**Alessandro Falaschi**

# *Trasmissione dei Segnali e Sistemi di Telecomunicazione*

Marzo 2014
Versione 1.2

*Web Edition*

*ai miei genitori*

*Titolo* Trasmissione dei segnali e sistemi di telecomunicazione

*Autore* Alessandro Falaschi - http://infocom.uniroma1.it/alef/wiki/

*Rilascio* Edizione 1.2, Marzo 2014

*ISBN* 9788891134882



*Sito Web* http://infocom.uniroma1.it/alef/wiki/Didattica.LibroTLC

*Strumenti*
- Lyx - http://www.lyx.org/
- Skencil - http://www.skencil.org/
- Inkscape - http://www.inkscape.org/
- Gimp - http://gimp.linux.it/www/
- Gnuplot - http://www.gnuplot.info/
- Octave - http://www.gnu.org/software/octave/

*Donazioni* http://infocom.uniroma1.it/alef/wiki/Main/Donazioni

*Copertina* Marco Sebastiani - http://www.marcosebastiani.it/

# Prefazione

Iniziai a autoprodurre questo testo, forse come forma di rivincita sulle fotocopie di appunti scritti a mano con cui da studente preparavo gli esami, e dopo qualche anno mi resi conto di essermi imbarcato in una impresa senza fine. Cionostante, ho continuato a migliorare la qualità di quanto scritto e ad integrare nuovi argomenti, incoraggiato dal buon successo di lettori che riscuoteva durante la sua permanenza on-line, ed a distanza di più di quindici anni dalla partenza, annuncio il rilascio dell'edizione *1.2*.

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

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

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

Oltre alla pubblicazione web del testo in formato HTML, ho deciso di produrre

questa versione stampabile *on-demand* tramite a *youcanprint.it*, acquistabile oltre che per posta, anche presso una delle più di mille librerie affiliate. Personalmente ritengo la versione fisica *insostituibile* per uno studio serio e ragionato, ma riconosco che anche la versione PDF *completamente navigabile* sia un formato di particolare valore intrinseco, e ne consento dunque il download a chi acquista la copia cartacea, o si rende disponibile a supportarne lo sviluppo.

## Note di rilascio all'edizione 1.2

Svolgiamo una breve sintesi di quanto ci sia di nuovo e di diverso tra questa edizione e la precedente. L'edizione 1.2 vede la nascita di un nuovo capitolo dedicato alla *codifica di sorgente multimediale*, che oltre alla codifica video e di immagine già introdotte nell'edizione 1.1, affronta anche la *codifica audio*, trattando di DPCM, ADPCM, LPC, *multipulse*, *quantizzazione vettoriale*, CELP, ACELP, *codifica psicoacustica*, MP3. La trattazione dei *collegamenti radio* è stata riorganizzata, ed approfonditi gli aspetti legati alle *comunicazioni mobili* ed al *fading piatto*, così come quelli relativi alle comunicazioni *spread spectrum*. Sempre nell'ambito delle trasmissioni numeriche, si è aggiunta la trattazione della *codifica differenziale*, ed un accenno alle questioni di *sincronizzazione* nel tempo ed in frequenza, come l'uso di *portanti pilota* nell'OFDM e di *trailer* ad inizio trama; migliorando quindi l'esposizione di PLL, QPSK, QAM, L-ASK e la discussione sulla equalizzazione *zero forcing*.

Per ciò che riguarda gli aspetti più di base si è aggiunta la *rappresentazione matriciale* per la DFT, oltre che la definizione di *v.a. gaussiana multidimensionale* e dimostrata la proprietà di indipendenza statistica in caso di incorrelazione, si è migliorata la trattazione relativa a *filtro adattato*, *segnalazione antipodale* e *ortogonale*, ed aggiunti dei chiarimenti sul *rapporto di verosimiglianza*.

Si sono poi attuate un'ampia serie di correzioni e miglioramenti, come per il *compromesso velocità distorsione*, la valutazione degli *errori su parole*, la definizione dei *codici polinomiali* ed il loro uso come CRC, l'uso di scale comparabili nella figura dell'*impulso a radice di coseno rialzato*, il valore iniziale della sommatoria di *Bernoulli*, il legame tra potenza e *varianza*, la definizione di $E_b/N_0$, le condizioni di applicabilità del *primo teorema di Shannon*, la figura relativa alle *funzioni di Bessel*, la trasformata del *triangolo*, e la sua energia.

Sono quindi stati aggiunti nuovi esempi, come per la *palette dei colori*, per il *codice di Huffman*, il grafico della *capacità di canale* in funzione della banda, la tabella del *tasso di codifica* per il codice di Hamming al crescere di $q$, ed un esempio di calcolo della *sindrome*.

Infine, si è provveduto come al solito ad una serie di piccoli aggiustamenti sintattici, chiarimenti, spostamenti parziali, reimpaginazione delle figure, e miglioramento del contrasto delle più sbiadite.

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

*Alessandro Falaschi, 2014*

# Indice

# Capitolo 1

# Introduzione

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

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

## 1.1 Trasmissione dell'informazione

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

**Distorsioni e disturbi** Il canale può ad esempio imporre una limitazione alla *banda di frequenze* del segnale in transito[1]; cause fisiche ineliminabili producono inoltre, al lato ricevente, l'insorgere di un segnale di disturbo additivo, comunemente indicato con il termine di *rumore*, che causa la ricezione di un segnale diverso da quello stesso presente all'uscita del canale. Pertanto, ci si preoccupa di caratterizzare il canale in modo da scegliere i metodi di trasmissione più idonei a rendere minima l'alterazione sul messaggio trasmesso.

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

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

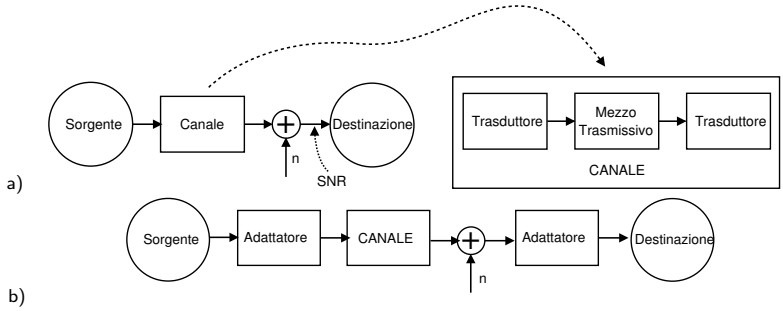
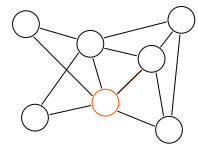



Figura 1.1: Elementi funzionali per la trasmissione dell'informazione

**Trasmissione** La Fig. 1.1a) evidenzia come il canale, nella realtà fisica, è costituito da un *mezzo trasmissivo* su cui si propaga un segnale di natura elettromagnetica, che viene convertito in tale forma da appositi *trasduttori* di trasmissione e ricezione[2]. Considerando per il momento i trasduttori come facenti parte del canale stesso, proseguiamo l'analisi concentrandoci sugli ulteriori aspetti del processo di comunicazione.

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

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



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

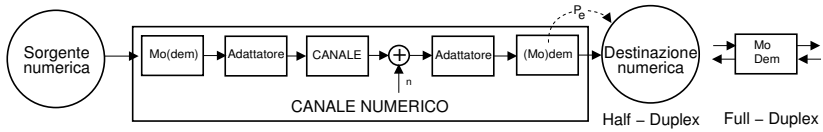
## 1.2 Trasmissioni numeriche

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

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

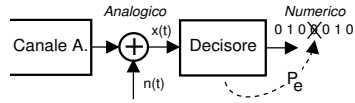
partenenti ad un insieme finito, come ad es. le lettere dell'alfabeto;

**Modem** i simboli che compongono un segnale *numerico* devono essere trasformati in un segnale analogico mediante l'utilizzo di dispositivi chiamati *modem*[3], come rappresentato dalla figura seguente, in cui è evidenziato come per una trasmissione unidirezionale[4] occorra solo *metà* delle funzioni del modem per entrambi i lati del collegamento, mentre nel caso di collegamento *full duplex* (in cui entrambi gli estremi possono essere contemporaneamente sorgente e destinazione) il modem opera allo stesso tempo nelle due direzioni.




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

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



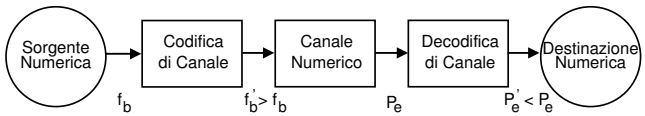

**Codifica di Canale** Nelle trasmissioni numeriche, si può introdurre una *ridondanza* nella sequenza trasmessa, inviando più simboli di quanti non ne produca la sorgente, e quindi di fatto *aumentando* il numero di simboli da trasmettere per unità di tempo; i simboli in più sono scelti in modo da essere in qualche modo *dipendenti* tra loro, e questa loro caratteristica rende possibile la riduzione della probabilità di errore di cui soffre il canale numerico. Infatti, grazie alla dipendenza (nota) tra i simboli trasmessi, il ricevitore è ora in grado di "accorgersi" che si è verificato un errore, in quanto la dipendenza prevista non è più rispettata; pertanto, il ricevitore può attuare delle contromisure. La ridondanza introdotta può essere così elevata da permettere la correzione di errori isolati[5], oppure il ricevitore può semplicemente richiedere la *ritrasmissione* del simbolo errato. Le trasformazioni del segnale ora descritte prendono il nome di *codifica*

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

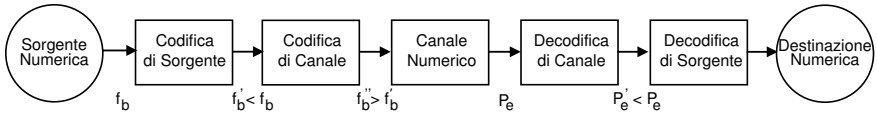
[4] Nelle trasmissioni unidirezionali, sorgente e destinazione non si scambiano i ruoli. La trasmissione stessa viene anche indicata con il termine di *half-duplex*.

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

*di canale*, e devono essere "rimosse" all'uscita dello stesso da un processo inverso di *decodifica*.




**Codifica di Sorgente** Possono essere introdotti due ulteriori blocchi, che operano una *codifica* (e relativa de-codifica) *di sorgente* sulla sequenza trasmessa, con uno scopo è per così dire "inverso" a quello della codifica di canale: infatti, la codifica di sorgente *rimuove le dipendenze* tra i simboli presenti nelle sequenze generate dalla sorgente, ottenendo di fatto un riduzione del numero di simboli da trasmettere per unità di tempo[6]. Un tipico esempio di codifica di sorgente è rappresentato dagli algoritmi di compressione esistenti per i file di computer (come i file *zippati*); in tal caso, il fattore di compressione ottenibile dipende dalla natura del file trattato, ed è tanto maggiore quanto più quest'ultimo presenta caratteristiche di ripetitività e quindi di predicibilità del suo contenuto. In altre parole, l'uscita di un codificatore di sorgente è una sequenza di simboli tendenzialmente *indipendenti* tra loro, nel senso che ogni forma di predicibilità di un simbolo a partire dai circostanti è stata rimossa.



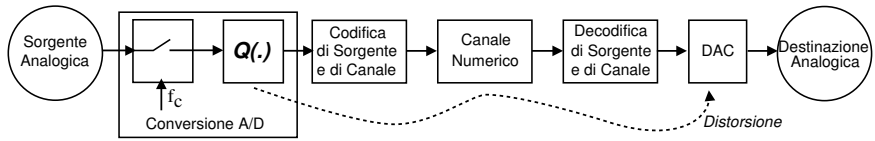

**Campionamento e Quantizzazione** Poniamoci ora il problema di utilizzare un *canale numerico*per effettuare una *trasmissione analogica*. Il vantaggio di tale "contorsione" è da ricercarsi nel migliore comportamento delle trasmissioni numeriche rispetto ai disturbi, nonché alla loro *generalità*[7].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

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

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

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

vicini tra loro. In definitiva, la distorsione risulta tanto minore, quanto maggiore è il flusso informativo[8] prodotto dal quantizzatore, espresso in bit/secondo.




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

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

## 1.3 Segnali analogici, certi ed aleatori

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

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

[8] Vedremo infatti al § 7.4 che un aumento della risoluzione del processo di quantizzazione corrisponde ad un aumento del numero di bit necessari a rappresentare ogni valore (o campione) di segnale.

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

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

### 1.3.1 Rappresentazione di segnali analogici

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

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

### 1.3.2 Rappresentazione di processi aleatori

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

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

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

### 1.3.3 Transito dei segnali attraverso sistemi fisici

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

**Risposta impulsiva e convoluzione** Per mezzo dell'integrale di *convoluzione* si esprime in forma chiusa l'uscita di un sistema (ad es. un circuito elettrico[11]) in base alla conoscenza dell'ingresso, e di una particolare caratteristica del sistema, la *risposta impulsiva*. Quest'ultima rappresenta la sua uscita quando in ingresso è presente una particolare funzione analitica, detta *impulso matematico*.

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

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

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

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

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

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

## 1.4 Segnali numerici

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

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

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

**Frequenza di simbolo** Il concetto di occupazione di banda, applicabile ai segnali analogici, è qui sostituito da quello di velocità di emissione, espressa in *simboli/secondo*, ed indicata come *frequenza di simbolo* o $f_s$. Una sequenza prodotta da una sorgente numerica si presta facilmente ad essere trasformata in un'altra, con un diverso alfabeto ed una differente frequenza di simbolo[12].

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

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

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

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

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

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

## 1.5 Teoria delle probabilità

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

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

## 1.6 Sistemi di telecomunicazione

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Elaborazione terminale** Questa categoria comprende tutti gli aspetti legati alle trasformazioni operate sull'informazione ai due estremi del collegamento[13]. Tra questi è possibile riconoscere

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

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

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

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

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

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

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

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

## 1.7 Segnali e sistemi

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

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

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

### 1.7.1 Caratteristiche dei sistemi

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

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

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

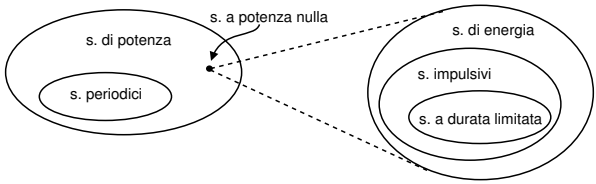



Figura 1.2: Visione insiemistica per le diverse classi di segnali

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

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

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

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

### 1.7.2 Caratteristiche dei segnali

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

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

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

Un segnale di potenza è inoltre detto

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

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

per qualsiasi valore di $t$, mentre si dice

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

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

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

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

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

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

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

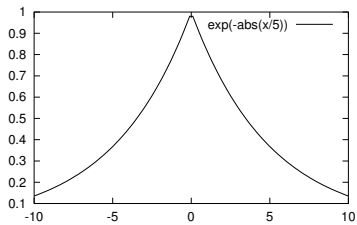
$$0 < \int_{-\infty}^{\infty} |s(t)| \, dt < \infty$$

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

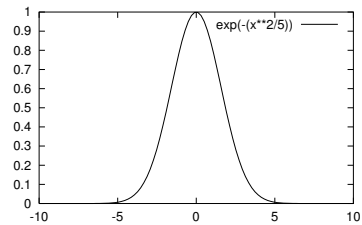
**Riassumendo**

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

**Esempi di segnali di energia**

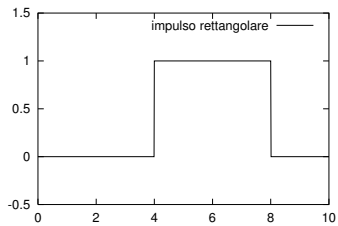



Impulso Esponenziale Bilatero

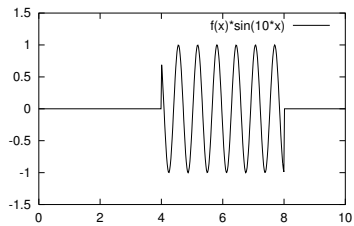



Impulso Gaussiano

**Esempi di segnale a durata limitata**




Impulso rettangolare tra 4 ed 8




Sinusoide troncata

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

### 1.7.3 Aspetti fisici delle grandezze energetiche

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

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

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

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

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

Nello stesso intervallo $T$, la resistenza *assorbe* una potenza $p_T(t) = \overline{e_T(t)} = \frac{1}{T} e_T(t)$ [Watt], che costituisce una media *a breve termine* dell'energia assorbita nell'intervallo[16].

