spaziale* (e non *geografica* come nel beamforming) della trasmissione del simbolo s_k^p destinato al terminale k, mentre $\boldsymbol{s}_p$ è il vettore di tutti i simboli trasmessi. Il terminale $j-$esimo riceve dunque (alla portante p)

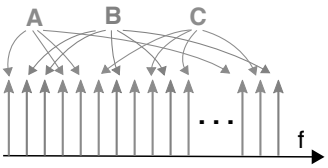
$$r_j^p = (\boldsymbol{h}_j^p)^T \boldsymbol{x}_p = (\boldsymbol{h}_j^p)^T \sum_{k=1}^{U} \boldsymbol{p}_k^p s_k^p + n = (\boldsymbol{h}_j^p)^T \boldsymbol{p}_j^p s_j^p + \sum_{\substack{k=1 \\ k \neq j}}^{U} (\boldsymbol{h}_j^p)^T \boldsymbol{p}_k^p s_k^p + n \tag{21.77}$$

in cui $\boldsymbol{h}_j^p$ è il vettore delle risposte in frequenza tra UE_k e le antenne della BS, che viene trasposto per la reciprocità del canale di DL; come per la (21.75) inoltre, il primo termine di (21.77) è quello desiderato, il secondo esprime l'interferenza dei simboli destinati agli altri utenti, ed il terzo è il rumore. Anche qui, una matrice di precodifica *zero forcing* ha l'effetto di annullare del tutto gli interferenti, fornendo $r_j = s_j + n = \tilde{s}_j$ da cui ottenere $\hat{s}_j = \arg\min_{s \in \mathcal{A}} (\hat{s}_j - s)^2$.

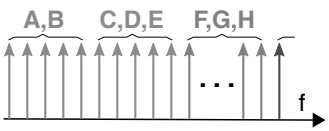
21.7.3.1 Ripartizione delle risorse

Oltre alla suddivisione dell'asse dei tempi per le due direzioni di trasmissione, nella trasmissione multiutente viene ripreso anche l'approccio dell'OFDMA (§ 16.8.12) di assegnare sotto-gruppi di portanti OFDM ad utenti differenti.

Una o più portanti per utente Qualora i dispositivi collegati alla BS non siano soggetti a spostamenti frequenti e non siano in numero troppo elevato si può procedere in modo *distribuito* come esemplificato in figura, in cui i singoli utenti sono rappresentati da colori diversi, utilizzando un numero di portanti possibilmente distanziate da un intervallo di frequenza maggiore della banda di coerenza in modo da poter sfruttare la relativa diversità in frequenza offerta dalla trasmissione OFDM. Il numero di portanti per utente dipende dalle sue esigenze trasmissive, mentre il relativo posizionamento in frequenza può dipendere dalla stima di canale effettuata dalla BS, in modo da usare frequenze per le quali si hanno buone condizioni di ricezione. Le portanti OFDM non utilizzate da un singolo utente vengono da queste poste a zero.



Uno o più utenti per blocco di coerenza Se invece si ha a che fare con uno scenario di utenza mobile, che determina la variabilità temporale delle condizioni di ricezione, le sottoportanti OFDM da assegnare ad uno (o più[98]) utenti possono essere *contigue in frequenza* per la durata di un ciclo UL-DL; il numero di portanti e la durata del ciclo dipendono dai valori di banda e tempo di coerenza B_c e T_c (§ 20.4.7) del collegamento. Anche se questi possono differire tra utenti diversi, prendendo *il caso peggiore* di entrambi si può suddividere il piano tempo-frequenza in unità denominate *Coherence Block* (CB) che raggruppano portanti e simboli contigui in intervalli entro i quali il canale può essere considerato stazionario e privo di distorsione lineare, e riservare i CB per la trasmissione (alternata) tra la BS ed i singoli UE.



Cosa si intende per blocco di coerenza Tralasciando per il momento l'aspetto MIMO, la parte superiore di fig. 21.8 rappresenta come il guadagno (in dB) della risposta in frequenza tra una antenna della BS e quelle di una coppia di utenti in movimento

[98] Distinguibili grazie ai vettori di precoding $\boldsymbol{p}_k$ e di combinazione $\boldsymbol{g}_k$.

possa variare sia rispetto alla frequenza che al tempo, guadagno rappresentato da una superficie di diverso colore per i due utenti. La parte inferiore mostra invece come i diversi CB vengano assegnati ora all'uno ora all'altro utente, in base a chi dei due subisce una minore attenuazione. La presenza di più antenne presso la BS *appiattisce* le superfici mostrate per quanto riguarda l'UL grazie alla possibilità di combinare in modo coerente i segnali ricevuti (eq. (21.75)), mentre il precoding (eq. (21.76)) attuato dalla BS in trasmissione ha il solo scopo di combattere l'interferenza tra gli utenti assegnati allo stesso CB[99], e dunque non modifica l'attenuazione subita nel DL.

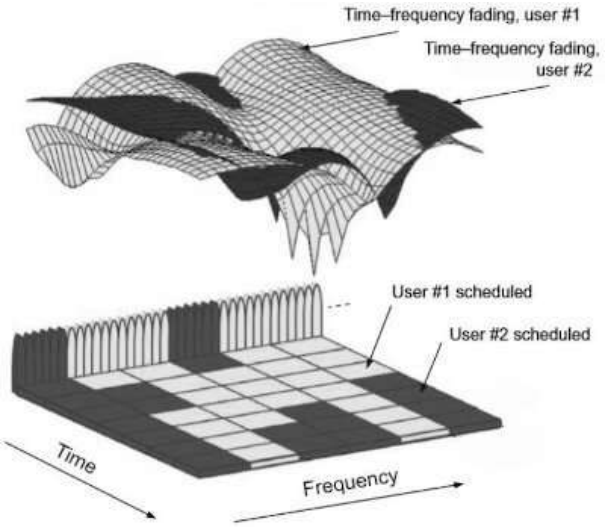


Figura 21.8: Assegnazione dei Coherence Block

La fig. 21.9 mostra[100] la suddivisione del piano tempo - frequenza in CB, evidenziando in giallo gli intervalli temporali dedicati alla trasmissione dell'uplink ed in rosa quelli per il downlink, mentre all'interno di un CB evidenzia la suddivisione in sottoportanti, ed i campioni complessi che si possono ottenere per ogni sottoportante, in numero totale (per ogni CB) τ_c proporzionale[101] al prodotto $B_c T_c$, per cui scriviamo

$$\tau_c = \alpha B_c T_c \quad \alpha < 1 \tag{21.78}$$

[99] Verso i quali si trasmette *in contemporanea* realizzando lo SDMA.

[100] Tratta dall'ottimo testo di E. BJÖRNSON, J. HOYDIS, L. SANGUINETTI, *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*, 2017, accessibile presso `https://massivemimobook.com`

[101] Considerando che un inviluppo complesso che occupa una banda *bilatera* B_c viene campionato a frequenza doppia di $B_c/2$, in T_c secondi si ottengono appunto $B_c T_c$ campioni. Una volta eliminato il prefisso ciclico (fig. 21.10) tale numero si riduce, ed il ricevitore OFDM ne effettua la FFT ottenendo altrettanti campioni complessi in frequenza.

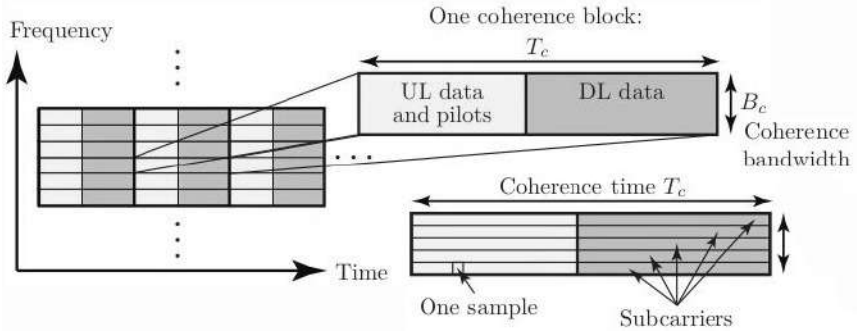


Figura 21.9: Suddivisione del piano tempo-frequenza in blocchi di coerenza e relativi intervalli temporali di trasmissione, sottoportanti, e campioni complessi

suddivisi tra tutte le portanti e tutti i simboli OFDM trasmessi nelle due direzioni nell'ambito di uno stesso CB.

Esempio In prima approssimazione valutiamo il tempo di coerenza T_c, legato alla velocità v del mobile, come quello necessario a spostarsi di $1/4$ di lunghezza d'onda λ, ovvero $T_c \simeq \lambda/4v$: quindi T_c è inversamente proporzionale alla frequenza portante $f_0 = c/\lambda$ *centrale* della trasmissione. D'altra parte la banda di coerenza può essere approssimata come $B_c \simeq 1/2\Delta\tau$ dove $\Delta\tau$ è la dispersione temporale tra la prima e l'ultima replica del multipath. Valutiamo ora due scenari per una frequenza $f_0 = 2$ GHz ovvero $\lambda = 15$ cm: il primo è un sistema *outdoor* con velocità $v = 37.5$ m/sec ossia 135 Km/h, ed un $\Delta\tau = 2.5$ μsec (cioè una differenza tra i percorsi di 750 metri). In tal caso si ottiene $T_c = 1$ msec e $B_c = 200$ KHz, e dunque $T_c B_c = 200$. Il secondo scenario è una trasmissione *indoor* caratterizzata da una mobilità $v = 0.75$ m/s (2.7 Km/h) ed un $\Delta\tau = 0.5$ μsec (differenza tra percorsi di 150 metri), a cui corrispondono $T_c = 50$ msec e $B_c = 1$ MHz, fornendo $T_c B_c = 50000$.

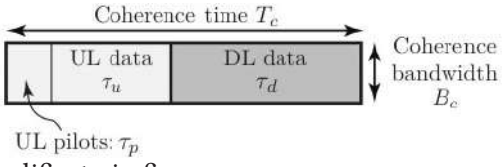
21.7.3.2 Stima di canale

Tutte le operazioni di ricezione delle trasmissioni sia MIMO che OFDM prevedono la conoscenza della risposta in frequenza del canale, che si suppone vari lentamente nel tempo rispetto alle costanti di tempo del sistema di trasmissione[102], e che fino ad ora abbiamo assunto nota senza chiederci come venirne a conoscenza: nel caso dell'OFDM la risposta in frequenza viene stimata in base alle condizioni di ricezione delle *portanti pilota* (§ 16.8.11, fig. 21.10) inserite in trasmissione.

Suddivisione del blocco di coerenza Lo schema di alternanza TDD che utilizza la stessa banda di frequenze per i due versi di trasmissione permette come già accennato di inserire le p. pilota nella sola direzione di UL ossia nella trasmissione da UE a BS, all'inizio di ogni blocco di coerenza come mostrato nella figura seguente, riservando per le stesse un numero di campioni complessi τ_p sul totale di τ_c (trasmessi e ricevuti, eq. (21.78)) in accordo alla ripartizione

$$\tau_c = \tau_p + \tau_u + \tau_d \qquad (21.79)$$

in cui i pedici u e d individuano rispettivamente i campioni riservati ai dati di uplink ed a quelli di downlink, come esemplificato in figura.



Ripartizione degli utenti La stima di canale dunque avviene per ciascun CB_h ($h = 1 \cdots n_{CB}$) in cui è suddiviso l'asse della frequenza ed a cui sono assegnati U_h utenti sul totale di U, ovvero $\sum_h U_h = U$, ad opera dalla sola BS, che ne usa il risultato sia per ricevere i dati di UL da parte di tutti gli UE_k, $k = 1 \cdots U_h$ assegnati allo stesso CB_h

[102] Per la modalità TDD su cui stiamo basando l'esposizione l'espressione *costante di tempo* individua la durata di un CB, mentre più in generale ci si riferisce non tanto al periodo di simbolo OFDM quanto ad un *periodo di trama*, che comprende un *preambolo* composto da qualche simbolo OFDM *accorciato* in cui trovano posto le portanti pilota, a cui fa seguito la sotto-trama dei simboli con i dati, per la cui ricezione si fa uso del risultato della stima di canale. Spesso alcune pilota sono presenti anche nei simboli dati, per favorire il mantenimento delle condizioni di sincronizzazione.

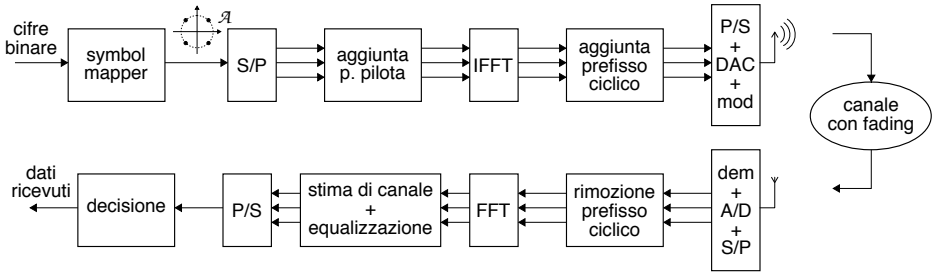


Figura 21.10: Schema simbolico di un *modem* OFDM per un collegamento SISO

(eq. (21.75)), sia per la trasmissione in DL verso gli stessi UE_k (eq. (21.77)). La relazione (21.79) pone un limite al numero massimo di campioni dedicati alle p. pilota, dato che tipicamente si adotta il vincolo $\tau_p < \tau_c/2$.

Sequenze pilota Per permettere alla BS di stimare il canale verso gli utenti, all'inizio del CB ogni UE_k trasmette una sequenza-pilota $\boldsymbol{\phi}_k$ di τ_p campioni complessi di *ampiezza unitaria*, ovvero tali che $\boldsymbol{\phi}_k^\dagger \boldsymbol{\phi}_k = \tau_p$, ed ortogonale[103] alle sequenze $\boldsymbol{\phi}_l$ trasmesse dagli altri utenti UE_l ($l \neq k$) assegnati allo stesso CB, ovvero $\boldsymbol{\phi}_k^\dagger \boldsymbol{\phi}_l = 0$. La sequenza $\boldsymbol{\phi}_k$ viene assegnata a ciascun utente k da parte dalla BS durante la fase di *richiesta di accesso* dell'utente alla rete, ed è scelta tra le colonne di una matrice $\boldsymbol{\Phi}$ di dimensione $\tau_p \times \tau_p$ detta *pilot-book* e che soddisfa la relazione $\boldsymbol{\Phi}^\dagger \boldsymbol{\Phi} = \tau_p \boldsymbol{I}_{\tau_p}$. Per $\boldsymbol{\Phi}$ può essere adottata una matrice di *Walsh-Hadamard*[104] che pone il vincolo su τ_p di essere una potenza di due, e che essendo composta da elementi ± 1 implica la trasmissione di simboli BPSK nella eq. (21.80); in alternativa per un valore τ_p qualunque la matrice $\boldsymbol{\Phi}$ può essere ottenuta come quella che definisce la DFT (eq. (4.12) a pag. 105) i cui elementi sono valori complessi equispaziati sul cerchio unitario e che dunque danno luogo a simboli L-PSK.

Ricezione delle pilota e stima di canale Gli elementi di $\boldsymbol{\phi}_k$ sono moltiplicati dai dispositivi UE_k per la costante $\sqrt{\mathcal{E}_k}$ in modo da presentare la stessa energia dei simboli che trasportano informazione, e sono quindi trasmessi come i simboli s_j in (21.74), con la differenza che ora la trasmissione è distribuita su tutte le portanti che fanno parte del CB e ripartita su più simboli OFDM. Per la definizione di CB la risposta in frequenza del canale è la stessa sia per le diverse portanti che per i diversi simboli, portando a descrivere il segnale ricevuto per il CB dalla BS in questa fase nella forma

$$\boldsymbol{Y}_h = \sum_{k=1}^{U_h} \sqrt{\mathcal{E}_k}\, \boldsymbol{h}_k^h\, \boldsymbol{\phi}_k^T + \boldsymbol{N}_h \tag{21.80}$$

in cui $\boldsymbol{h}_k^h$ è il vettore delle n_{BS} risposte in frequenza tra l'antenna dell'utente k assegnato al CB h e tutte le antenne della BS, $\boldsymbol{Y}_h$ è la matrice complessa di dimensione $n_{BS} \times \tau_p$

[103] Notiamo che per uno spazio descritto da una base vettoriale τ_p-dimensionale si possono individuare non più di τ_p diversi vettori *ortogonali* $\boldsymbol{\phi}_k$.

[104] Vedi ad es. `https://en.wikipedia.org/wiki/Walsh_matrix`

utilizzata per stimare il canale, e $\boldsymbol{N}_h$ è la matrice di eguali dimensioni dei campioni complessi ed indipendenti di rumore, a media nulla e varianza σ^2_{UL}.

Dato che la BS conosce la sequenza $\boldsymbol{\phi}_k$ utilizzata da ciascun utente k, può effettuare la stima del canale $\boldsymbol{h}^h_j$ relativo ad uno specifico utente j moltiplicando (correlando) la (21.80) per la sequenza coniugata $\boldsymbol{\phi}^*_j$ di quella assegnata all'utente j, ottenendo

$$\boldsymbol{y}^h_j = \boldsymbol{Y}_h\,\boldsymbol{\phi}^*_j = \sum_{k=1}^{U_h} \sqrt{\mathcal{E}_k}\,\boldsymbol{h}^h_k\,\boldsymbol{\phi}^T_k\boldsymbol{\phi}^*_j + \boldsymbol{N}_h\,\boldsymbol{\phi}^*_j = \sqrt{\mathcal{E}_k}\,\tau_p\,\boldsymbol{h}^h_j + \boldsymbol{N}_h\boldsymbol{\phi}^*_j$$

in quanto per l'ortogonalità delle sequenze pilota assegnate ad utenti differenti, il prodotto $\boldsymbol{\phi}^T_k\boldsymbol{\phi}^*_j$ è nullo per $k \neq j$ e pari a τ_p per $k = j$. Dato che il prodotto $\boldsymbol{N}_h\boldsymbol{\phi}^*_j$ è un vettore complesso gaussiano circolare a media nulla e covarianza $\sigma^2_{UL}\tau_p\boldsymbol{I}_{n_{BS}}$, la stima di massima verosimiglianza (e dunque di minima distanza) per $\boldsymbol{h}^h_j$ si ottiene come

$$\hat{\boldsymbol{h}}^h_j = \frac{\boldsymbol{y}^h_j}{\sqrt{\mathcal{E}_k}\,\tau_p} = \boldsymbol{h}^h_j + \frac{\boldsymbol{N}_h\boldsymbol{\phi}^*_j}{\sqrt{\mathcal{E}_k}\,\tau_p} \tag{21.81}$$

e dunque le componenti $\hat{h}_{ij}$ di $\hat{\boldsymbol{h}}^h_j$ sono v.a. gaussiane a media h_{ij} e varianza $\frac{\sigma^2_{UL}}{\sqrt{\mathcal{E}_k}\tau_p}$: in altre parole, l'utilizzo di una sequenza-pilota di lunghezza τ_p determina un *guadagno di processo* pari a τ_p ovvero un miglioramento dell'SNR di τ_p volte quello che si otterrebbe per una pilota singola.

Numero di utenti per blocco di coerenza Per garantire l'ortogonalità delle sequenza pilota assegnate ai diversi utenti che ricadono nello stesso CB_h il loro numero U_h non può superare il valore della lunghezza τ_p delle sequenze. Dato che aumentare τ_p determina una riduzione dei campioni utili alla trasmissione dati (eq. (21.79)) si manifesta un compromesso tra il numero di utenti e capacità trasmissiva.

21.7.3.3 Rete cellulare

Con questo termine si indica una suddivisione del territorio del tipo di fig. 21.11, in cui più BS sono contemporaneamente attive. In tale contesto per ogni CB le BS ricevono le trasmissioni, oltre che degli UE ad esse associati, anche degli UE nelle celle vicine ed assegnati al CB nella stessa regione di frequenza, così come i diversi UE ricevono il segnale delle altre BS oltre a quello della propria. Ciò comporta che nelle espressioni (21.74) e (21.77) dei segnali ricevuti rispettivamente dalla BS e dagli UE di una singola cella compaiano termini di interferenza che originano dalle celle limitrofe, la cui entità può essere ridotta mediante operazioni di combinazione lineare e precoding estendendo la procedura di stima di canale da parte di ciascuna BS anche alle risposte in frequenza relative agli UE associati ad altre BS ossia *residenti* in altre celle.

E' ovvio che per quanto grande possa essere la dimensione del pilot book $\boldsymbol{\Phi}$, è impossibile assegnare sequenze differenti a tutti gli utenti che possono interferire tra loro, sicché la stima di canale svolta da parte delle BS è gravata da un errore detto *pilot contamination*, e che consiste in un termine di errore nella stima (21.81) che origina da

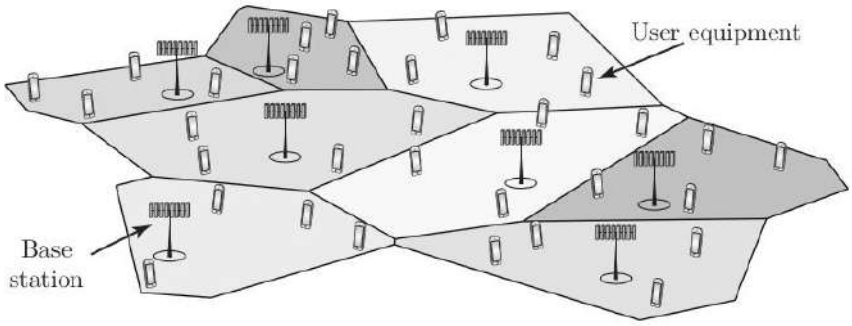


Figura 21.11: Suddivisione del territorio in celle in cui gli utenti (UE) si *associano* alla base station (BS) il cui segnale è più forte

tutti gli UE nelle altre celle che condividono la stessa sequenza pilota $\boldsymbol{\phi}$ dello UE di cui si sta stimando la risposta in frequenza.

Senza avere la minima pretesa di approfondire a sufficienza l'argomento, al cui riguardo si rimanda il lettore al testo indicato alla nota 100, ci si limita a citare che in questo caso la problematica viene affrontata tenendo conto della diversa correlazione *spaziale* $\mathcal{R}_h^k$ che lega statisticamente le componenti del vettore delle risposte in frequenza $\boldsymbol{h}_k$ di ciascun generico utente k nei confronti delle antenne di una BS.

A differenza di un collegamento MIMO punto-punto, in cui tra le diverse antenne da ambo i lati si manifestano vettori $\boldsymbol{h}$ descritti da matrici di correlazione $\mathcal{R}_h$ sostanzialmente simili, nel caso multiutente ogni utente k è in genere distante da tutti gli altri per un buon multiplo di lunghezze d'onda λ, e dunque manifesta una diversa matrice $\mathcal{R}_h^k$, ovvero ogni vettore $\boldsymbol{h}_k$ risulta essere una realizzazione di una diversa d.d.p. gaussiana multivariata a media nulla e covarianza $\mathcal{R}_h^k$. Ciò consente di impostare la stima di canale secondo un approccio MMSE, capace di sfruttare l'informazione legata alla conoscenza delle matrici $\mathcal{R}_h$ dei diversi utenti, e in grado di subire la *pilot contamination* solo nel caso di utenti che, oltre alla stessa sequenza $\boldsymbol{\phi}$, hanno in comune anche il medesimo *orientamento* rispetto alla BS nei cui confronti si sta stimando il vettore $\boldsymbol{h}$. Per una esemplificazione di quanto illustrato, si veda la fig. 21.12.

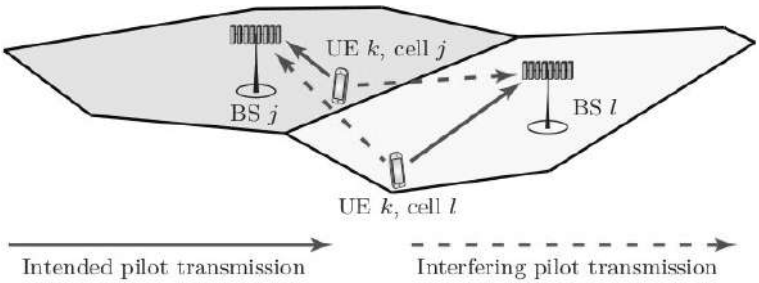


Figura 21.12: Mentre lo UE nella cella l causa una forte interferenza nella ricezione presso la BS$_j$ della sequenza pilota $\boldsymbol{\phi}$ che ha in comune con lo UE nella cella j, lo stesso non avviene tra la cella j e quella l in virtù della diversa matrice di correlazione $\mathcal{R}_h$, che a sua volta dipende dagli elementi riflettenti prossimi alla BS e che determinano i cammini multipli, simili per i due UE per la BS j, e differenti per quella l

Per quanto riguarda le matrici $\mathcal{R}_h$, esse vengono stimate a partire dalle stime dei vettori $\boldsymbol{h}$ collezionate nel tempo, e mantenute aggiornate nel caso di mobilità adattandone la stima a partire dai dati via via raccolti.

21.8 Single frequency network - SFN

Il caso della diffusione televisiva digitale terrestre DVB-T[105] rientra nella casistica di un sistema MISO in quanto prevede la ricezione da parte dell'apparecchio televisivo di segnali trasmessi da più antenne dislocate sul territorio, ognuna delle quali effettua una trasmissione OFDM

1. centrata sulla stessa frequenza portante, i cui simboli
2. contengono esattamente gli stessi bit e
3. sono trasmessi esattamente allo stesso istante.

Notiamo come i requisiti 1) e 3) siano comuni a quelli necessari a realizzare il collegamento di uplink di un sistema MU-MIMO (§ 21.7.3), conseguiti in tale contesto per mezzo dei protocolli di controllo peculiari dei relativi sistemi di accesso multiplo, che non abbiamo approfondito[106]. In questa sezione illustriamo invece come sia affrontato il problema per i sistemi di diffusione DVB-T, non prima però di aver discusso le differenze rispetto alla tecnica *multiple frequency network* (MFN) adottata dalla diffusione televisiva *analogica*.

Rete a frequenza multipla o MFN La televisione analogica (§ 25.1) si basa sulla assegnazione di una diversa portante (o canale) a ciascuna emittente (o broadcaster), su cui trasmettere il proprio segnale con modulazione AM-BLR, tipicamente ricevuto dall'apparecchio TV mediante una antenna direttiva posta sul tetto ed orientata verso il ripetitore più vicino, dunque in condizione di visibilità. Nel caso di antenna indoor si verifica(va) di frequente la comparsa di immagini sdoppiate (dette *ghost* o fantasma[107]) dovute all'insorgenza di cammini multipli, con conseguente ricezione di copie ritardate dello stesso segnale. La presenza di più ripetitori, necessari a coprire differenti regioni del territorio per le emittenti nazionali, richiede(va) quindi l'adozione di frequenze portanti differenti per ciascuna regione, pena la comparsa di *ghost* legati alla ricezione del segnale trasmesso (per la stessa emittente) dai diversi ripetitori, ognuno posto ad una differente distanza dalla TV, qualora utilizzanti la stessa portante.

Fantasmi televisivi ma non radiofonici Cogliamo l'occasione per investigare sul meccanismo di formazione dei *ghost* televisivi in una SFN: sono causati dalla *sincronizzazione di quadro* (§ 25.1.2) che l'apparecchio analogico consegue con riferimento al segnale *più forte* ricevuto. In virtù della linearità della AM, la demodulazione della somma delle

[105] Vedi ad es. `https://en.wikipedia.org/wiki/DVB-T`

[106] In effetti nella trasmissione di *uplink* di un sistema MU-MIMO il requisito 3) è ancora più stringente, dato che in tal caso è richiesto che i simboli trasmessi dagli UE *arrivino* alla BS allo stesso istante, cosa resa possibile dalla conoscenza della distanza tra ciascuno di essi e la BS; ciò non è possibile in un sistema *broadcast*, data la presenza di *molteplici* ricevitori ad antenna singola, al posto di una singola BS.

[107] Vedi ad es. `https://en.wikipedia.org/wiki/Ghosting_(television)`

repliche produce un segnale di immagine (luminanza e crominanza) che è la somma delle relative immagini, sfalsate dei corrispondenti ritardi nel tempo di arrivo. Nel caso della radio AM-BLD-PI (§ 12.1.1.2) manca invece la necessità di sincronizzazione, la frequenza portante ridotta rende trascurabile la differenza di fase tra le repliche ritardate, e la minore occupazione spettrale rende trascurabile l'entità della distorsione lineare. Infine nella radio FM il PLL del ricevitore (§ 12.3.2.1) si aggancia alla portante del segnale più forte.

Vantaggi di una SFN Il vantaggio più evidente rispetto ad una MFN è una migliore efficienza spettrale, in quanto lo stesso programma diffuso via MFN in regioni limitrofe necessita di altrettanti canali radio, contro l'unico canale necessario con una architettura SFN. A questa considerazione si può aggiungere che mentre nel caso di TV analogica basata su MFN ogni emittente necessita di un proprio canale, nel DVB i canali radio sono usati per trasmettere interi *transport stream* MPEG o TS (§ 10.3.2.1), ognuno dei quali ospita un multiplex di *elementary stream* (§ 10.3.2) provenienti da più emittenti, che così condividono la medesima SFN e dunque un unico canale radio.

Un secondo vantaggio è indicato come *network gain* e consiste nel vantaggio tipico della trasmissione MISO di poter sommare in modo coerente più repliche dello stesso messaggio, purché giunte nei limiti del tempo di guardia per come descritto appresso. A questo corrisponde una migliore qualità del segnale in termini di SNR, ovvero una minore criticità nei confronti della figura di rumore del ricevitore. I dispositivi che ne traggono maggior vantaggio sono i ricevitori mobili o nomadici, che non utilizzando antenne direttive possono facilmente trovarsi nelle condizioni di ricevere segnali da più fonti.