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

| | | |
|---|---|---|
| Energia: | $\mathcal{E}_x = \int_{-\frac{T}{2}}^{\frac{T}{2}} \lvert x\rvert^2(t)\, dt = e_T\left(\frac{T}{2}\right)$ | [Volt$^2$·sec] *o* [Ampere$^2$·sec] |
| Potenza: | $\mathcal{P}_x = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} \lvert x\rvert^2(t)\, dt = \frac{1}{T} e_T\left(\frac{T}{2}\right) = p_T\left(\frac{T}{2}\right)$ | [Volt$^2$] *o* [Ampere$^2$] |

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

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

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

[16] Anticipando una notazione che verrà usata nel corso del testo, il pedice $_T$ indica l'estesione temporale a cui è riferita la grandezza che presenta il pedice, mentre la sopralineatura di una grandezza che dipende dal tempo, indica una media temporale della grandezza stessa.

# Capitolo 2

# Serie di Fourier

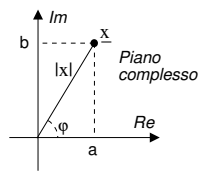
Sono qui impostate alcune relazioni trigonometriche fondamentali, per poi definire la rappresentazione di segnali periodici mediante lo sviluppo in serie di Fourier. Sono quindi introdotte le proprietà di simmetria e di approssimazione, e sviluppati i concetti di ortogonalità. E' infine enunciato il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

## 2.1 Prerequisiti trigonometrici

### 2.1.1 Numeri complessi

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

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




E' spesso utile ricorrere ad una rappresentazione di $\underline{x}$ nel piano complesso, che mette in luce l'espressione alternativa[1] di $\underline{x}$ nei termini di modulo $|x|$ e fase $\varphi$:

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

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

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

mentre le relazioni inverse risultano

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

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

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

### 2.1.2 Formule di Eulero

L'esponenziale $e^{j\varphi}$ è un particolare numero complesso con modulo pari ad uno[2], e che quindi si scompone in parte reale ed immaginaria come

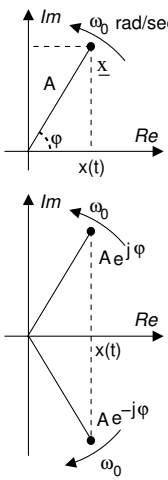
$$e^{\pm j\varphi} = \cos\varphi \pm j \sin\varphi$$

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

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

e che possono tornare utili nel semplificare i calcoli, trasformando i prodotti tra funzioni trigonometriche in somme di angoli[3].

### 2.1.3 Fasori




Un segnale del tipo $x(t) = A\cos(2\pi f_0 t + \varphi)$ è completamente rappresentato dal numero complesso $\underline{x} = Ae^{j\varphi}$ detto *fasore*, la cui conoscenza permette di riottenere il segnale originario mediante la relazione $x(t) = \Re\left\{\underline{x} \cdot e^{j2\pi f_0 t}\right\}$, che una volta sviluppata[4] risulta infatti pari a

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

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

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

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

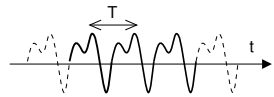
[2] L'espressione più generale $e^{\gamma}$ con $\gamma = \alpha + j\varphi$ è ancora un numero complesso, di modulo $e^{\alpha}$. Infatti $e^{\gamma} = e^{\alpha+j\varphi} = e^{\alpha}e^{j\varphi} = e^{\alpha}(\cos\varphi + j\sin\varphi)$.

[3] L'affermazione nasce dalla relazione $e^{\alpha}e^{\beta} = e^{\alpha+\beta}$. Ad esempio quindi, il prodotto $\cos\alpha \cdot \sin\beta$ diviene
$= \frac{1}{4j}\left(e^{j\alpha} + e^{-j\alpha}\right)\left(e^{j\beta} - e^{-j\beta}\right) = \frac{1}{4j}\left[e^{j\alpha}e^{j\beta} - e^{j\alpha}e^{-j\beta} + e^{-j\alpha}e^{j\beta} - e^{-j\alpha}e^{-j\beta}\right]$
$= \frac{1}{4j}\left[e^{j(\alpha+\beta)} - e^{-j(\alpha+\beta)} - e^{j(\alpha-\beta)} + e^{-j(\alpha-\beta)}\right] = \frac{1}{4j}\left[2j\sin(\alpha+\beta) - 2j\sin(\alpha-\beta)\right]$
$= \frac{1}{2}\left[\sin(\alpha+\beta) - \sin(\alpha-\beta)\right]$

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

## 2.2 Serie di Fourier

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




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

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

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

$$X_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \ \mathrm{e}^{-j2\pi nFt} dt \tag{2.2}$$

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

$$x(t) = \sum_{n=-\infty}^{\infty} X_n \, \mathrm{e}^{j2\pi nFt} \tag{2.3}$$

Osserviamo che:

- La conoscenza di $\{X_n\}$ *equivale* a quella di $x(t)$ *e viceversa*, esistendo il modo di passare dall'una all'altra rappresentazione;
- Le funzioni della base di rappresentazione $\mathrm{e}^{j2\pi nFt}$ sono funzioni trigonometriche a frequenza multipla (*n-esima*) della fondamentale, detta anche *n-esima* armonica[5]
- I temini $X_n \mathrm{e}^{j2\pi nFt}$ sono chiamati *componenti armoniche* di $x(t)$ a frequenza $f = nF$;
- Il coefficiente $X_0 = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)\, dt$ rappresenta la componente continua (o *valor medio*) di $x(t)$;
- La serie di Fourier dà valori esatti in tutti i punti in cui $x(t)$ è continuo, mentre in corrispondenza di discontinuità di prima specie fornisce un valore pari alla media dei valori agli estremi, cosicché il valore dell'energia di un periodo è preservato;
- I coefficienti di Fourier $X_n$ possono essere calcolati anche per un segnale di estensione finita $T$. Antitrasformando, il segnale diventa periodico!

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

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

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

essendo

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

### 2.2.1 Segnali reali

#### 2.2.1.1 Simmetria coniugata

I coefficienti della serie di Fourier possono essere calcolati anche per segnali complessi; nel *caso particolare di $x(t)$ reale* i coefficienti di Fourier risultano godere della proprietà di *simmetria coniugata*, espressa come

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

e che significa che i coefficienti con indice $n$ negativo possiedono una parte reale uguale a quella dei coefficienti con (uguale) indice positivo, e parte immaginaria cambiata di segno[6]. Ciò comporta una proprietà analoga per il modulo e la fase di $\{X_n\}$, e dunque possiamo scrivere:

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

Tali relazioni evidenziano che

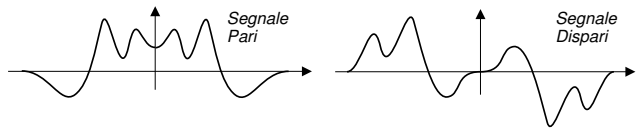
> Se $x(t)$ è reale, i coefficienti $X_n$ risultano avere modulo pari e fase dispari, ovvero parte reale pari e parte immaginaria dispari.

Un corollario di questo risultato è che[7]

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

[6] La dimostrazione di questa proprietà si basa sul fatto che, scomponendo l'esponenziale complesso che compare nella formula per il calcolo degli $X_n$ come $e^{-j2\pi nFt} = \cos 2\pi nFt - j\sin 2\pi nFt$, ed essendo $x(t)$ reale, l'integrale stesso si suddivide in due, ognuno relativo al calcolo indipendente della parte reale e quella immaginaria: $X_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)\cos 2\pi nFtdt - \frac{j}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)\sin 2\pi nFtdt$. Essendo il coseno una funzione pari, il primo integrale fornisce gli stessi risultati per $n$ cambiato di segno; il secondo integrale invece cambia segno con $n$, essendo il seno una funzione dispari.

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

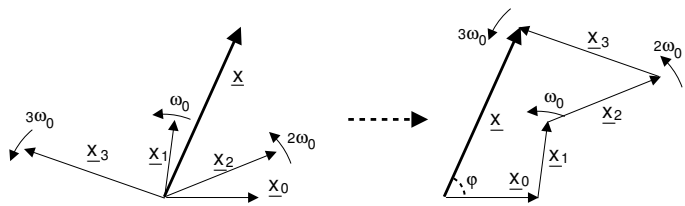



#### 2.2.1.2 Interpretazione degli $X_n$ come fasori

Confrontando la formula di ricostruzione

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

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




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

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

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

- per $n = 0$, osserviamo che $\mathrm{e}^{-j2\pi nFt}\big|_{n=0} = \mathrm{e}^0 = 1$, e dunque
$$X_0 = \int_{-\frac{T}{2}}^{\frac{T}{2}} \mathrm{e}^{j2\pi Ft} dt = 0$$
poiché in un intervallo $T$ entra esattamente un ciclo di (co)sinusoide a frequenza $F$, risultando in un valor medio nullo.
- per $n = 1$, si ha che $\int_{-\frac{T}{2}}^{\frac{T}{2}} \mathrm{e}^{j2\pi Ft}\,\mathrm{e}^{-j2\pi Ft} dt = \int_{-\frac{T}{2}}^{\frac{T}{2}} \mathrm{e}^0 dt = T$, mentre $\int_{-\frac{T}{2}}^{\frac{T}{2}} \mathrm{e}^{-j2\pi Ft}\,\mathrm{e}^{-j2\pi Ft} dt = \int_{-\frac{T}{2}}^{\frac{T}{2}} \mathrm{e}^{-j2\pi 2Ft} dt = 0$, dato che in un periodo T entrano due cicli esatti della funzione periodica integranda, ottenendo quindi
$$X_1 = \frac{A}{2}\mathrm{e}^{j\varphi}$$

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

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

#### 2.2.1.3 Serie trigonometrica

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

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

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

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

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

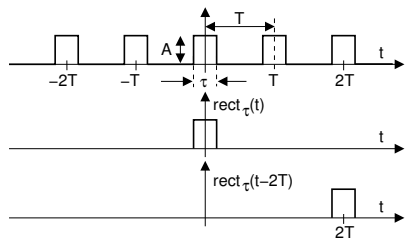
in cui

$$R_0 = M_0 = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \, dt \qquad \text{e} \qquad \begin{cases} R_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \, \cos\left(2\pi nFt\right) dt \\ I_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \, \sin\left(2\pi nFt\right) dt \end{cases}$$

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

#### 2.2.1.4 Serie di Fourier di un'onda rettangolare

La figura a lato mostra un segnale *ad onda quadra* con un *duty cycle*[8] del 33%, la cui espressione analitica può essere scritta come




$$x(t) = \sum_{n=-\infty}^{\infty} A \, rect_{\tau}\left(t - nT\right)$$

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

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

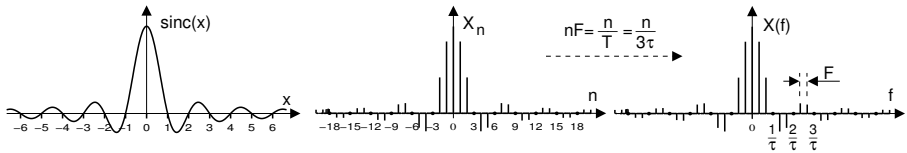



Figura 2.1: Funzione $\mathrm{sinc}\,(x)$ e coefficienti di Fourier dell'onda quadra

**Esercizio** Il calcolo dei coefficienti di Fourier per il segnale in questione non presenta particolari difficoltà, ma l'esito si presta ad alcune utili considerazioni. Applicando infatti un risultato noto[9], si ottiene

$$X_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x\,(t)\,\mathrm{e}^{-j2\pi nFt}dt = \frac{1}{T}\int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} A\mathrm{e}^{-j2\pi nFt}dt =$$

$$= \frac{A}{T}\,\frac{\mathrm{e}^{-j2\pi nFt}}{-j2\pi nF}\Big|_{-\frac{\tau}{2}}^{\frac{\tau}{2}} = \frac{A}{\pi nFT}\,\frac{\mathrm{e}^{j2\pi nF\frac{\tau}{2}} - \mathrm{e}^{-j2\pi nF\frac{\tau}{2}}}{2j} = \tag{2.4}$$

$$= A\frac{\tau}{T}\frac{\sin\,(\pi nF\tau)}{\pi nF\tau} = A\frac{\tau}{T}\mathrm{sinc}\,(nF\tau) \tag{2.5}$$

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

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

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

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

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

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

**Relazione tra i coefficienti della serie ed i parametri dell'onda quadra** La parte in alto di Fig. 2.2 mostra quattro possibili modi di variare l'onda quadra di partenza: la colonna di sinistra rappresenta il caso in cui il periodo $T$ si mantenga costante, mentre la durata $\tau$ di un singolo ciclo *raddoppia* (prima riga) o si *dimezza*

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

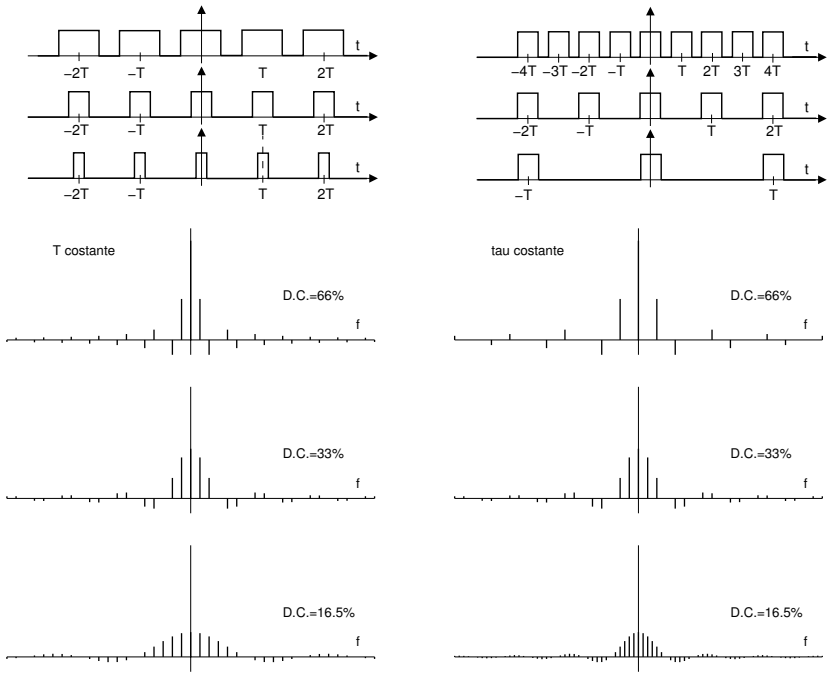



Figura 2.2: Modifiche allo spettro di ampiezza per variazioni della forma d'onda

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

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

### 2.2.2 Serie di Fourier a banda limitata

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

$$x\,(t) = \sum_{k=-\infty}^{\infty} rect_{\frac{T}{2}}\,(t - kT)$$

rappresentata mediante una serie troncata di Fourier in cui si considerano solo i coefficienti $X_n$ con indice $-N \leq n \leq N$. Sappiamo che $X_n = \frac{\tau}{T}\frac{\sin(\pi nF\tau)}{\pi nF\tau}$ e, per $\tau = \frac{T}{2}$, si

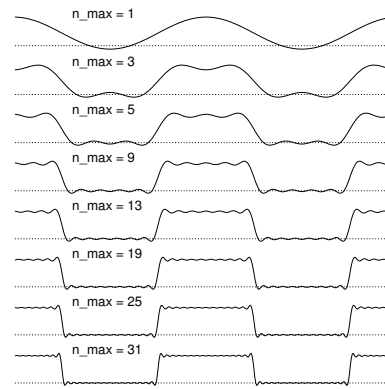
ottiene $X_n = \frac{1}{2}\frac{\sin\left(n\frac{\pi}{2}\right)}{n\frac{\pi}{2}} = \frac{1}{2}\text{sinc}\left(\frac{n}{2}\right)$, che risulta diverso da zero solo con $n$ *dispari*, e dunque:

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

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

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

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




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

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

Il caso mostrato è emblematico della inaccuratezza che si commette considerando contributi frequenziali ridotti rispetto a quelli propri della forma d'onda[10], a causa (ad esempio) di un *filtraggio* del segnale.

## 2.3 Teorema di Parseval

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

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

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

Sviluppiamo i calcoli che danno luogo al risultato mostrato:

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

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

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

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

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

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

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

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

$$\mathcal{P}_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}\left[X_n\,\mathrm{e}^{j2\pi nFt}\right]\left[X_n^*\,\mathrm{e}^{-j2\pi nFt}\right]dt = \frac{|X_n|^2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}dt = |X_n|^2$$

e quindi osserviamo che

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

Si presti attenzione che il risultato è una diretta conseguenza dell'ortogonalità della base di rappresentazione: infatti, la potenza di una somm*a non è* in generale pari alla somma delle potenze[11]; l'uguaglianza ha luogo solo nel caso di in cui gli addendi siano ortogonali.

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

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

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

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

## 2.4 Appendici

### 2.4.1 Algebra vettoriale

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

Lo *spazio prodotto interno* (o *spazio normato*) è quello spazio lineare, in cui è definito il *prodotto scalare* $(\overline{x}, \overline{y})$ tra generici vettori $\overline{x}$ ed $\overline{y}$[12]. In tal caso, si può definire la *norma* $\|\overline{x}\|$ di un vettore $\overline{x}$ come

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

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

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

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

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

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

In tal caso, il prodotto scalare tra due vettori $\overline{x}$ ed $\overline{y}$ ha espressione[13]

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

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

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

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

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

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

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

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

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

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

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

Considerando quindi il sottospazio dei segnali con periodo $T = \frac{1}{F}$ costituito dall'insieme $\left\{e^{j2\pi nFt}\right\}$, la validità di (2.6) di pag. 24 prova che gli $\left\{e^{j2\pi nFt}\right\}$ costituiscono una base ortonormale per i segnali periodici di periodo $T$; in particolare si riconosce che l'espressione (2.2) di pag. 17 rappresenta la proiezione[14] del segnale lungo i vettori della base, mentre la formula di ricostruzione (2.3) costituisce la rappresentazione del segnale nei termini delle sue componenti ortogonali.

**Ri-definizione dei coefficienti di Fourier** Moltiplicando il segnale periodico per $e^{-j2\pi mFt}$ ed eseguendo l'integrale tra due istanti $t_1$ e $t_2$ presi a distanza di un multiplo intero di periodi (ossia $t_2 - t_1 = kT$), si ottiene

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

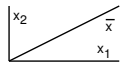
in quanto per $n \neq m$ la funzione integranda ha valor medio nullo, dato che nell'intervallo $(t_1, t_2)$ (dovunque collocato dell'asse dei tempi) presenta un numero intero di periodi. Pertanto, il calcolo dei coefficienti può ottenersi a partire da un qualunque intervallo esteso su un numero intero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

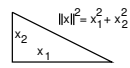
$$X_n = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} x(t)\, e^{-\cdot j2\pi nFt} dt$$

---

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

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

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

diventa 

.

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

**Disuguaglianza di Schwartz** Consiste nel risultato

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

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

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

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

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

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

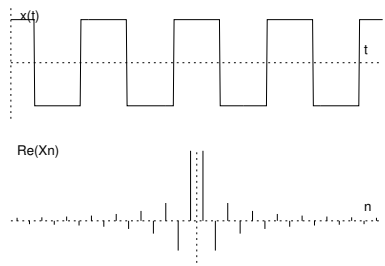
### 2.4.2 Esempi di Sviluppo in serie

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

**Onda quadra simmetrica**

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

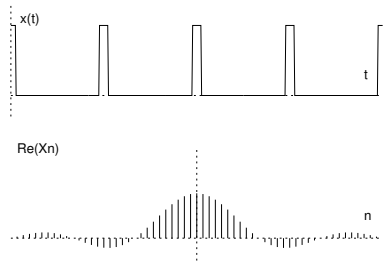
$$X_n = \begin{cases} \operatorname{sinc}\left(\frac{n}{2}\right) & n \neq 0 \\ 0 & n = 0 \end{cases}$$




**Treno di impulsi rettangolari**

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

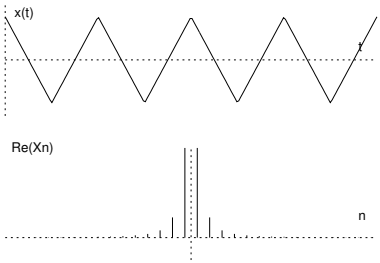
$$X_n = \frac{\tau}{T}\operatorname{sinc}\left(\frac{n\tau}{T}\right)$$



**Onda triangolare simmetrica**

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

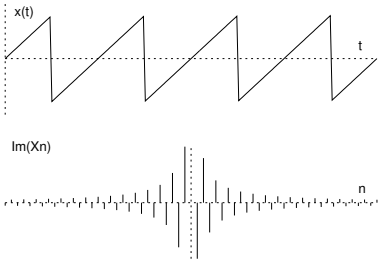
$$X_n = \begin{cases} \operatorname{sinc}^2\left(\frac{n}{2}\right) & n \neq 0 \\ 0 & n = 0 \end{cases}$$




**Dente di sega simmetrico**

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

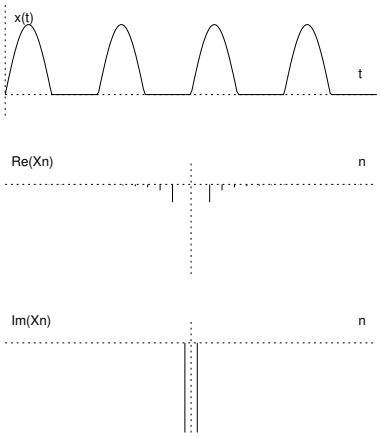
$$X_n = \begin{cases} j\frac{(-1)^n}{n\pi} & n \neq 0 \\ 0 & n = 0 \end{cases}$$




**Rettificata a singola semionda**

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

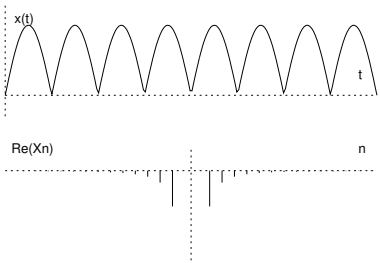
$$X_n = \begin{cases} \frac{1}{\pi\left(1-n^2\right)} & n \text{ pari} \\ -j\frac{1}{4} & n = \pm 1 \\ 0 & \text{altrimenti} \end{cases}$$




**Rettificata a onda intera**

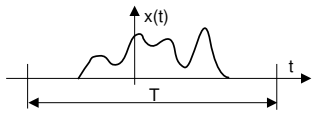
$$x(t) = |\sin\omega_0 t|$$

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



# Capitolo 3

# Trasformata di Fourier




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

## 3.1 Definizione

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

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

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

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

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

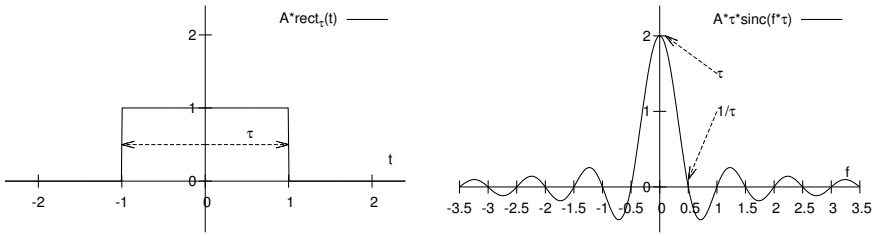



Figura 3.1: $\mathcal{F}$-trasformata di un rettangolo di base $\tau = 2$ ed ampiezza $A = 1$

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

La formula di ricostruzione, se messa a confronto con la serie di Fourier, può essere pensata come una somma integrale di *infinite* componenti $X(f)\,df e^{j2\pi ft}$ di ampiezza (complessa) infinitesima, evidenziando come ora siano presenti *tutte* le frequenze e non solo le armoniche. Una seconda analogia con la serie di Fourier deriva dal considerare un segnale $x(t)$ di durata limitata $T$, e calcolare $X(f) = \mathcal{F}\{x(t)\}$ per $f = \frac{n}{T} = nF$. In tal caso, è facile verificare[1] che risulta

$$X(f = nF) = T \cdot X_n \tag{3.1}$$

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

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

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

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

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

[1]

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

## 3.2 Energia incrociata e densità di energia

Similmente al caso dei segnali periodici, viene ora stabilita una relazione tra *l'energia* di un segnale, e la *distribuzione* della stessa nel dominio della frequenza. In base alle considerazioni geometriche esposte al § 2.4.1, definiamo come prodotto scalare tra i segnali di energia $x(t)$ e $y(t)$ (detto anche *energia incrociata*) il valore

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

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

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

Il risultato

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

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

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

Esaminando quest'ultima espressione, possiamo indicare

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

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

## 3.3 Prime proprietà della trasformata di Fourier

**Simmetria coniugata** Nel caso in cui $x(t)$ sia reale, risulta[2]

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

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

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

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

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

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

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

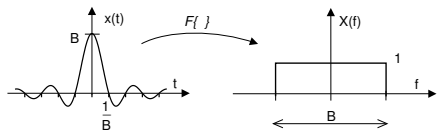
Allora, ricordando che

$$\mathcal{F}\{rect_\tau(t)\} = \tau sinc(f\tau)$$

scriviamo direttamente

$$\mathcal{F}\{B \cdot sinc(tB)\} = rect_B(f)$$

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




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

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

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

$$m_x = \int_{-\infty}^{\infty} x(t)\,dt = X(f=0) \quad \text{e, per dualita'} \quad x_0 = x(t=0) = \int_{-\infty}^{\infty} X(f)\,df$$

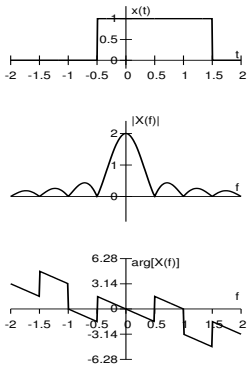
dove l'ultima relazione esprime la proprietà del *valore iniziale*.

**Esempio** Come applicazione, troviamo subito che

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

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

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




la cui dimostrazione è fornita sotto[3].

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

$$X(f) = \mathcal{F}\{rect_\tau(t)\} = \tau sinc(f\tau)$$

[3] La dimostrazione della proprietà di traslazione nel tempo si basa su di un semplice cambio di variable: $Z(f) = \int x(t-T)\,\mathrm{e}^{-j2\pi ft}dt = \int x(\theta)\,\mathrm{e}^{-j2\pi f(T+\theta)}d\theta = \mathrm{e}^{-j2\pi fT}\int x(\theta)\,\mathrm{e}^{-j2\pi f\theta}d\theta = X(f)\,\mathrm{e}^{-j2\pi fT}$

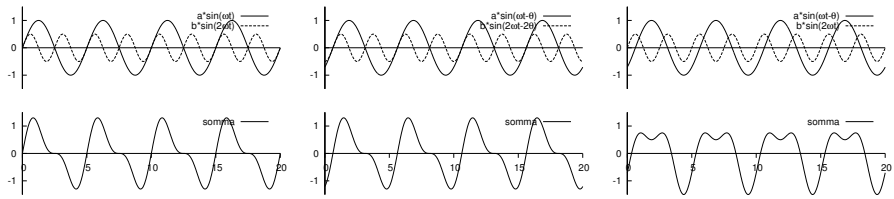



Figura 3.2: Confronto tra diversi spettri di fase

a cui si aggiunge un contributo di fase *lineare*

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

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

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

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

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

**Conseguenze della linearità di fase** La circostanza che un *ritardo temporale* del segnale $x(t)$ si traduca in una alterazione *lineare* della fase[4] della sua trasformata $X(f)$, comporta una conseguenza notevole anche nel passaggio da $f$ a $t$, ossia:

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

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

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

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

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

Ponendo $\omega T = \theta$, otteniamo

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

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

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

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

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

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

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

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

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

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

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

**Coniugato** Deriva direttamente[6] dalla definizione di $\mathcal{F}$-trasformata:

$$\mathcal{F}\{x^*(t)\} = X^*(-f)\,; \qquad \mathcal{F}^{-1}\{X^*(f)\} = x^*(-t) \tag{3.2}$$

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

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

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

in cui scegliendo ad es. $a > 1$, si ottiene una *accelerazione* temporale ed una *allargamento* dello spettro (ed il contrario, con $a < 1$). La dimostrazione (per $a > 0$) è riportata alla nota[7]. Un corollario di questa proprietà, è che se $a = -1$, allora

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

## 3.4 Impulso matematico

Prima di esporre altre proprietà della trasformata di Fourier, occorre definire ed analizzare le proprietà della "funzione" *impulso matematico*, indicato con $\delta(.)$. Questo è

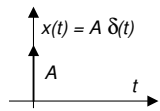
[6] Infatti $\mathcal{F}\{x^*(t)\} = \int_{-\infty}^{\infty} x^*(t)\, e^{-j2\pi f t} dt = \int_{-\infty}^{\infty} \left[x(t)\, e^{j2\pi f t}\right]^* dt$.

[7] $\int x(at)\ \mathrm{e}^{-j2\pi f t} dt = \frac{1}{a} \int x(at)\ \mathrm{e}^{-j2\pi \frac{f}{a} at} d(at) = \frac{1}{a} \int X(\beta)\ \mathrm{e}^{-j2\pi \frac{f}{a}\beta} d\beta = \frac{1}{a} X\left(\frac{f}{a}\right)$

definito come un segnale $\delta(t)$[8] che vale zero ovunque, tranne per $t = 0$ dove vale $\infty$; per contro, l'area di $\delta(t)$ è unitaria:

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

Da un punto di vista analitico, $\delta(f)$ non è una funzione, ma una *distribuzione*, definita come il limite a cui tende una serie di funzioni, come mostrato in appendice 3.9.2. E' prassi rappresentare graficamente $A \cdot \delta(f)$ come una freccia (vedi figura) con scritto accanto il valore dell'area A.




### 3.4.1 Trasformata di una costante

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

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

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

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

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

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

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

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

### 3.4.2 Trasformata per segnali periodici

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

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

Applicando la proprietà di linearità, il risultato per la trasformata di una costante, e ricordando la proprietà della traslazione in frequenza, troviamo[9] che la $\mathcal{F}$-trasformata di $x(t)$ vale:

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

[8] L'impulso matematico è noto anche con il nome di *Delta di Dirac*, e per questo è rappresentato dal simbolo $\delta$.