Tempo di guardia per una SFN Come anticipato, la diffusione in una rete SFN si avvale di una trasmissione OFDM *sincronizzata* da parte dei diversi trasmettitori Tx_i, che grazie all'inserimento del *prefisso ciclico* di durata T_g (o *tempo di guardia*, pag. 520) rende il ricevitore *esente* da interferenza intersimbolica qualora T_g sia maggiore del massimo ritardo legato al multipath (che indichiamo con $\delta\tau$), sommato (vedi fig. 21.13) alla massima differenza dei ritardi $\Delta\tau$ con cui viene ricevuto il segnale trasmesso dalle antenne Tx_i.

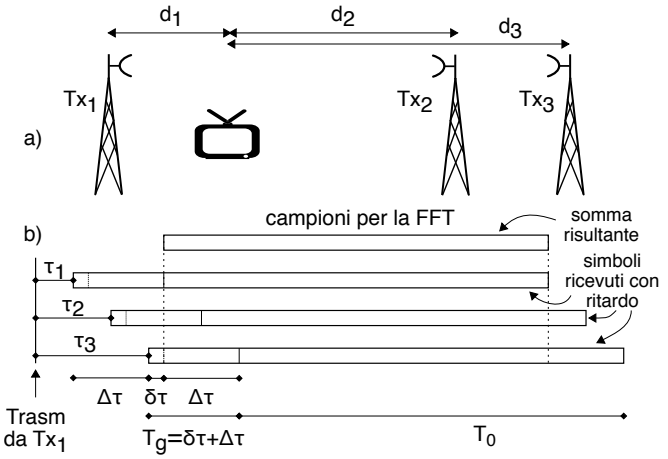


Figura 21.13: Ricezione SFN: a) - tre trasmettitori a diversa distanza dal ricevitore b) - tempo di guardia T_g somma di due ritardi *massimi*: $\delta\tau$ del mutipath, e $\Delta\tau$ tra prima e ultima replica

Dalla formula $\tau = {}^{d}/_{c}$ che fornisce il ritardo τ con cui si riceve un segnale ra-

dio trasmesso da una distanza d e che si propaga alla velocità $c = 3 \cdot 10^8$ metri/sec si ottiene che per $d = 100$ metri τ ha un valore di 330 nsec, mentre per 100 Km si ottiene $\tau = 330$ μsec: la scelta del valore di T_g impone quindi quello del *massimo* distanziamento tra antenne che prendono parte allo stesso sistema SFN.

Valori di $T_g \simeq 200$ μsec sono adeguati per le reti a diffusione nazionale con ripetitori distanziati di una cinquantina di Km, mentre reti di ambito regionale o locale possono adottare valori inferiori.

Legame tra T_g ed efficienza spettrale Come illustrato in fig. 21.13, dalla somma delle repliche ricevute si individua una finestra di durata T_0 secondi (il periodo principale) i cui campioni di inviluppo complesso vengono inviati (vedi fig. 21.10) al computo di una FFT in grado di ri-ottenere i punti di costellazione con cui sono modulate le sottoportanti. A questo riguardo il sistema DVB-T prevede di poter operare nelle modalità dette 2K ed 8K[108], corrispondenti ad una FFT di dimensione N pari a 2048 e 8192 (vedi tab. 21.2) di cui solamente un sottoinsieme di portanti $\tilde{N}$ sono *attive* ovvero effettivamente modulate (§ 16.8.2), in parte con dati di servizio[109], ed in larga misura con i dati relativi alla trasmissione[110]. Notiamo inoltre che i parametri di tab. 21.2 si riferiscono ad un distanziamento tra canali SFN (ovvero TS differenti) pari ad 8 MHz, mantenendo quindi una separazione (o banda di guardia) di 580 KHz tra due canali SFN.

Rimarchiamo ora il fatto che solamente la porzione di simbolo OFDM relativa al periodo principale (di durata T_0) trasporta informazione, e dunque una volta determinato il valore T_g necessario a supportare le caratteristiche territoriali della SFN che si intende sviluppare è opportuno far si che la somma $T_g + T_0$ sia la massima possibile, in quanto minore è il rapporto T_g/T_0 e maggiore sarà l'efficienza del collegamento[111], e dunque la velocità binaria effettivamente trasferita.

Sempre con riferimento alla tab. 21.2 osserviamo che, fissata la durata di T_0 per le due modalità 8K e 2K, il numero delle possibili scelte per il rapporto T_g/T_0 è ristretto a 4 alternative, ovvero 1/4, 1/8, 1/16 ed 1/32; in particolare, i valori più elevati di T_g (associati a ripetitori più distanti) sono possibili solo in modalità 8K. Per lo stesso

[108] Le cui sigle non hanno nulla a che vedere con la qualità dell'immagine!

[109] Come *dati di servizio* si intendono le portanti pilota di sincronizzazione in frequenza e nel tempo, di ausilio alla stima di canale, e che trasportano dati di segnalazione a riguardo (tra l'altro) del tipo di modulazione (M-QAM con $M = 4, 16$ o 64) e di codifica di canale, parametri questi ultimi che determinano la velocità binaria conseguita dal flusso dati trasmesso, variabile in un intervallo da 5 a 31 Mbps.

[110] Che consistono nel *transport stream* MPEG che multiplexa le emittenti che prendono parte alla SFN, a cui sono stati applicati stadi di codifica di canale e di scrambling, vedi ad es. Lo standard DVB-T, Centro Ricerche RAI, da cui è anche tratta la tab. 21.2

[111] Indicando con $\tilde{N}_d$ il numero di portanti che trasportano informazione, con R_c il tasso di codifica FEC, e con M il numero di bit/portante della costellazione adottata, se il tempo di guardia fosse nullo si otterrebbe una velocità $f_b^{max} = \tilde{N}_d \cdot R_c \cdot M/T_0$. Dato però che il periodo di simbolo è pari a $T_s = T_g + T_0$, la velocità si riduce a $f_b = f_b^{max} \frac{T_0}{T_g+T_0}$ e dunque l'efficienza è pari a $\eta = \frac{T_0}{T_g+T_0} = \frac{1}{1+T_g/T_0}$. Con i valori di T_g/T_0 in fig. 21.2 si ottiene $\eta = 0.8, 0.88, 0.94$ e 0.97.

	modo 8k				modo 2k			
dimensione FFT N	8192				2048			
portanti attive $\tilde{N}$ (di cui dati)	6817 (6048)				1705 (1512)			
periodo principale T_0	896 μsec				224 μsec			
intervallo tra portanti $\Delta = 1/T_0$	1116 Hz				4464 Hz			
banda occupata $B = \tilde{N}\Delta$	7.61 MHz							
rapporto T_g/T_0	1/4	1/8	1/16	1/32	1/4	1/8	1/16	1/32
durata T_g (μsec)	224	112	56	28	56	28	14	7

Tabella 21.2: Parametri di una trasmissione DVB-T per una canalizzazione di 8 MHz

motivo la modalità 2K, che impone l'uso di valori di T_g inferiori, può essere adottata solamente in regioni in cui la densità delle antenne è particolarmente elevata.

Rete di distribuzione e sincronizzazione in frequenza Descriviamo per ultima la soluzione che si è adottata per soddisfare le condizioni elencate a pag. 754 e che permettono la realizzazione di una SFN, ovvero come fare in modo che le diverse antenne trasmettano allo *stesso istante e frequenza* simboli OFDM che contengono *gli stessi bit*. Questi ultimi fanno parte di un *flusso di trasporto* MPEG o TS (§ 10.3.2.1)[112], che deve poi essere ulteriormente elaborato aggiungendo una protezione FEC mediante due stadi di codifica di canale associati ad interleaving (§ 17.4.2.6). La trasmissione OFDM del risultato viene anche indicata come trasmissione COFDM (§ 16.8.10).

La soluzione cercata è ottenuta grazie all'adozione della architettura di fig. 21.14, che *centralizza* la creazione del flusso di bit da trasmettere in un unico luogo e lo inoltra ai ripetitori che realizzano la SFN per mezzo di una differente rete *di distribuzione*, tipicamente in fibra ottica, oppure realizzata con ponti radio, o via satellite.

[112]Costituito da una serie di pacchetti di 188 byte il cui contenuto è descritto dal campo PID presente nell'header, che può anche specificare un pacchetto *nullo* inserito per adattare la velocità variabile dei contenuti multimediali con quella fissa offerta dal canale a disposizione.

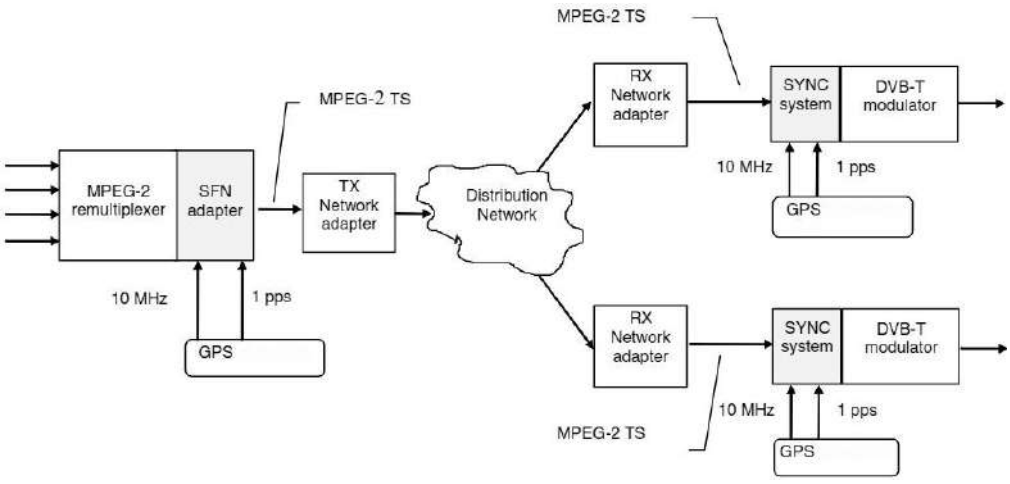


Figura 21.14: Distribuzione ai ripetitori del segnale DVB-T mediante adattatori SFN

La sincronizzazione *in frequenza* tra le diverse antenne è conseguita grazie al comune utilizzo da parte delle antenne di una medesima *sorgente esterna* di sincronismo, come quella offerta dal segnale GPS[113] che provvede a fornire due riferimenti temporali, uno (detto PPS o *pulse per second*) che scandisce intervalli di 1 secondo, e l'altro a 10 MHz che suddivide lo stesso secondo in 10^7 parti ovvero individua intervalli di 100 nanosecondi. Lo stadio di modulazione finale di ciascuna antenna può pertanto lavorare ad una frequenza ottenuta per sintesi digitale (§ 12.4.3) a partire dal comune riferimento a 10 Mhz.

Pacchetto MIP e *megaframe* Lo stesso segnale GPS, ricevuto oltre che dalle antenne anche dal nodo di controllo, permette a quest'ultimo di orchestrare la *sincronizzazione temporale* dei simboli OFDM trasmessi (e del loro contenuto) grazie all'inserimento nel flusso MPEG di uno speciale pacchetto TS (con PID 0x15) denominato *mega frame initialization packet* o MIP, contenente oltre ad un *timbro temporale* ottenuto con riferimento al PPS e con risoluzione 100 nsec, anche le direttive a riguardo dei parametri per lo stadio FEC e per il tipo di costellazione M-QAM da applicare alle portanti.

L'inserimento del pacchetto MIP avviene in un punto qualunque di una struttura sintattica denominata *mega frame* e che viene definita dallo SFN *adapter* del multiplexer, delimitando una sequenza di pacchetti TS che sono a loro volta contenuti in 8 *frame*[114] DVB-T in modalità 8K (oppure in 32 frames in modalità 2K). La durata temporale di un megaframe dipende solamente dal rapporto T_g/T_0 e dalla banda del canale, ma non da T_0 né dalla scelta della costellazione o dal tipo di FEC, variando tra circa 0.5 e 0.6 secondi per un canale di 8 MHz.

Sincronizzazione temporale Dato che la rete di distribuzione presenta ritardi di attraversamento differenti per raggiungere le diverse antenne, la sincronizzazione temporale è ottenuta inserendo in un punto qualunque del megaframe m il pacchetto MIP contenente (tra le altre) le informazioni

- un *puntatore* che individua l'inizio del megraframe seguente $m + 1$, espresso come il numero di pacchetti TS che separano il MIP dal megaframe successivo;
- un *synchronization time stamp* (STS) che misura il tempo che intercorre[115] tra l'impulso PPS che *precede* il megaframe $m + 1$, e la trasmissione del primo bit del primo pacchetto di tale megaframe, evento indicato come **A** in figura;
- il valore *maximum delay*[116] che deve essere maggiore del massimo ritardo di attraversamento della rete di distribuzione, per un qualsiasi percorso.

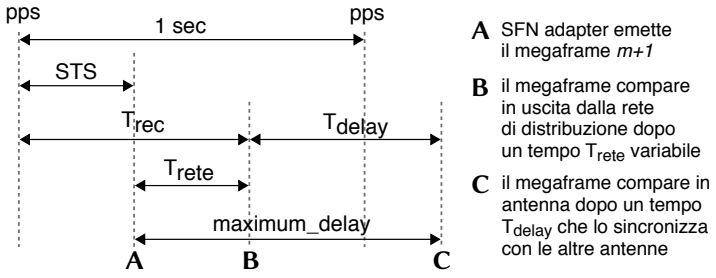
[113] Vedi `https://it.wikipedia.org/wiki/Sistema_di_posizionamento_globale`

[114] La definizione esatta di *frame* DVB-T trascende lo scopo di questa sezione, e verrà (forse) affrontata in una prossima edizione. Il lettore interessato può affrontare la lettura di ETSI EN 300 744 V1.6 - Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television

[115] Espresso come numero di colpi di clock da 10 MHz, ovvero come multiplo di 0.1 μsec.

[116] sempre misurato in multipli di 0.1 μsec

L'evento indicato come **B** alla figura seguente rappresenta l'arrivo del megaframe presso una delle antenne che prendono parte alla SFN, ed il valore T_{rec} individua il ritardo con cui si verifica **B** rispetto al PPS che precede l'inizio del megaframe. Il compito del *SYNC system* (mostrato in fig. 21.14) di ciascuna antenna della SFN è dunque quello di



- localizzare l'inizio del megaframe $m+1$ a partire dall'informazione fornita dal campo *puntatore* presente nel MIP contenuto nel megaframe m;
- calcolare il valore

$$T_{delay} = (STS + maximum\,delay - T_{rec}) \text{ modulo } 10^7$$

di cui ritardare la trasmissione del megaframe $m+1$ (evento **C**) rispetto all'istante **B** in cui lo si è ricevuto per mezzo della rete di distribuzione.

In tal modo l'inizio della trasmissione del megaframe sarà lo stesso per tutte le antenne che partecipano alla SFN. Come anticipato, il MIP convoglia anche altre informazioni, come il tipo di codifica FEC e la costellazione da adottare: questi (eventualmente nuovi) parametri vengono inseriti nelle portanti pilota dei simboli OFDM utilizzati per trasmettere il contenuto del megaframe $m+1$, e resi operativi a partire dal megaframe $m + 2^{117}$, in modo da permettere ai dispositivi riceventi di adattare a loro volta lo stadio di demodulazione alla nuova configurazione del modem OFDM.

Il sistema di diffusione DVB-T presenza molteplici aspetti qui solamente accennati o del tutto trascurati; il lettore interessato può approfondirli presso ETSI TR 101 190 V1.3 - Digital Video Broadcasting (DVB); Implementation guidelines for DVB terrestrial services; Transmission aspects.

21.9 Appendice

21.9.1 Entropia di variabile gaussiana complessa multivariata

Estendiamo il risultato trovato al § 9.3.2 per l'entropia differenziale $h_G(X) = \frac{1}{2}\log_2\left(2\pi e\sigma^2\right)$ di una v.a. gaussiana reale X con varianza σ^2 a due altri casi, quello in cui $\boldsymbol{X}$ sia una v.a. *vettoriale* a componenti gaussiane (§ 6.5) con matrice di covarianza $\boldsymbol{\Sigma}$, e quello in cui oltre a ciò, le v.a. marginali assumano valori *complessi*.

Caso reale La v.a. $\boldsymbol{X}$ (vettore colonna ad n elementi) è descritta dalla d.d.p. $n-$dimensionale

$$p_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right\} \tag{21.82}$$

[117] Vedi ETSI TS 101 191 V1.4 - Digital Video Broadcasting (DVB); DVB mega-frame for Single Frequency Network (SFN) synchronization

alla quale applichiamo la definizione di entropia $H_X = E\left\{-\log_2 p_X(\boldsymbol{x})\right\}$, ponendo $\boldsymbol{m} = 0$ in quanto l'entropia differenziale è indipendente dalle traslazioni (pag. 267), e tenendo conto che $\log_2 p(x) = \frac{\ln p(x)}{\ln 2}$, in modo da ottenere

$$h_{GM} = \frac{1}{\ln 2} E\left\{\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln(\det(\Sigma)) + \frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right\} = \tag{21.83}$$
$$= \frac{1}{\ln 2}\left(\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln(\det(\Sigma)) + \frac{1}{2}E\left\{\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right\}\right)$$

All'ultimo termine si applica ora *il trucco della traccia*[118] che permette due sostituzioni: la prima è[119] $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} = \mathrm{tr}\left(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right)$, e la seconda che si può scrivere[120] $\mathrm{tr}(UVW) = \mathrm{tr}(WUV)$, dunque

$$E\left\{\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right\} = E\left\{\mathrm{tr}\left(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right)\right\} = \mathrm{tr}\left(E\left\{\boldsymbol{x}\boldsymbol{x}^T\right\}\Sigma^{-1}\right) = \mathrm{tr}\left(\Sigma\Sigma^{-1}\right) = n$$

dato che $E\left\{\boldsymbol{x}\boldsymbol{x}^T\right\}$ è una matrice con elementi $E\left\{x_i x_j\right\} = \sigma_{ij}$ ossia esattamente gli elementi della covarianza Σ, il cui prodotto per Σ^{-1} è dunque la matrice identità di dimensione n. Pertanto risulta

$$h_{GM} = \frac{1}{\ln 2}\left(\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln(\det(\Sigma)) + \frac{n}{2}\right) \tag{21.84}$$

e dato che $\ln(\mathrm{e}) = 1$ scriviamo

$$h_{GM} = \frac{1}{\ln 2}\left(\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln(\det(\Sigma)) + \frac{n}{2}\ln\mathrm{e}\right) =$$
$$= \frac{1}{2}\frac{1}{\ln 2}\ln\left((2\pi\mathrm{e})^n\det(\Sigma)\right) = \frac{1}{2}\log_2\left((2\pi\mathrm{e})^n\det(\Sigma)\right)$$

che può anche essere scritta come $h_{GM} = \frac{1}{2}\log_2(\det(2\pi\mathrm{e}\,\Sigma))$.

Caso complesso In questo caso $\boldsymbol{Z}$ è un vettore ad n componenti complesse $z_k = x_k + jy_k$ note come v.a. *gaussiana centrata a simmetria circolare*[121], le cui parti reale ed immaginaria x_k, y_k sono v.a. gaussiane indipendenti (in quanto $E\left\{x_k y_k^*\right\} = 0$) a media nulla e varianza $\frac{1}{2}\sigma_k^2$ (in modo che $E\left\{z_k z_k^*\right\} = E\left\{x_k x_k^*\right\} + E\left\{y_k y_k^*\right\} = \sigma_k^2$), con matrice di covarianza $\sum$ *Hermitiana*[122] ad elementi $\sigma_{i,j} = E\left\{z_i z_j^*\right\}$. Per calcolare l'entropia h_{GC} di questo tipo di sorgente informativa[123] si può procedere in (almeno) due modi: il primo prende in considerazione la d.d.p. di $\boldsymbol{Z}$ che ora si scrive

$$p_X(\boldsymbol{z}) = \frac{1}{\pi^n\det(\Sigma)}\exp\left\{-\boldsymbol{z}^\dagger\Sigma^{-1}\boldsymbol{z}\right\} \tag{21.85}$$

[118] Vedi ad es. `https://sgfin.github.io/`

[119] L'operatore matriciale $tr(A)$ restituisce la somma degli elementi sulla diagonale della matrice *quadrata* A. Dal momento che $\boldsymbol{x}\Sigma\boldsymbol{x}^T$ è un numero, ovvero una matrice 1×1, la sua traccia è pari al numero stesso.

[120] Nota come proprietà *ciclica*, vedi `https://en.wikipedia.org/wiki/Trace_(linear_algebra)`

[121] Vedi ad es. `https://en.wikipedia.org/wiki/Complex_normal_distribution`

[122] Ossia $\sum$ coincide con la propria trasposta coniugata (o aggiunta); dunque se la matrice è ad elementi reali, è simmetrica. Vedi ad es. `https://it.wikipedia.org/wiki/Matrice_hermitiana`

[123] O molto più banalmente, il campionamento dell'inviluppo complesso del rumore osservato su n canali equivalenti di bassa frequenza.

in cui (a parte il resto) al posto del trasposto ^{T} che compare nella (21.82) ora si usa il trasposto coniugato † o *Hermitiano*. A partire dalla (21.85), applicando sviluppi analoghi alle (21.83) - (21.84) si può ottenere

$$h_{GC}(\boldsymbol{Z}) = \log\left(\det\left(\pi \mathrm{e} \boldsymbol{\Sigma}\right)\right)$$

Un secondo modo di procedere prende in considerazione il fatto che l'entropia $h_{GC}(\boldsymbol{Z})$ è pari a quella $h(\boldsymbol{X}, \boldsymbol{Y})$ di un vettore $[\boldsymbol{X}, \boldsymbol{Y}] = [\Re\{\boldsymbol{Z}\}, \Im\{\boldsymbol{Z}\}]$ ottenuto concatenando i vettori delle parti reale ed immaginaria di $\boldsymbol{Z}$. Dato che gli elementi di $\boldsymbol{X}$ ed $\boldsymbol{Y}$ sono statisticamente indipendenti a coppie perché incorrelati, si ha $h(\boldsymbol{X}, \boldsymbol{Y}) = h(\boldsymbol{X}) + h(\boldsymbol{Y})$. Dato che inoltre sia $\boldsymbol{X}$ che $\boldsymbol{Y}$ sono vettori gaussiani a media nulla e covarianza $\frac{1}{2}\boldsymbol{\Sigma}_z$ l'entropia differenziale di ciascuno di essi è pari a $\frac{1}{2}\log_2\left(\det(2\pi \mathrm{e}\, \frac{1}{2}\boldsymbol{\Sigma}_z)\right)$. Mettendo tutto assieme si ottiene

$$\begin{aligned} h_{GC}(\boldsymbol{Z}) &= h(\boldsymbol{X}, \boldsymbol{Y}) = h(\boldsymbol{X}) + h(\boldsymbol{Y}) = \\ &= \tfrac{1}{2}\log_2\left(\det(2\pi \mathrm{e}\, \tfrac{1}{2}\boldsymbol{\Sigma}_z)\right) + \tfrac{1}{2}\log_2\left(\det(2\pi \mathrm{e}\, \tfrac{1}{2}\boldsymbol{\Sigma}_z)\right) = \\ &= \log_2\left(\det(\pi \mathrm{e}\, \boldsymbol{\Sigma}_z)\right) \end{aligned}$$

Parte IV

Sistemi di Telecomunicazione

Prefazione alla quarta parte

Tutto ciò che è *avanzato* dalla riorganizzazione degli argomenti realizzata a partire dall'edizione 1.5 è stato raggruppato in questa quarta parte del testo, dove nelle prossime edizioni *potrebbero* essere sviluppati argomenti come il WiFi, la telefonia mobile, il digitale terrestre, i sistemi di accesso e di comunicazione personale.

Il capitolo 22 affronta il tema della teoria del traffico e dei sistemi di servizio, che costituiscono il modello probabilistico per le reti a commutazione *di circuito*, intese come un sistema di servizio orientato *alla perdita*, e per quelle a commutazione *di pacchetto* che invece, essendo basate su code di ingresso-uscita, realizzano un sistema di servizio orientato *al ritardo*. Dopo aver caratterizzato le tipologie di pacchetto dati, viene poi sviluppata una *tassonomia* che permette di inquadrare le diverse architetture di rete in un contesto unitario, che ne mette in luce le differenti peculiarità.

Forti delle nuove basi teoriche acquisite, il capitolo 23 affronta il funzionamento della rete Internet, articolandone la descrizione nei termini degli strati di cui si compone, corrispondenti a diversi livelli di indirizzamento, da quello mnemonico del DNS, a quelli di trasporto del TCP/UDP e di rete dell'IP, giù fino agli indirizzi fisici o *Ethernet*. Quindi, sono descritte le modalità di funzionamento di una rete a pacchetto del tutto differente, l'ATM, che anche se non più molto diffusa se non nella sezione di accesso, rappresenta un classico esempio di rete a *circuito virtuale*.

Le reti a *commutazione di circuito*, dove le risorse sono assegnate agli utenti in modalità permanente e garantita, sono discusse al capitolo 24. Dopo aver richiamato i concetti fin qui esposti, al fine di individuare gli elementi costitutivi della rete telefonica, viene ripercorsa la sua evoluzione storica in una sorta di cammino a tappe forzate, coinvolgendo la descrizione della trama PCM, della relativa gerarchia plesiocrona PDH, e di quella sincrona o SDH. Sono quindi forniti dei cenni relativi alla topologia della rete italiana (sicuramente obsoleti, ma solo gli operatori ne conoscono il reale sviluppo attuale), a cui seguono alcune fondamentali definizioni relative ai principi di commutazione.

Il capitolo 25 ospita una breve descrizione di tre sistemi di trasmissione diffusiva che nelle edizioni precedenti erano adottati come esempi di applicazione delle teorie esposte,

e precisamente: la televisione analogica, le trasmissioni radio FM, e la televisione satellitare. In qualche futura edizione, in questa sede verrà sviluppata la trattazione del *digitale terrestre* e del DAB.

E le reti di telefonia cellulare? Sebbene diverse tecnologie su cui si basano sono state descritte ai capitoli precedenti, un capitolo che metta tutte queste informazioni assieme ancora non è stato sviluppato; ciononostante, si è voluto tenere *occupato il posto* con il capitolo 26, che semplicemente ospita un argomento sul GSM che precedentemente compariva altrove come nota.

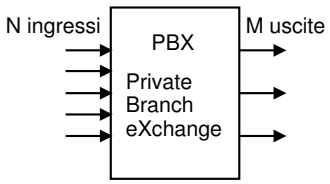
Capitolo 22

Sistema di servizio, teoria del traffico, e delle reti

Qui trovano spazio argomenti di *ingegneria delle reti*, a carattere prevalentemente teorico, mentre le realizzazioni pratiche sono sviluppate nei capitoli che seguono. Tali aspetti si basano sui concetti di probabilità affrontati al capitolo 6, e dopo aver caratterizzato in tal senso il *traffico* informativo, sono fornite 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. Nell'ultima sezione poi, dopo aver discusso brevemente dei compromessi da affrontare nella progettazione di un sistema di comunicazione a pacchetto, viene sviluppata una tassonomia che permette di inquadrare le diverse architetture di rete in un contesto unitario, che ne mette in luce le differenti peculiarità.

22.1 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

[1] Il 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, vedi pag. 643. Nel caso presente, una buona qualità corrisponde a una bassa probabilità di occupato.

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

$$p^k q^{N-k}$$

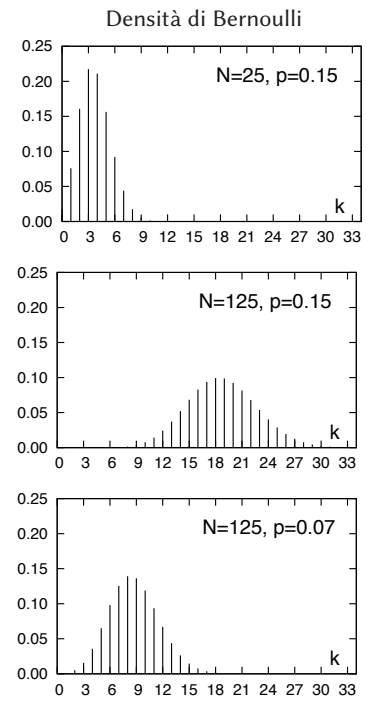
in cui $q = 1 - p$. Dato che il numero di differenti modi di scegliere k oggetti tra N è pari al *coefficiente binomiale*

$$\binom{N}{k} = \frac{N!}{k!\,(N-k)!} = \frac{N\,(N-1)\cdots(N-k+1)}{k!}$$

allora la probabilità di avere k (qualsiasi) persone al telefono è pari a

$$p_B\,(k) = \binom{N}{k} p^k q^{N-k} \tag{22.1}$$

Risultando che $\sum_{k=0}^{N} p_B\,(k) = 1$, la funzione $p_B\,(k)$ rappresenta una densità di probabilità di v.a. discreta, detta anche variabile aleatoria di *Bernoulli* o *binomiale*[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 produce una concentrazione di $p_B\,(k)$ attorno a valori k inferiori; valori di p ancora più piccoli producono una d.d.p. che descresce monotonamente per $k > 0$. Infine, osserviamo che non si possono avere più di N utenti al telefono.

Per conoscere il numero di linee necessarie 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 aleatorio[3].