[9] $X(f) = \mathcal{F}\left\{\sum_{n=-\infty}^{\infty} X_n\, e^{j2\pi nFt}\right\} = \sum_{n=-\infty}^{\infty} X_n\, \mathcal{F}\left\{1 \cdot e^{j2\pi nFt}\right\} = \sum_{n=-\infty}^{\infty} X_n \cdot \delta(f - nF)$

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

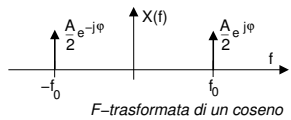
Un modo alternativo di calcolare la trasformata di segnali periodici è illustrato alla sezione 3.8.1.

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

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

la $\mathcal{F}$-trasformata risulta:

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



F–trasformata di un coseno

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

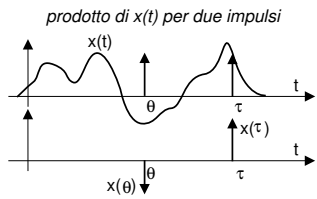
**Potenza di un coseno** Cogliamo l'occasione per calcolare la potenza di una sinusoide. Applicando il teorema di Parseval si ottiene:

$$\mathcal{P}_x = |X_1|^2 + |X_2|^2 = 2\frac{A^2}{4} = \frac{A^2}{2}$$

### 3.4.3 Proprietà di *setacciamento*

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

$$x(t)\,\delta(t-\tau) = x(\tau)\,\delta(t-\tau)$$




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

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

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

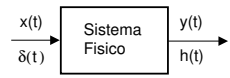
## 3.5 Risposta impulsiva e convoluzione

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

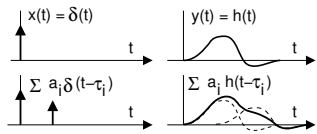
### 3.5.1 Risposta impulsiva

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

Se il sistema è *lineare e permanente*[10], applicando un ingresso costituito da *più impulsi*, ognuno con area differente $a_i$ e centrato ad un diverso istante $\tau_i$, ovvero




$$x(t) = \sum_{i=1}^{N} a_i \delta(t - \tau_i) \tag{3.3}$$




si ottiene una uscita pari a

$$y(t) = \sum_{i=1}^{N} a_i h(t - \tau_i) \tag{3.4}$$

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

### 3.5.2 Integrale di convoluzione

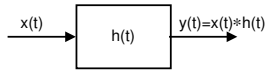
Consideriamo ancora lo stesso sistema fisico, al cui ingresso sia posto un generico segnale $x(t)$ che, grazie alla proprietà di setacciamento, rappresentiamo come scomposto in infiniti termini, ossia come somma integrale di impulsi centrati in $\tau$ (variabile) ed area $x(\tau)\,d\tau$ (infinitesima):

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

Questa espressione, formalmente simile alla (3.3), è equivalente alla proprietà di setacciamento, dato che $\delta(t)$ è una funzione pari.

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

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




*Calcolo dell'uscita per un ingresso qualunque*

in cui $x(\tau)\,d\tau$ è l'area degli impulsi che costituiscono l'ingresso, e $h(t - \tau)$ è l'uscita all'istante $t$ causata dall'impulso in ingresso centrato all'istante $\tau$. Il risultato ottenuto, formalmente simile a (3.4), prende il nome di *integrale di convoluzione*, e viene indicato in forma simbolica

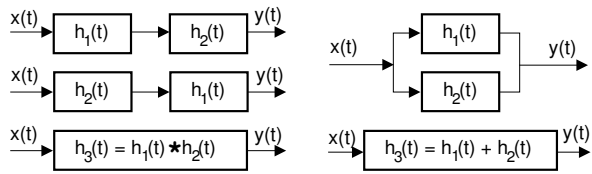
[10] Il significato di questa classificazione si trova al Capitolo 1, a pag. 11.

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

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

**Proprietà commutativa** Se un segnale con andamento $h(t)$ è posto in ingresso ad un sistema con risposta impulsiva $x(t)$, si ottiene ancora la stessa uscita, ossia l'integrale di convoluzione è commutativo[11]:

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

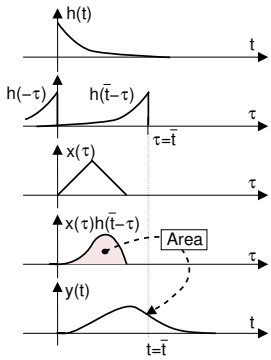



*Risposta impulsiva equivalente per sistemi in serie e parallelo*

Questa proprietà, assieme a quella di linearità, consente di stabilire le equivalenze mostrate in figura, dove si mostra come l'attraversamento *in serie* ed *in parallelo* di più sistemi lineari può essere ricondotto all'attraversamento di un sistema equivalente, con risposta impulsiva pari rispettivamente alla convoluzione ed alla somma delle singole risposte impulsive.

### 3.5.3 La risposta impulsiva come funzione memoria

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




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

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

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

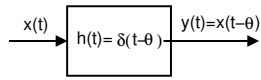
Il calcolo dell'area di $x(\tau)\, h\left(\bar{t}-\tau\right)$ ha il significato di sommare le risposte causate da tutti i valori di ingresso, ogni risposta presa con il rispettivo ritardo $\bar{t}-\tau$ tra gli istanti (passati) $\tau \leq \bar{t}$ di applicazione dei valori di ingresso, e l'istante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

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

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

### 3.5.4 Convoluzione con l'impulso traslato

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



*Elemento di ritardo*

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

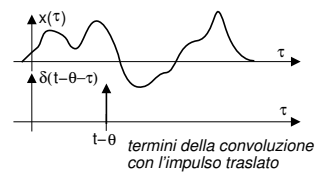
$$\begin{aligned} y(t) &= x(t) * h(t) = x(t) * \delta(t - \theta) = \\ &= \int_{-\infty}^{\infty} x(\tau)\ \delta(t - \theta - \tau)\ d\tau = x(t - \theta) \end{aligned}$$

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

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

## 3.6 Moltiplicazione in frequenza e nel tempo

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




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

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

La dimostrazione è riportata alla nota[12]. Sussiste inoltre anche la proprietà *duale*, ovvero ad un *prodotto nel tempo* corrisponde una *convoluzione in frequenza*, che si scrive

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

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

---

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

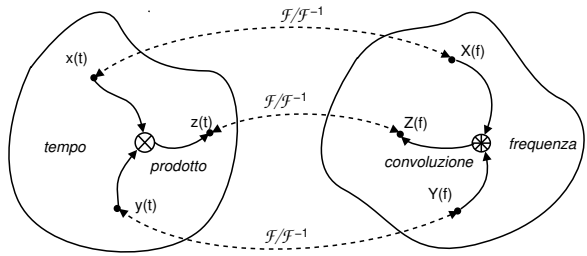



Figura 3.3: Isomorfismo tra gli spazi di segnale nel tempo e nella frequenza

### 3.6.1 Moltiplicazione in frequenza (*filtraggio*)

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

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

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

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

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

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

**Autovettori e misura di $H(f)$** Ricordando come in algebra lineare, applicando una trasformazione lineare ad un proprio autevettore, si ottiene l'autovettore stesso, moltiplicato per il rispettivo autovalore, osserviamo che la stessa definizione è ora perfettamente applicabile alle funzioni esponenziali complesse $\mathrm{e}^{j2\pi f_0 t}$, che risultano essere gli autovettori (o autofunzioni) di un sistema con risposta in frequenza $H(f)$, ed alle quali risulta associato l'autovalore $H(f_0)$. Questa particolarità consente di misurare $H(f)$ alle diverse frequenze, come illustrato a pag. 223, semplicemente osservando l'uscita del sistema, quando in ingresso è posto un segnale sinusoidale.

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

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

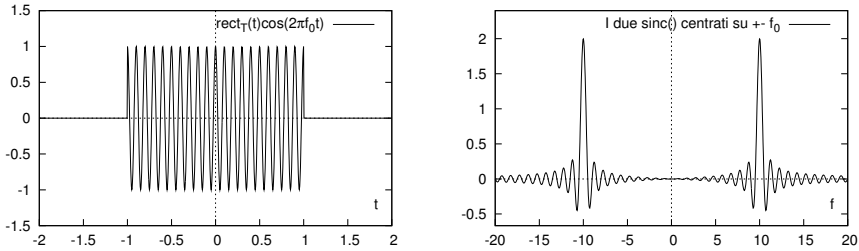



Figura 3.4: Trasformata di un coseno finestrato con $T = 2$, $f_0 = 10$

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

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

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

D'altra parte, scrivendo l'integrale di convoluzione, e ricordando che $h(t) = \mathcal{F}^{-1}\left\{\mathrm{e}^{-j2\pi f\theta}\right\} = \delta(t-\theta)$, avremmo ottenuto $y(t) = \int_{-\infty}^{\infty} x(\tau)\,\delta(t-\theta-\tau)\,d\tau = x(t-\theta)$, ritrovando la proprietà della convoluzione per un impulso traslato. Un sistema siffatto è indicato come *canale perfetto* a pag. 333, in quanto privo di distorsioni lineari (vedi § 14.5).

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

### 3.6.2 Moltiplicazione nel tempo *(modulazione e finestratura)*

La relazione

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

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

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

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

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

Il risultato dell'esempio, mostrato a destra in fig. 3.4, coincide[13] con quello previsto: l'espressione di $Z(f)$ infatti è anche pari alla convoluzione tra $\mathcal{F}\{rect_T(t)\}$, ed i due impulsi traslati $\mathcal{F}\{\cos 2\pi f_0 t\} = \frac{1}{2}(\delta(f-f_0) - \delta(f+f_0))$: determinando quindi la replica dello spettro del *rect*, traslata alla frequenza del coseno.

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

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

## 3.7 Derivazione ed integrazione nel tempo

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

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

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

e più in generale $\mathcal{F}\left\{\frac{d^n}{dt^n}x(t)\right\} = (j2\pi f)^n \cdot X(f)$. Per segnali di energia, la dimostrazione è svolta nella nota[14]. L'andamento dello spettro originario $X(f)$ risulta *esaltato* alle frequenze più elevate, in quanto il suo modulo è moltiplicato per $2\pi|f|$. La fase,

[13] Dalla figura si può anche arguire come, per $T$ crescente, $Z(f)$ tenda sempre più ad assomigliare ad una coppia di impulsi, ossia al risultato noto per un un coseno di durata *infinita*. Qualora si consideri invece solo un *breve intervallo* di un segnale, il suo spettro si modifica, a seguito della convoluzione con lo spettro della finestra di analisi.

L'operazione di estrazione di una porzione di segnale di durata limitata, a partire da un altro comunque esteso, è indicata come una operazione di *finestratura* (WINDOWING). In appendice (3.9.3) sono svolte considerazioni relative alla scelta di una finestra rettangolare o con *altro andamento*.

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

invece, subisce un incremento di $\frac{\pi}{2}$ a tutte le frequenze (il numero immaginario puro $j2\pi f = 2\pi f\, e^{j\frac{\pi}{2}}$ ha fase $\frac{\pi}{2}$).

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

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

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

Si ottiene:

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

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

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

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

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

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

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

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

In questo caso, le basse frequenze del segnale originario sono esaltate seguendo un andamento $1/2\pi|f|$, mentre la fase subisce una variazione (un ritardo) costante pari a $-\frac{\pi}{2}$. Notiamo come questo risultato determini una singolarità per $f = 0$ in presenza di componenti continue per $x(t)$: in tal caso infatti il suo integrale tende a divergere, ed il risultato non è più di energia.

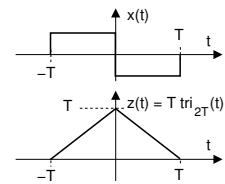
**Esempio:** Trasformata di un triangolo.

Consideriamo un segnale

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

ed il suo integrale

$$z(t) = \int_{-\infty}^{t} x(\theta)\, d\theta = T\, tri_{2T}(t)$$




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

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

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

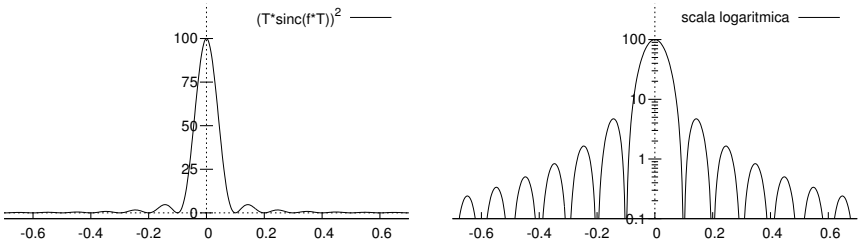



Figura 3.5: Andamento di $(T\mathrm{sinc}\,(fT))^2$in scala lineare e logaritmica; $T = 10$.

Dividendo quindi per $j2\pi f$ si ottiene

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

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

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

$$z\,(t) = T \cdot tri_{2T}\,(t) = rect_T\,(t) * rect_T\,(t) \tag{3.5}$$

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

$$Z\,(f) = \mathcal{F}\left\{T \cdot tri_{2T}\,(t)\right\} = \left[\mathcal{F}\left\{rect_T\,(t)\right\}\right]^2 = \left[Tsinc\,(fT)\right]^2 \tag{3.6}$$

Il risultato fornito da (3.6), è anche pari alla densità di energia $\mathcal{E}_y\,(f)$ di un segnale rettangolare $y\,(t) = rect_T\,(t)$: infatti per il teorema di Parseval, si ha $\mathcal{E}_y\,(f) = Y\,(f)\,Y^*\,(f)$, in cui $Y\,(f) = \mathcal{F}\left\{rect_T\,(t)\right\} = Tsinc\,(fT)$, e pertanto

$$\mathcal{E}_y\,(f) = \left[Tsinc\,(fT)\right]^2$$

## 3.8 Trasformata di segnali periodici

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

### 3.8.1 Treno di impulsi

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

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

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

### 3.8.2 Segnale periodico

Consideriamo un segnale periodico di periodo $T$ espresso come

$$x(t) = \sum_{m=-\infty}^{\infty} g(t - mT) \tag{3.7}$$

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

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

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

### 3.8.3 Trasformata del treno di impulsi

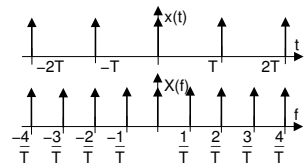
L'approccio che conviene seguire è di pensare a $\pi_T(t)$ come ad un segnale periodico, e svilupparlo in serie di Fourier. I coefficienti si calcolano allora come:

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

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

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

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



*treno di impulsi e sua trasformata*

### 3.8.4 Trasformata di segnale periodico

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

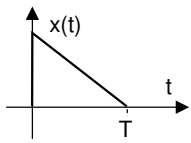
$$X(f) = G(f) \cdot \frac{1}{T}\pi_{\frac{1}{T}}(f)$$

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

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

## 3.9 Appendici

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




Sia dato il segnale

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

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

1. Si calcola $X(f) = \mathcal{F}\{x(t)\} = \int_{-\infty}^{\infty} x(t)\,e^{-j2\pi ft}dt$ e quindi $\mathcal{E}_x(f) = |X(f)|^2$;
2. Anticipando un risultato del capitolo 7, è possibile calcolare $\mathcal{R}_x(f) = \int_{-\infty}^{\infty} x(t)\,x(t+\tau)\,dt$, e quindi $\mathcal{E}_x(f) = \mathcal{F}\{\mathcal{R}_x(f)\}$;
3. Notando che $x(t) = y(t)\cdot z(t)$ con $y(t) = \mathrm{tri}_{2T}(t)$ e $z(t) = \mathrm{rect}_T\left(t-\frac{T}{2}\right)$, possiamo scrivere $X(f) = Y(f) * Z(f)$, e quindi procedere come in 1);
4. Notiamo che la derivata[15] di $x(t)$ vale $g(t) = \frac{d}{dt}x(t) = \delta(t) - \frac{1}{T}\mathrm{rect}_T\left(t-\frac{T}{2}\right)$; questo ci permette di calcolare $G(f)$ come $G(f) = \mathcal{F}\{g(t)\} = 1-\mathrm{sinc}(fT)$. Otteniamo quindi $X(f) = \frac{G(f)}{j2\pi f}$, e procediamo come in 1).

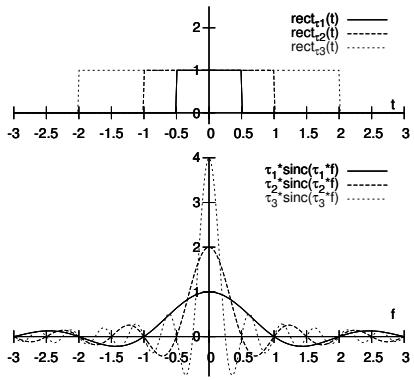
### 3.9.2 Sulla trasformata di una costante

Svolgiamo alcune considerazioni sul risultato mostrato a pag. 35, illustrando come l'impulso $\delta(.)$ permetta di rappresentare particolari situazioni. Consideriamo pertanto il segnale costante $x(t) = A$, che trattiamo come un segnale periodico con periodo $T$ tendente ad $\infty$[16], ed esprimiamo $x(t)$ nei termini dei coefficienti di Fourier. L'integrale $X_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} A\,\mathrm{e}^{-j2\pi nFt}dt$ per $T\to\infty$ fornisce zero per tutti gli $n$ tranne che per $n=0$, e quindi si ottiene $X_n = \begin{cases} A & \text{con} \quad n=0 \\ 0 & \text{con} \quad n\neq 0 \end{cases}$.

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

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

[16] In effetti, un segnale costante è un segnale periodico, con periodo $T$ qualsiasi.




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

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

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

Al tendere di $\tau$ ad $\infty$, l'energia diviene infinita, mentre la potenza vale

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

L'espressione $\lim_{\tau\to\infty} \frac{|X_\tau(f)|^2}{\tau}$ rappresenta dunque[17] lo *spettro di densità di potenza* $\mathcal{P}_x(f)$ della costante A, che finalmente scriviamo come $\mathcal{P}_x(f) = A^2\delta(f)$, in cui $\delta(f)$ è la funzione *impulso matematico* introdotta in 3.4. In tal modo infatti, è facile verificare che risulta $\mathcal{P}_x = \int_{-\infty}^{\infty} A^2\delta(f)\,df = A^2$, e $\mathcal{P}_x(f) = \begin{cases} \infty & \text{con } f = 0 \\ 0 & \text{altrove} \end{cases}$. In altre parole, il formalismo dell'impulso matematico rende possibile trattare questo caso, dove la potenza (finita) è tutta concentrata in un unico punto ($f = 0$) dando luogo ad una densità infinita.

### 3.9.3 Finestratura e stima spettrale

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

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

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

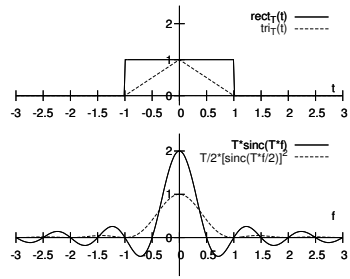
che è tanto più diverso dalle due linee spettrali del coseno (vedi Fig. 3.4), quanto più è piccolo $T$.

[17] vedi anche §9.3.1 a pag. 215

Valutiamo ora gli effetti derivanti dall'uso di una funzione finestra diversa da quella rettangolare. Se ad esempio si sceglie di adottare una finestra triangolare di eguale durata $T$, il risultato mostrato a pag. 43 dalle (3.5) e (3.6) permette di ottenere

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

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




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

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

Considerazioni analoghe a quelle ora svolte, possono essere intraprese per diverse scelte[18] di funzione finestra, in dipendenza dal particolare obiettivo della stima spettrale.

### 3.9.4 Trasformata di un gradino

Definiamo la funzione gradino come $g(t) = \begin{cases} 1 & \text{per} \quad t > 0 \\ \frac{1}{2} & \text{per} \quad t = 0 \\ 0 & \text{per} \quad t < 0 \end{cases}$ che, fornendo $\int_{-\infty}^{\infty} |g(t)|\, dt = \infty$, non dovrebbe avere una trasformata. Come già visto per la costante, il calcolo può essere condotto a termine come limite a cui tende la trasformata di una diversa funzione, il cui limite tende al gradino. Scegliamo quindi $g_\alpha(t) = e^{-\alpha t}$ per $t > 0$, per

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

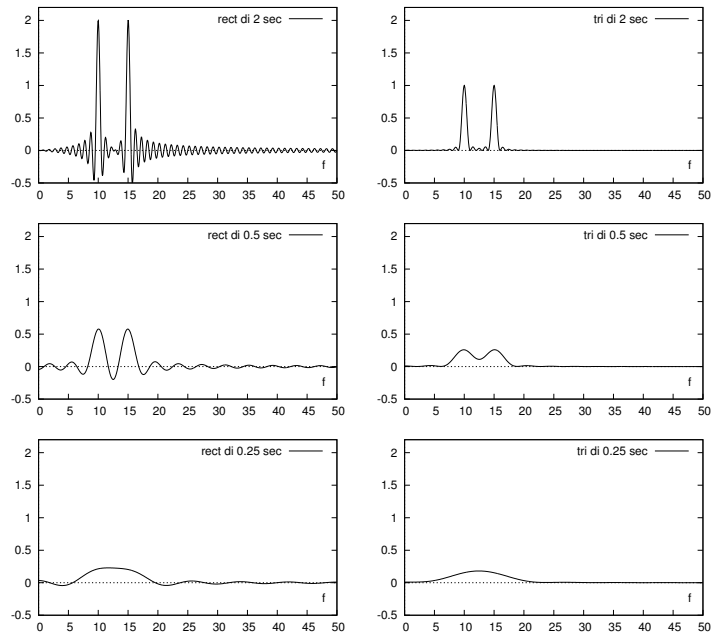



Figura 3.6: Trasformata di due toni a 10 e 15 Hz, con finestra temporale $\mathrm{rect}_T$ e $\mathrm{tri}_T$ di durata 2, .5 e .25 secondi

la quale risulta $\lim_{\alpha\to 0} g_\alpha(t) = g(t)$, e troviamo

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

Si può mostrare che

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

mentre risulta

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

ottenendo[19] in definitiva

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

### 3.9.5 Sintesi delle proprietà della trasformata di Fourier

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

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

### 3.9.6 Trasformate di segnali

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

# Capitolo 4

# Campionamento ed elaborazione numerica

In questo capitolo è esposta la teoria che consente di rappresentare un segnale per mezzo dei suoi *campioni* temporali, permettendone la rappresentazione in forma numerica: i campioni infatti, sono una sequenza di *numeri*[1]. Molte operazioni che si svolgono sul segnale originario, possono quindi essere eseguite direttamente sulla sua rappresentazione numerica, dando luogo alla *elaborazione numerica* dei segnali.

## 4.1 Teorema del campionamento

> *Un segnale con spettro nullo a frequenze maggiori di $W$, è univocamente definito a partire dai valori che assume agli istanti $t = \frac{n}{2W}$, con $n$ intero.*

La frequenza $2W$ è chiamata *frequenza di Nyquist*. In virtù del teorema, l'andamento di un segnale $x(t)$ limitato in banda tra $-W$ e $W$ può essere ricostruito in base ai suoi campioni, presi a frequenza doppia della sua banda a frequenze positive, per mezzo della formula:

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2W}\right) \cdot \mathrm{sinc}\left(2W\left(t - \frac{n}{2W}\right)\right) \tag{4.1}$$

in cui la funzione $sinc(2Wt) = \frac{\sin 2\pi Wt}{2\pi Wt}$ è mostrata in Fig. 4.1, assieme ad una sua replica traslata.

Per dimostrare il risultato, adottiamo lo schema simbolico mostrato in Fig. 4.2, che

[1] DIGITS in inglese, da cui il termine *digitale* come sinonimo di *numerico*.

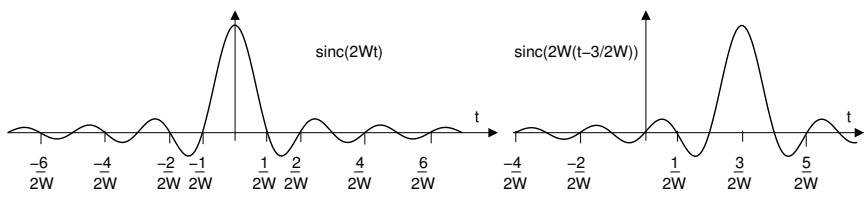



Figura 4.1: La funzione $\mathrm{sinc}(2Wt)$ centrata in 0 e traslata in $\frac{3}{2W}$

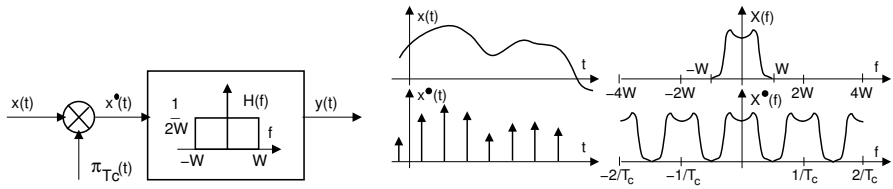
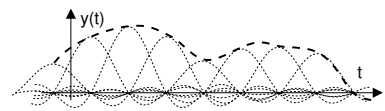



Figura 4.2: Circuito di campionamento e restituzione, e spettri dei segnali

(come mostreremo) realizza le operazioni della formula di ricostruzione. Innanzitutto, viene generato il segnale $x^{\bullet}(t) = \sum_{n=-\infty}^{\infty} x(nT_c)\,\delta(t - nT_c)$ mediante moltiplicazione di $x(t)$ limitato in banda per un treno di impulsi con periodo $T_c = \frac{1}{2W}$. Lo spettro di ampiezza di $X^{\bullet}(f) = \mathcal{F}\{x^{\bullet}(t)\}$ risulta quindi

$$\begin{aligned} X^{\bullet}(f) &= \mathcal{F}\{x(t)\cdot\pi_{T_c}(t)\} = X(f) * \frac{1}{T_c}\Pi_{\frac{1}{T_c}}(f) = X(f) * \frac{1}{T_c}\sum_{n=-\infty}^{\infty}\delta\left(f - \frac{n}{T_c}\right) = \\ &= 2W\cdot\sum_{n=-\infty}^{\infty} X(f) * \delta\left(f - \frac{n}{T_c}\right) = 2W\cdot\sum_{n=-\infty}^{\infty} X\left(f - \frac{n}{T_c}\right) \end{aligned}$$

dove il penultimo passaggio scambia l'integrale di una somma con una somma di integrali, e l'ultimo passaggio tiene conto della proprietà di convoluzione con un impulso. In definitiva si è mostrato che $X^{\bullet}(f)$ è costituito dalle repliche di $X(f)$ centrate a multipli della frequenza di campionamento. Pertanto, il filtro passa-basso $H(f)$ (chiamato anche con il nome di *filtro di restituzione*) lascia passare solo *una* delle repliche spettrali, e dunque è evidente come $Y(f) = \frac{1}{2W}2W\cdot X(f) = X(f)$.



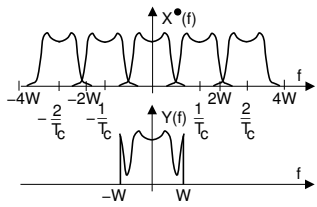

Anche la formula di ricostruzione (4.1) che fa uso dei campioni $x\left(\frac{n}{2W}\right)$ e delle funzioni $sinc(2Wt)$ deriva dallo schema descritto, e può essere illustrata con l'aiuto della figura a fianco. Infatti, $y(t)$ è il risultato della convoluzione tra $x^{\bullet}(t)$ e $h(t) = \mathcal{F}^{-1}\{H(f)\} = \mathcal{F}^{-1}\left\{\frac{1}{2W}\mathrm{rect}_{2W}(f)\right\} = \mathrm{sinc}(2Wt)$, e dunque ogni impulso di cui è composto $x^{\bullet}(t)$, quando convoluto con $h(t)$, trasla la forma d'onda $h(t)$ all'istante $nT_c$ a cui è centrato l'impulso. In formule:

$$\begin{aligned} y(t) &= \left[x(t)\cdot\sum_n \delta(t - nT_c)\right] * h(t) = \sum_n x(nT_c)\,\delta(t - nT_c) * \mathrm{sinc}(2Wt) = \\ &= \sum_n x(nT_c)\,\mathrm{sinc}(2W(t - nT_c)) \end{aligned}$$

Questo risultato mostra come il teorema del campionamento definisca essenzialmente una formula di interpolazione: i valori del segnale ricostruito hanno l'esatto valore dei campioni di segnale negli istanti di campionamento, mentre negli istanti intermedi il valore si forma dalla somma di tutte le "code" dei *sinc* adiacenti. Il processo di costruzione grafica ora descritto è riportato nella figura precedente.

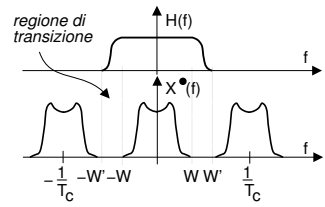
### 4.1.1 Aliasing

Questo termine ha origine dalla parola inglese[2] *alias* (copia, clone) e sta ad indicare il fenomeno che si produce nell'applicare il teorema del campionamento quando i requisiti non sono soddisfatti, e cioè quando la frequenza di campionamento è inferiore alla frequenza di Nyquist, ossia $f_c = \frac{1}{T_c} < 2W$ (ovvero $T_c > \frac{1}{2W}$). In questo caso le repliche spettrali che compongono $\overset{\bullet}{X}(f)$ sono più ravvicinate, e si sovrappongono (l'aliasing è indicato anche come fold-over, *ripiegamento*).




Quando questo avviene, il filtro passa-basso di restituzione non è più in grado di estrarre la replica centrata in $f = 0$ (vedi figura a lato), e dunque alla sua uscita è presente $y(t) \neq x(t)$, che si differenzia da $x(t)$ in particolar modo per i contenuti energetici nella regione delle frequenze più elevate. In un segnale audio, ad esempio, ci si accorge che c'è aliasing quando è udibile una distorsione (rumore) congiuntamente ai passaggi con maggior contenuto di alte frequenze.

Il fenomeno dell'aliasing può insorgere, oltre che nel caso in cui si commetta il banale errore di adottare $T_c > \frac{1}{2W}$, anche a causa di una imperfetta limitazione in banda di $x(t)$ (che viene in genere filtrato proprio per accertarsi che sia $X(f) \approx 0$ con $|f| > W$).




Altri problemi possono essere causati dal filtro di restituzione $H(f)$, che difficilmente si riesce a realizzare ideale. Questo può presentare infatti una regione di transizione tra banda passante e banda soppressa di larghezza non nulla (vedi figura). In questo caso occorre sovracampionare con periodo $Tc = \frac{1}{2W'} < \frac{1}{2W}$, in modo che le repliche spettrali siano più distanziate tra loro, e quindi il filtro di ricostruzione possa isolare la replica centrale.