---

[2] Infatti i termini $\binom{N}{k}$ sono pari ai coefficienti 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 a lato.

```
          1
        1   1
      1   2   1
    1   3   3   1
  1   4   6   4   1
1   5  10  10   5   1
  6  15  20  15   6   1
```

[3] Infatti si applica ad un qualunque fenomeno aleatorio rappresentato dalla ripetizione di un secondo fenomeno aleatorio *soggiacente*, come ad esempio il lancio ripetuto di monete o di dadi: in questi casi,

Intensità di traffico Il valor 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 *l'intensità di 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 *traffico* 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 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. Provando a chiamarla 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

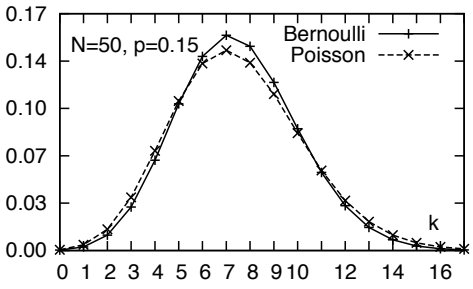
Indichiamo con $p = 0.2$ la probabilità di successo di un singolo tentativo, e con $q = 1 - p = 0.8$ quella di fallimento, identificando così il problema nel contesto delle *prove ripetute*.

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 $p + p \cdot q + p \cdot q \cdot q = .2 + .2 \cdot .8 + .2 \cdot .8 \cdot .8 = 0.488 = 48.8$ %.
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 $p_0 = \binom{12}{0} p^0 q^{12} = \frac{12!}{12!} .8^{12} = 0.0687195$, 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 $\binom{12}{3} p^3 q^9 = \frac{12 \cdot 11 \cdot 10}{3 \cdot 2} \cdot .2^3 \cdot .8^9 = 0.23$.

22.2 Distribuzione di Poisson

Al crescere del numero N di utenti, l'utilizzo della distribuzione Binomiale può risultare disagevole, per via dei fattoriali, e si preferisce trattare il numero di conversazioni attive

ha senso chiedersi con che probabilità una funzione della v.a. soggiacente acquisisce un certo valore, per un certo numero di volte. *Esempio*: si voglia calcolare la probabilità di osservare 3 volte testa, su 10 lanci di una moneta. Applicando la (22.1), si ottiene $p_B(3) = \binom{10}{3} p^3 q^7 = 120 \cdot .5^3 \cdot .5^7 = 0.117$, 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, come illustrato al § 18.3.2.

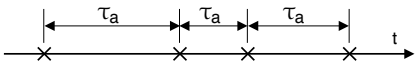


k come una variabile aleatoria di POISSON, la cui densità di probabilità ha espressione

$$p_P(k) = e^{-\alpha}\frac{\alpha^k}{k!} \tag{22.2}$$

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.

Più in generale, la densità (22.2) 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 adottato finora 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 una *frequenza media di interarrivo* λ che rappresenta la velocità (come *richieste per unità di tempo*) con cui si presentano le nuove chiamate[5]. L'inverso di λ rappresenta un tempo, ed esattamente $\overline{\tau}_a = 1/\lambda$ 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 chiamate[6] pari a $\alpha = \lambda T$. Pertanto, possiamo scrivere la d.d.p. 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([7]).

[4] Usando il modello Poissoniano la probabilità che (ad esempio) si stiano svolgendo *meno* di 4 conversazioni contemporanee è pari pertanto a $p_P(0) + p_P(1) + p_P(2) + p_P(3) = e^{-\alpha}\left(1 + \alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{6}\right)$.

[5] La 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...

[6] Esempio: se da un cavalcavia osserviamo (mediamente) $\lambda = 3$ auto/minuto, nell'arco di $T = 2$ minuti, transiteranno (in media) 3*2 = 6 autovetture.

[7] *Esempio*: 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)*** $p_P(2)|_{10} = e^{-\frac{10}{8}}\frac{\left(\frac{10}{8}\right)^2}{2} = 0.224$ pari al 22.4%

22.2.1 Variabile aleatoria esponenziale negativa

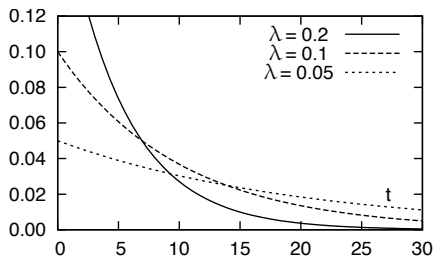
La descrizione statistica che la ddp di Poisson fornisce per *il numero* di eventi che si verificano in un (generico) tempo t è strettamente legata al considerare gli eventi come *indipendenti*, *identicamente distribuiti*, e per i quali *l'intervallo di tempo* tra l'occorrenza degli stessi è una determinazione di variabile aleatoria *completamente casuale*[8], descritta da una densità di probabilità *esponenziale negativa*[9], 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[10] $m_E = \frac{1}{\lambda}$ 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

$$Pr(t > t_0) = \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_{t_0}^{\infty} = e^{-\lambda t_0} \tag{22.3}$$

e questo risultato ci permette di verificare il legame con la Poissoniana[11].



Densità di probabilità *esponenziale*

[8]Da 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 di scrivere

$$Pr(t > t_0 + \theta / t > t_0) = Pr(t > \theta)$$

ossia 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 (22.3):

$$\begin{aligned} Pr(t > t_0 + \theta / t > t_0) &= \frac{Pr(t > t_0 + \theta; t > t_0)}{Pr(t > t_0)} = \frac{Pr(t > t_0 + \theta)}{Pr(t > t_0)} = \\ &= \frac{e^{-\lambda(t_0+\theta)}}{e^{-\lambda t_0}} = e^{-\lambda\theta} = Pr(t > \theta) \end{aligned}$$

[9]La 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.

[10]Per quanto riguarda il valor medio $m_E = \int_0^{\infty} t\lambda e^{-\lambda t} dt$ possiamo procedere *per parti*, ossia applicando la regola $\int_a^b f'(t) g(t) dt = f(t) g(t)|_a^b - \int_a^b f(t) g'(t) dt$, avendo posto $f'(t) = e^{-\lambda t}$ e $g(t) = \lambda t$: si ottiene allora

$$m_E = -\frac{1}{\lambda} e^{-\lambda t} \cdot \lambda t \Big|_0^{\infty} - \int_0^{\infty} -\frac{1}{\lambda} e^{-\lambda t} \cdot \lambda dt = -0 + 0 - \frac{1}{\lambda} e^{-\lambda t}\Big|_0^{\infty} = \frac{1}{\lambda}$$

essendo $\lim_{t\to\infty} e^{-\lambda t} \cdot \lambda t = 0$. Per $\sigma_E^2 = \int_0^{\infty} t^2 \lambda e^{-\lambda t} dt - (m_E)^2$, il primo integrale (sempre procedendo per parti) fornisce $\int_0^{\infty} t^2 \lambda e^{-\lambda t} dt = \frac{2}{\lambda^2}$, e dunque $\sigma_E^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

[11]Consideriamo un ospedale in cui nascono *in media* 6 bimbetti al giorno (o 0.25 nascite l'ora), e consideriamo l'intervallo tra questi eventi come una v.a. completamente casuale. 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,

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?

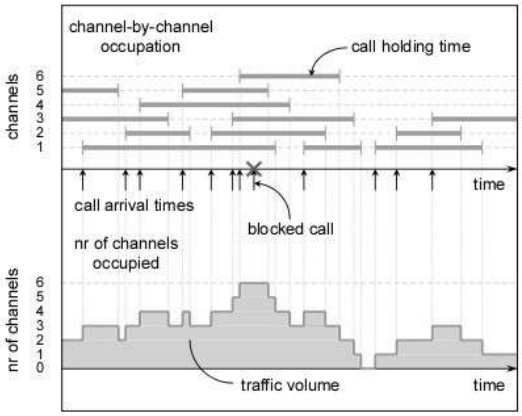
Risposta: ci viene fornito un tempo di attesa medio τ_a, a cui corrisponde una frequenza di servizio $\lambda = \frac{1}{\tau_a}$, e quindi la soluzione risulta $Pr\,(t > 20) = \int_{20}^{\infty} \frac{1}{\tau_a} e^{-t/\tau_a} dt = e^{-20/5} = 0.0183 = 1.83\%$.

Un corollario[12] della (22.3) è che, se $t_0 \to 0$, 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

$$Pr\,(t \leq t_0)|_{t_0 \to 0} = \lambda t_0 + o\,(t_0) \tag{22.4}$$

22.3 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 dispone[13]. 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* (che affrontiamo al § 22.4), oppure rifiutate (vedi la figura a fianco), come avviene per i sistemi *orientati alla perdita*. Scopo della presente sezione è quindi quello di determinare il numero di serventi 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.



richieste di servizio e occupazione serventi nel tempo

22.3.1 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 serventi (indicato come *processo di servizio*) è

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.

[12] La dimostrazione della (22.4) si basa sulla considerazione che $Pr\,(t \leq t_0) = 1 - Pr\,(t > t_0)$, e sulla espansione in serie di potenze $e^x = 1 + x + x^2/2 + x^3/3! + \cdots$ che si riduce a $e^x = 1 + x + o\,(t_0)$ se $x \to 0$. Pertanto la (22.3) diviene $Pr\,(t > t_0)|_{t_0 \to 0} = 1 - \lambda t_0 + o\,(t_0)$, e quindi $Pr\,(t \leq t_0) = 1 - 1 + \lambda t_0 + o\,(t_0) = \lambda t_0 + o\,(t_0)$.

[13] Gli 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*....

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 esponenziale[14], ovvero che le durate degli eventi "nuova richiesta" e "servente occupato" siano *completamente casuali*[15].

22.3.2 Intensità media di traffico

Il rapporto $A_o = \frac{\lambda}{\mu}$ è indicato come *intensità media* del traffico *offerto*[16] e descrive quanti serventi (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 serventi 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).

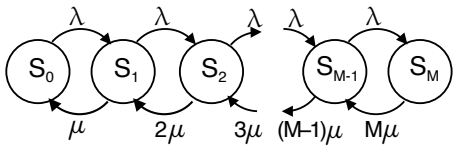
22.3.3 Probabilità di rifiuto

La teoria che porta a determinare la probabilità che una nuova richiesta di servizio non possa essere accolta a causa dell'esaurimento dei serventi si basa sulla descrizione di un cosiddetto *processo di nascita e morte*, che rappresenta da un punto di vista statistico 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 assommi 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.

[14] L'ipotesi permette di valutare la probabilità che l'intervallo temporale tra due eventi di ingresso sia superiore a θ, in base alla (22.3), 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).

[15] Le ipotesi poste fanno sì che i risultati a cui giungeremo siano conservativi, ovvero il numero di serventi 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 serventi; nella realtà, ci si troverà in situazioni intermedie.

[16] Si noti che il pedice $_o$ è una "o" e non uno "0", ed identifica appunto l'aggettivo *offerto*.



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, 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 numero[17]. Se λ e μ non variano nel tempo, una volta esaurito un transitorio iniziale il sistema di servizio si troverà in *condizioni stazionarie*, permettendoci di scrivere le *equazioni di equilibrio statistico*

$$\lambda p_k = \mu (k+1) p_{k+1} \quad \text{con} \quad k = 0, 1, 2, \ldots, M-1 \tag{22.5}$$

che eguagliano la frequenza media con cui il sistema evolve dallo stato k verso $k+1$, alla frequenza media con cui avviene la transizione inversa[18]. La (22.5) può essere riscritta come

$$p_{k+1} = \frac{\lambda}{\mu (k+1)} p_k = \frac{A_o}{(k+1)} p_k$$

che applicata ricorsivamente, porta a scrivere

$$p_k = \frac{A_o^k}{k!} p_0 \tag{22.6}$$

Non resta ora che trovare il modo per dare un valore a p_0, e questo è oltremodo semplice, ricordando che deve risultare[19] $1 = \sum_{m=0}^{M} p_m = p_0 \sum_{m=0}^{M} \frac{A_o^m}{m!}$, e quindi

$$p_0 = \left(\sum_{m=0}^{M} \frac{A_o^m}{m!} \right)^{-1} \tag{22.7}$$

Nei due casi distinti in cui i serventi siano in numero finito (e pari ad M) od infinito ($M = \infty$) otteniamo rispettivamente il caso cercato, ed un caso limite. Se poniamo

[17] Pensiamo 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 deriva 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 $M\mu$.

[18] E' 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 $\mu_r = \sum_{k=1}^{M} \mu \cdot k \cdot p_k$

[19] Usiamo il pedice m anziché k per non creare confusione nella (22.8)

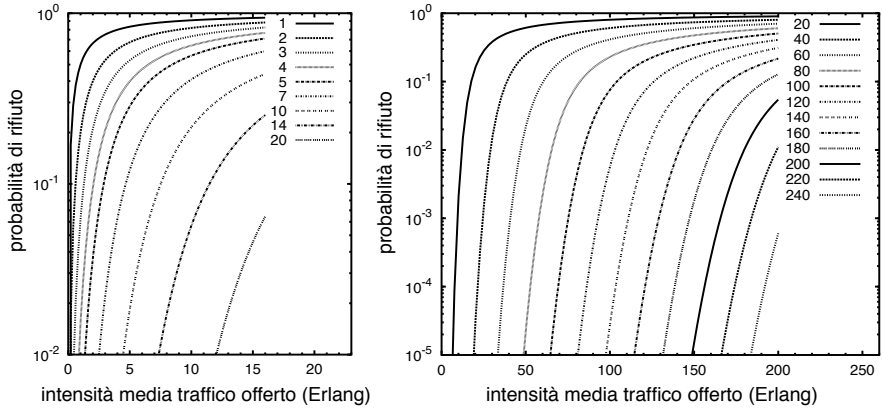


Figura 22.1: Valori della probabilità di blocco P_B in un sistema orientato alla perdita, al variare di A_o, per il numero di serventi indicato sulle curve

$M = \infty$, tenendo conto dell'espansione in serie $\sum_{m=0}^{\infty} \frac{A_o^m}{m!} = e^{A_0}$, si ottiene che la (22.7) fornisce appunto $p_0 = e^{-A_0}$, e la (22.6) diviene $p_k = e^{-A_0} \frac{A_o^k}{k!}$, che come riconosciamo immediatamente è proprio la ddp di Poisson (22.2) con valore medio A_0[20]. Se invece poniamo M finito, la sommatoria che compare in (22.7) non corrisponde ad una serie nota, e dunque rimane come è, fornendo il risultato

$$p_k = Pr\left(S_k\right) = \frac{\frac{A_o^k}{k!}}{\sum_{m=0}^{M} \frac{A_o^m}{m!}}$$

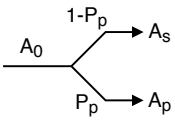
Notiamo ora che p_M è la probabilità che tutti i serventi siano occupati, pari dunque alla probabilità che una nuova richiesta di servizio sia rifiutata. Chiamiamo allora questo valore *probabilità di Blocco, di Rifiuto* o *di Perdita,* la cui espressione prende il nome di Formula B di Erlang, del primo tipo, di ordine M ed argomento A_o:

$$P_B = Pr\left(S_M\right) = p_M = \frac{\frac{A_o^M}{M!}}{\sum_{m=0}^{M} \frac{A_o^m}{m!}} = E_{1,M}\left(A_o\right) \tag{22.8}$$

L'andamento di P_B in funzione di M e di A_o è graficato in Fig. 22.1, e mostra come (ad esempio) per una intensità di traffico offerto pari a 40 Erlang, siano necessari più di 50 serventi per mantenere una P_B minore dell'1%, che salgono a più di 60 per una $P_B = 10^{-3}$.

22.3.4 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 $\left(1 - P_p\right) \cdot 100$ % delle richieste è smaltito, e quindi A_o si ripartisce tra l'intensità media di *traffico smaltito* $A_s = A_o\left(1 - P_p\right)$, e l'intensità media



[20]Questo risultato è in perfetto accordo con le la (22.2), quando abbiamo sostituito alla ddp di Bernoulli quella di Poisson, mantenendo inalterato il numero medio di serventi occupati, che ora indichiamo con A_o, come definito al § 22.3.2.

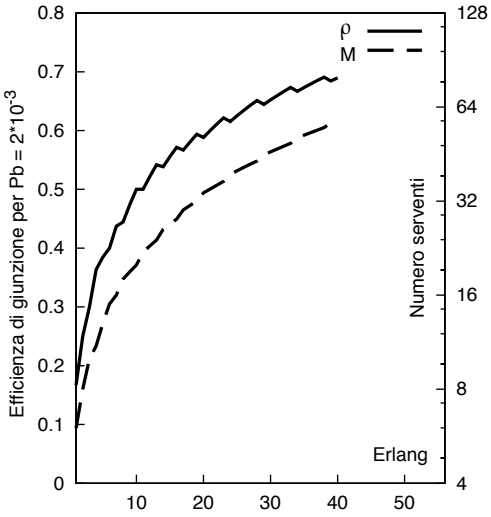


Figura 22.2: Efficienza di giunzione

di *traffico perso* $A_p = A_o P_p$. Possiamo definire un coefficiente di utilizzazione, o efficienza

$$\rho = \frac{A_s}{M} = \frac{A_o}{M}\left(1 - P_p\right)$$

che rappresenta la percentuale di impegno dei serventi, e di cui la figura 22.2 mostra l'andamento al variare di A_o, per una P_B assegnata e pari a $2 \cdot 10^{-3}$, assieme al numero di serventi necessario a garantire tale probabilità di blocco.

Come si può osservare, una volta fissato il grado di servizio, all'aumentare del numero di serventi il traffico smaltito cresce più in fretta di quanto non crescano i serventi[21], 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.

22.3.5 Validità del modello

Le considerazioni esposte si riferiscono ad una ipotesi di traffico completamente casuale con tempi di interarrivo e di servizio esponenziali[22], 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.

[21] ovvero, 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 serventi, per una intensità doppia $A_o = 20$ il numero di serventi necessario a mantenere la stessa P_B risulta poco più di 32.

[22] In 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.

[23] Un 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.

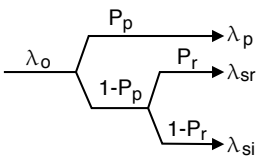
1. Determinare la f_b di una sorgente nelle fasi di attività;
2. determinare il numero massimo di sorgenti contemporaneamente attive;
3. determinare il grado di servizio (Probabilità di rifiuto) ottenibile con il mezzo trasmissivo indicato;
4. indicare la capacità da aggiungere al collegamento per garantire un grado di servizio cento volte migliore.

Risposte .

1. $f_b = \frac{bit}{campione} \cdot \frac{campioni}{secondo} = 12 \cdot 21.33 \cdot 10^3 = 256$ Kbps;
2. 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;
3. 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. 22.1 si trova una probabilità di rifiuto pari a circa 10^{-2}:
4. 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. 22.1 si osserva che ciò richiede (a parità di A_o) almeno 120 (circa) serventi, 20 in più, pari ad una capacità aggiuntiva di $20 \cdot 256 \cdot 10^3 = 5.12$ Mbps.

22.4 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 lo schema che esemplifica 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:

$$\lambda_p = P_p \lambda_o; \qquad \lambda_{sr} = P_r \left(1 - P_p\right) \lambda_o; \qquad \lambda_{si} = \left(1 - P_r\right) \left(1 - P_p\right) \lambda_o$$

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

$$A_o = A_p + A_{sr} + A_{si} \quad \text{e quindi} \quad A_{sr} = \frac{\lambda_{sr}}{\mu},\ A_{si} = \frac{\lambda_{si}}{\mu}$$

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 verificare[24] che ora la P_p risulta inferiore alla P_B del caso senza coda, e pertanto l'intensità di traffico smaltito $A_s = A_{sr} + A_{si} = \left(1 - P_p\right) A_o$ aumenta, a parità di offerta.

22.4.1 Risultato di Little

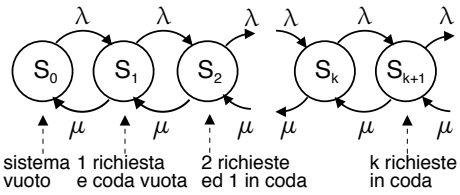
Si tratta di un risultato molto generale, valido per qualsiasi distribuzione dei tempi di interarrivo e di servizio, la cui applicazione può tornare utile nell'analisi, e che recita:

> *Il numero medio $\overline{N}$ 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 $\overline{N} = \lambda_s \cdot \tau_p$. 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.

22.4.2 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. Da un punto di vista statistico un tale sistema è descritto 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. 774 dove si 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 $p_k = Pr\left(S_k\right)$ la probabilità che il sistema contenga k richieste, l'equilibrio statistico si traduce nell'insieme di equazioni

$$\lambda_o p_k = \mu p_{k+1} \quad \text{con} \quad k = 0, 1, 2, \ldots, \infty \tag{22.9}$$

[24] Se 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 $\lambda_o \left(P_p + P_r \left(1 - P_p\right)\right)$, ed eguagliando le due espressioni si ottiene $P_p = \frac{P_B - P_r}{1 - P_r}$, che è sempre minore di P_B.

Infatti, in base alle stesse considerazioni svolte nella prima parte della nota 18 di pag. 774, $\lambda_o p_k$ 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 (22.9) è di natura ricorsiva, e può esprimersi come

$$p_k = \left(\frac{\lambda_o}{\mu}\right)^k p_0 = A_o^k p_0$$

Per determinare il valore $p_0 = Pr(S_0)$, uguale alla probabilità che il sistema sia vuoto, ricordiamo[25] 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 a[26]

$$P_r = 1 - p_o = 1 - (1 - A_o) = A_o$$

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 $\overline{L}$: risulta essere semplicemente il valore atteso del numero di richieste presenti nel sistema, ovvero[27]

$$\overline{L} = E\{k\} = \sum_{k=0}^{\infty} k p_k = (1 - A_o) \sum_{k=0}^{\infty} k A_o^k = \frac{A_o}{1 - A_o}$$

da cui risulta che per $A_o \to 1$ 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

[25] Nella 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 $A_o = \frac{\lambda_o}{\mu} < 1$; infatti se il servente è unico 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.

[26] Ricordiamo che p_0 è la probabilità che il sistema sia vuoto, e dunque $1 - p_0$ quella che *non* sia vuoto.

[27] si fa uso della relazione $\sum_{k=0}^{\infty} k\alpha^k = \alpha \sum_{k=0}^{\infty} k\alpha^{k-1} = \alpha \frac{\partial}{\partial \alpha} \sum_{k=0}^{\infty} \alpha^k = \alpha \frac{\partial}{\partial \alpha} \frac{1}{1-\alpha} = \frac{\alpha}{(1-\alpha)^2}$

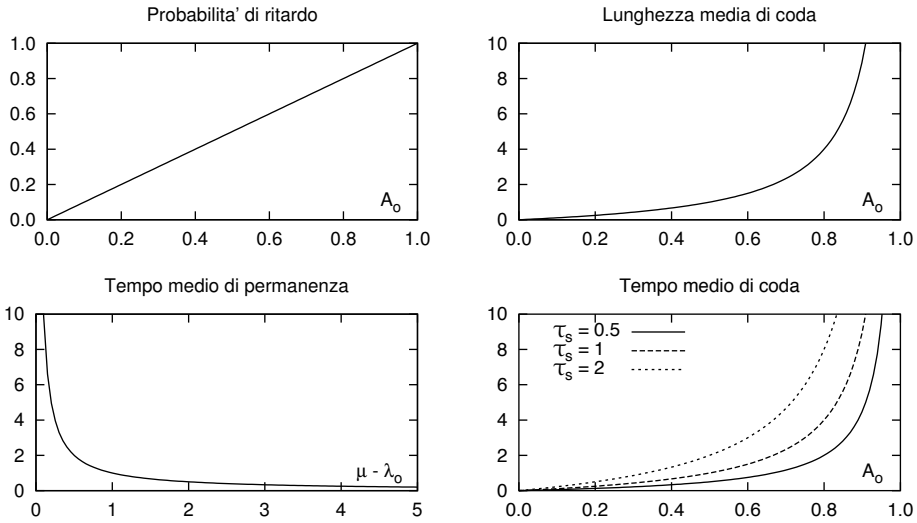


Figura 22.3: Grandezze di interesse per il sistema a coda infinita ed unico servente

applicare qui il risultato di Little $\overline{N} = \lambda_s \cdot \tau_p$, che esprime la relazione tra numero medio $\overline{N}$ di richieste presenti, frequenza di smaltimento (qui pari a quella di offerta[28]), e tempo medio di permanenza; infatti accade che $\overline{N} = \overline{L}$, ed utilizzando il risultato $\overline{L} = \frac{A_o}{1-A_o}$ si ottiene

$$\tau_p = \frac{\overline{N}}{\lambda_s} = \frac{\overline{L}}{\lambda_o} = \frac{A_o}{1-A_o}\frac{1}{\lambda_o} = \frac{\lambda_o}{\mu}\frac{1}{\lambda_o}\frac{1}{1-\lambda_o/\mu} = \frac{1}{\mu-\lambda_o}$$

da cui si osserva che, se la frequenza di offerta tende al valore della frequenza di servizio, il tempo medio di permanenza tende ad ∞.

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 $\rho \longrightarrow \infty$.

La fig. 22.3 mostra l'andamento delle grandezze appena calcolate.

22.4.3 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.

[28] Non 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.

Probabilità di k richieste nel sistema

$$p_k(A_o) = \begin{cases} \frac{A_o^k}{k!\alpha(A_o)} & 0 \le k \le M \\ \frac{A_o^k}{M^{k-M}M!\alpha(A_o)} & M \le k \le M+L \end{cases}$$

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 $0 \le k \le M$ 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

$$P_r = \sum_{k=M}^{M+L} p_k(A_o) = p_M(A_o) \frac{1-\rho^{L+1}}{1-\rho} \quad \text{in cui} \quad \rho = \frac{A_o}{M}$$

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 22.4 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 L, 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 esponenziale[29], i risultati ottenuti costituiscono una *stima conservativa* delle prestazioni del sistema (che potranno cioè essere migliori).

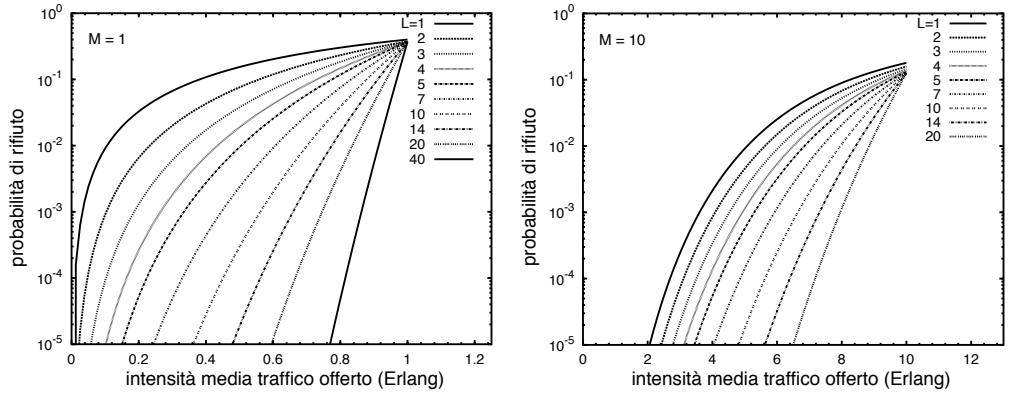


Figura 22.4: Probabilità di perdita per un sistema a coda finita con uno o dieci serventi

[29] In una trasmissione a pacchetto, operata a frequenza binaria f_b e con pacchetti di lunghezza media $\overline{L}_p$ bit, il tempo *medio* di servizio per un singolo pacchetto è pari a quello medio necessario alla sua trasmissione, e cioè $\tau_S = \overline{L}_p / f_b$.

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 multiplazione di pacchetti di dimensione media di 8 KByte[30] su collegamenti con velocità binaria $f_b = 100$ Mbps[31]

1) Determinare il tempo medio di servizio di ogni singolo pacchetto;

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: $\tau_S = \frac{1}{\mu} = durata\ di\ un\ bit \cdot \frac{bit}{pacchetto} = \frac{1}{10^8}\left[\frac{secondi}{bit}\right] \cdot 1024\left[\frac{byte}{pacchetto}\right] \cdot 8\left[\frac{bit}{byte}\right] \simeq 655\ \mu$sec;

2) $\tau_a = \frac{1}{\lambda} = \frac{1}{1200} = 833\ \mu$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 $\tau_p = \frac{1}{\mu - \lambda_o} = \frac{1}{\frac{10^6}{655} - \frac{10^6}{833}} = \frac{1}{326} \simeq 3$ 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 $3 \cdot 10^{-3}\ [sec] \cdot 1200\left[\frac{pacch}{sec}\right] \cdot 1024\left[\frac{byte}{pacch}\right] \simeq 3.7$ Kbyte.

22.5 Reti per trasmissione dati

In questa sezione illustriamo le particolarità legate alla *trasmissione dati*, e come possa essere vantaggiosamente sfruttata per conseguire la *maggiore efficienza* che i sistemi di

[30] 1 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 $10^3 = 1000$.

[31] In virtù di quanto esposto alla nota precedente, in questo caso $1M = 10^6 = 1000000$.

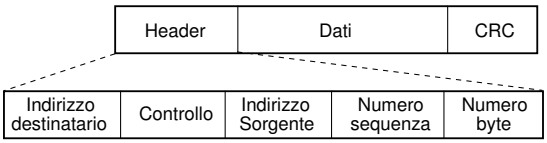
servizio a coda presentano rispetto a quelli orientati alla perdita. Le particolari *modalità* e *funzioni* legate alle trasmissioni dati saranno classificate secondo uno schema che ne consente il confronto in termini di prestazioni e vincoli sulla realizzazione della rete. Infine, verranno formalizzati modelli operativi volti 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 scontato, 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.

22.5.1 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 suddividendo il messaggio in pacchetti è necessario prevedere un aumento delle informazioni da trasmettere, dato che ognuni pacchetto dovrà contenere informazioni addizionali 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.