### 4.1.2 Energia di un segnale campionato

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

$$\int_{-\infty}^{\infty} \operatorname{sinc}(2W(t - kT_c)) \operatorname{sinc}(2W(t - hT_c))\, dt = \begin{cases} 0 & \text{se} \quad h \neq k \\ \frac{1}{2W} & \text{se} \quad h = k \end{cases}$$

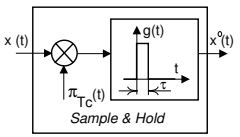
Pertanto, il valore dell'energia di un segnale limitato in banda è calcolabile a partire dai suoi campioni, e vale:

$$\begin{aligned} \mathcal{E}_x &= \int_{-\infty}^{\infty} x(t)\, x^*(t)\, dt = \sum_k \sum_h x_k x_h^* \int_{-\infty}^{\infty} \operatorname{sinc}(2W(t - kT_c)) \operatorname{sinc}(2W(t - hT_c))\, dt \\ &= \sum_k \sum_h x_k x_h^* \frac{1}{2W} \delta(h, k) = \frac{1}{2W} \sum_k |x_k|^2 \end{aligned}$$

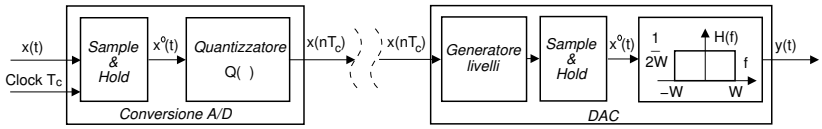
[2] In realtà *alias* è di origine latina !!!

### 4.1.3 Uso pratico

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

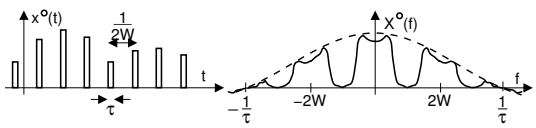



Per coglierne i lati pratici, consideriamo innanzitutto che *non viene* generato il segnale $X^{\bullet}(f)$, in quanto i campioni di $x(t)$ sono in realtà ottenuti mediante un circuito *Sample and Hold* (S&H, ovvero *campiona e mantieni*) che produce una uscita costante (per un tempo $\tau$), pari al valore assunto dall'ingresso negli istanti di campionamento (detti di *clock* = orologio); l'uscita del S&H viene quindi *quantizzata* (vedi il § 7.4), ovvero misurata e convertita in un valore numerico dal dispositivo $Q(\cdot)$ .




Campionamento e restituzione nel *mondo reale*

I valori così ottenuti possono essere memorizzati, oppure trasmessi. Per ricostruire il segnale originario si adotta un DAC (*Digital to Analog Converter*, ossia Convertitore Digitale-Analogico) che può essere realizzato dai tre componenti mostrati in figura, ossia un dispositivo che per ogni diverso valore numerico genera un segnale di ampiezza pari ad uno dei livelli di quantizzazione, un S&H ed un filtro passa-basso di restituzione[3].




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

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

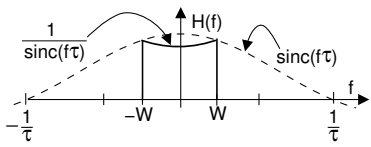
e dunque

$$X^{\circ}(f) = X^{\bullet}(f) \cdot \tau \,\mathrm{sinc}(f\tau) \tag{4.2}$$

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

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

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




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

## 4.2 Trasformata discreta di Fourier

L'analisi in frequenza di un segnale discussa al cap. 3 può essere condotta mediante programmi di elaborazione su computer[4], utilizzando i campioni $x_m = x(mT_c)$ estratti dallo stesso, prelevati ad intervalli fissi $T_c$.

Disponendo di una sequenza di $N$ valori $x_m$, $m = 0, 1, ..., N-1$, indichiamo come DISCRETE FOURIER TRANSFORM (**DFT**) la nuova sequenza

$$X_n = \sum_{m=0}^{N-1} x_m \mathrm{e}^{-j2\pi \frac{m}{N} n} \tag{4.3}$$

univocamente definita per $n = 0, 1, ..., N-1$, e che costituisce una approssimazione[5] della sequenza di campioni della trasformata $X(f) = \mathcal{F}\{x(t)\}$, calcolata per $f = \frac{n}{NT_c}$, e divisa per $T_c$:

$$X_n \simeq \frac{1}{T_c} X\left(f = \frac{n}{NT_c}\right) \tag{4.4}$$

Notiamo subito che la (4.3) è valida per qualsiasi $n$, ed ha un andamento periodico con periodo $N$, a cui corrisponde una frequenza $f = \frac{1}{T_c}$, in accordo con la separazione

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

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

$$\begin{aligned} X_w\left(f = \frac{n}{NT_c}\right) &= \int_0^{(N-1)T_c} x(t)\, e^{-j2\pi \frac{n}{NT_c} t} dt \\ &\simeq \sum_{m=0}^{N-1} x_m \cdot \int_0^{(N-1)T_c} \mathrm{sinc}\left(\frac{t - mT_c}{T_c}\right) e^{-j2\pi \frac{n}{NT_c} t} dt \end{aligned}$$

in cui la seconda eguaglianza utilizza l'interpolazione $x(t) = \sum_{m=-\infty}^{\infty} x_m \cdot \mathrm{sinc}\left(\frac{t-mT_c}{T_c}\right)$ fornita dal teorema del campionamento, ed introduce una seconda fonte di approssimazione legata all'intervallo *finito* di variazione per $m$ (infatti, benchè l'integrale abbia estensione limitata, i valori di $x(t)$ che cadono entro tale estensione, dovrebbero dipendere da *tutti* i suoi campioni). L'ultimo integrale è a sua volta una approssimazione (a causa degli estremi di integrazione limitati, e peggiore per i sinc centrati in prossimità dei confini della finestra) della trasformata (calcolata in $f = \frac{n}{NT_c}$) di $\mathrm{sinc}\left(\frac{t-mT_c}{T_c}\right)$, pari quest'ultima a $T_c \mathrm{rect}_{\frac{1}{T_c}}(f)\, e^{-j2\pi f m T_c}$, che quando valutata per $f = \frac{n}{NT_c}$, fornisce il risultato

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

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

a - modulo della DFT di una sinusoide a 10 Hz, $f_c = 100$ Hz, con finestre di diversa lunghezza

b - modulo della $\mathcal{F}$-trasformata di una sinusoide a 10 Hz, con finestre di diversa lunghezza

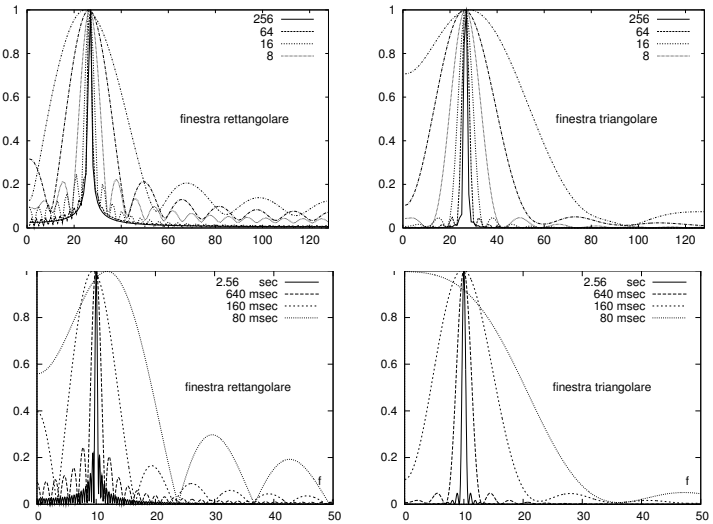



Figura 4.3: Confronto tra DFT ed $\mathcal{F}$-trasformata con uguale estensione temporale

tra le repliche spettrali prevista dal teorema del campionamento; per questo motivo, qualora il segnale originario $x(t)$ contenga componenti a frequenze maggiori di $\frac{1}{2T_c}$, gli $X_n$ con indici prossimi ad $\frac{N}{2}$ presentano errore di aliasing[6]. Notiamo inoltre che la (4.3) può essere espressa in forma matriciale: ad esempio, per $N = 4$ si ottiene

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

in cui notiamo le caratteristiche di simmetria per la matrice dei coefficienti.

Allo scopo di concretizzare le differenze tra la trasformata di Fourier ed i valori forniti dalla DFT, in fig. 4.3-a sono riportati i valori $|X_n|$ per la DFT di una sinusoide, adottando due diverse finestre di analisi, prelevando alla medesima frequenza di campionamento (100 Hz) un numero variabile di campioni (mostrato in figura), e ponendo i rimanenti a zero, per calcolare in tutti i casi la medesima DFT a 256 punti[7].

Il risultato è quindi confrontato (fig. 4.3-b) con quello ottenibile per via analitica calcolando la $\mathcal{F}$-trasformata dello stesso segnale, adottando le medesime finestre tem-

[6] Come osservato al § 4.1.1, lo spettro $X^\bullet(f)$ di un segnale campionato a frequenza $f_c$ è costituito dalle repliche del segnale originario, distanziate di multipli di $f_c$: $X^\bullet(f) = \sum_{n=-\infty}^{\infty} X(f - nf_c)$, e coincide con $X(f)$ per $-f_c/2 < f < f_c/2$, se $X(f)$ è limitata in banda tra $\pm W$ ed $f_c \geq 2W$. Al contrario, se $f_c < 2W$, allora le repliche $X(f - nf_c)$ si sovrappongono, e la (4.4) si riscrive come $X_n \simeq \frac{1}{T_c} X^\bullet\left(f = \frac{n}{NT_c}\right)$.

[7] Il metodo esposto di porre a zero i campioni fino al raggiungimento di una potenza di due, in modo da utilizzare per il calcolo una FFT, è detto ZERO PADDING. Il calcolo della DFT su di un numero di punti pari ai campioni di segnale disponibili, non avrebbe dato luogo all'effetto finestra, ma avrebbe fornito in tutti i casi andamenti simili a quello osservabile per 256 punti. Infine, notiamo che nelle figure sono mostrati solo i primi 128 valori, essendo i rimanenti speculari.

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

Il passaggio dai campioni $x_m$ alla sequenza $X_n$ è invertibile[8], ricorrendo alla INVERSE DISCRETE FOURIER TRASFORM (**IDFT**)

$$x_m = \frac{1}{N}\sum_{n=0}^{N-1} X_n \mathrm{e}^{j2\pi\frac{n}{N}m} \tag{4.5}$$

che per $m$ esterno a $[0, N-1]$ continua a valere, ed assume valori periodici, coerentemente a quanto accade per lo sviluppo in serie di Fourier. Infatti il legame tra DFT e serie di Fourier è molto stretto, in quanto i valori $X_n$ rappresentano una approssimazione[9] dei rispettivi coefficienti della serie di Fourier $X_n^{SF} = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x_T(t)\, e^{-j2\pi\frac{n}{T}t}dt$, calcolati a partire dal segmento $x_T(t)$ estratto da $x(t)$, e moltiplicati per $N$:

$$X_n \simeq N \cdot X_n^{SF} \tag{4.6}$$

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

### 4.2.1 Relazione tra DFT e trasformata z

Così come per i segnali analogici sussiste una relazione (vedi pag. 229) tra la trasformata di FOURIER e quella di LAPLACE, così nel contesto delle sequenze, esistono legami tra DFT e *trasformata zeta*, definita come

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)\, \mathrm{z}^{-n} \tag{4.7}$$

che, nel caso in cui la serie converga per $|z| = 1$, permette di definire la *trasformata di Fourier per sequenze* $X\left(\mathrm{e}^{j\omega}\right)$, ottenuta calcolando $X(z)$ sul cerchio unitario $z = \mathrm{e}^{j\omega}$

$$X\left(\mathrm{e}^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x(n)\, \mathrm{e}^{-j\omega n} = X(z)|_{z=e^{j\omega}}$$

che, se la sequenza $x(n)$ è ottenuta per campionamento, con periodo $T \leq \frac{1}{2W}$, di un segnale $x(t)$ limitato in banda tra $\pm W$, coincide (per $-\pi \leq \omega < \pi$) con la trasformata $X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi ft}dt$ calcolata in $-\frac{1}{2T} \leq f < \frac{1}{2T}$.

---

[8] Sostituendo la (4.3) nella (4.5), otteniamo

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

ma, dato che $\sum_{n=0}^{N-1}\mathrm{e}^{j2\pi\frac{m-k}{N}n} = \begin{cases} N & \text{se } k = m + lN \\ 0 & \text{altrimenti} \end{cases}$ con $l$ intero, allora nella sommatoria esterna sopravvive solo il termine $x_m$, essendo $x$ costituito solo da $N$ valori.

[9] La relazione (4.6) si dimostra combinando le relazioni (3.1) e (4.4): $X_n \simeq \frac{1}{T_c}X\left(\frac{n}{NT_c}\right) = \frac{1}{T_c}X\left(\frac{n}{T}\right) = \frac{1}{T_c}TX_n^{SF} = NX_n^{SF}$

Al di fuori di tale intervallo, $X\left(\mathrm{e}^{j\omega}\right)$ è periodica in $\omega$ con periodo $2\pi$, analogamente a ciò che risulta per la trasformata di Fourier $X^{\bullet}(f)$ di un segnale campionato; in particolare, se $x(m)$ è ottenuta campionando con periodo $T > \frac{1}{2W}$, allora $X\left(\mathrm{e}^{j\omega}\right)$ corrisponde proprio ad $X^{\bullet}(f)|_{f=\omega\frac{\pi}{W}}$, affetta da aliasing.

Se la $X(z)$, ottenuta per una sequenza $x(n)$ aperiodica, è campionata in $N$ punti equispaziati e disposti sul cerchio unitario, ossia per $z = \mathrm{e}^{-j2\pi\frac{k}{N}n}$, con $k = 0, 1, \ldots, N-1$, si ottiene una sequenza periodica[10]

$$\tilde{X}(k) = \sum_{n=-\infty}^{\infty} x(n)\,\mathrm{e}^{-j2\pi\frac{k}{N}n} = X(z)|_{z=e^{-j2\pi\frac{k}{N}n}} = X\left(\mathrm{e}^{j\omega}\right)\Big|_{\omega=2\pi\frac{k}{N}} \tag{4.8}$$

a cui è possibile applicare la (4.5) per ottenere una nuova sequenza di valori nel tempo, periodica di periodo $N$, espressa come

$$\tilde{x}(n) = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{X}(k)\,\mathrm{e}^{j2\pi\frac{k}{N}n} \tag{4.9}$$

I valori $\tilde{x}(n)$ dipendono da quelli $x(n) = x(t)|_{t=nT}$ del segnale originario $x(t)$, campionato agli istanti $t = nT$, mediante la relazione[11]

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

e quindi i primi $N$ valori di $\tilde{x}(n)$ coincidono con i campioni di $x(t)$ solo se quest'ultimo ha durata limitata, con estensione minore di $NT$, ossia se $N$ è sufficientemente elevato in modo che $NT$ copra tutta la durata di $x(t)$, e la (4.7) si riconduca alla somma di un numero finito di termini. D'altra parte, se $x(t)$ ha durata maggiore di $NT$, ovvero $X(z)$ è stata campionata su di un numero di campioni troppo ristretto, allora l'applicazione della IDFT (4.9) ad $\tilde{X}(k)$ provoca il fenomeno di *aliasing temporale*.