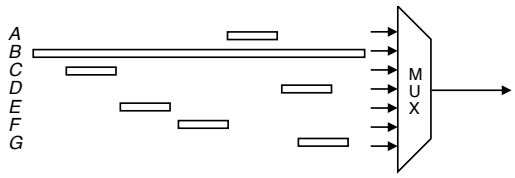


In termini generali, un pacchetto è composto da una *intestazione* (HEADER), dalla parte di messaggio che trasporta (*dati*), e da un campo *codice di parità* (CRC) necessario a rivelare l'occorrenza di errori di trasmissione[32]. L'*header* a sua volta può essere suddiviso in campi, in cui trovano posto (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 aggiuntive[33], la trasmissione a pacchetto consegue una efficienza maggiore di quella a circuito, in quanto è attuata mediante sistemi a coda.

[32] La 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 § 15.6.3.3) 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.

[33] L'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).

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 (15.27) a pag. 472 e la discussione al § 22.6.2.3), e dunque l'uso di dimensioni contenute riduce le necessità di ritrasmissione.

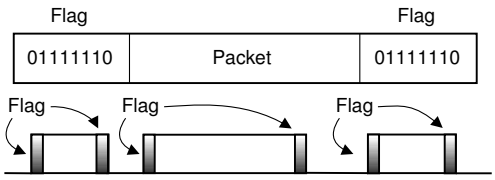
22.5.2 Modo di trasferimento delle informazioni

È definito in base alla specificazione di 3 caratteristiche che lo contraddistinguono: *lo schema di multiplazione*, il *principio di commutazione* e *l'architettura protocollare*.

22.5.2.1 Schema di multiplazione

Al § 24.3.1 è descritto uno schema a divisione di tempo che prevede l'uso di una *trama* in cui trovano posto diverse comunicazioni vocali[34], e che necessita di un funzionamento sincronizzato (o quasi) 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; pertanto si rendono necessarie soluzioni idonee alla *delimitazione* dei pacchetti.

Ad esempio, i protocolli HDLC ed x.25[35] presentano pacchetti di dimensione variabile, e fanno uso di un byte di *flag* (vedi pag. 486) costituito dalla 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.



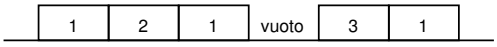
Nel caso in cui il pacchetto invece abbia una *dimensione fissa*[36], ci si trova ad

[34] Come nel PCM telefonico, vedi § 24.3.1

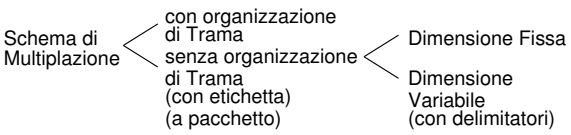
[35] A riguardo di questi due protocolli ormai *fuori moda*, si veda ad es. `https://it.wikipedia.org/wiki/High-Level_Data_Link_Control` e `https://it.wikipedia.org/wiki/X.25`

[36] Un modo di trasferimento con pacchetti di dimensione fissa è l'ATM (*Asynchronous Transfer Mode*) che viene descritto al § 23.2.

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.

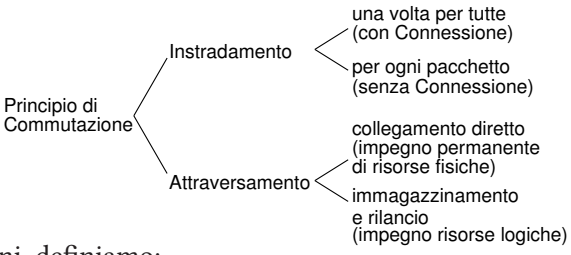


22.5.2.2 Principio di commutazione

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 *demultiplare* le informazioni in ingresso e multiplarle di nuovo su uscite diverse: ciò può avvenire mediante un *collegamento diretto* o per *immagazzinamento e rilancio*.



Sulla base di queste considerazioni, definiamo:

Commutazione di circuito: l'instradamento avviene una volta per tutte prima della comunicazione, e l'attraversamento impegna in *modo permanente* ed *esclusivo* le *risorse fisiche* dei nodi della rete; è il caso della telefonia, sia POTS che PCM[37].

Commutazione di pacchetto a *circuito virtuale*: L'instradamento è determinato *una volta per tutte* prima dell'inizio della trasmissione, durante una fase di *setup* delle risorse necessarie, conseguente alla *richiesta di connessione* effettuata da parte del

[37] Nel caso del POTS (vedi § 24.9.1) si creava un vero e proprio circuito elettrico (vedi anche pag. 830), e le risorse fisiche impegnate sono gli organi di centrale ed i collegamenti tra centrali, assegnati per tutta la durata della comunicazione in esclusiva alle due parti in colloquio. Nel caso del PCM (vedi § 24.3.1), le risorse allocate cambiano natura (ad esempio consistono anche nell'intervallo temporale assegnato al canale all'interno della trama) ma ciononostante vi si continua a far riferimento come ad una rete a *commutazione di circuito*.

nodo sorgente. Dopodiché i pacchetti di uno stesso messaggio seguono tutti lo stesso percorso, e l'attraversamento si basa sull'impegno di *risorse logiche*[38] ed avviene per *immagazzinamento e rilancio*. La trasmissione ha luogo dopo aver contrassegnato ogni pacchetto con un *identificativo di connessione* (IC) che individua un *canale virtuale*[39]tra coppie di nodi di rete, che ne identifica l'appartenenza ad uno dei collegamenti in transito.

L'intestazione del pacchetto può essere ridotta, al limite, a contenere il solo IC del canale virtuale. L'attraversamento avviene consultando apposite tabelle (di *routing*), generate nella fase di setup che precede quella di trasmissione, in cui è indicata la porta di uscita per tutti i pacchetti appartenenti ad uno stesso messaggio. Facciamo un esempio, riferendoci allo schema di fig. 22.5: una sorgente, a seguito della fase di instradamento, invia i pacchetti con identificativo IC = 1 al primo nodo individuato dal routing. Consultando la propria tabella, il nodo trova che il canale virtuale *1* sulla *porta di ingresso* (PI) *A* si connette al c.v. *3* sulla *porta di uscita* (PU) *C*. Ora i pacchetti escono da *C* con IC = 3 ed una volta giunti al nodo seguente sulla PI *A*, escono dalla PU *B* con IC = 2 e giungono finalmente a destinazione. Notiamo che su di un collegamento *tra due nodi*, i numeri dei canali virtuali identificano in modo univoco il collegamento a cui appartengono i pacchetti, mentre uno stesso numero di canale virtuale può essere riutilizzato su porte differenti[40].

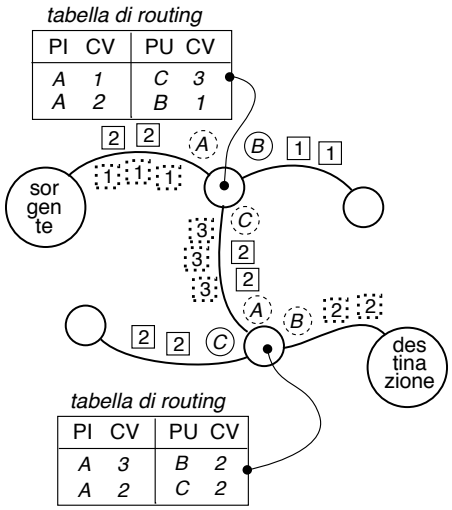


Figura 22.5: Commutazione di pacchetto a circuito virtuale

La concatenazione dei canali virtuali attraversati viene infine indicata con il termine *Circuito 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

[38]Le risorse impegnate sono dette *logiche* in quanto corrispondono ad entità concettuali (i *canali virtuali* descritti nel seguito).

[39]Il termine *canale virtuale* simboleggia il fatto che, nonostante i pacchetti di più comunicazioni viaggino "rimescolati" su di uno stesso mezzo, questi possono essere distinti in base alla comunicazione a cui appartengono, grazie ai differenti *IC* (numeri) con cui sono etichettati; pertanto, è come se i pacchetti di una stessa comunicazione seguissero un proprio *canale virtuale* indipendente dagli altri.

[40]I numeri di c.v. sono negoziati tra ciascuna coppia di nodi durante la fase di instradamento, e scelti tra quelli non utilizzati da altre comunicazioni già in corso. Alcuni numeri di c.v. inoltre possono essere *riservati*, ed utilizzati per propagare messaggi di segnalazione inerenti il controllo di rete.

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 anche i canali virtuali delle altre connessioni 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 pacchetti[41]. 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*, il quale (chiamato ora *datagramma*) deve 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 utilizzate[42]. 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, di tipo sia 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 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

[41] In 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*.

[42] Per questo motivo, il collegamento è detto *senza connessione*.

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*[43] 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 congestione[44].

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. 805).

Proseguiamo la descrizione delle reti per dati con l'ultima caratteristica di un modo di trasferimento:

22.5.2.3 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:

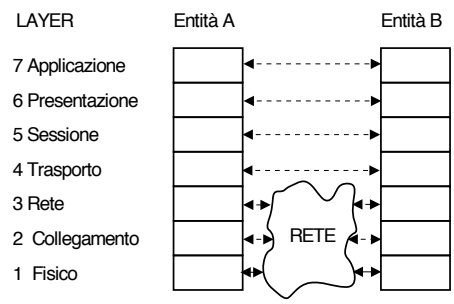
- il *controllo di flusso*, che impedisce la saturazione dei buffer;
- la *consegna ordinata*, per riassemblare messaggi frammentati su più datagrammi;
- la *segmentazione e riassemblaggio*, che definisce le regole per frammentare un messaggio in pacchetti e ricomporli, ad esempio in corrispondenza dei "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;

[43] *Il termine* buffer *ha traduzione letterale "respingente, paracolpi, cuscinetto" ed è a volte espresso in italiano dalla locuzione* memoria tampone.

[44] La 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.

- il *controllo di errore*, che provvede a riscontrare le unità informative, a rilevare gli errori di trasmissione, a gestire le richieste di trasmissione;
- 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)[45], 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.



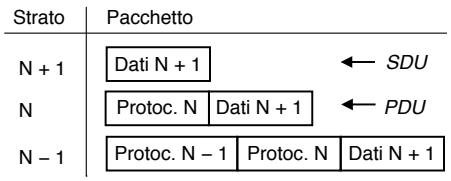
La relazione gerarchica individuata stabilisce tra due strati contigui un rapporto di tipo *utente-servizio*; ovvero lo svolgimento delle funzioni di strato superiore necessita dei servizi offerti dallo strato inferiore. A titolo di esempio, si pensi all'invio di un documento mediante un corriere espresso: ci si affida allora ad uno strato di trasporto che offre all'utente (strato di sessione) un servizio (appunto) 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)[46].

[45] In virtù dell'intreccio di sigle, il modello di riferimento prende il nome (palindromo) di modello *ISO-OSI*.

[46] Il modo di trasferimento è completamente definito dopo che sia stato specificato in quale strato siano svolte le funzioni di commutazione e multiplazione. 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

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.

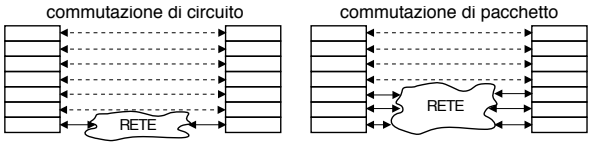
In particolare, se ci riferiamo all'interfaccia tra gli strati di trasporto e di rete, lo strato di rete può realizzare con il suo pari collegamenti con o senza connessione, mentre quello di trasporto offre allo stesso tempo (ma in modo indipendente) agli strati superiori un servizio con o senza connessione, dando luogo alle seguenti 4 possibilità:

	Servizio di trasporto	
Servizio di rete	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

modo trasparente sorgente e destinazione. Nella commutazione di pacchetto, invece, le funzioni di multiplazione 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.



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 riassemblaggio 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.

22.6 Protocolli a richiesta automatica

Le trasmissioni ARQ (pag. 471) prevedono l'esistenza di un canale di ritorno, mediante il quale chiedere la ri-trasmissione delle trame ricevute con errori[47]; 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.

22.6.1 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 22.6 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 che le trame devono essere etichettate con un numero di sequenza, in modo da permettere

[47] Queste 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.

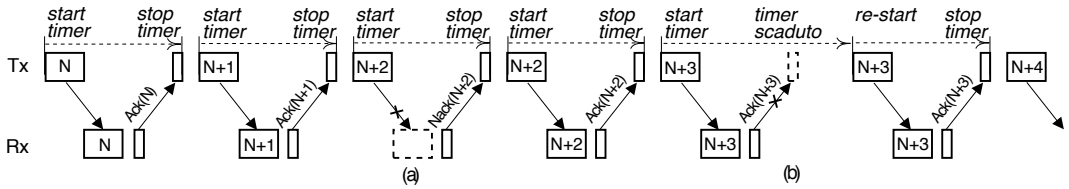


Figura 22.6: Richieste di ritrasmissione per un protocollo send and wait

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:

a) un cavo ritorto di 1 km,

b) una linea dedicata di 200 km,

c) 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 $t_p = \frac{spazio}{velocità}$ 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:

a) 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 $U \simeq 1$;

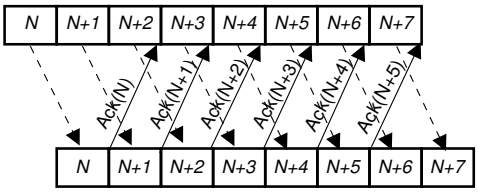
b) per $f_b = 1$ kbps si ottiene $a = 10^{-3}$ e quindi $U \simeq 1$, per $f_b = 1$ Mbps risulta $a = 1$ e quindi $U = 0.33$;

c) 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.

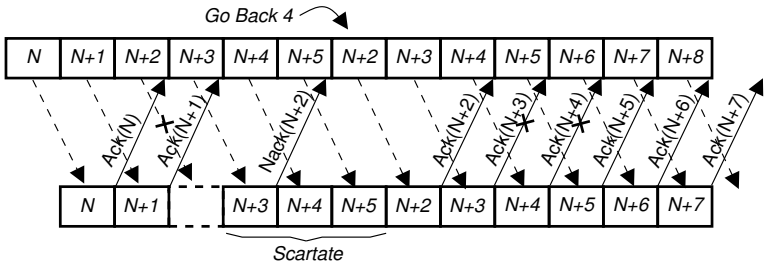
22.6.2 Continuous RQ



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*.

22.6.2.1 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 $N + i$, rilevato[48] 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.

[48] Sottolineiamo nuovamente l'importanza dei numeri di sequenza, che permettono al ricevitore di capire il numero della trama corrotta, grazie alla discontinuità dei numeri stessi.

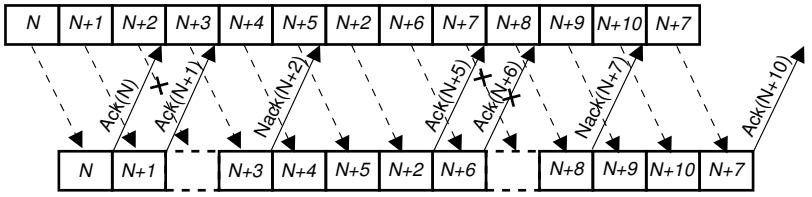


Figura 22.7: Protocollo di ritrasmissione *selective repeat*

22.6.2.2 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 alla figura 22.7, a seguito della ritrasmissione della trama $N+2$ per cui si è ricevuto il NACK, il trasmettitore continua ad inviare nuove trame, fino al numero massimo previsto; in assenza di ulteriori ACK, un timer determina la ritrasmissione delle trame non riscontrate. Quando al ricevitore perviene la trama $N+2$, questo emette un ACK$(N+5)$, consentendo al trasmettitore di rilasciare tutta la memoria occupata dalle trame in sospeso, 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 diverso 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.

22.6.2.3 Efficienza dei protocolli a richiesta automatica

Mentre con le tecniche FEC siamo obbligati ad aumentare la f_b dei dati trasmessi per poter inserire i bit di ridondanza, potrebbe sembrare che nel caso ARQ ciò non si verifichi. Ma anche se la ridondanza introdotta è effettivamente inferiore, la necessità di dover ritrasmettere l'intero pacchetto ricevuto errato è anch'essa fonte di riduzione della velocità trasmissiva. Valutiamo di quanto.

Indichiamo con p la probabilità di dover chiedere la ritrasmissione di una trama[49]

[49] Nel caso in cui l'integrità della trama sia protetta da un codice a blocco (n,k) con $d_m = l+1$ (§ 15.6.2.1), la probabilità che la trama contenga più di l errori e che quindi venga accettata dal ricevitore anche se errata, vale approssimativamente $P(l+1,n) = \binom{n}{l+1} P_e^i$ (vedi formula (15.27)). 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 $P(l+1,n) \ll P(0,n)$ (vedi eq. 15.27), 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 $nP_e \ll 1$).

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 discreta, caratterizzata dalle probabilità

$$\begin{aligned} p_M(1) &= Pr(m=1) = 1-p \\ p_M(2) &= p(1-p) \\ p_M(3) &= p^2(1-p) \\ &\vdots \\ p_M(m) &= p^{m-1}(1-p) \end{aligned}$$

che descrivono come sia possibile ricevere la stessa trama come errata $m-1$ volte, finché all'$m-$esima trasmissione non si rilevano più errori. Pertanto, il numero medio di trasmissioni per una stessa trama è pari a

$$\begin{aligned} \overline{m} &= E\{m\} = \sum_{m=1}^{\infty} m p_M(m) = \sum_{m=1}^{\infty} m p^{m-1}(1-p) = \\ &= (1-p)\sum_{n=0}^{\infty}(n+1)\cdot p^n = (1-p)\frac{1}{(1-p)^2} = \frac{1}{1-p} \end{aligned}$$

in cui alla quarta eguaglianza si è posto $n = m-1$, ed alla quinta si è utilizzato il risultato noto della serie geometrica[50] per ottenere

$$\begin{aligned} \sum_{n=0}^{\infty}(n+1)p^n = \sum_{n=1}^{\infty} n p^{n-1} = \sum_{n=0}^{\infty} n p^{n-1} = \sum_{n=0}^{\infty}\frac{\partial}{\partial p}p^n = \\ = \frac{\partial}{\partial p}\sum_{n=0}^{\infty}p^n = \frac{\partial}{\partial p}\frac{1}{1-p} = \frac{1}{(1-p)^2} \end{aligned}$$

Quindi, per trasmettere una frequenza binaria di f_b bps (comprensivi di CRC e *overhead* dei numeri di sequenza, vedi § 22.6.3.3), occorre in realtà inviare dati ad una velocità *media* pari a $f_b/(1-p)$ bps[51]. Questo risultato approssimato si applica ad un protocollo di tipo *selective repeat*, e trascurando gli errori sul canale a ritroso.

22.6.3 Controllo di flusso

Si è illustrato come 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 per 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 conduttura viene ristretta al fine di ridurre il flusso di

[50] $\sum_{k=0}^{\infty}\alpha^k = \frac{1}{1-\alpha}$ se $|\alpha| < 1$

[51] Dato che p aumenta con n (vedi pag. 471), l'efficienza del protocollo ARQ *peggiora* con l'aumentare della dimensione delle trame. Questo risultato determina l'esigenza di ricercare una soluzione di compromesso, dato che l'incidenza dell'overhead sulla dimensione complessiva della trama invece *si riduce* all'aumentare di n.

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.

22.6.3.1 Round trip time

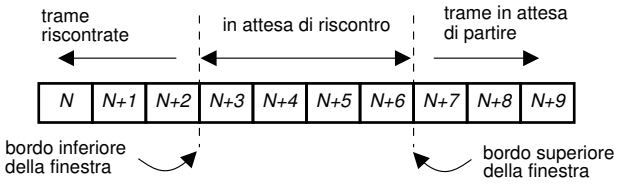
Potrebbe esssere tradotto come *tempo di girotondo*, ed è il tempo che intercorre tra l'inizio della trasmissione di una trama e l'arrivo del relativo ACK. La sua valutazione spesso si avvale della ipotesi di poter trascurare il tempo di trasmissione dell'ACK, e quindi si ottiene

$$RTT \simeq t_{Tx} + 2t_p$$

dove t_{Tx} è il tempo necessario a trasmettere il pacchetto, e t_p è il ritardo di propagazione.. 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*[52]. 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.

22.6.3.2 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 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

	fin. di trasm.	*fin. di ric.*	N_{Max}
send and wait	1	1	2
go back N	W	1	W+1
selective repeat	W	W	2W+1

[52] L'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.

realizzare un protocollo ARQ,
che quindi può essere ri-classificato in questi termini come mostrato in tabella, dove la colonna *finestra di ricezione* indica anche i requisiti di memoria al lato ricevente[53]. Notiamo quindi 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.

22.6.3.3 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 simile[54] 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.

[53] La ricezione di una sequenza di trame corrette determina l'avanzamento alternato dei due bordi della finestra al ricevitore: questa inizialmente è vuota, poi 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.

[54] Se 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.

Capitolo 23

Reti a pacchetto

Alfine è giunto il momento di parlare della *rete delle reti*, ossia di Internet! Il tema è sviluppato con riferimento ai vari strati funzionali che sono coinvolti nella sua operatività, iniziando da una visione di insieme che descrive la concatenazione di indirizzi su cui si basa la trasmissione, per approfondire l'analisi a partire dallo strato di trasporto, giù fino allo strato fisico. Per una visione ancora più ampia sugli aspetti che *sovrastano* lo strato di trasporto, il lettore può far riferimento ad un altro testo dello stesso autore, *Lo strato applicativo di Internet*. Sempre in questo capitolo, sono discussi anche i principi e le pratiche su cui si basa l'ATM, una architettura di rete nata quasi in contemporanea ad Internet, e che pur non avendone eguagliato il successo, rappresenta un caso di scuola per la categoria di reti basate sul paradigma del circuito virtuale. Viceversa, la discussione sulle reti orientate *alla perdita* ovvero a commutazione di circuito è rimandata al capitolo 24.

Affermiamo fin da subito che il modello a strati ISO-OSI (pag. 789) 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, realizzazioni come Internet si sono sviluppate precedentemente alla definizione di tale modello, mentre altre (come ATM) seguono filosofie che solo successivamente sono state incorporate nel modello di riferimento. Pertanto, utilizzeremo le classificazioni ISO-OSI come riferimento culturale e terminologico, mediante il quale analizzare le funzioni delle reti reali.

23.1 La rete Internet

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*[1]. 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

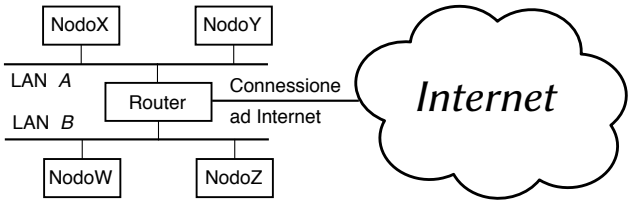
[1] Migliore sforzo, ossia la rete dà il massimo, senza però garantire nulla.

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 DoD[2] 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 *Mosaic*, 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.

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 (LAN[3]), la quale a sua volta è interconnessa ad Internet mediante dei nodi detti *router*[4] 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*[5], che individua i singoli nodi in modo univoco su scala mondiale.

Come anticipato a pag. 790, lo strato di rete (o strato IP) realizza un modo di trasferimento a datagramma e non fornisce garanzie sulla qualità di servizio (QoS, QUALITY O SERVICE) in termini di ritardi, errori e pacchetti persi. La situazione è

[2]Department of Defense.

[3]LOCAL AREA NETWORK, ossia *rete locale*. Con questo termine si indica un collegamento che non si estende oltre (approssimativamente) un edificio.

[4]La 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 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* .

[5]IP = Internet Protocol.

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.

23.1.1 Gli indirizzi

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.

Strato	Indirizzo
Applicazione	*protocollo://nodo.dominio.tld*
Trasporto	*socket TCP o porta*
Rete	*indirizzo IP x.y.w.z*
Collegamento	*indirizzo Ethernet a:b:c:d:e:f*

23.1.1.1 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 un gruppo di 4 byte[6] e che si scrive *x.y.w.z* con ognuna delle 4 variabili separate da punti 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 esempio[7] 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 configurato 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.

23.1.1.2 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

[6] Con 4 byte si indirizzano (in linea di principio) $2^{32} = 4.29 \cdot 10^9$ diversi nodi (più di 4 miliardi). E' stato sviluppato il cosiddetto IPv6, che estende 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*. Per approfondire, vedi ad es. `https://it.wikipedia.org/wiki/IPv6`

[7] Evidentemente 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.

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 destinatario[8] ed inviargli il pacchetto direttamente. In caso contrario, X invierà il pacchetto al proprio *default gateway* verso Internet.

23.1.1.3 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.

23.1.1.4 Domain Name Service (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 Service (*servizio 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 domain[9]), 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 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 indirizzo che termina per dominio.tld.

Quando un nodoX generico deve comunicare con nodoY.dominio.tld, interroga il proprio DNS[10] per conoscerne l'IP. Nella rete sono presenti molti DNS, alcuni dei quali detengono informazioni *autorevoli*[11] 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*

[8] Mostriamo in seguito che questo avviene mediante il protocollo ARP.

[9] I 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.

[10] Il "proprio" DNS viene configurato per l'host in modo fisso, oppure in modo dinamico dai Service Provider raggiungibili via ADSL, e convenientemente corrisponde ad un nodo situato "vicino" al nodo che lo interroga.

[11] Chi 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.

nodoY, allora[12] 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*[13] 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 DNS[14].

23.1.1.5 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 implementa uno specifico *servizio* 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*[15]), identificate da numeri[16], 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 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 sempre il chiamante non ha già aperto altre connessioni differenti.

23.1.2 TCP

Discutiamo ora del TCP[17], che offre ai processi applicativi un servizio di trasporto a

[12] In 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.

[13] Il 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.

[14] CACHE è un termine generico, che letteralmente si traduce come *nascondiglio dei viveri*, e che viene adottato ogni volta si debba indicare una memoria che contiene copie di riserva, o di scorta...

[15] *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.

[16] Spesso 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.

[17] TCP = *Transport Control Protocol*.

circuito virtuale, *attaccato* ad una porta[18] 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.

23.1.2.1 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 ricevere[19], 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; 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 valore 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

[18] Il 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 serventi sono in attesa di richieste di connessione, è presente in tutte le distribuzioni Linux nel file */etc/services*.

[19] Il 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.

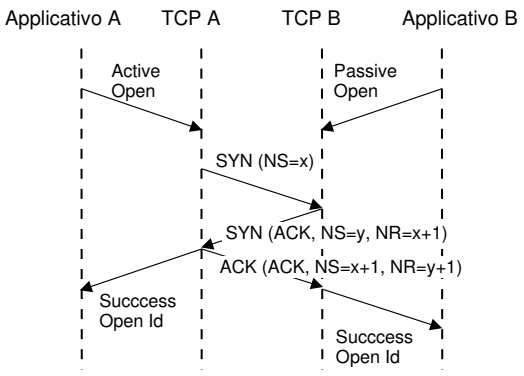
al ricevente per verificare se si sia verificato un errore, il *Puntatore Urgente* contiene il numero di sequenza dell'ultimo byte di una sequenza di dati urgenti, e le *Opzioni* (di lunghezza variabile) sono presenti solo raramente, ed utilizzate a fini di controllo, ad esempio per variare la dimensione 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.

23.1.2.2 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 19.

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 HANDSHAKE[20]. 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 iniziato[21]. 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.



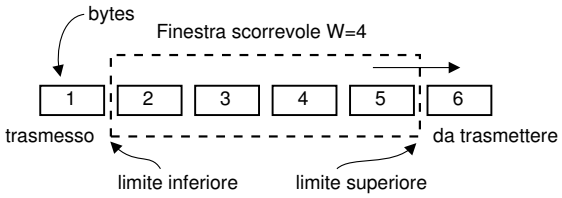
23.1.2.3 Protocollo a finestra

La funzione di controllo di flusso (ossia il dosaggio del ritmo con cui trasmettere i pacchetti) viene attuata dal TCP sfruttando la conoscenza del numero di riscontro NR inviato dal ricevente.

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

[20] HANDSHAKE = *stretta di mano.*

[21] Per 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.



non 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 pervenuto[22].

Nel caso in cui sia settato il bit URG[23] significa che 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 siano andati persi, e li re-invia[24]. Il valore del *timeout* viene calcolato dinamicamente dal TCP in base alle sue misure di *round-trip delay*[25], 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-spedire 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 verifichino errori, la rete

[22] Il 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*.

[23] In tal caso, il campo *Puntatore Urgente* contiene il numero di sequenza del byte che delimita superiormente i dati che devono essere consegnati urgentemente.

[24] Il mancato invio del riscontro può anche essere causato dal verificarsi di un *cecksum* errato dal lato ricevente, nel qual caso quest'ultimo semplicemente evita di inviare il riscontro, confidando nella ritrasmissione per timeout.

[25] Con licenza poetica: *il ritardo del girotondo*, che qui raffigura un percorso di andata e ritorno senza soste.

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 la dimensione della finestra di trasmissione può essere ridotta, riducendo così il carico della rete.

23.1.2.4 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 pacchetti 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.

23.1.3 IP

L'Internet Protocol costituisce *l'ossatura* di 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 dimensione[26] 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 garantita[27], 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.

23.1.3.1 Intestazione IP

Codifica le informazioni mostrate nella figura a lato. 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 un *Type of Service* per differenziare ad esempio la QoS[28] richiesta. L'*identificazione* riporta lo stesso valore

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	

[26] L'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.

[27] Si 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).

[28] La 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

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 *Flags* 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 rete[29], 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'header[30].

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.

23.1.3.2 Indirizzamento e Routing

A pagina 802 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 sottorete[31]. 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

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.

[29]Lo 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.

[30]In 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.

[31]Possiamo 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.

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.

23.1.3.3 Subnetting e Supernetting

Osserviamo ora che la maschera di sottorete presentata a pag. 802 *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 B[32]. 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.