### 4.2.2 Relazione tra DFT e DCT

Anche per la DFT risulta valida la proprietà di simmetria coniugata (§ 3.3) e quindi, se i valori della sequenza $x_m$ di lunghezza $N$ che compare nella (4.3) sono reali anziché complessi, allora i coefficienti di DFT $X_n$ presentano parte reale pari e parte immaginaria dispari. In particolare, se immaginiamo di *estendere* la lunghezza della sequenza a $2N$ punti, ottenuti ribaltando sugli indici negativi la sequenza di partenza come $x_{-m} = x_m$ (vedi prima riga di fig. 4.4), allora siamo nelle condizioni di sequenza *reale pari*, che determina una trasformata solo reale (e pari), con parte immaginaria nulla.

[10] Indichiamo qui ed al prossimo §, una sequenza periodica mediante la tilde ˜.

[11] Infatti, sostituendo la (4.8) in (4.9), otteniamo

$$\tilde{x}(n) = \frac{1}{N}\sum_{k=0}^{N-1}\sum_{m=-\infty}^{\infty} x(m)\,\mathrm{e}^{-j2\pi\frac{k}{N}m}\mathrm{e}^{j2\pi\frac{k}{N}n}$$

che, scambiando l'ordine delle sommatorie, riscriviamo come

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

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

Per arrivare a definire la DISCRETE COSINE TRANSFORM (DCT) si calcola una DFT *bilatera* sulla sequenza lunga $2N$ ottenuta traslando quella descritta in modo da renderla effettivamente pari (seconda riga di fig. 4.4). Considerando che per segnali reali pari le componenti sinusoidali della base della DFT non danno contributi al risultato[12], e adottando un nuovo cambio di variabile, si ottiene in definitiva la formula di calcolo della DCT come

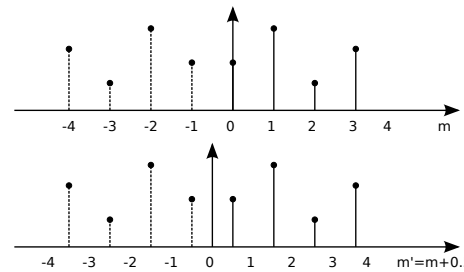


Figura 4.4: Estensione pari di sequenza reale

$$X_n = \sum_{m=0}^{N-1} x_m \cos\left[\frac{\pi}{N}\left(m+\frac{1}{2}\right)n\right] \tag{4.10}$$

a cui è associata la trasformazione inversa IDCT

$$x_m = \frac{1}{2}X_0 + \sum_{n=1}^{N-1} X_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)m\right] \tag{4.11}$$

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

### 4.2.3 Filtraggio numerico via DFT

La definizione di DFT illustrata al § 4.2 ben si presta a calcolare il risultato relativo ad un integrale di convoluzione, a patto di seguire alcune accortezze.

**Convoluzione discreta** Dati due segnali $x(t)$ e $h(t)$ limitati in banda tra $-W$ e $W$, anche il risultato della convoluzione $y(t) = x(t) \star h(t)$ è limitata in banda, ed i

[12] Scriviamo la (4.3) come

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

in cui $x_{m'}$ è quella disegnata per seconda in fig. 4.4. La quarta eguaglianza tiene conto del fatto che il termine immaginario si annulla, in quanto sommatoria bilatera di una funzione dispari (ottenuta come prodotto di $x_{m'-1/2}$ pari e $\sin\left(2\pi\frac{m'}{2N}n\right)$ dispari), e del fatto che essendo i termini coseno pari, la sommatoria può essere ristretta ai soli indici positivi, raddoppiati. La penultima eguaglianza rappresenta il semplice cambio di variabile $m = m' - 1/2$, mentre l'ultima è (a parte il fattore 2) la definizione della DCT data in (4.10).

suoi campioni $h(n) = h(nT_c)$ (con $T_c > \frac{1}{2W}$) possono essere calcolati[13] a partire da quelli di $x(t)$ e $h(t)$ come $y(n) = \sum_{m=-\infty}^{\infty} x(m)\, h(n-m)$

**Convoluzione circolare** Date due sequenze $x(n)$ ed $h(n)$ di durata finita $N$, il prodotto $Y(k) = X(k)\, H(k)$ delle rispettive DFT $X(k) = \sum_{n=0}^{N-1} x(n)\, \mathrm{e}^{-j2\pi \frac{n}{N} k}$ ed $H(k) = \sum_{n=0}^{N-1} h(n)\, \mathrm{e}^{-j2\pi \frac{n}{N} k}$ possiede antitrasformata $\tilde{y}(n) = \mathcal{DFT}^{-1}\{Y(k)\}$ periodica in $n$ di periodo $N$, e pari a

$$\tilde{y}(n) = \sum_{m=0}^{N-1} \tilde{x}(m)\, \tilde{h}(n-m) \tag{4.12}$$

in cui $\tilde{x}(n)$ e $\tilde{h}(n)$ sono le sequenze periodiche di periodo $N$ ottenute replicando infinitamente le sequenze originali $x(n)$ ed $h(n)$ ([14]). La convoluzione (4.12) è detta *circolare* perché è possibile immaginare le sequenze $x(n)$ ed $h(n)$ *incollate* su due cilindri concentrici, e la somma svolta sui prodotti degli elementi coincidenti. Ad ogni valore di $n$, corrisponde una diversa rotazione relativa (con angolo multiplo di $2\pi/N$) dei cilindri, ed il campione di $h(.)$ che era allineato ad $x(N-1)$ *rientra* dall'altro lato, per corrispondere con $x(0)$.

**Convoluzione tra sequenze limitate** Sappiamo che la convoluzione produce un risultato di durata pari alla somma delle durate degli operandi; per l'esattezza, nel caso di due sequenze $x(n)$ ed $h(n)$ di durata $N$ ed $M$, il risultato della convoluzione discreta $y(n) = \sum_{m=-\infty}^{\infty} x(m)\, h(n-m)$ ha estensione $N+M-1$. Pertanto, perché la (4.12) produca lo stesso effetto di una convoluzione discreta, occorre costruire delle sequenze $x'(n)$ e $h'(n)$ di lunghezza almeno pari ad $N+M-1$, ottenute a partire dai valori di $x(n)$ ed $h(n)$, a cui si aggiungono $M-1$ ed $N-1$ valori nulli, rispettivamente. In tal modo, il prodotto $X'(k)\, H'(k)$ tra le DFT ad $N+M-1$ punti di queste due nuove sequenze, può essere antitrasformato, per fornire il risultato corretto.

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

Ciononostante, disponendo di un numero di campioni sufficientemente elevato, si può assumere che la trasformata della finestra si attenui, fino a rendersi trascurabile, oltre ad una certa frequenza. Inoltre, l'adozione di una frequenza di campionamento

[13] Il risultato può essere ottenuto esprimendo l'integrale $x(t) \star h(t)$ nei termini dei campioni di $x(t)$ e $h(t)$, e sfruttando la proprietà di ortogonalità (vedi § 4.1.2) di $\mathrm{sinc}(.)$.

[14] Infatti, ad $x(n)$ ed $h(n)$ corrispondono le DFT periodiche $\tilde{X}(k)$ ed $\tilde{H}(k)$, che hanno per antitrasformata $\tilde{x}(n)$ ed $\tilde{h}(n)$. Il prodotto $\tilde{X}(k)\, \tilde{H}(k)$, espresso in termini di $\tilde{x}(n)$ ed $\tilde{h}(n)$, risulta pari a $\tilde{Y}(k) = \tilde{X}(k)\, \tilde{H}(k) = \sum_{m=0}^{N-1} \sum_{r=0}^{N-1} \tilde{x}(m)\, \tilde{h}(r)\, \mathrm{e}^{-j2\pi \frac{m+r}{N} k}$, ed applicando a questo la (4.5), otteniamo:

$$\tilde{y}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{Y}(k)\, \mathrm{e}^{j2\pi \frac{n}{N} k} = \frac{1}{N} \sum_{m=0}^{N-1} \sum_{r=0}^{N-1} \tilde{x}(m)\, \tilde{h}(r) \left( \sum_{k=0}^{N-1} \mathrm{e}^{j2\pi \frac{n-m-r}{N} k} \right)$$

Dato che $\sum_{k=0}^{N-1} \mathrm{e}^{j2\pi \frac{n-m-r}{N} k} = \begin{cases} N & \text{se } r = (n-m) + lN \\ 0 & \text{altrimenti} \end{cases}$, con $l$ intero, risulta allora $\tilde{y}(n) = \sum_{m=0}^{N-1} \tilde{x}(m)\, \tilde{h}(n-m)$, come anticipato.

più elevata, provoca un allontanamento delle repliche spettrali del segnale campionato. In queste due ipotesi, è lecito ritenere l'elaborazione condotta sui campioni di segnale equivalente a quella da svolgere sul segnale originario.

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

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

in modo che $x(n) = \sum_{q=-\infty}^{\infty} x_q(n)$, ed operare una successione di convoluzioni discrete $y_q(n) = x_q(n) \star h(n)$, in modo da ottenere $y(n) = x(n) \star h(n) = \sum_{q=-\infty}^{\infty} x_q(n) \star h(n)$ per la linearità della convoluzione.

Osserviamo ora che ognuno dei termini $y_q(n)$ risulta di estensione $N = M + L - 1$ punti, e può essere calcolato mediante una DFT inversa ad $N$ punti del prodotto $X'_q(k) H'(k)$ tra le DFT ad $N$ punti delle versioni *allungate con zero* (ZERO PADDED) di $x_q(n)$ ed $h(n)$.

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

### 4.2.4 Riassumendo

la DFT (4.3) e la IDFT (4.5) costituiscono una coppia di relazioni invertibili che permettono di passare da una sequenza complessa di lunghezza N ad un'altra di pari lunghezza. Ma:

- calcolando la DFT su di una finestra di campioni $x_m$ di un segnale $x(t)$ limitato in una banda $W < 1/2T_c$, si ottengono delle stime $X_n$ dei campioni della sua trasformata di Fourier $X(f)$ per $f = \frac{n}{NT_c}$, ossia $X\left(\frac{n}{NT_c}\right) \simeq T_c \cdot X_n$ con $n = 0, 1, \ldots, N-1$
- calcolando la IDFT degli $X_n$ si ri-ottengono i campioni di $x(t)$ di partenza
- sia gli $X_n$ che gli $x_m$ sono in realtà sequenze periodiche di periodo $N$
- i calcoli indicati dalle (4.3) e (4.5) sono effettuati mediante un diverso algoritmo, chiamato *Fast Fourier Transform* o FFT, che ha il vantaggio di richiedere una complessità $O(N \log_2 N)$ ridotta rispetto a quella della DFT, che è $O\left(N^2\right)$
- la FFT deve essere calcolata su di un numero di campioni $N_{FFT}$ che sia una potenza di due, ossia deve essere $N_{FFT} = 2^M > N$. La finestra di analisi viene quindi estesa ponendo a zero gli $N_{FFT} - N$ valori finali

Se la sommatoria (4.3) venisse applicata, anzichè ad un numero finito $N$ di campioni $x_m$, ad un loro numero infinito, allora

- si otterrebbe una sequenza periodica $\tilde{X}(k)$ di periodo $N$, corrispondente al campionamento dello spettro periodico $X^{\bullet}(f)$

- l'applicazione della IDFT a $\tilde{X}(k)$ produrrebbe una sequenza periodica $\tilde{x}(m)$ di valori uguali agli $x_m$ solo se questi ultimi erano in numero limitato, inferiore ad $N$
- segmentando un segnale $x(t)$ in sotto-intervalli disgiunti, si può eseguire la convoluzione tra $x(t)$ ed una $h(t)$ di durata finita, operando esclusivamente nel dominio digitale, e sommando tra loro le IDFT dei prodotti tra la DFT dei campioni di $h(t)$ e le DFT dei segmenti di $x(t)$

L'interpretazione dei valori che risultano dalla applicazione della DFT su dei campioni di segnale, come stima della trasformata di Fourier del segnale, deve tenere conto oltre che delle fonti di approssimazioni evidenziate nella nota 5, anche dei corretti valori da assegnare alla scala delle frequenze e delle ampiezze, ossia:

#### 4.2.4.1 Le frequenze della DFT

Occorre tener presente il valore della frequenza di campionamento, e se il segnale di partenza $x(t)$ è reale, della periodicità degli $X_n$. Infatti i valori $X_n$ per $n = 0, 1, \ldots N-1$ corrispondono ai campioni di $X^{\bullet}(f)$ per $f = \frac{n}{NT_c}$, ma se $x(t)$ è reale, $X^{\bullet}(f)$ oltre ad essere periodico presenta simmetria coniugata, e dunque per valori $f > W = \frac{1}{2T_c}$, $X^{\bullet}(f)$ assume valori *speculari* a quelli risultanti per $f < W = \frac{1}{2T_c}$. Per fissare le idee, procediamo con un esempio: se $N = 512$ (come nel caso di una FFT), i primi 256 valori (da 0 a 255, ossia per $n = 0, 1, 2, \ldots, {}^{N}/_{2} - 1$) sono da mettere in corrispondenza con quelli di $X(f)$ con $f = 0, \frac{1}{NT_c}, \frac{2}{NT_c}, \ldots \frac{N/2-1}{NT_c}$; mentre i restanti 256 valori (da 256 a 511, ossia per $n = {}^{N}/_{2}, {}^{N}/_{2}+1, \ldots, N-1$, e corrispondenti a $f = \frac{1}{2T_c}, \frac{N/2+1}{NT_c}, \ldots, \frac{N-1}{NT_c}$) esibiscono un comportamento speculare a quello dei precedenti, essendo relativi a frequenze maggiori di quella *di Nyquist*.

#### 4.2.4.2 Le ampiezze della DFT

Come espresso dalla (4.6), i valori $X_n$ rappresentano una approssimazione dei coefficienti della serie di Fourier calcolati sulla finestra temporale da cui provengono i campioni di segnale, e moltiplicati per il numero di campioni utilizzati nel calcolo: $X_n \simeq N \cdot X_n^{SF}$. Pertanto, i valori ottenuti dalla DFT devono essere normalizzati, dividendoli per $N$.

# Capitolo 5

# Trasmissione dati

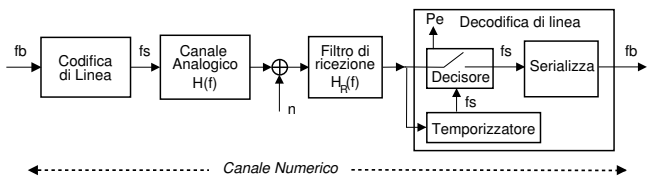
Dopo che i segnali sono stati convertiti in sequenze numeriche mediante campionamento (Cap. 4), e queste ultime convertite in cifre binarie mediante quantizzazione (vedi § 7.4), non esiste alcuna differenza formale tra segnali analogici-numerizzati, e dati nativamente numerici, come documenti di computer. Pertanto, in questo capitolo si affrontano gli argomenti legati alla trasmissione dei dati, comprendendo in questo termine entrambi i casi.

## 5.1 Trasmissione su canale numerico

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

### 5.1.1 Trasmissione numerica di banda base

La figura a lato rappresenta uno schema di canale numerico con evidenziati i principali elementi che lo compongono. Notiamo innanzitutto come *il ritmo* con cui le informazioni numeriche sono emesse dal *codificatore di linea* (§ 5.1.2) sia descritto da una *velocità di simbolo* $f_s$ potenzialmente inferiore al ritmo binario $f_b$ che è proprio del messaggio[1]. Come vedremo al § 5.1.2.4, il caso tipico




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

per cui deve essere $f_s < f_b$ si verifica quando il canale analogico presenta una risposta in frequenza $H(f)$ limitata in banda, al punto da dover ridurre la banda del segnale trasmesso, fino a farla *rientrare* in quella del canale, in modo da poter ri-ottenere in uscita dal canale lo stesso segnale trasmesso[2]. Il lato ricevente del canale numerico, costituito dal decodificatore di linea, *ricostruisce* la sequenza delle informazioni trasmesse campionando (con ritmo $f_s$) il segnale dati ricevuto, e confrontando i valori così ottenuti con delle soglie di decisione.