23.1.3.4 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 poco[33]. 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.

23.1.3.5 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

[32] In questo caso, l'Università di Roma "La Sapienza" è intestataria della rete 151.100.

[33] Ad esempio, organizzazioni con poco più di un migliaio di nodi erano costrette a richiedere una intera classe B con capacità di 65536 nodi.

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"[34].

23.1.3.6 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*[35] 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.

23.1.3.7 Multicast

Tornando all'esame della tabella di pag. 808, in cui la classe E costituisce evidentemente una "riserva" di indirizzi per poter effettuare sperimentazioni, la classe D individua invece dei canali *multicast*[36]. Quando un nodo decide di aderire ad un canale multicast, invia un messaggio[37] 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 multicast[38]

[34] Sebbene 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.

[35] Vedi ad es. `https://didattica-2000.archived.uniroma2.it/rt/deposito/rt04-12.pdf`

[36] Il termine *multicast* è ispirato alle trasmissioni *broadcast* effettuate dalle emittenti radio televisive.

[37] Mediante 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.

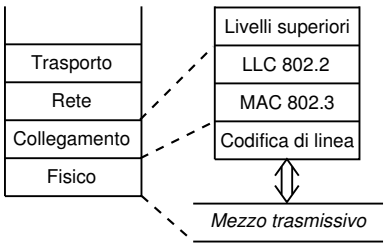
[38] Data l'impossibilità a stabilire un controllo di flusso con tutti i destinatari, il traffico multicast viaggia all'interno di pacchetti UDP.

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.

23.1.4 Ethernet

Ci occupiamo qui di un caso particolare di realizzazione dei primi due livelli del modello ISO-OSI. Come anticipato a pag. 801, 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 casuale[39]: se infatti questi fossero usati come indirizzi a livello di rete, le tabelle di instradamento dovrebbero essere a conoscenza di *tutti* i nodi esistenti[40]. Puntualizziamo inoltre che un nodo di Internet può essere connesso alla rete anche in svariati altri metodi come mediante modem telefonico, rete cellulare, WiFi: qui ci limitiamo 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 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.

23.1.4.1 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 router[41], questo invia su quella LAN un pacchetto *broadcast*, su cui viaggia una richiesta ARP (ADDRESS RESOLUTION

[39] E rappresenta quindi ciò che viene detto uno *spazio di indirizzi piatto* (FLAT ADDRESS SPACE).

[40] Al contrario, il partizionamento dell'indirizzo IP in rete+nodo permette di utilizzare tabelle di routing di dimensioni gestibili.

[41] Ad ogni porta del router è associata una coppia sottorete/maschera (vedi pag. 802) 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.

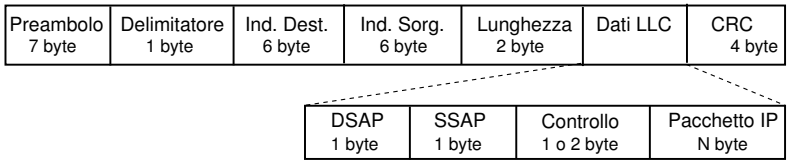


Figura 23.1: Formato di un pacchetto (o *trama*) Ethernet

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 tabella[42].

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*.

23.1.4.2 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, con il risultato finale mostrato in fig. 23.1 . 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*.

L'LLC da parte sua inserisce (in testa al pacchetto IP) due indirizzi *SAP* (SERVICE ACCESS POINT) di sorgente e destinazione, da utilizzare per indicare il codice che identifica il tipo di rete e/o protocollo del pacchetto incapsulato (ad es., IP od ARP). Nel campo di *controllo* possono essere anche ospitati numeri di sequenza, per i casi che lo possano richiedere, ed infine troviamo il pacchetto IP originario.

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*)[43] 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

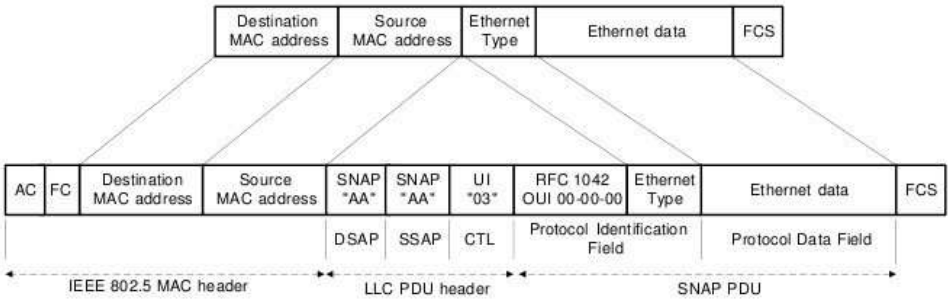
[42] Dato che i nodi possono essere spostati, possono cambiare scheda di rete e possono cambiare indirizzo IP assegnatogli, la corrispondenza IP-Ethernet è tutt'altro che duratura, ed ogni riga della tabella ARP indica anche quando si sia appresa la corrispondenza, in modo da poter stabilire una scadenza, 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.

[43] `http://en.wikipedia.org/wiki/Subnetwork_Access_Protocol`

byte seguenti (indicati come *protocol ID*) debbano essere interpretati come un codice *Ethertype*[44], lo stesso usato nel formato *Ethernet II* discusso appresso, permettendo quindi di specificare finalmente il protocollo incapsulato.

Infine, viene molto frequentemente usato un formato di trama ancora diverso, detto *Ethernet II* o DIX[45], 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 si 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.



23.1.4.3 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 propagazione[46] 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.

[44] `http://en.wikipedia.org/wiki/EtherType`

[45] Vedi `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

[46] Su di un cavo coassiale *tick* da 50 Ω, la velocità di propagazione risulta di $231 \cdot 10^6$ 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 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.

Qualora sia rilevata una contesa, i due nodi smettono di trasmettere, e riprovano solo dopo una attesa di durata casuale.

23.1.4.4 Trasmissione

Il segnale relativo al pacchetto Ethernet viene trasmesso adottando una codifica di linea di tipo Manchester differenziale (§ 15.2.1). 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 collegamento di banda base a 10 Mbps, su cavo *tick* e *thin*[47], con estensione massima 500 e 200 metri[48].

23.1.5 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.

23.1.5.1 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 UTP[49] per le due direzioni di trasmissione, rendendo eventualmente la comunicazione *full-duplex*[50]. In quest'ambito sono definiti i sistemi 10BASET e 100BASET, 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 o *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 verificasi.

23.1.5.2 LAN Switch

D'altra parte, i dispositivi detti BRIGDE 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

[47] TICK = *duro* (grosso), THIN = *sottile*. Ci si riferisce al diametro del cavo.

[48] Le sigle indicano infatti la velocità, se in banda base o meno, e la lunghezza della tratta.

[49] UNSHIELDED TWISTED PAIR (UTP), ossia la coppia ritorta non schermata.

[50] La 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.

detezione di collisione, si è dovuta ridurre di pari misura la massima distanza tra nodi trasmittenti.

23.1.5.3 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'*esplora risorse di rete*. I 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 uno stesso cablaggio. Per interconnettere le LAN, occorre attraversare un dispositivo router.

23.1.5.4 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*[51]. 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 μm	laser 850 nm
	550 m	f.o. multimodo ϕ 50 μm	laser 850 nm
1000BASE-LX	550 m	f.o. multimodo ϕ 50 o 62.5 μm	laser 1300 nm
	3000 m	f.o. monomodo ϕ 9 μm	laser 1300 nm
1000BASE-CX	25 m	cavo STP (*shielded twisted pair*)	
1000BASE-T	25-100 m	4 coppie di cavo UTP	

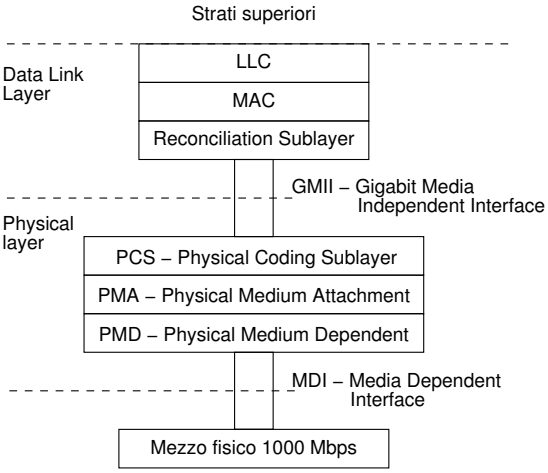
23.1.5.5 Packet bursting

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.

[51] Vedi ad es. `https://it.wikipedia.org/wiki/Fibre_Channel`

23.1.5.6 Architettura di Gigabit Ethernet

La figura a lato 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.

23.1.5.7 Ripetitore full-duplex e controllo di flusso

Se tutte le porte di un ripetitore sono di tipo full-duplex, allora non può più verificarsi contesa di accesso al mezzo; semmai 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.

23.1.5.8 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.

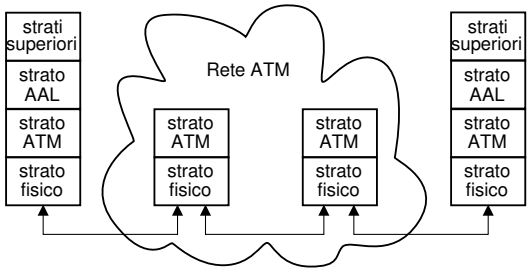
23.2 ATM

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-time[52], che per questo motivo è indicata anche come B-ISDN[53]. 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.

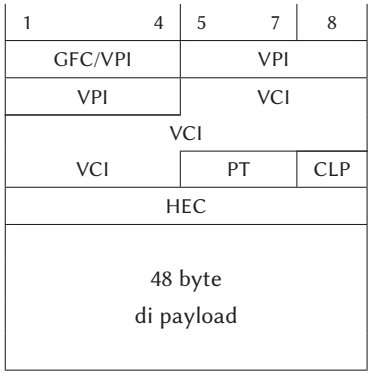
23.2.1 Architettura ATM

La rete ATM viene anche definita 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 23.1 riporta le principali funzioni svolte dai tre strati, e pone in evidenza come in uno stesso strato siano identificabili diverse sotto-funzioni.



23.2.2 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 (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.



Formato della cella ATM

In funzione del mezzo trasmissivo, può variare la *struttura di trama*[54] (mostrata in figura) in cui

[52] Per traffico real-time si intende sia quello telefonico, sia più in generale quello di natura multimediale.

[53] 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.

[54] 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).

strato	sotto-strato	funzioni
ATM Adaptation Layer (AAL)	• Convergenza (CS)	Definisce il servizio offerto agli strati superiori
	• Segmentazione e Riassemblaggio (SAR)	Suddivide i dati in modo compatibile con la dimensione di cella, e li ricostruisce in ricezione
ATM layer		Multiplazione e demultiplazione delle celle Traslazione delle etichette VPI/VCI Generazione/estrazione dell'HEADER della cella Gestione del controllo di flusso GFC
Physical Layer (PL)	• Convergenza di trasmissione (TC)	Delimitazione delle celle Inserimento celle IDLE per adattamento velocità Generazione e verifica dell'HEC (controllo di errore) Generazione della trama di trasmissione
	• Mezzo Fisico (PM)	Temporizzazione e sincronizzazione Gestione del mezzo

Tabella 23.1: Stratificazione delle funzioni in una rete ATM

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 uno[55]. Nel caso in cui la sorgente produca dati a velocità inferiore a quella del collegamento sono inserite celle aggiuntive di tipo IDLE, rimosse al ricevitore[56]. 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.

23.2.3 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)[57].

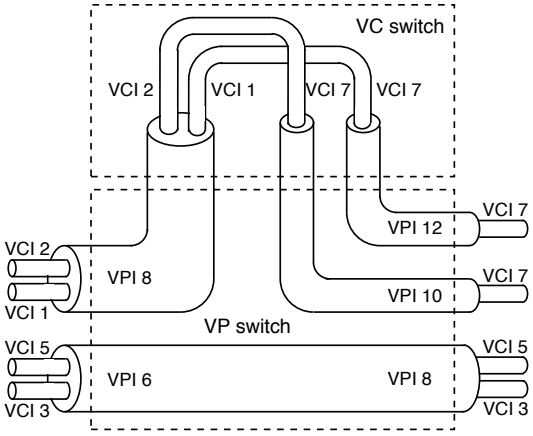
Il motivo della suddivisione risiede nella possibilità di raggruppare logicamente diversi circuiti virtuali che condividono lo stesso percorso nella rete. Nei collegamenti di cui è composto il percorso comune, viene usato uno stesso VPI per tutte le celle, mentre le diverse connessioni su quel percorso sono identificate mediante diversi VCI.

[55] 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.

[56] 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.

[57] Mentre per VCI sono riservati 16 bit, per VPI si usano 12 bit all'interno della rete, e 8 bit ai suoi bordi, riservando 4 bit indicati come *Generic Flow Control* (GFC) per regolare il flusso delle sorgenti.

L'instradamento congiunto delle celle con uguale VPI è effettuato nei nodi (VP *switch*), che si occupano solo[58] di scambiare il VPI delle celle, e di porle sulla porta di uscita corretta, come indicato dalle tabelle di instradamento.



La sequenza dei nodi attraversati dall'instradamento è indicata come *Virtual Path Connection* (VPC), è composta da zero o più VP switch, ed è delimitata tra due nodi (VC o VP/VC switch[59]) che elaborano anche i VCI. La sequenza dei VC switch che elaborano i VCI, e che si estende tra due nodi che terminano lo strato di adattamento, è indicata invece con il termine *Virtual Channel Connection* (VCC)[60] e comprende uno o più VPC, coincidendo spesso[61] con il percorso tra ingresso ed uscita[62] della rete ATM.

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)[63]; 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.

23.2.4 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

[58] Questa 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.

[59] Nel caso in cui venga invece scambiato solo il VCI, si ottiene uno switch VC puro.

[60] La 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.

[61] Può accadere infatti di incontrare uno switch VC puro, in cui è scambiato solo il VCI, ed al quale fanno capo due diverse VCC.

[62] I nodi di ingresso ed uscita sono indicati come *ingress* ed *egress* nella terminologia ATM.

[63] Nella 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.

nelle VCC siano conformi al rispettivo contratto, svolgendo un *Usage Parameter Control* (UPC) detto anche *policing*[64]. Prima di proseguire, forniamo però alcune definizioni.

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 voce[65] 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 indipendenti[66].

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 Tansfer 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

[64] Letteralmente: POLIZIOTTAMENTO. Il controllo può anche essere effettuato su di una intera VPC.

[65] La classe CBR si presta bene a trasportare traffico telefonico PCM. In questo caso, può trasportare solo gli intervalli temporali realmente occupati.

[66] La 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.

essere ridotta adottando degli *shaper*[67], 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 contrattuali[68] può intraprendere svariate azioni, tentando di non scartare immediatamente la cella, ma provvede comunque a segnalare l'anomalia, ponendo pari ad uno il bit *Cell Loss Priority* (CLP) dell'header. Ciò fa si che la cella divenga *scartabile*[69] 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.

23.2.5 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 (AFI). Nel primo caso, detto *formato NSAP*[70], 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 multiplare più entità presso il terminale, ed è ignorato dalla rete.

Rete Privata				
AFI	ICD/DCC	HO-DSP	ESI	SEL
Rete Pubblica				
AFI	E.164	HO-DSP	ESI	SEL

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.

23.2.6 Strato di adattamento

Come mostrato in tab. 23.1, l'AAL è suddiviso in due componenti, *Segmenting and Reassembly* (SAR) e *Convergence Sublayer* (CS); le funzioni di quest'ultimo sono ulteriormente

[67] Un *sagomatore* è composto in prima approssimazione da un buffer di memoria, il cui ritmo di svuotamento *non è mai* superiore ad un valore costante.

[68] Ad 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.

[69] Alcune 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 posso 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.

[70] Il formato NSAP si ispira al *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 23.2: Classi di servizio della rete ATM

ripartite tra una *Common Part* (CPCS) ed un *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 demultiplazione 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. 23.2 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, ed è 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*, 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 coda[71], 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 informazioni[72], 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.

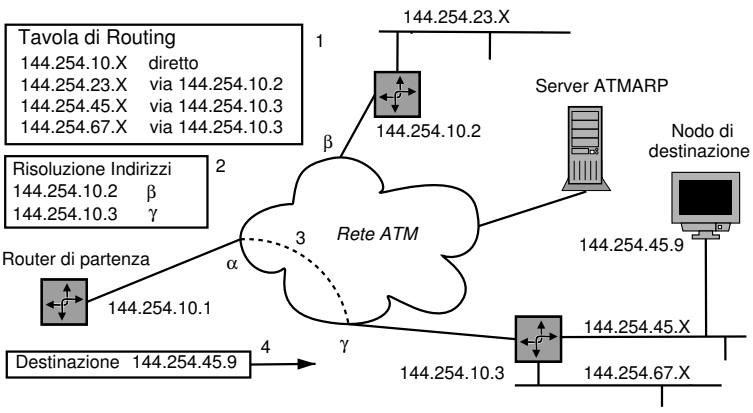
[71] Questi ultimi 4 byte contengono l'indicazione (2 bit) se si tratti della prima, ultima od intermedia cella di una stessa CS-PDU, la lunghezza dei dati validi se è l'ultima (6 bit), un numero di sequenza (4 bit), un controllo di errore (10 bit), ed una etichetta (10 bit) che rende possibile interallacciare temporalmente le celle di diverse CS-PDU.

[72] In 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.

23.2.7 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 distanza[73] 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 LIS[74], che risponde comunicando 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). Una tale 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 PVC[75], il pacchetto IP viene incapsulato in un header LLC IEEE 802.2

[73] Quando 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 cittadine. Per estensioni ancora maggiori si parla di *reti in area geografica*.

[74] Tutti i nodi della LIS hanno configurato manualmente l'indirizzo ATM del server ATMARP.

[75] Un 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 una stesso sistema

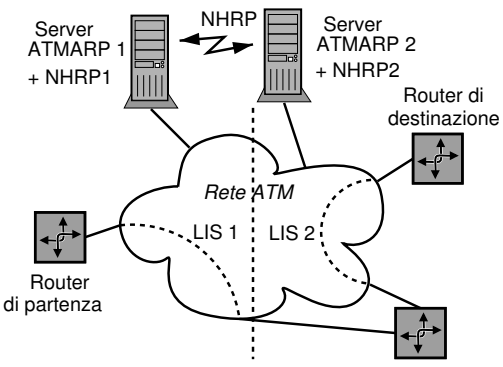
prima di essere consegnato all'AAL5. In tal modo, il router ricevente esamina l'header LLC del pacchetto ricevuto dal nodo ATM di *egress,* per consegnare il pacchetto al protocollo appropriato, realizzando così un *trasporto multiprotocollo* su ATM.

23.2.8 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 passaggi[76].

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 stessa, e risparmiando al router intermedio il compito di riassemblare 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 *<ind. IP; ind. ATM>* 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



autonomo, eliminando la necessità di sviluppare in proprio un impianto di TLC tra le sedi.

[76]La 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 ATM interrogando il LES, incapsula le trame IP con un header LLC IEEE 802.2 proprio come nel caso classico.

di *passa-parola*[77], 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[78]).

23.2.9 MPLS

Il *Multi Protocol Label Switching* (MPLS) è un 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)[79], 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 dalla *label*[80]. 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

[77] I 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.

[78] Il metodo si basa su di un meccanismo indicato come *flow detection*, attuato dal ruoter 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 ruoter 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*.

[79] Nel routing IP tradizionale, una FEC coincide con l'instradamento individuato dal *longest match*.

[80] Nel 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.

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 destinazione[81] in base al punto di ingresso, così come le FEC posso 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*[82], 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 istradamento 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 riassemblaggio 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.

[81] Il 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.

[82] Infatti, 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 24

Reti a commutazione di circuito

Ora facciamo *un passo indietro* rispetto al capitolo precedente, andando a riannodare la storia a partire dal punto in cui le reti telefoniche hanno iniziato a svilupparsi, trovandosi per prime a dover coniugare gli aspetti legati alla teoria del traffico con le tecnologie che via via si rendevano disponibili.

Dopo aver definito natura e scopo delle reti *di accesso* e *di trasporto*, ed aver sottolineato le differenze tra la multiplazione temporale di tipo *statistico* delle reti per dati e quella *deterministica* dei sistemi basati su trame e sulla commutazione di circuito, si ripercorre lo sviluppo delle reti dette appunto *a circuito*, dalla rete *plesiocrona* e la gerarchia PDH fino a quella *sincrona* o SDH, assieme alle diverse modalità di gestire la segnalazione, l'inserimento di tributari, la sincronizzazione, la struttura di trama ed i formati dei contenitori. Si affrontano quindi argomenti di natura *topologica*, ovvero di come i nodi della rete sono disposti sul territorio, definendo la natura dei dispositivi che operano su reti SDH basate su fibra ottica. Per ultimo viene affrontato il tema della *commutazione*, ovvero la funzione primaria dei nodi della rete che devono smistare le comunicazioni in transito attraverso instradamenti differenti, utilizzando matrici multistadio di commutazione spaziale, eventualmente corredate da elementi di commutazione temporale nei sistemi di multiplazione organizzati in trame.

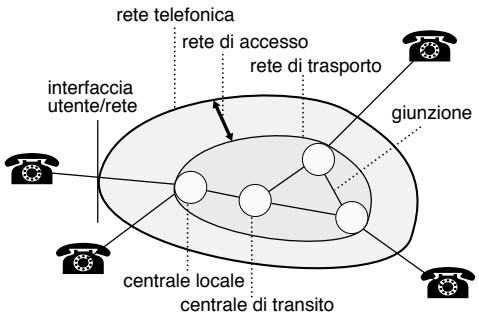
24.1 Introduzione

Rispetto ai collegamenti *punto-punto* in cui una unica sorgente informativa comunica con un unico destinatario, è assai più frequente il caso in cui i soggetti coinvolti affidano la comunicazione ad una *rete* di collegamento, consegnando il messaggio al *nodo* di commutazione a cui hanno accesso. Una volta individuato un percorso di *attraversamento* che coinvolga i nodi della rete più opportuni il messaggio giunge al destinatario, grazie anche all'impiego di informazioni addizionali dette *di segnalazione*.

Le nozioni che seguono fanno esplicito riferimento alle *reti di telefonia* per come si sono evolute a partire dalla fine dell'800, dette anche a *commutazione di circuito*. Gli aspetti legati alle *reti di trasmissione dati* che si sono sviluppate *durante e dopo* tale evoluzione sono sviluppati al cap. 23.

24.1.1 Elementi della rete telefonica

Con riferimento alla figura, discuteremo innanzitutto dei metodi di *multiplazione* 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, al § 24.8 esamineremo 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.



24.1.2 La rete di accesso

E' la parte più rilevante e *pesante* della rete, e consiste nel *doppino* in rame (pag. 650) 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 § 24.9.1;
- ISDN (*Integrated Service and Data Network*) - vedi § 24.9.2;
- ADSL (*Asimmetric Digital Subscriber Line*) - che in realtà usa POTS solo come tramite per raggiungere una *rete IP* vedi § 24.9.4.

Oltre a questi, nella rete di accesso sono contemplate forme di collegamento anche diverse dal cavo, come

- accesso ottico - come per l'FTTH (vedi § 19.3.4.3), permettendo
  - di interconnettere un insieme più numeroso di collegamenti POTS già multiplati assieme, come nel caso di un grosso centralino aziendale, sovrapponendosi 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
  - GSM[1] - noto anche come sistema cellulare di seconda generazione[2], 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 (pag. 300), che per il meccanismo di accesso multiplo al mezzo trasmissivo[3], che adotta una organizzazione in

[1] `http://it.wikipedia.org/wiki/Global_System_for_Mobile_Communications`

[2] La prima generazione si riferisce al sistema analogico TACS `http://it.wikipedia.org/wiki/TACS`

[3] `http://en.wikipedia.org/wiki/Time_division_multiple_access`

trame; inoltre, ha introdotto la comunicazione dei messaggi SMS[4];

– GPRS[5] e UMTS[6] - 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;
– WiFi[7] e WiMax[8] - 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. Entrambi 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

- 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.

24.2 Multiplazione

Il principio di 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, nel caso di sorgenti non continuamente attive (vedi § 22.3.4);
- semplificare la gestione e la manutenzione dei collegamenti a lunga distanza, essendo questi minori in numero.

Le tecniche di multiplazione possono operare secondo le modalità di

- *divisione di frequenza e di lambda* - ogni comunicazione usa una banda di frequenze diversa, come descritto al § 11.1.1 nel contesto dello studio dei segnali modulati, oppure al § 19.3.3.2 in quello della trasmissione su fibra ottica;
- *divisione di tempo* - ciascuna comunicazione avviene in intervalli di tempo diversi da quelli delle altre, facendo uso di canali numerici: viene affrontata nel resto di questo capitolo;

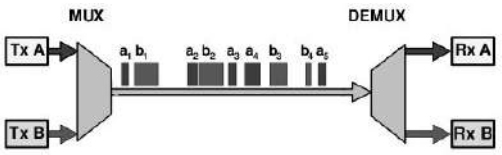
[4] `http://it.wikipedia.org/wiki/SMS`
[5] `http://it.wikipedia.org/wiki/General_Packet_Radio_Service`
[6] `http://it.wikipedia.org/wiki/Universal_Mobile_Telecommunications_System`
[7] `http://it.wikipedia.org/wiki/Wi-Fi`
[8] `http://it.wikipedia.org/wiki/WiMAX`

- *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, in virtù della *ortogonalità* tra i codici utilizzati (vedi § 7.6.2), come nella trasmissione *a spettro espanso* (§ 16.9);
- *divisione di spazio* - come nei sistemi *multiantenna* (cap. 21) che sfruttano la diversità di risposta in frequenza che si verifica sotto opportune condizioni.

24.2.1 Multiplazione a divisione di tempo

E' una modalità praticabile solo per segnali di natura *numerica*. Mentre per segnali campionati l'approccio è stato quello illustrato ai §§ successivi, per i segnali *dati* la tecnica di multiplazione è sempre stata basata sull'uso di un *pacchetto dati* (vedi § 22.5.1), attuando uno schema detto

Multiplazione statistica e commutazione di pacchetto In questo caso il mezzo trasmissivo non è impegnato in modo esclusivo, ma la trasmissione può avvenire in forma sporadica, 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 multiplazione dei pacchetti avviene in modo *statistico*, senza riservare risorse a questo o quel tributario: il multiplatore si limita ad inserire i pacchetti ricevuti in apposite code, da cui li preleva (con un *bit rate* maggiore) per poterli trasmettere in sequenza, attuando una modalità di trasferimento *orientata al ritardo* (vedi § 22.4). La presenza di code comporta

- il determinarsi di un ritardo variabile ed impredicibile
- la possibilità che la coda sia piena, ed il pacchetto in ingresso venga scartato

D'altra parte ogni pacchetto contiene le informazioni necessarie al suo recapito, facilitando l'instradamento (vedi § 24.7). A seconda dell'adozione di un principio di commutazione di tipo *a circuito virtuale* oppure a *datagramma* (vedi § 22.5.2.2), può essere presente o meno una *fase di setup* precedente l'inizio della comunicazione.

Multiplazione deterministica e commutazione di circuito La modalità usata nella rete telefonica è invece basata su di uno schema di multiplazione *con organizzazione di trama* (vedi § 22.5.2.1) e 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 veniva effettuata *manualmente* da parte di un centralinista umano, che creava un vero e proprio *circuito elettrico* 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* (§ 24.3.1), in cui viene riservato un intervallo temporale per ognuno dei flussi tributari.

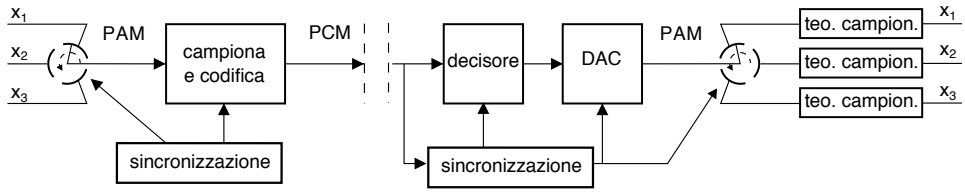
Ad ogni buon conto, si noti che un risultato della teoria del traffico (pag. 782) mostra come l'adozione di una strategia *orientata al ritardo* migliora notevolmente l'efficienza di utilizzo del mezzo stesso.

24.3 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 multiplazione sincrona (§ 24.4). 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.

24.3.1 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 § 4.3.2), e trasmessi (8 bit alla volta[9]) a turno su di un unico collegamento, producendo un segnale binario che prende il nome di PCM (*Pulse Code Modulation*[10]). In figura è evidenziato inoltre un blocco di sincronizzazione (§ 24.3.3) necessario a ricostruire la corretta sequenza ricevuta, in modo da redistribuire correttamente i campioni ai filtri di restituzione.



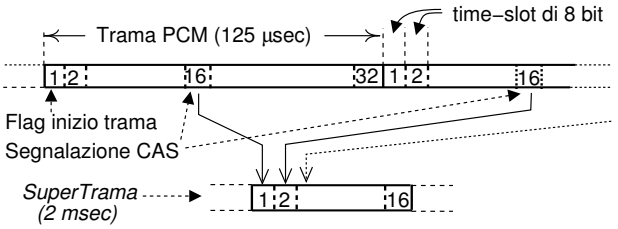
[9]La tecnica di multiplare un *blocco di bit* (in questo caso 8) alla volta prende il nome di *word interleaving*, distinto dal *bit interleaving*, in cui l'alternanza è a livello di bit.

[10]Il segnale PCM ispira il suo nome dal PAM (vedi § 24.9.5) in quanto ora, anziché trasmettere le *ampiezze* degli impulsi, si inviano i *codici* binari dei livelli di quantizzazione.

La struttura temporale ripetitiva che ospita i campioni dei singoli tributari prende il nome di *trama* (FRAME[11]), ed è composta da 32 intervalli temporali detti *time-slot*. Trenta di questi ospitano a turno i bit di un campione proveniente da un numero massimo di 30 tributari[12], mentre i rimanenti due intervalli convogliano le informazioni di segnalazione[13], che indicano lo stato dei singoli collegamenti (il 16^o intervallo) e forniscono il sincronismo relativo all'inizio della trama stessa (il primo).

Le trame si susseguono ad una velocità pari alla frequenza di campionamento di ciascun tributario e quindi abbiamo 8000 $\frac{trame}{sec}$; ognuno dei 32 $\frac{intervalli}{trama}$ ospita 1 $\frac{campione}{intervallo}$, a sua volta formato da 8 $\frac{bit}{campione}$, ottenendo così una velocità binaria complessiva di $8000 * 32 * 1 * 8 = 2048000$ $\frac{bit}{sec}$; per questo motivo ci si riferisce all'insieme come alla *trama PCM a 2 Mbit*. D'altra parte, essendo pari al periodo di campionamento la durata della trama è di 1/8000 = 125 μsec.

Il primo time-slot della trama contiene una configurazione di bit sempre uguale, chiamata FLAG (*bandiera*, vedi § 22.5.2.1), che ha lo scopo di indicare ai circuiti di sincronismo l'inizio della trama stessa.

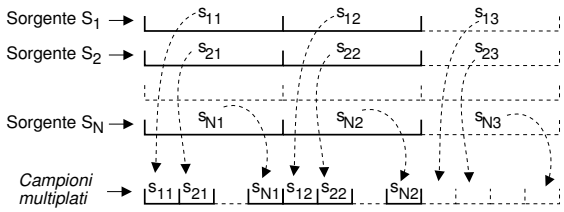


I dati di segnalazione contenuti nel 16^o intervallo devono essere *diluiti* su più trame, per poter rappresentare tutti i 30 tributari[14]. Si è stabilito che occorra prelevare il 16^o intervallo di 16 trame successive, per ricostruire una struttura detta *supertrama* (di 16*8 = 128 bit) che rappresenta le informazioni di tutti i tributari (disponendo così di 4 bit/tributario/supertrama), e che si ripete ogni 16*125 = 2000 μsec = 2 msec.

In effetti nel 16^o 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".

[11] FRAME significa più propriamente *telaio*, e in questo caso ha il senso di individuare una struttura, da "riempire" con il messaggio informativo.

[12] In 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).



[13] Vedi anche le sezioni 24.3.2 e 24.9.1.

[14] Gli 8 bit del 16^o intervallo sono infatti insufficienti a codificare lo stato dei 30 tributari che contribuiscono al segnale TDM.

24.3.2 Messaggi di segnalazione

Come illustrato al § 24.9.1 la rete di accesso è sede di uno scambio di informazioni tra terminale e centrale locale, detta *segnalazione di utente*, 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. L'inoltro di queste informazioni da parte della centrale di accesso verso la rete (e viceversa) può essere gestito 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^o 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[15]. 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 24.1 mostra come la numerazione venga recepita da un organo di *controllo centrale*, che provvede a impostare il dispositivo di commutazione (§ 24.8), 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.

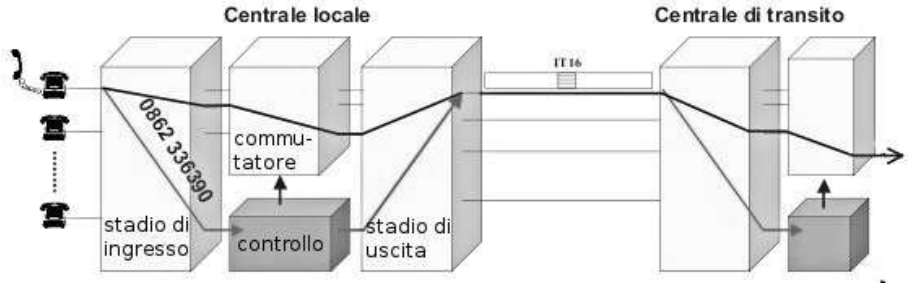


Figura 24.1: Controllo di centrale nel caso di segnalazione *associata al canale*

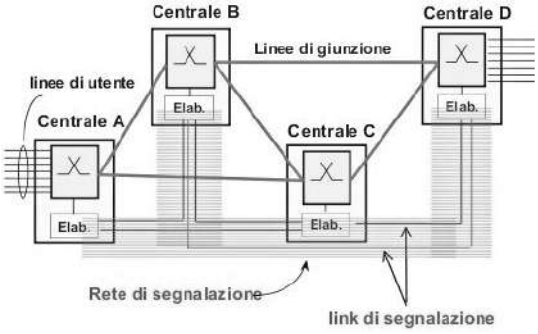
Segnalazione a canale comune Il primo passo evolutivo è stato quello di provvedere ad un *canale comune* di segnalazione direttamente collegato agli organi di controllo (vedi fig. 24.2), su cui poter convogliare la segnalazione relativa *a tutte* le comunicazioni in transito tra le due centrali.

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.

[15]Vedi ad es. `https://en.wikipedia.org/wiki/In-band_signaling`

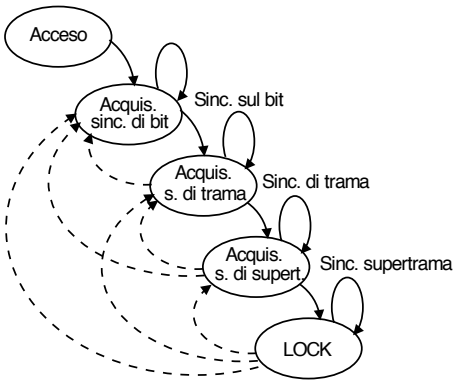
Figura 24.2: Separazione della segnalazione in un *canale comune*

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 § 24.9.3). Questo 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....



24.3.3 Sincronizzazione di centrale

Alla figura seguente 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[16]; 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. Per ogni stato esiste poi la possibilità (fortunatamente remota) di perdere il sincronismo ed *indietreggiare* (linee tratteggiate) nel diagramma di stato,



[16]Nel caso specifico, l'HDB3, pag. 448.

perdendo le comunicazioni in corso.

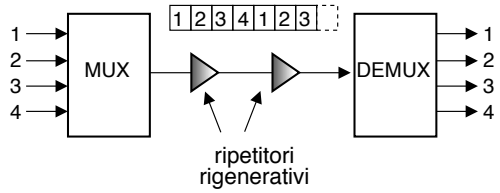
24.3.4 Multiplazione asincrona e PDH

L'argomento di questo paragrafo non va confuso con la *trasmissione* asincrona (quella START-STOP mostrata al § 15.7.1), che descrive una modalità di *inviare* informazioni numeriche; qui invece si tratta di *multiplare*, ossia come *mettere assieme* più comunicazioni.

Man mano che i nodi vicini alla *periferia* della rete di trasporto instradano il traffico verso quelli di livello gerarchico superiore, associati ad aree di influenza geografica più estesa (vedi § 24.5), i collegamenti di giunzione iniziano a trasportare un numero di tributari sempre più elevato, ottenuti *mettendo assieme* tutte le conversazioni contemporaneamente dirette verso la stessa destinazione. Considerando allo stesso tempo la necessità di dover svolgere nelle centrali la funzione di *commutazione*, è importante individuare metodi efficienti per *raggruppare* assieme più tributari, anche a velocità diverse, rendendo allo stesso tempo relativamente agevole *inserire* o *rimuovere* i singoli tributari. Rimandiamo al § 24.4 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 (§ 24.3.1), 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) mediante appositi dispositivi *multiplatori* (o MULTIPLEXER, o MUX). Il collegamento tra nodi può inoltre prevedere dispositivi detti *ripetitori rigenerativi* (uno o più di uno, vedi § 18.3.2) 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 loro velocità possono essere lievemente differenti l'una dall'altra[17], 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 CCITT[18], nota come *Plesiochronous Digital Hierarcy* (PDH), secondo la quale ad esempio 4 flussi da 2 Mbps (detti E1) sono multiplati in uno

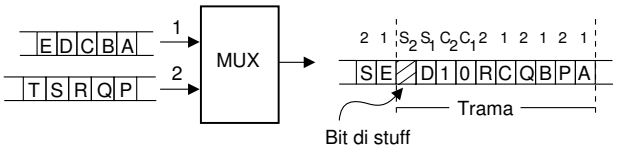
[17] Un oscillatore con precisione di una parte su milione, produce un ciclo in più o in meno ogni 10^6; ad una velocità di 2 Mbps, ciò equivale a un paio di bit in più od in meno ogni secondo.

[18] *Comité Consultif International pour la Telephonie et Telegraphie.* Questo organismo non esiste più. ed ora l'ente di standardizzazione ha nome ITU-T , vedi `https://it.wikipedia.org/wiki/ITU-T`.

da 8 Mb/sec (E2): notiamo che sebbene siano teoricamente sufficienti 8192 Mb/sec, in realtà il multiplexer ne produce di più (8448). La quantità in eccesso ha lo scopo di permettere la multiplazione di segnali non necessariamente sincroni, mediante la tecnica del *bit stuffing*[19].

24.3.4.1 Bit stuffing

Consideriamo l'esempio in figura, con 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 (ABC e PQR) vengono comunque alternativamente trasmessi, mentre il 4^o bit di ciascun flusso 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) contiene un dato valido oppure sia solo un *bit di stuff*, 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 il *destuffing*, e riottenere i flussi originari. Nella realtà le informazioni di controllo sono molto ridondate, perché se scambiassimo un bit di stuff per uno buono (o viceversa), distruggeremmo anche la struttura di trama del tributario che ha subito l'errore.

24.3.4.2 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 multiplato di ordine elevato, ovvero realizzare una funzione detta *Add and Drop*. In questo caso è infatti necessario eseguire un'operazione inversa a quella di multiplazione, ovvero (vedi fig. 24.3) demultiplare l'intero flusso, compresi tutti gli altri tributari, e successivamente ri-multiplare di nuovo il tutto.

Questa caratteristica limita notevolmente la flessibilità delle configurazioni di rete ottenibili con questa tecnologia, e per i tributari passanti comporta l'aggiunta di un

[19] Da: TO STUFF = *riempire.*

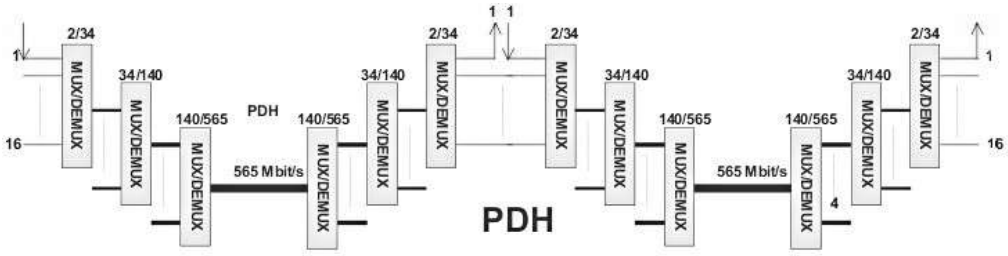


Figura 24.3: Gerarchia di multiplazione PDH e complessità di un ADD AND DROP MULTIPLEXER

ritardo temporale causato dalle operazioni di demultiplazione e ri-multiplazione. Nella pratica vengono usati solo flussi di tipo E1, E3 ed E4, che sono quelli più adatti per essere trasportati nella gerarchia sincrona SDH, multiplando direttamente sedici tributari a 2 Mbit/s all'interno di un unico flusso a 34 Mbit/s.

24.3.5 Sincronizzazione di rete

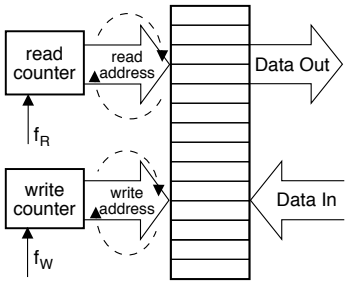
Se tutti i nodi della rete operassero alla stessa velocità, non sussisterebbero problemi nella multiplazione di più tributari. Nel caso in cui la sincronizzazione tra nodi sia completamente affidata ad un orologio di centrale di elevata precisione, si verifica il caso di funzionamento *plesiocrono*, 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 i 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 segnale proveniente da un satellite in orbita terrestre.

24.3.5.1 Elastic store

Si tratta di un accorgimento[20] idoneo ad *assorbire* le fluttuazioni della velocità di trasmissione, come ad esempio nel caso della sincronizzazione mutua. Mentre il *bit stuffing* (§ 24.3.4.1) è adottato nella multiplazione 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 (§ 24.8).

E' 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 incrementa[21] 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



[20] Letteralmente: *magazzino elastico*.

[21] Il 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.

trama, e nulla più[22].

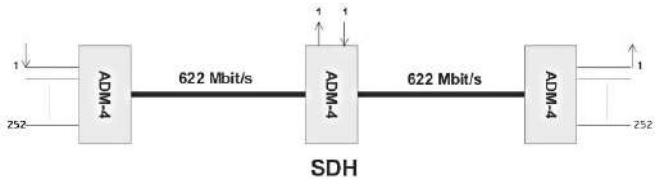
24.4 Gerarchia digitale sincrona

Definizione dei livelli gerarchici Come anticipato, la *Synchronous Digital Hierarcy* (SDH[23]) è una metodologia di multiplazione che presuppone un funzionamento perfettamente sincrono degli elementi di rete, ed ha solo una variante (nel Nord 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 24.1 elenca le velocità del *payload*[24] e di trasmissione associate ai diversi livelli della gerarchia di multiplazione SDH/SONET; la sigla STM sta per *synchronous transport module* ed il numero che segue indica il numero di flussi STM-1 che sono aggregati.

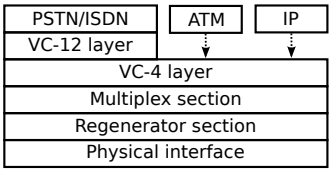
SONET	SDH	payload (kbps)	v. trasm. (kbps)
STS-1	-	48.960	51.840
STS-3	STM-1	150.336	155.520
STS-12	STM-4	601.344	622.080
STS-24	STM-8	1.202.688	1.244.160
STS-48	STM-16	2.405.376	2.488.320
STS-96	STM-32	4.810.752	4.976.640
STS-192	STM-64	9.621.504	9.953.280
STS-768	STM-256	38.486.016	39.813.120
STS-1536	STM-512	76.972.032	79.626.120

Tabella 24.1: Nomenclatura della gerarchia ottica e relative velocità

Multiplexer Add and Drop La differenza strutturale rispetto al 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 L'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.



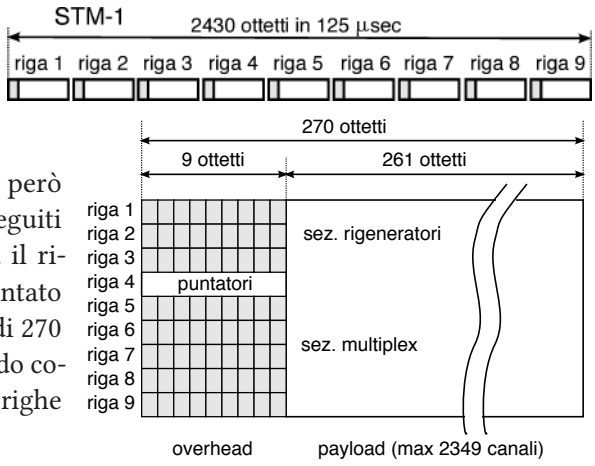
[22] Infatti il sincronismo di trama viene preservato; inoltre l'evento di sovrapposizione dei puntatori può essere rilevato, e segnalato ai dispositivi di demultiplazione, in modo che tengano conto dell'errore che si è verificato.

[23] `http://it.wikipedia.org/wiki/Synchronous_Digital_Hierarchy`

[24] Con il termine *payload* si indica il *carico pagante*, ossia i dati che vengono trasportati

Struttura di trama La gerarchia SDH si basa su di una struttura di trama di durata di 125 μsec, 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.

Synchronous Transport Module STM-1 Il livello più basso della gerarchia è indicato come STM-1, opera ad una velocità di 155.52 Mbit/s, può trasportare 63 flussi PDH E1 (ovvero 63 flussi * 32 timeslot/flusso = 2016 canali PCM) multiplati mediante una trama composta da 2430 ottetti, di cui 81 di segnalazione (in grigio) e 2349 di dati[25], 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^o, 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 multiplazione, ossia finalizzate all'espletamento di funzioni OAM (*Operation*, *Administration*, *Maintenance*), che sono ora associate ad un annidamento di sezioni di trasmissione: *path*, *multiplazione* e *rigenerazione*, vedi fig. 24.4. Il percorso (*path*) compiuto da un

[25]Notiamo 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*.

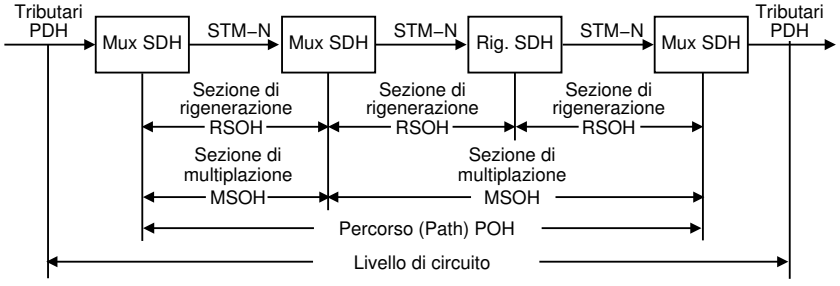


Figura 24.4: Definizione delle sezioni di path, multiplazione e rigenerazione di un multiplex SDH

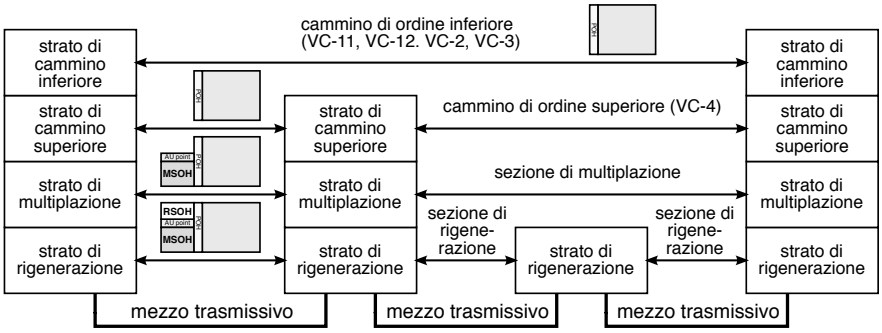
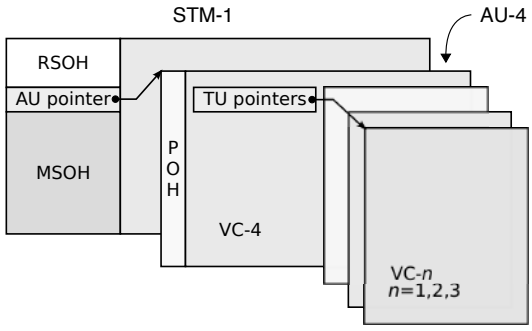


Figura 24.5: Stratificazione funzionale della segnalazione per multiplazione sincrona

singolo tributario si snoda infatti tra un unico multiplatore di ingresso ed un unico demultiplatore di uscita, ma ad ogni multiplatore *add and drop* (o commutatore) incontrato, viene definita una nuova *sezione di multiplazione*. Allo stesso modo, per ogni ripetitore rigenerativo incontrato (§ 18.3.2), è definita una nuova *sezione di rigenerazione*. Per ognuna di queste sezioni, è definito un *overhead* (OH) specifico per le operazioni OAM associate.

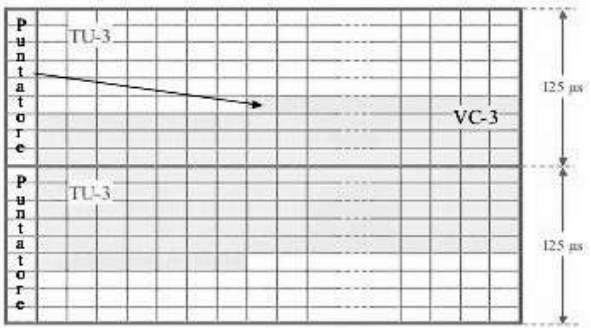
Dato che un ADM è anche rigeneratore, e che i dispositivi di ingresso / uscita del tributario sono anche ADM, si determina la *stratificazione funzionale* per la segnalazione schematizzata in fig. 24.5, in cui è evidenziato come l'overhead associato alle sezioni più esterne venga *impilato* su quello delle sezioni interne. Ma a differenza dell'incapsulamento (pag. 22.5.2.3) proprio 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 multiplazione. 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.

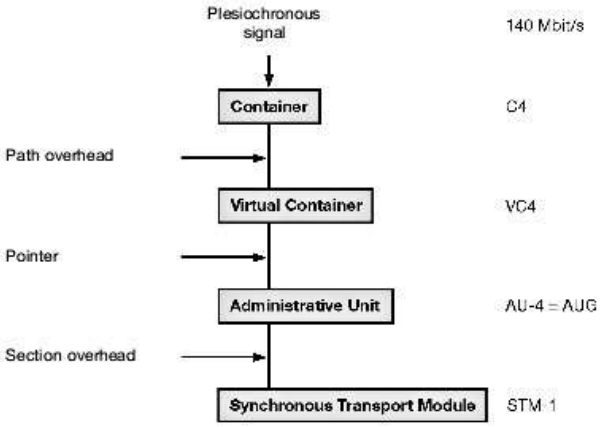


La presenza del puntatore AU deriva dalla volontà di ridurre al minimo l'uso di buffer e di evitare l'introduzione di ritardi di consegna; pertanto i dati da trasmettere *non* vengono inseriti nella struttura di trama all'inizio della stessa, bensì *al primo ottetto possibile* al momento della disponibilità dei dati stessi. Quindi, è più che normale il

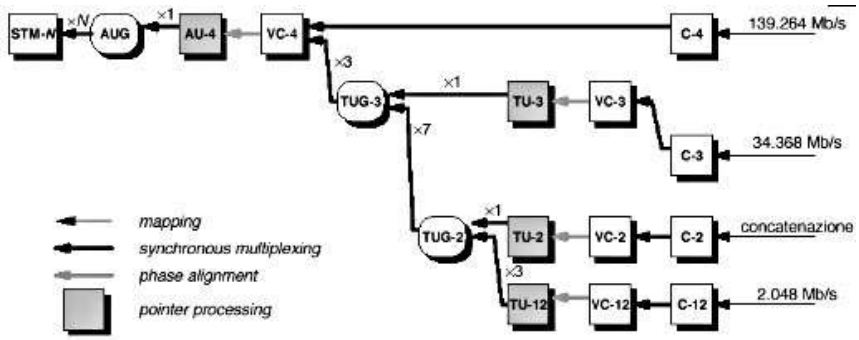
caso in cui la AU inizi a metà di una trama, e termini a metà della trama successiva, come illustrato nella figura che a lato. La coppia AU ed AU Pointer prende quindi il nome di *Administrative Unit Group* (AUG).



Virtual Container e Tributary Unit Il riempimento della AU con i dati da trasmettere, avviene (vedi figura seguente) mediante una serie di passi successivi. Viene per prima creata la 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.



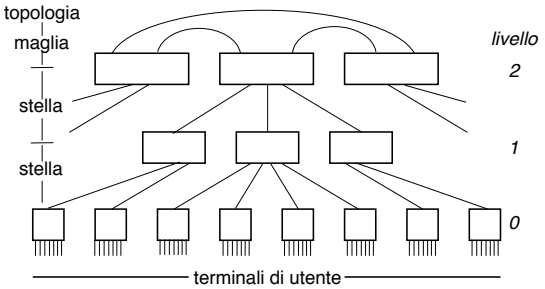
Non approfondiamo oltre questo argomento, che richiede una buona dose di pazienza per essere analizzato a fondo, e ci limitiamo ad inserire un diagramma che mostra le possibilità di combinazione di tributari differenti, in accordo alle specifiche di ETSI.



Esercizio Quanti 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∗3∗7∗3 = 1890 canali voce. In alternativa, qualora VC4 contenga un unico flusso E4 a 140 Mb, 1920 canali volce.

24.5 Topologia di rete



Nella figura a lato 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 discusso al capitolo 22, il numero dei possibili collegamenti contemporaneanei 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 realizzano gli instradamenti necessari alla comunicazione 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 consentono anche di instradare i collegamenti verso altre reti (ad es. di altre nazioni od operatori). Le centrali di 2^o 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.

Nello specifico caso italiano[26], la denominazione attribuita alle centrali dei diversi livelli segue lo schema mostrato nella parte sinistra della figura 24.6: i terminali di

[26] Sto parlando di qualcosa la cui odierna evoluzione mi è ignota, ma la storia merita di essere raccontata.

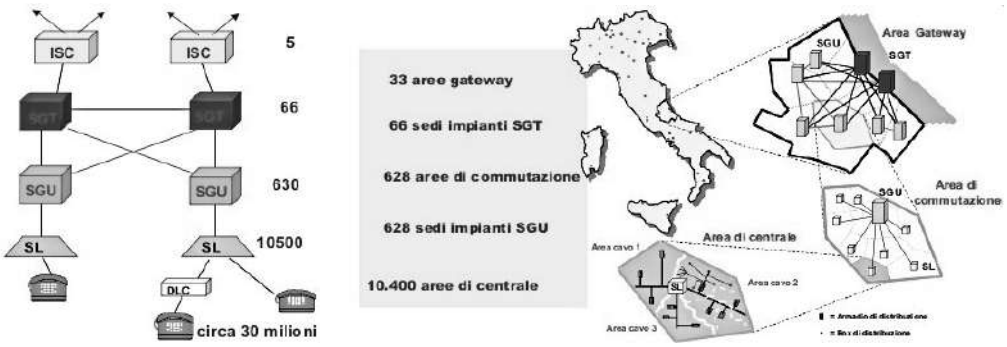


Figura 24.6: Topologia della rete telefonica Italiana

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.

24.6 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 § 19.3; nel seguito illustriamo i dispositivi utilizzati, la topologia risultante, ed i sistemi di protezione.

24.6.1 Dispositivi SDH

Come anticipato, la trasmissione SDH si avvale di elementi (vedi Fig. 24.7) che possono essere descritti in termini funzionali secondo la seguente classificazione:

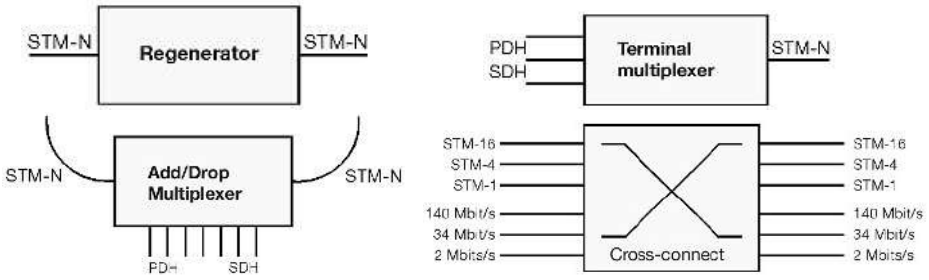


Figura 24.7: Dispositivi SDH

Rigeneratori Sono gli elementi di base, che consentono di suddividere in più tratte i collegamenti più lunghi, e che eliminano dal segnale in transito gli effetti del rumore e della dispersione temporale.

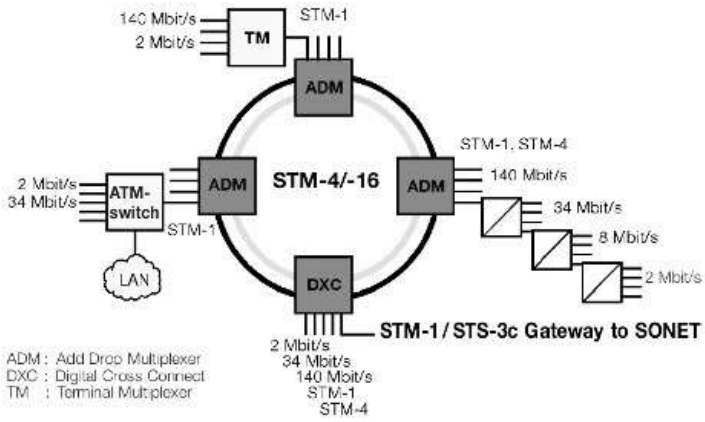
Multiplatori Combinano tributari PDH ed SDH, in modo da inserirli in flussi a velocità più elevate.

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 da un flusso entrante all'interno di un diverso flusso uscente, realizzando così la funzione di commutazione.

24.6.2 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.



24.6.2.1 Rete di trasporto

Al 2002, l'interconnessione dei collegamenti SDH nazionali risultava permessa dalla struttura su tre livelli riportata in Fig. 24.8.

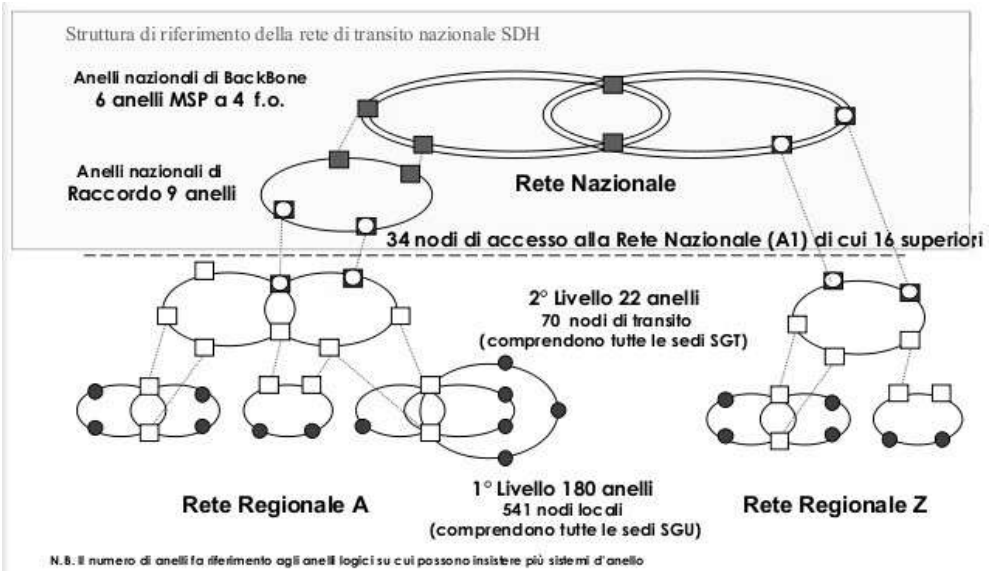


Figura 24.8: Struttura collegamenti SDH nazionali - anno 2002

24.6.2.2 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. 24.9 mostra alcuni casi pratici di accesso in fibra ottica. Iniziando da destra, sono mostrate delle reti Gigabit Ethernet (pag. 815) residenziali, interconnesse mediante *switch di livello 2* ad un POP (*Point of Presence*), il cui router (12000) si interconnette ad un anello SDH a 622 Mbps, sul quale sono instradati i pacchetti IP diretti verso

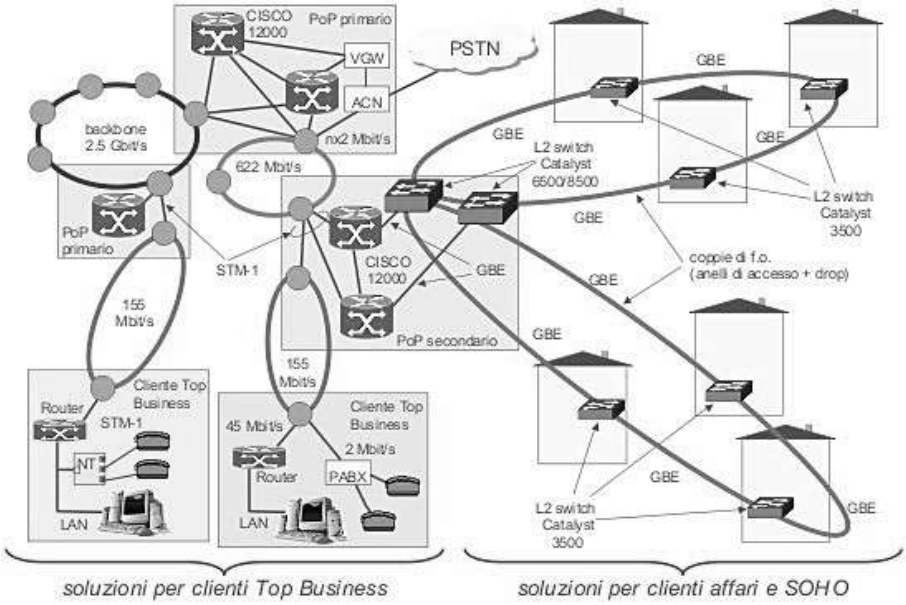


Figura 24.9: Diffusione della fibra ottica nella rete di accesso

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.

24.6.3 Sistemi di protezione automatica

L'abbondanza di ottetti OAM nella multiplazione SDH permette un monitoraggio costante della qualità del collegamento e di eventuali malfunzionamenti, al punto che gli stessi apparati di commutazione sono in grado di svolgere compiti di rimpiazzo automatico tra la linea andata fuori servizio, ed una riserva presente, come indicato nei seguenti schemi.

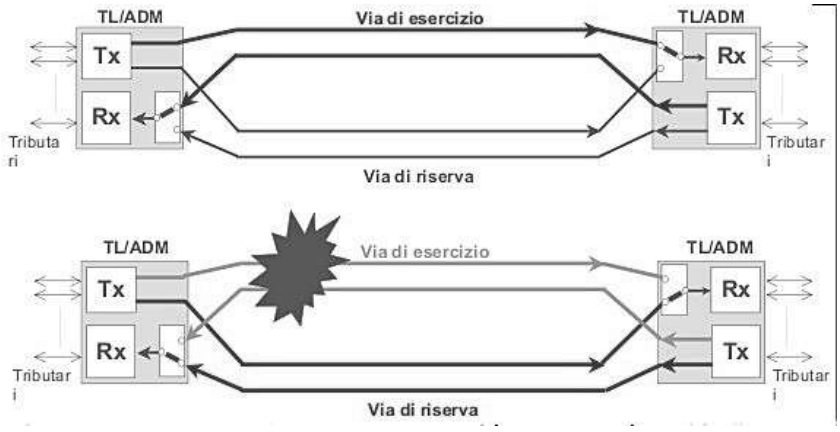


Figura 24.10: Sistema di protezione 1+1

Protezione 1+1 In questo caso, ogni collegamento (vedi fig. 24.10) 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 nel caso precedente, per ognuna delle due direzioni di trasmissione è impegnata una diversa fibra ottica, ma in questo caso la via di ritorno (vedi fig. 24.11) si sviluppa investendo l'altra metà dell'anello, percorso 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.

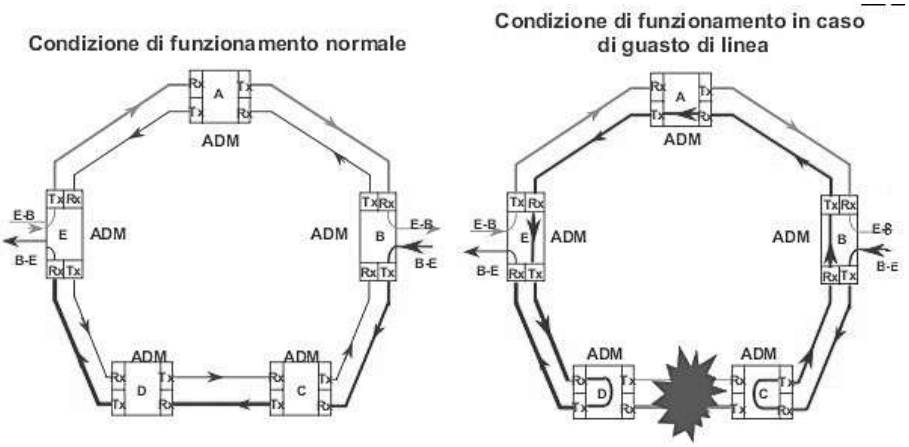


Figura 24.11: Configurazione ridondante ad anello

24.7 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 interni alla centrale e diretti verso le 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 dinamiche, 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);

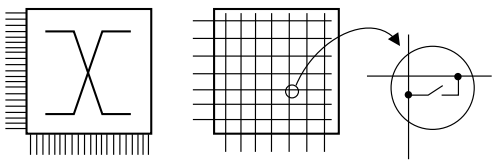
Tramite ccs (*Common Channel Signaling*, segnalazione a canale comune): le decisioni sull'instradamento sono demandate ad una rete di segnalazione parallela ed indipendente da quella del traffico smaltito, e che collega tutte le centrali ad un unico organo di controllo (il *canale comune*), il quale 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.

24.8 Commutazione

Illustriamo ora l'architettura dei 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. L'altra modalità di commutazione, quella di *pacchetto*, è stata illustrata al cap. 22.

24.8.1 Reti a divisione di spazio

Individuano 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*[27], 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 pensato come una matrice binaria in cui ogni elemento (1 o 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...

24.8.2 Reti multistadio

...di cui alla figura 24.12 è riportato un esempio a 3 stadi, con gli N ingressi che 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 CLOS[28]). Una connessione da sinistra a destra ha ora la

[27] come 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*.

[28] E' una condizione *sufficiente* a scongiurare il blocco anche nella condizione *peggiore*. Tale circostanza si verifica quando:

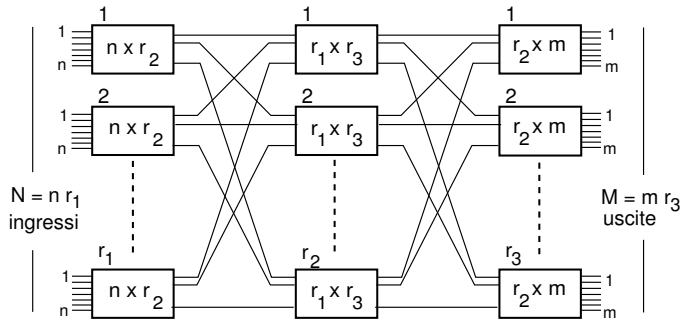


Figura 24.12: Rete a divisione di spazio a tre stadi

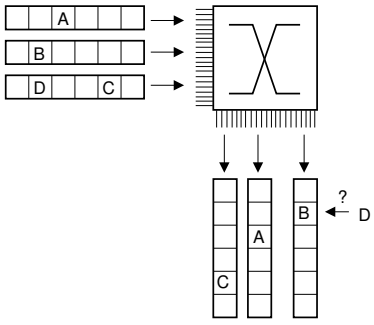
possibilità di scegliere attraverso quale matrice intermedia passare.

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*.

24.8.3 Commutazione numerica a divisione di tempo

Consideriamo il caso in cui si debbano commutare le comunicazioni associate ai singoli *time-slot* presenti in diversi flussi[29] numerici organizzati in trame. Avendo a disposizione 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*.



24.8.3.1 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

▹ una matrice del primo stadio (i) ha $n - 1$ terminazioni occupate

▹ una matrice del terzo stadio (j) ha $m - 1$ terminazioni occupate e

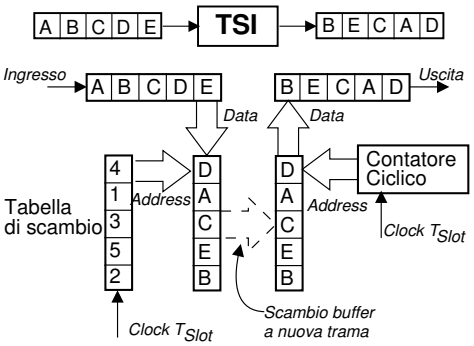
▹ tali terminazioni non sono connesse tra loro, anzi le connessioni associate impegnano ognuna una diversa matrice intermedia e

▹ si richiede la connessione tra le ultime due terminazioni libere di i e j

$\Rightarrow$ in totale si impegnano allora $m - 1 + n - 1 + 1 = m + n - 1$ matrici intermedie.

[29] Le comunicazioni presenti in uno stesso flusso, ovvero appartenenti alla stessa trama, condividono la stessa origine/destinazione.

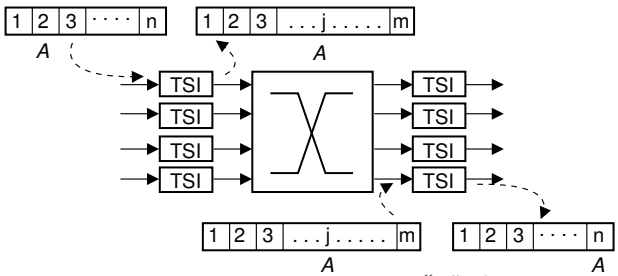
averne cambiato l'ordine temporale. In figura è mostrato un possibile schema di funzionamento: una trama entrante viene scritta, agli indirizzi ottenuti leggendo sequenzialmente la tabella di scambio, in un buffer di memoria (*es.: entra* E *e lo scrivo al* 4^o *posto, poi entra* D *e va al* 1^o *posto, etc.*). Prima dell'inizio di una nuova trama, il primo buffer è copiato in un secondo[30], 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.



24.8.3.2 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 riescono 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.

Notiamo subito come in questo schema il numero di intervalli temporali *in uscita* dai TSI di ingresso è maggiore di quelli *in ingresso* ossia $m > n$ ([31]): 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^o 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 $m = 2n - 1$([32]).



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 !

[30] La tecnica prende il nome di *double buffering*.

[31] Ovviamente, $m - n$ intervalli sono lasciati vuoti, in ordine *sparso* tra gli m.

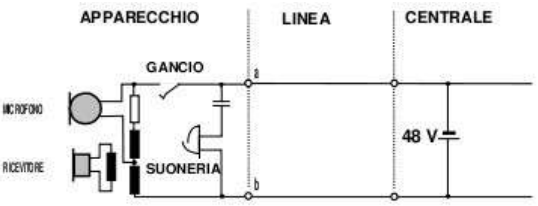
[32] Si confronti questo risultato con la condizione di Clos, fornita al § 24.8.2.

24.9 Appendici

Dopo aver illustrato le tecniche realizzative più generali, approfondiamo il funzionamento di una serie di metodi di accesso che hanno partecipato alla evoluzione delle reti a commutazione di circuito, fino alla moderna ADSL.

24.9.1 Plain old telephony services (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[33], 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 *allertiamo* 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*[34], 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 combinatore è stato poi soppiantato dalla attuale tastiera numerica DTMF (*Dual Tone Multi Frequency*), in cui ad ogni tasto (vedi lato sinistro della fig. 24.13) sono 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 Hz[35], le cui durate sono descritte in basso a sinistra nella figura.

[33] In inglese si dice andare OFF-HOOK, con riferimento storico al gancio su cui riporre la cornetta, presente nei primi modelli di telefono.

[34] `http://en.wikipedia.org/wiki/Hybrid_coil`

[35] corrispondente al *la* centrale del pianoforte. Ho provato a verificare, e... a me arriva un la bemolle!

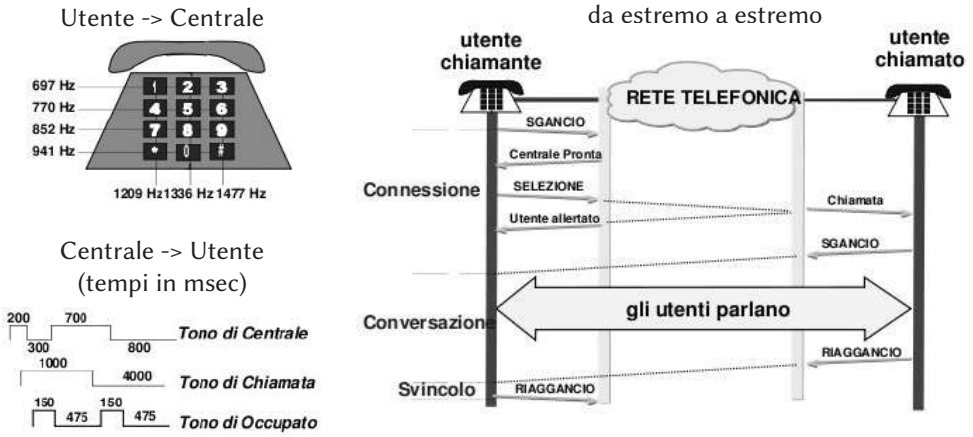


Figura 24.13: Segnalazione di utente

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 *istradamento*, 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. 24.13, 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.

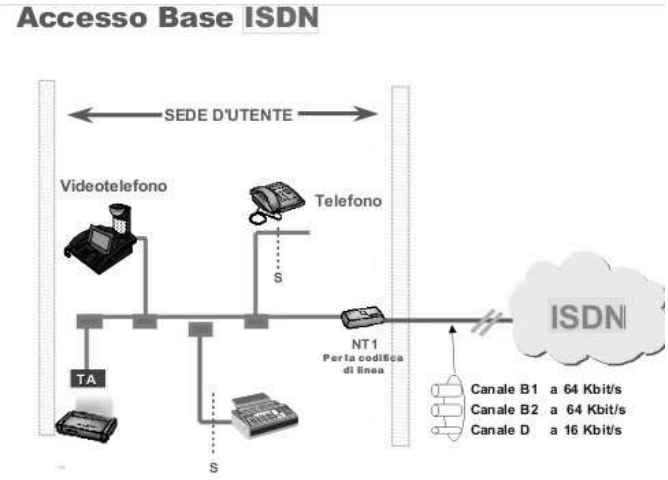
Il passaggio dalla telefonia analogica a quella numerica, in cui il segnale vocale è campionato e quantizzato come PCM, non ha di fatto alterato la presenza di queste tre fasi.

24.9.2 ISDN

La *Integrated Service Data Network*[36] è una modalità di accesso *numerico* alla rete telefonica, definito da una serie di standard reperibili presso l'ITU[37]. In ISDN la conversione A/D avviene all'interno del terminale di utente, il quale può collegare allo stesso bus

[36] `https://it.wikipedia.org/wiki/Integrated_Services_Digital_Network`
[37] `http://www.itu.int/rec/T-REC-I/e`

ISDN (interfaccia S[38] a quattro fili, utilizzante 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*[39], su cui è trasmesso un segnale a quattro livelli noto come 2B1Q[40], 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 trama 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 segnalazione[41], come il protocollo Q.931[42].

Nella modalità di *Accesso Primario* (PRI, *Primary Rate Interface*), adatta al collegamento di centralini, si hanno a disposizione 30 canali B (voce) a 64kbps, ed un canale D (dati) di segnalazione a 64 kbps. Pertanto, PRI viene direttamente interconnesso al primo livello (E1) della gerarchia PDH descritta al § 24.3.4.

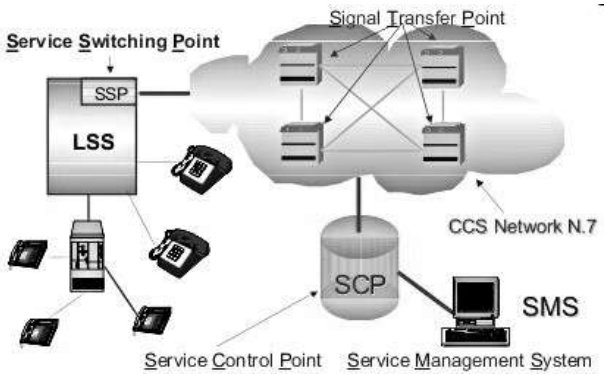
Dato che l'accesso ISDN preserva il flusso binario inviato sui canali B da estremo a estremo della rete, su quegli stessi canali possono essere inviate anche informazioni niente affatto vocali, ma bensì nativamente numeriche, purché il ricevente condivida le stesse modalità di interpretazione dei bit in arrivo. Sfruttando tale possibilità, sono stati (ad esempio) definiti i primi standard di videotelefonia H.320[43].

24.9.3 Sistema di segnalazione numero 7

Il *Signaling System #7* (SS7[44]) è un insieme di protocolli di segnalazione telefonica a canale comune, usato per controllare la maggior parte delle chiamate telefoniche della PSTN

[38] http://hea-www.harvard.edu/~fine/ISDN/n-isdn.html
[39] http://www.ralphb.net/ISDN/ifaces.html
[40] https://it.wikipedia.org/wiki/2B1Q
[41] http://www.rhyshaden.com/isdn.htm
[42] https://en.wikipedia.org/wiki/Q.931
[43] https://it.wikipedia.org/wiki/H.320
[44] https://en.wikipedia.org/wiki/Signalling_System_No._7

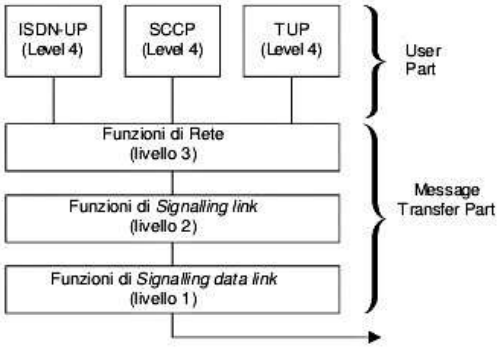
mondiale, che in questo caso prende il nome di *Intelligent Network* (IN[45]). Oltre ad gestire la fasi di instaurazione e abbattimento della chiamata, permette altri servizi come reindirizzamento, carte prepagate, SMS, numero verde, conferenza, richiamata su occupato...



L'SS7 è descritto dalla serie di raccomandazioni ITU-T Q.700[46], 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* (SSP[47]), che termina la segnalazione di utente, ed invia una query all'SCP per determinare come gestire la richiesta di servizio;
- *Signal Transfer Point* (STP[48]), che instrada i messaggi SS7 tra le diverse entità della IN;
- *Service Control Point* (SCP[49]), che interroga un *Service Data Point* (SDP[50]), 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* (INAP[51]) che opera sopra il *Transaction Capabilities Application Part* (TCAP) della pila protocollare SS7.

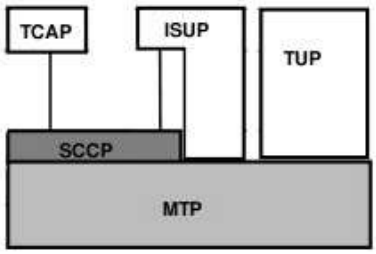
Oltre alle entità che prendono parte alla architettura, SS7 è definito anche nei termini della gerarchia protocollare che descrive la stratificazione delle funzioni necessarie allo svolgimento dei servizi richiesti. Il semplice scambio dei messaggi tra le entità è basato su di una rete a commutazione di pacchetto, ed avviene in base alle procedure collettivamente in-



[45] https://en.wikipedia.org/wiki/Intelligent_Network
[46] https://www.itu.int/rec/T-REC-Q.700/en
[47] https://en.wikipedia.org/wiki/Service_switching_point
[48] https://en.wikipedia.org/wiki/Signal_Transfer_Point
[49] https://en.wikipedia.org/wiki/Service_control_point
[50] https://en.wikipedia.org/wiki/Service_data_point
[51] https://en.wikipedia.org/wiki/INAP

dicate come *Message Transfer Part* (MTP[52]), 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.

Al *disopra* della MTP possono operare diversi protocolli detti di *User Part*, come ad esempio il *Signalling Connection Control Part* (SCCP[53]), 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* (TCAP[54]).

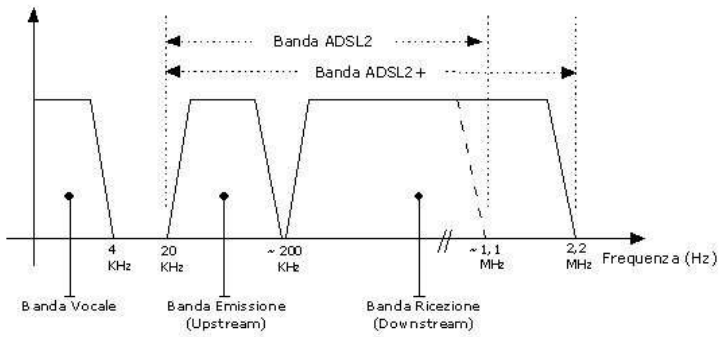


Altri esempi di User Part sono la *Telephone User Part* (TUP[55]) e la ISDN *User Part* (ISUP[56]). TUP è stata la prima UP ad essere definita, e fornisce il supporto all'offerta di servizi PSTN mediante la rete SS7. Attualmente è quasi ovunque rimpiazzato da ISUP, che offre altri servizi, come ad esempio 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 SIGTRAN[57] o *Signalling Transport*.

24.9.4 ADSL

L'*Asymmetric Digital Subscriber Line*[58] 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 doppino telefonico (*subscriber loop*) in rame, già utilizzato per il normale servizio telefonico POTS. 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 fi-

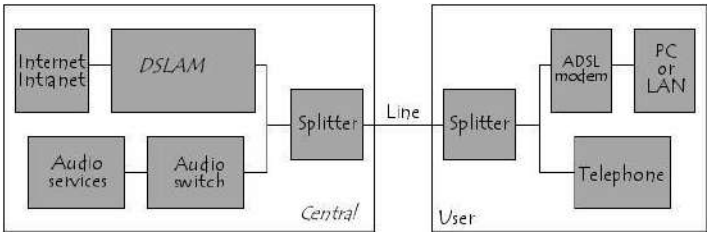


[52] https://en.wikipedia.org/wiki/Message_Transfer_Part
[53] https://en.wikipedia.org/wiki/Signalling_Connection_Control_Part
[54] https://en.wikipedia.org/wiki/Transaction_Capabilities_Application_Part
[55] https://en.wikipedia.org/wiki/Telephone_User_Part
[56] https://en.wikipedia.org/wiki/ISDN_User_Part
[57] https://en.wikipedia.org/wiki/SIGTRAN
[58] https://en.wikipedia.org/wiki/Asymmetric_digital_subscriber_line

gura, dove sono rappresentati gli intervalli di frequenza riservati alla telefonia PSTN, ai dati in uscita (*upstream*) ed in ingresso (*downstream*).

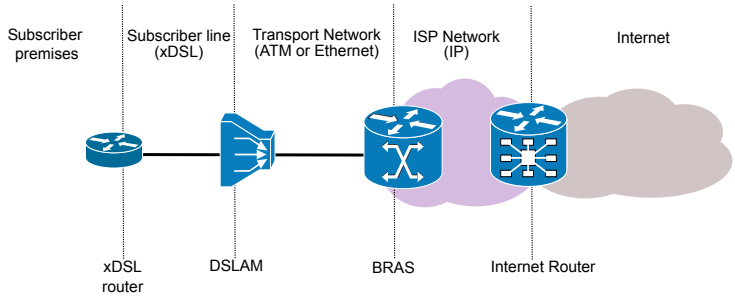
La massima velocità di trasmissione è stata inizialmente posta rispettivamente pari ad 1 ed 8 Mbps per i due versi trasmissivi, anche in funzione della lunghezza del collegamento utente - centrale[59]; successivamente, la massima velocità di ricezione è stata elevata rispettivamente a 12 e 20 Mbit/sec per gli standard ADSL2 e ADSL2+.

I due segnali (vocale e numerico) sono poi separati su linee differenti[60] inserendo, a valle della presa telefonica casalinga, un doppio filtro passa-alto e passa-basso detto *splitter*[61]. 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 trasporto[62] basata su ATM o ETHERNET, che termina il traffico sul *Broadband Remote Access Server* (BRAS)[63] dell'ISP, utilizzato da quest'ultimo anche per terminare il protocollo di strato di collegamento PPP[64], svolgere le funzioni di autenticazione dell'accesso, ed applicare eventuali *policy* a livello IP. Quindi, l'ISP provvede ad interconnettere il traffico del



[59] All'aumentare della lunghezza del collegamento, oltre a ridursi la potenza ricevuta e quindi peggiorare l'SNR, aumenta l'entità della diafonia (di tipo NEXT, pag. 648) tra utenti differenti, determinando un ulteriore peggioramento di SNR, che la tecnica di modulazione affronta riducendo la velocità trasmissiva.

[60] Ciò permette di non ascoltare nel telefono il *fruscio* della trasmissione ADSL, e di ridurre il rischio che le comunicazioni vocali determino *la caduta* della connessione ADSL.

[61] I modem più recenti incorporano il passa alto al loro interno, e sono venduti assieme a splitter con la sola funzione passa basso per il canale vocale.

[62] 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.

[63] `https://en.wikipedia.org/wiki/Broadband_remote_access_server`

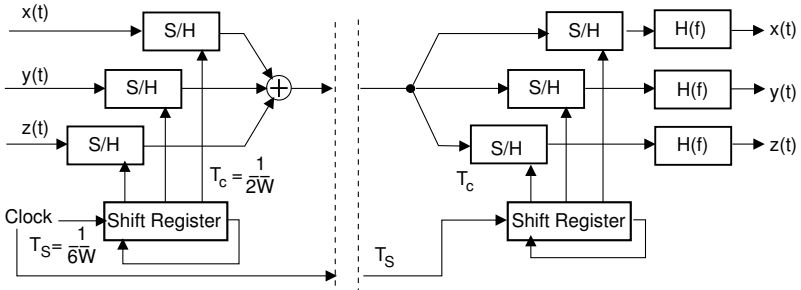
[64] `https://en.wikipedia.org/wiki/Point-to-Point_Protocol`

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 clienti[65], che in questo caso producono direttamente traffico IP, inoltrato verso la *core network* dell'ISP usando la sua stessa connettività.

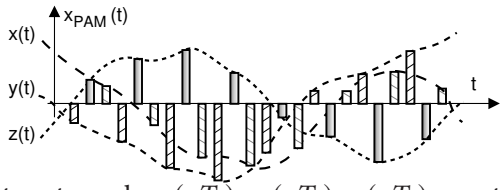
DMT Il modem ADSL utilizza una tecnica di modulazione numerica multi-portante detta *Discrete Multi Tone*, in cui il flusso binario viene ripartito su più canali di frequenza contigui, ed il segnale analogico sintetizzato direttamente nel dominio della frequenza mediante il calcolo di una FFT, come previsto dalla tecnica di trasmissione OFDM (vedi § 16.8). 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.

24.9.5 TDM mediante modulazione di ampiezza degli impulsi

Al tempo in cui la realizzazione del componente di quantizzazione (vedi § 4.3.1.1) presentava discrete difficoltà circuitali, si pensò[66] di sfruttare il teorema del campionamento (vedi § 4.1) per inviare su di un unico collegamento più comunicazioni multiplate a divisione di tempo (TDM = *Time Division Multiplex*). E' sufficiente infatti sommare alla funzione $x^{\circ}(t)$ introdotta al § 4.2.4 altri segnali simili, ad esempio $y^{\circ}(t)$, $z^{\circ}(t)$ come mostrato alla figura seguente, ognuno campionato a frequenza f_c, ma sfasato rispetto agli altri.



Da questa modalità di multiplazione 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 multiplati. 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 è mostrato a lato, e può 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



[65] Come ad es, nel caso dell'Unbundling: `https://it.wikipedia.org/wiki/Unbundling_local_loop`

[66] 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.

facendone passare gli impulsi campionati a frequenza f_c in un filtro di ricostruzione con banda $W \leq f_c/2$.

24.10 Riferimenti

Per questo capitolo un po' particolare, si elencano in modo distinto alcune fonti on-line a cui ci si è ispirati, e dalle quali sono state tratte alcune illustrazioni. Purtroppo mi avvedo che ora (2022) la quasi totalità degli indirizzi risulta *non più raggiungibile! :-(*

- 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/CorsiBO/RetiLB/lucidi/ISDN.pdf` di *Giorgio Corazza* `http://www-tlc.deis.unibo.it/Didattica/CorsiBO/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 25

Broadcast

Ritroviamo qui alcuni argomenti che nelle precedenti edizioni erano inseriti come appendici di alcuni capitoli, e come casi di applicazione delle teorie esposte alla realtà quotidiana che ci circonda. Nel frattempo la realtà si è evoluta, cosicché le tecniche televisive da analogiche sono diventate numeriche. Ad ogni modo, sono raggruppati qui i casi di *radiodiffusione* audio e video, come la radio FM, la televisione analogica, e quella satellitare, senza particolari miglioramenti od aggiunte rispetto a quanto già sviluppato a suo tempo.

25.1 Trasmissione televisiva analogica

Illustriamo molto brevemente le modalità di codifica e trasmissione del segnale televisivo mediante broadcast *analogico*, con riferimento a standard e tecnologie ormai *dismessi*[1], e che hanno rappresentato una delle più diffuse applicazioni della modulazione a *banda laterale ridotta* discussa al § 12.1.3.

25.1.1 Codifica di 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 volte[2] ogni $\frac{1}{25} = 0.04$ secondi (seppure in modo alternato) portando così a 50 semiquadri/secondo[3] la frequenza di rinfresco, in modo da impedire i fenomeni di *sfarfallamento* (FLICKER) ottico[4].

[1] Lo *switch off* al digitale terrestre in Italia è avvenuto nel luglio 2012, ed anche i *tubi catodici* sono pressoché estinti..

[2] La riproduzione di metà quadro alla volta è chiamata *scansione interallacciata* dell'immagine. Nulla vieta al costruttore del ricevitore di prevedere una *memoria di quadro* e di riprodurre le immagini in modo non interallacciato; il segnale trasmesso invece presenta sempre le righe in formato interallacciato.

[3] La 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.

[4] Il *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.

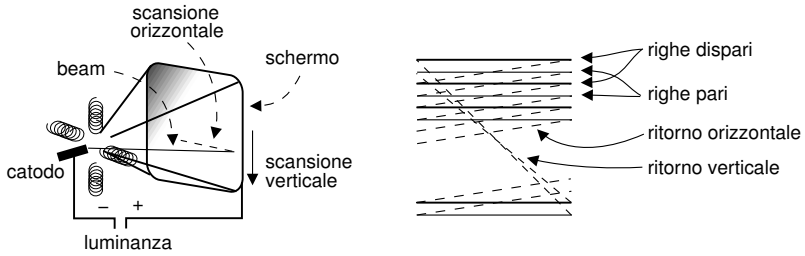


Figura 25.1: Modalità di scansione interallacciata dell'immagine televisiva

La riproduzione di un quadro avviene (vedi fig. 25.1) mediante un *tubo catodico*, il quale dispone posteriormente di un catodo che emette elettroni, accelerati da un segnale di luminanza positivo applicato all'anodo, e che terminano la 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 $\frac{linee}{quadro} \cdot 25 \frac{quadri}{secondo} = 15625$ Hz (*frequenza di riga*), sia verticalmente con velocità di 50 $\frac{semiquadri}{secondo}$.

25.1.2 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 un semiquadro ogni riga dispone di $\frac{1}{15625} = 64$ μsecondi. Il segnale modulante è sempre positivo (vedi fig. 25.2), ed associa ai valori più piccoli la maggiore luminanza[5], trasmettendo in logica negata, in modo che gli impulsi di sincronismo orizzontale siano di ampiezza superiore al *livello del nero*, pari quest'ultimo al 70 % dell'ampiezza massima. Il tempo dedicato alla trasmissione della luminanza di una riga è di 52 μsec, mentre nei restanti 12 il segnale oltrepassa il livello del nero (in modo da rendere invisibile il beam) e quindi un impulso rettangolare determina il ritorno orizzontale. In figura è anche mostrato il *burst colore* che è presente nelle trasmissioni a colori per sincronizzare la *portante di colore* (vedi di seguito).

[5]In questo modo si riduce mediamente la potenza trasmessa, dato che sono più frequenti scene chiare che scure.

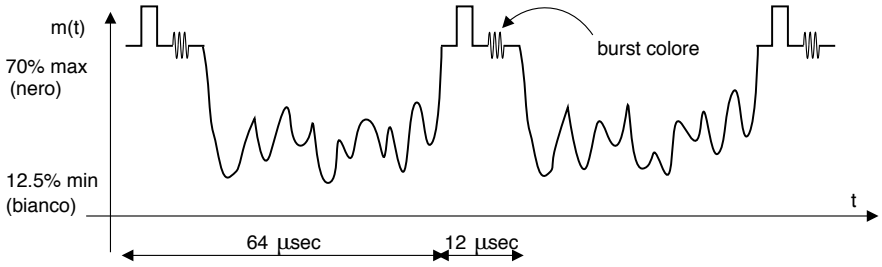


Figura 25.2: Forma d'onda del segnale televisivo analogico

25.1.3 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, mentre 25 linee per ogni semiquadro cadono al di fuori dello schermo[6].

25.1.4 Occupazione spettrale

Diverse considerazioni[7] 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 superiore *ridotta* (§ 12.1.2), conseguendo un risparmio di banda e contemporaneamente preservando le componenti del messaggio a frequenze più basse.

La figura 25.3 mostra la situazione in forma schematica, 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 $\underline{H}(f)+\underline{H}^*(-f) = cost$ (vedi nota 29 a pag. 377).

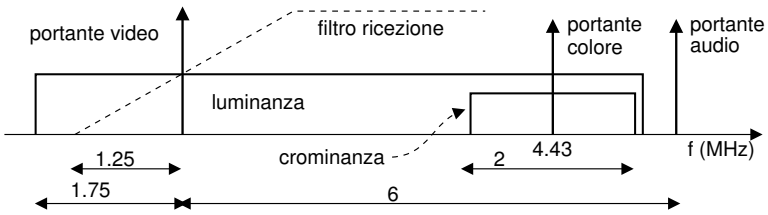


Figura 25.3: Occupazione spettrale di un segnale televisivo

Segnale audio In figura 25.3 è rappresentata anche una portante audio, che viene trasmessa *oltre* la banda occupata dal segnale video, mediante una modulazione FM con Δf_{Max} = 25 KHz.

25.1.5 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 dei ricevitori in bianco e nero.

Un risultato di colorimetria è che ogni colore è scomponibile nella somma di tre colori fondamentali (verde, blu e rosso o GBR che sta per *green*, *blue* and *red*),

[6] Nel tempo destinato alle linee che non sono mostrate vengono comunque trasmesse altre informazioni, come ad esempio i dati che compaiono nelle pagine del televideo.

[7] Ad 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 $625 \cdot \frac{4}{3} = 833$ $\frac{punti}{linea}$, e quindi $833 \cdot 625 = 520625$ $\frac{punti}{quadro}$, ossia circa $13 \cdot 10^6$ $\frac{punti}{secondo}$. Per il teorema del campionamento, il segnale deve allora avere una banda minore od uguale di $\frac{f_c}{2}$ = 6.5 MHz.

effettivamente operata dagli apparati di acquisizione. La somma[8] 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 ottenuto a partire da una coppia di *segnali differenza* di rosso e blu rispetto alla luminanza, ossia $\begin{cases} \Delta_R = R - L \\ \Delta_B = B - L \end{cases}$, usati poi per modulare in ampiezza, portante soppressa, una portante di colore, usando Δ_R come componente in fase e Δ_B come componente in quadratura[9]. Una analisi più precisa è fornita al § 10.2.2.

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.

25.1.6 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 (fig. 25.2), 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.

25.1.7 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 ridotto[10] alla crominanza. Quest'ultima quindi, prima di essere demodulata, 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 μsec, che determina uno spettro con energia concentrata alle armoniche di f_r = 15625 Hz. Pertanto la portante di colore

[8] In 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.

[9] Le 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.

[10] Possiamo 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 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 righine bianche e nere, divenire cangianti ?

viene collocata *nel mezzo* a due armoniche del segnale di luminanza[11], in modo che le densità spettrali risultino, pur se sovrapposte, intercalate. L'uso di *filtri a pettine*[12] nel ricevitore può quindi ridurre notevolmente l'interferenza.

25.1.8 Video composito o separato

Il segnale video in *banda base* realizzato come ora descritto, e privo della componente audio, può essere trasferito tra dispositivi e/o distribuito via cavo coassiale, e prende il nome di *segnale video composito*[13], contrapponendosi al *separate video* o S-VIDEO[14], in cui invece luminanza e crominanza vengono mantenute separate, ed utilizzato ad esempio nei cavi con attacco SCART[15].

25.2 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. Attualmente tale tecnica è affiancata a quella di natura digitale nota come DAB[16].

Ad ogni emittente radio FM è concessa una deviazione massima della frequenza istantanea rispetto alla portante assegnata pari a Δf = 75 KHz, con il trasmettitore che viene *tarato* mediante un messaggio $m(t)$ sinusoidale a frequenza di 15 KHz, mentre il fattore k_f è regolato in modo da ottenere Δf = 75 KHz. In queste condizioni si ottiene un indice di modulazione $\beta = \frac{k_f}{w} = \frac{75}{15} = 5$, e la regola di Carson (pag. 389) fornisce

$$B_C = 2\left(k_f + w\right) = 2\left(75 + 15\right) = 180\ kHz$$

Un esame degli andamenti riportati in Fig. 12.5 mostra che per $\beta = 5$, le funzioni di Bessel per cui risulta $\mathcal{J}_n(\beta) \neq 0$ sono le prime 8, e dunque la "*vera*" banda ha una estensione

$$B = 2 \cdot 8w = 16 \cdot 15 \cdot 10^3 = 240\ kHz$$

mostrando l'approssimazione della regola di Carson. D'altra parte, risulta che

$$2\sum_{n=6}^{8} |\mathcal{J}_n(5)|^2 = 2\left[(.13)^2 + (.05)^2 + (.02)^2\right] = 2 \cdot 0.0198 = 0.0396$$

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 $\pm W = \pm 15$ 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 data al § 12.3.3.4 per

[11] La portante di colore si colloca a 4,43361975 MHz per il sistema PAL.

[12] Introdotto al § 5.2, l'argomento può essere approfondito presso `https://it.wikipedia.org/wiki/Filtro_comb`

[13] Vedi `https://it.wikipedia.org/wiki/Video_composito`

[14] Vedi `https://it.wikipedia.org/wiki/S-Video`

[15] Vedi `https://it.wikipedia.org/wiki/SCART`

[16] `https://it.wikipedia.org/wiki/Digital_Audio_Broadcasting`

l'indice di modulazione *per processi*

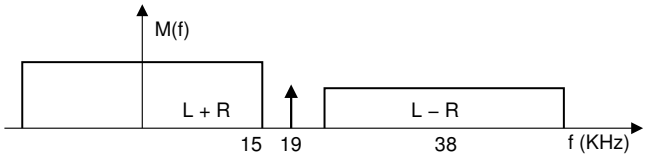
$$\beta_p = \frac{\sigma_f}{W} = \frac{k_f\sqrt{P_m}}{W} = \beta\sqrt{P_m} = \beta\frac{1}{\sqrt{2}} = 0.707 \cdot \beta$$

a cui corrisponde una banda *efficace*

$$B = 2W\left(\beta_p + 1\right) = 2 \cdot 15 \cdot 10^3 \cdot (0.707 \cdot 5 + 1) = 136\ kHz$$

Nell'FM *stereo* il segnale trasmesso deve essere compatibile con i ricevitori mono; pertanto il segnale *modulante* viene realizzato come un segnale multiplato FDM, e *composto* da tre *canali*, schematizzati come riportato in figura:

- 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;
- una portante posta a 19 KHz, ed 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 ciò non avviene, per due motivi:

- quando L+R è *grande*, significa 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.

25.3 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 25.4, assieme all'*ipsogramma*[17] relativo.

25.3.1 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 (§ 10.3.1.4) del segnale televisivo, ottenendo un flusso numerico chiamato PS (*Program Stream*);
- più PS sono pacchettizzati e multiplati (§ 10.3.2.1) 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. 471) a tre stadi (§ 17.4.2.6), in cui è prima applicato un codice di *Reed-Solomon* (§ 17.4.1.4), poi un *interleaver*, (§ 15.6.2.3) e quindi un codificatore *convoluzionale* (§ 17.4.2), rendendo il segnale particolarmente robusto nei confronti degli errori di trasmissione sia singoli che a burst;
- il nuovo flusso numerico è modulato QPSK (a due bit per simbolo, § 16.2.1) 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 § 15.2.2.3).

25.3.2 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 irradiata[18], che si riduce notevolmente nella trasmissione da terra a satellite. Considerando una portante di 2 GHz e la quota di un satellite in orbita geostazionaria[19] (36.000 Km da terra), l'attenuazione di spazio

[17] Dal greco *hypsos* 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 come varia il livello di potenza in funzione della portata di un collegamento.

[18] Più 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.

[19] Un 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,

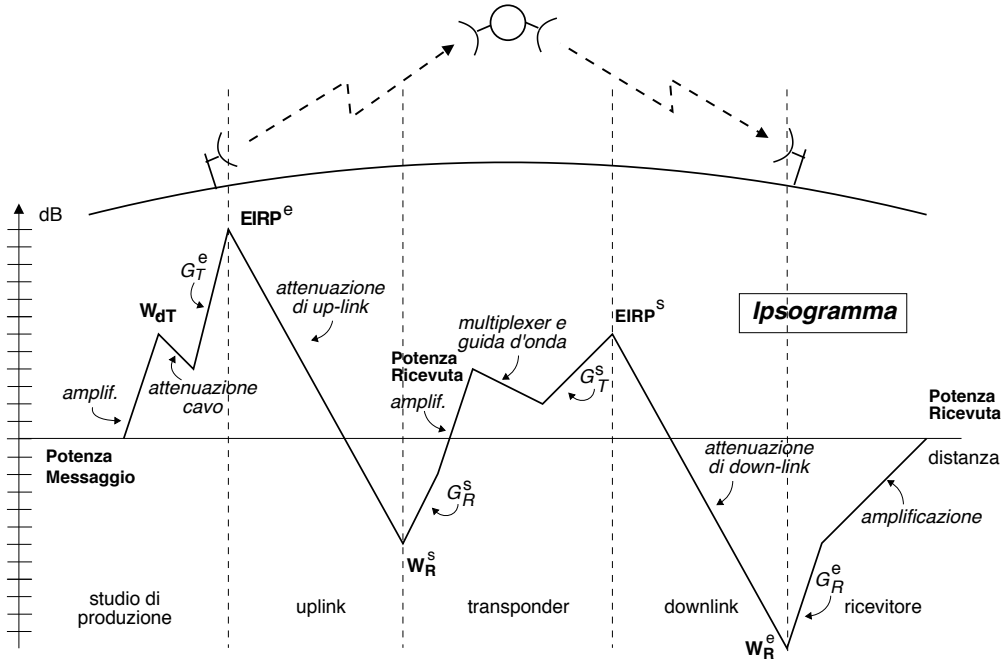


Figura 25.4: Andamento del livello di potenza in dB per un collegamento satellitare

libero dell'Up-Link (eguale a quella del Down-Link da satellite a terra) è di circa 190 dB.

25.3.3 Transponder

Il segnale ricevuto, di potenza W_R^s, è captato dall'antenna ricevente del satellite di guadagno G_R^s, e quindi ulteriormente amplificato, dopodichè si verificano alcune perdite di potenza nel collegamento con l'antenna *trasmittente* del satellite, di guadagno G_T^s, determinando così il valore della EIRP^s all'uscita del *trasponder* 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 insostenibile termine di *interferenza* per il lato ricevente del satellite, nonostante l'elevata direttività delle antenne, dando così luogo ad un fenomeno di *diafonia*[20]. La Fig. 25.5 mostra come il segnale a banda larga (che trasporta molteplici canali) ricevuto da terra viene prima filtrato alla banda del segnale utile, quindi amplificato

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.

[20]Le considerazioni sulla diafonia si applicano altrettanto bene anche al caso di ripetitori terrestri.

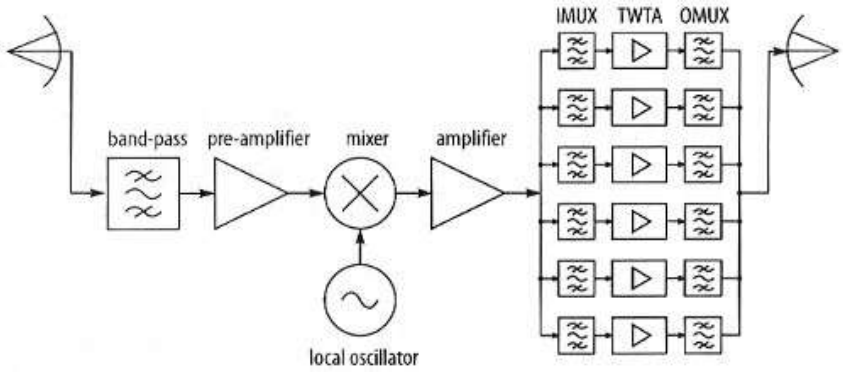
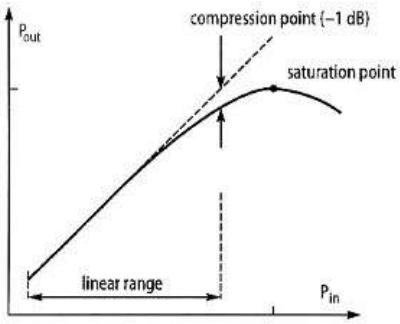


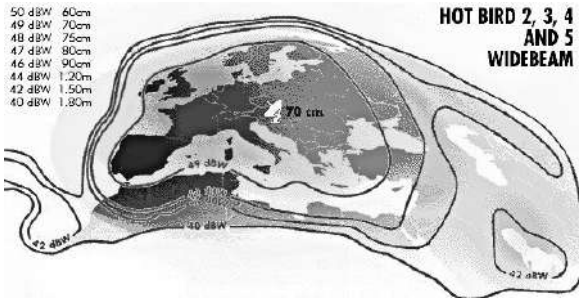
Figura 25.5: Elaborazione di bordo per un trasponder DVB satellitare

una prima volta, poi miscelato con un oscillatore locale[21], ed infine amplificato una seconda volta[22]. I singoli canali FDM che compongono il segnale sono quindi separati tra loro mediante il banco di filtri passa-banda indicati come IMUX (*input multiplexer*), e amplificati individualmente mediante dei TWTA[23] che, se spinti alla massima potenza, presentano una caratteristica ingresso-uscita non lineare (vedi § 8.3), mostrata nella figura a lato. Nel caso di trasmissioni modulate angolarmente la distorsione in ampiezza è ben tollerata (§ 13.3.3), e quindi l'entità del *back-off* di potenza (§ 13.3.1) può essere mantenuta limitata; 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.



25.3.4 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



[21] Come descritto al § 12.2.7, 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.

[22] La suddivisione della amplificazione in due stadi a frequenza diversa previene fenomeni di reazione positiva.

[23] *Travelling Wave Tube Amplifier*, ovvero tubi amplificatori ad onda progressiva: `https://it.wikipedia.org/wiki/Travelling_wave_tube`.

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 realizzare un *phased array*[24]. 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.

25.3.5 Temperatura di antenna

Come illustrato a pag. 669 una antenna ricevente è schematizzabile come un generatore controllato, ed al § 8.4.2.1 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 $\mathcal{W}_{dn}(f) = \frac{1}{2}kT_g$, in cui T_g ora viene detta *temperatura di antenna* T_a, e assume un valore inferiore ai 290 ^{o}K, e precisamente compreso tra i 15 ed i 60 ^{o}K. La fonte diretta di rumore, in questo caso, è il *rumore galattico*, la cui temperatura si abbatte a 10 ^{o}K sopra i 2,5 GHz, mentre i *lobi laterali* del diagramma di radiazione captano il rumore legato alla temperatura terrestre[25].

25.3.6 Ricevitore a terra

La figura 25.6 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 coassiale[26] che collega l'antenna al ricevitore casalingo. Quindi, si ritrova uno 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.

25.3.7 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

[24] `https://en.wikipedia.org/wiki/Phased_array`

[25] Per 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.

[26] Come descritto nel paragrafo che discute dell'*effetto pelle* (pag. 646), l'attenuazione in dB del cavo aumenta con l'aumentare della radice della frequenza.

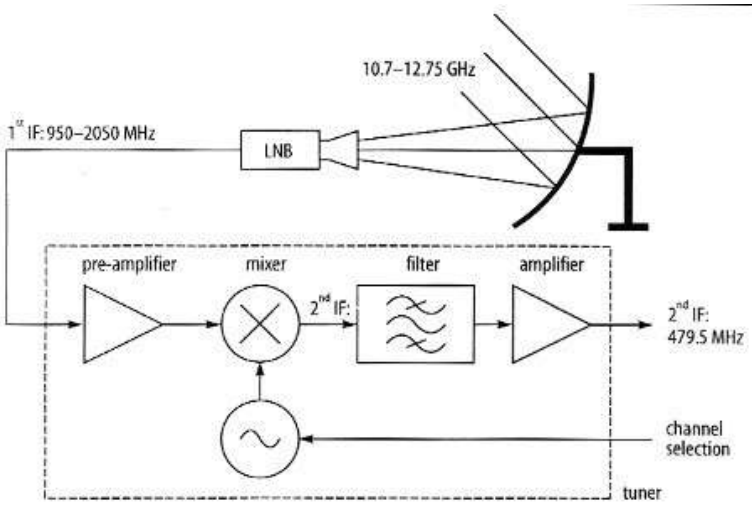


Figura 25.6: Ricevitore satellitare DVB

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 impredicibile 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.

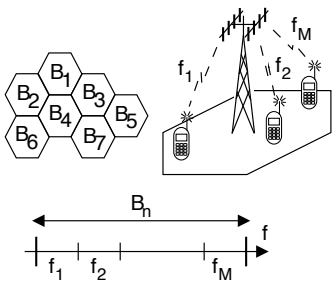
Capitolo 26

Telefonia mobile

Sicuramente questo è tutto fuorché un capitolo! La sua presenza è solamente un auspicio di sviluppo futuro: anche se l'argomento è realmente vasto ed in continua evoluzione, proprio per questo si sente il bisogno di una narrazione che, forte del bagaglio culturale fino ad ora esposto, riesca a descrivere il quadro delle tecnologie che si sono accavallate nel tempo. Al momento il capitolo ospita solo un accenno relativo al sistema GSM, presente nelle precedenti edizioni sotto forma di nota a piè di pagina nel § 11.1.1.3. In attesa del resto della storia, possiamo sempre approfondire gli argomenti grazie come al solito a *Wikipedia*: `https://it.wikipedia.org/wiki/Telefonia_cellulare`.

26.1 La trama del GSM

Con riferimento a quanto accennato al § 11.1.1.3, ecco un minimo di approfondimento. Senza entrare nei dettagli, specifichiamo semplicemente che celle limitrofe utilizzano bande di frequenza B_i differenti, onde evitare fenomeni di interferenza tra celle. Inoltre per il sistema GSM le trasmissioni che (in una stessa cella) adottano la stessa portante f_i si avvicendano nel tempo in accordo ad una struttura di trama, in modo da permettere l'utilizzo dello stesso canale da parte di più terminali mobili contemporaneamente, multiplati a divisione di tempo.



Per ciascun radiomobile la scelta della portante su cui comunicare avviene in base alla valutazione delle reciproche condizioni di ricezione tra radiomobile e stazione radio base (BS) della cella, dato che per effetto dei cammini multipli (§ 20.3.3) ci si trova verosimilmente nella condizione di *selettività in frequenza* del canale, e dunque di ricevere *meglio* certe portanti piuttosto che altre.

A questo punto la BS assegna a ciascun terminale che condivide con gli altri la stessa portante un opportuno *time-slot* nell'ambito della struttura di trama che permette l'alternanza temporale della trasmissione da parte dei singoli terminali, attuata

mediante una modulazione numerica di tipo GMSK (pag. 541) che ha dunque luogo solo per brevi periodi, in corrispondenza del time-slot di propria pertinenza.

Dato che i singoli terminali si trovano a distanze diverse dalla BS, diversi sono i tempi di propagazione dei segnali di sincronismo (di trama e di time-slot), e dunque l'intervallo temporale che viene "riempito" da ogni terminale giunge alla BS 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.

Download del formato PDF navigabile

Come anticipato nella prefazione, il download del formato PDF *navigabile* dell'edizione *2.0* nella versione *completa* di tutte le quattro parti è consentito a tutti coloro che acquistano una copia in formato cartaceo od ebook, o che effettuano una *donazione*[1] per sostenere il progetto di cultura libera, visto che senz'altro acquistare il libro significa sostenerlo. Chi lo desidera può inviare *una foto* del volume in suo possesso ad `alef@teoriadeisegnali.it`, e riceverà al suo indirizzo le istruzioni per il download.

Restiamo in contatto

Presso `teoriadeisegnali.it` puoi trovare anche

- esercizi di esame svolti;
- il testo intitolato *Lo strato applicativo di Internet* assieme ad altro materiale didattico;
- le recensioni dei lettori, a cui sei invitato ad aggiungere la tua;
- un blog per restare aggiornati sugli sviluppi del sito e dei suoi contenuti;
- la possibilità di iscriverti alla *Newsletter* con gli annunci delle nuove edizioni;
- il link di invito al gruppo Telegram con il *backstage* degli sviluppi;
- una mappa interattiva con i luoghi di provenienza delle più recenti visite al sito.

[1] Vedi `https://teoriadeisegnali.it/donazione.html`

Bibliografia

A.A. V.V., *Wikipedia, L'enciclopedia libera e collaborativa*, `http://it.wikipedia.org`

A. Abrardo, *Comunicazioni Radiomobili,* `http://www.dii.unisi.it/~abrardo/comunicazioni_radiomobili.pdf`, URL consultato il 25/01/2016

J. B. Anderson, *Digital Transmission Engineering*, 2nd Edition, August 2005, Wiley-IEEE Press

G. E. Agrawal, *Coherent Optical Communications*, Third Edition, 2002 John Wiley & Sons, Inc.

S. Barbarossa, T. Bucciarelli, *Teoria dei Segnali,* Ingegneria 2000, Roma

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

E. Björnson, J. Hoydis, L.Sanguinetti *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*, 2017, Foundations and Trends in Signal Processing: Vol. 11, No. 3-4, pp 154–655. DOI: 10.1561/2000000093. Accessibile presso `https://massivemimobook.com`

G. Cancellieri, *Telecomunicazioni*, 2000 Pitagora editrice - Bologna

A. B. Carlson, *Communication Systems*, 3^{rd} Edition, 1986 Mc Graw Hill

M. Chiodi, *Richiami di algebra elementare delle matrici*, `https://www.marcellochiodi.com/mylessons/MLA2020matriciA4.pdf`, URL consultato il 26/09/2021

M. Cover, J.A. Thomas, *Elements of information theory*, 2006 by John Wiley & Sons

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

L. E. Franks, *Signal Theory*, 1969 Prentice-Hall, Englewood Cliff, N.J.

R. L. Freeman, *Telecommunication System Enginnering*, 2^{nd} Edition, John Wiley & Sons

A. Goldsmith, *Wireless Communications*, Draft of Second Edition, Marzo 2020, `http://web.stanford.edu/class/ee359/doc/WirelessComm_Chp1-16_March32020.pdf`, URL consultato il 12/09/2021

T. T. Ha, T*heory and Design of Digital Communication Systems*, Cambridge University Press 2011

F. Halsall, *Multimedia Communications*, 2001 Pearson Education Limited

C. W. Helstrom, *Probability and Stocastic Processes for Engineers*, 2^{nd}, 1991 Macmillan Publishing Company

N. S. Jayant, P. Noll, *Digital Coding of Waveforms*, 1984 Prentice-Hall, N.J.

E. Krouk, S. Semenov, *Modulation and coding techniques in wireless communications*, 2011 John Wiley & Sons Ltd.

D. Leon, W. Couch, *Fondamenti di telecomunicazioni*, 2004 Apogeo

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

M. Luise, *Lezioni di comunicazioni digitali, DRAFT*, 2022 Università di Pisa, `http://docenti.ing.unipi.it/m.luise/InfTheory/LCD_Luise_draft.pdf`, Url consultato il 15/01/2023

P. Mandarini, *Teoria dei Segnali*, 1979 Editrice La Goliardica - Roma

A. V. Oppenheim, R. W. Shafer, *Digital Signal Processing*, 1975 Prentice Hall, NJ

C. D. Pagani, S. Salza, *Analisi Matematica 1*, 2015 Zanichelli

A. Papoulis, *Probability, Random variables, and Stochastic Processes*, 1991 McGraw-Hill Int.Eds.

B. Peroni, *Comunicazioni Elettriche*, Ed. Scientifiche Siderea, 1973 Roma

A. Perotti, *Introduzione alle Comunicazioni Radiomobili*, `Researchgate`, consultato il 15/01/2023

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

T.S. Rappaport, *Wireless Communications: Principles and Practice, Second Edition*, Prentice Hall, 2002

U. Reimers, *Digital Video Broadcasting*, Springer-Verlag Berlin Heidelberg 2001

F. Rocca, *Elaborazione Numerica dei Segnali*, Edizioni CUSL Milano 1998

A. Roveri, *Reti di telecomunicazione*, Appunti delle lezioni

S. Sardellitti, *Trasmissione Numerica*, Appunti delle lezioni

M. Lops, *Information Sources*, Materiale didattico di Teoria dell'inforazione, 2020, `https://www.docenti.unina.it/marco.lops`

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