La presenza di un processo di *rumore* $n(t)$ in ingresso al decisore fa si che i valori ottenuti per campionamento possano occasionalmente superare le soglie di riferimento del decisore, determinando *un errore* (con probabilità $P_e$) nella decisione di quale valore sia stato trasmesso. La valutazione di tale probabilità viene svolta al § 7.5.2, ma anticipiamo subito che $P_e$ è tanto maggiore quanto più è grande la potenza del rumore in ingresso al decisore, che viene resa minima grazie al filtro $H_R(f)$ di ricezione, che ha appunto questo scopo (§ 12.1.1), o in alternativa quello di realizzare un *filtro adattato* (§ 9.4.4), od un dispositivo di *equalizzazione* (§ )14.5.6.

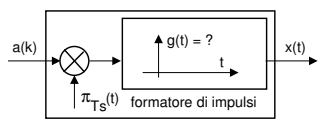
Lo schema di figura mostra inoltre come il decodificatore di linea debba svolgere anche una funzione di *temporizzazione*, per consentire al dispositivo decisore di operare al passo con il ritmo $f_s$ dei simboli in arrivo; al § 5.5 sono descritte alcune tecniche per affrontare questo aspetto. Infine, nel caso in cui ogni simbolo sia rappresentativo di più di un bit (vedi § 5.1.2.4), è evidenziata la presenza di un *serializzatore* che provvede ad emettere uno dopo l'altro i bit corrispondenti ai diversi simboli ricevuti.

### 5.1.2 Segnale dati e codifica di linea

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

$$x(t) = \sum_k a_k \cdot g(t - kT_s) \tag{5.1}$$

e idealizzare il codificatore stesso come illustrato nella figura a lato[3], in cui $\pi_{T_s}(t)$ è un treno di impulsi (§ 3.8.1) con periodo $T_s = 1/f_s$. Notiamo come tale figura sia del tutto simile al S&H introdotto al § 4.1.3, tranne che ora $g(t)$ è generico. Come sarà approfondito al § 9.2.4, si può mostrare che (sotto condizioni abbastanza generali) lo spettro di densità di potenza di $x(t)$ può essere espresso come



Costruzione del SEGNALE DATI

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

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

[2] Se non fosse preso questo provvedimento, e si trasmettesse un segnale con una occupazione spettrale maggiore della banda del canale, nel segnale di uscita 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 § 5.1.2.2).

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

**Onda pam** Il segnale dati (5.1) è a volte indicato come *onda* PAM, forzando la semantica di *Pulse Amplitude Modulation* per indicare la modulazione delle ampiezze degli impulsi che realizzano un segnale periodico nella forma di eq. (3.7) (pag. 45) in base ai valori $a_k$. Un ulteriore punto di vista è illustrato al § 6.9.5.

#### 5.1.2.1 Segnale binario e onda rettangolare

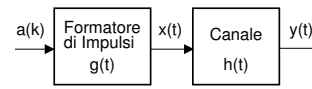
Come applicazione della (5.2) osserviamo che, nel caso in cui si effettui una trasmissione binaria (ossia $a_k = \{0, 1\}$ ed $f_s = f_b$) e si scelga un impulso dati rettangolare $g(t) = \text{rect}_\tau\left(t - \frac{\tau}{2}\right)$ (con $\tau \leq T_b$), allora la densità spettrale di $x(t)$ presenta[4] un andamento $\text{sinc}^2(f\tau)$, con il primo zero per $f = \frac{1}{\tau}$, ed occupazione di banda (approssimativamente) pari ad alcuni multipli di tale valore[5].

#### 5.1.2.2 Effetto della limitazione in banda e ISI

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

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

come mostrato alla nota[6]. L'effetto della convoluzione tra $g(t)$ ed $h(t)$, è quello di *disperdere* nel tempo la forma d'onda $g(t)$, che quindi anche se delimitata entro un periodo di simbolo, *invade* gli intervalli temporali riservati ai simboli adiacenti, dando luogo al fenomeno della *interferenza intersimbolica* (ISI, *InterSymbolic Interference*).




#### 5.1.2.3 Diagramma ad occhio

In Fig. 5.1 è riportato un esempio di segnale dati binario ad onda rettangolare con

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

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

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

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

[6] $y(t) = \left[\sum_k a_k \cdot g(t - kT_s)\right] * h(t) = \left[g(t) * \sum_k a_k \cdot \delta(t - kT_s)\right] * h(t) =$
$= g(t) * h(t) * \sum_k a_k \cdot \delta(t - kT_s) = g'(t) * \sum_k a_k \cdot \delta(t - kT_s) =$
$= \sum_k a_k \cdot g'(t - kT_s)$

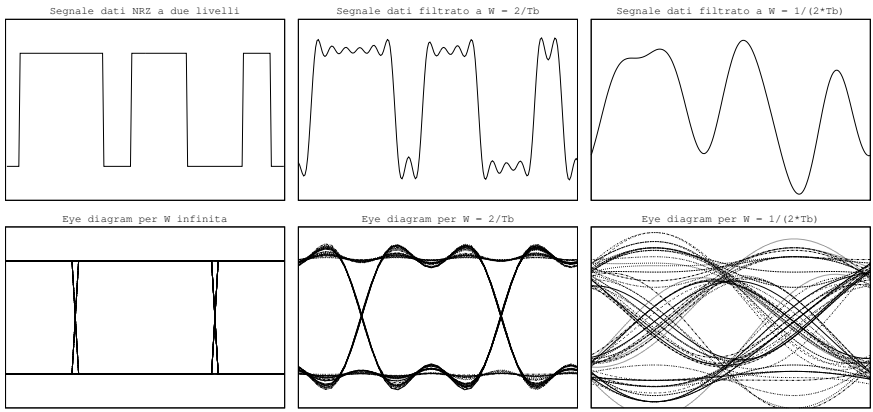



Figura 5.1: Confronto tra segnale dati a banda infinita e limitata, mediante diagramma ad occhio

base pari a $T_b$, a cui è applicata una limitazione di banda mediante un filtro passa-basso ideale. Nella riga superiore è mostrato l'andamento del segnale, mentre alla riga inferiore si riporta un grafico noto come *diagramma ad occhio* (EYE DIAGRAM)[7], che si ottiene visualizzando il segnale dati mediante un oscilloscopio con la base dei tempi sincronizzata al periodo di simbolo. Come si può vedere, in presenza di una limitazione di banda, il valore del segnale dati in corrispondenza di un determinato simbolo viene a dipendere anche dal valore dei simboli circostanti.

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

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

Nel seguito della sezione, assumiamo vere queste ipotesi.

#### 5.1.2.4 Trasmissione multilivello

Nel caso in cui la banda a disposizione per la trasmissione sia insufficiente, una soluzione di semplice attuazione è quella di ricorre ad una trasmissione non più *binaria*, ma che impieghi $L$ possibili diversi simboli, o *livelli*[8]. A tale scopo, occorre raggruppare $M$ bits del messaggio $a_n$ (che arrivano a velocità $f_b$ bits/sec) in una unica parola binaria.

Scegliendo $M = \log_2 L$, occorrono $T_s = MT_b$ secondi per accumulare $M$ bit, ed emettere uno tra $L = 2^M$ possibili valori $b_m$, usati quindi come ampiezze degli impulsi (generati a ritmo $T_s$) necessari a produrre un segnale dati[9] $x(t)$ ad $L$ livelli, e caratterizzato da una frequenza di simbolo $f_s = 1/T_s = 1/MT_b = f_b/M$.

[7] Il nome deriva dalla forma del diagramma, che in corrispondenza del *centro* degli impulsi mostra una apertura simile per l'appunto ad un occhio, e la cui analogia apparirà più evidente in seguito all'adozione di un impulso $g(t)$ a banda limitata (figura 5.4), ed in presenza di rumore (figura 7.5).

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

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

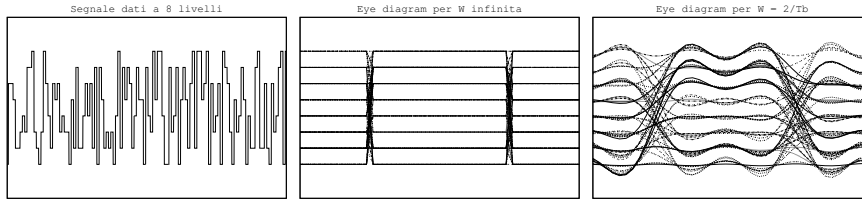
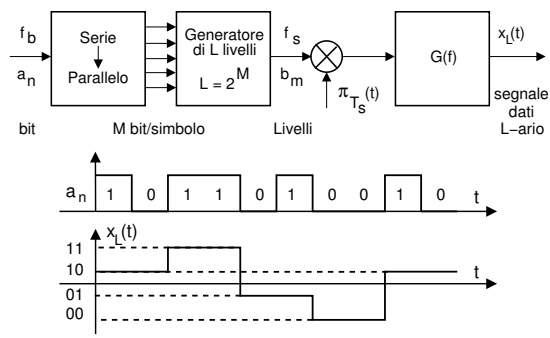



Figura 5.2: Segnale dati ad 8 livelli e diagramma ad occhio per banda infinita o limitata

La figura a lato mostra il segnale dati binario ad onda rettangolare, e quello che invece si ottiene con $L = 4$.




L'aumento del periodo di simbolo $T_s$ corrisponde ad un aumento della durata di $g(t)$, e di conseguenza ad una *contrazione* di $G(f)$, determinando quindi una eguale riduzione della banda occupata da $x(t)$, come si verifica tenendo conto della (5.2). Pertanto, l'occupazione di banda del segnale dati può essere ridotta a piacere, semplicemente aumentando il numero $M$ di bit raggruppati in una singola parola. D'altra parte al § 7.5 si mostra come, a meno che la potenza del segnale non venga aumentata, si assisterà ad un peggioramento della probabilità di errore del ricevitore, in quanto a parità di potenza e quindi di ampiezza, i livelli risultano ora ravvicinati. Questo fenomeno è raffigurato mediante l'esempio di figura 5.2 dove a sinistra è mostrato un segnale dati ad 8 livelli, al centro il diagramma ad occhio corrispondente, ed a destra una versione del segnale limitata in banda.

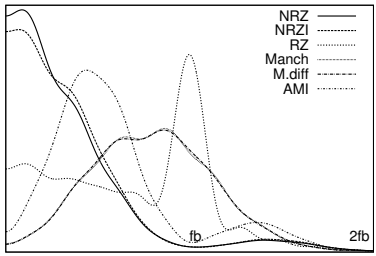
## 5.2 Generazione del segnale dati

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

### 5.2.1 Codici di linea a banda infinita

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

La figura seguente mostra lo spettro di densità di energia $\mathcal{E}_G(f)$ corrispondente alla scelta di alcune delle forme d'onda $g(t)$ discusse nel seguito, e mostrato per frequenze




fino al doppio di $f_b$. Gli andamenti riportati in figura sono ottenuti generando i valori (0 o 1) per 200 simboli binari $a_k$ in modo pseudo-casuale[10], campionando il segnale dati (5.1) con 16 campioni per periodo di bit, e calcolando una FFT smussata sui campioni così ottenuti. Ogni particolare $g(t)$ dà origine alla definizione di un *codice di linea* corrispondente, usato nella pratica per trasmettere informazioni di natura binaria. Elenchiamo di seguito caratteristiche e proprietà di tali segnali, aiutandoci con gli esempi riportati in figura.

**Codici unipolari** Sono associati a segnali sbilanciati, e presentano un valore nullo o diverso da zero per i due livelli logici 0 ed 1.

**NRZ - No Return to Zero** La sigla NRZ individua il fatto che il segnale "non torna a zero" per tutto il periodo di bit, essendo $g(t) = rect_{T_b}(t)$; pertanto ha spettro di tipo $\mathrm{sinc}(fT_b)$, con il primo zero a $f = 1/T_b$, e presenta una componente continua. Rimane costante per dati costanti e ciò complica la sincronizzazione del clock del ricevitore, data l'assenza in questo caso di transizioni; la presenza di uno zero ad $f = 1/T_b$ aggrava inoltre la situazione.

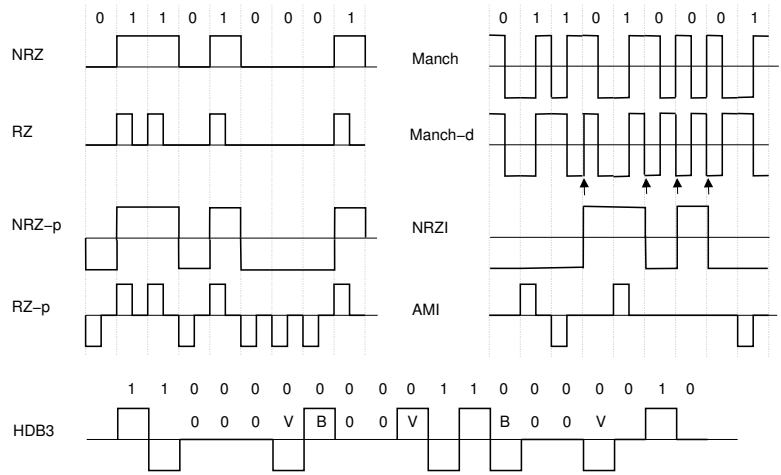
**RZ - Return to Zero** In questo caso l'impulso $g(t)$ ha durata pari a $T_b/2$, il segnale presenta (a parità di ampiezza) minore energia di NRZ, mentre lo spettro presenta una componente pronunciata esattamente a frequenza $f_b$, agevolando la sincronizzazione sul bit ma occupando una banda maggiore. Il segnale si mantiene però costante per lunghe sequenze di *zeri*.

**Codici bipolari** Usano segnali bilanciati, e sono ricevuti mediante uno stadio di ingresso differenziale, riducendo la sensibilità al rumore. In funzione del tipo di codice, è possibile garantire l'assenza di una componente continua nel segnale.

**NRZ polare, RZ polare** Realizzano l'impulso con polarità negativa quando associato ad un bit pari a zero, e presentano media nulla solo se i valori 0 ed 1 sono equiprobabili. RZ polare non è mai costante, facilitando il compito della sincronizzazione.

**Manchester o Diphase NRZ** Realizza una codifica *di fase*, in quanto usa un impulso RZ a piena dinamica, in salita od in discesa, corripondentemente ai bit 1 e 0. Per questo motivo il segnale risulta sempre a media nulla. L'occupazione spettrale è intermedia tra il caso NRZ ed RZ, dato che la durata media dell'impulso può essere $T_b$ o $T_b/2$. L'uso del codice Manchester è prescritto dallo standard IEEE 802.3 per le LAN a bus con contesa di accesso CDMA/CD (vedi § 8.6.1.6).

[10] 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.




**Alternate Mark Inversion (AMI)** Codifica gli 1 con polarità alternate, mediante un impulso $g(t)$ rettangolare di estensione $T_b$ o $T_b/2$, e gli zeri con assenza di segnale, garantendo assenza di valore medio. Qualora la logica sia invertita (impulsi per gli zeri, e silenzio per gli uni) il codice prende il nome di *pseudoternario*. Da un punto di vista spettrale, l'AMI esibisce una occupazione di banda[11] ridotta rispetto a RZ, per via dei periodi silenti corrispondenti agli zeri. Se il periodo silente è prolungato, l'assenza di transizioni può compromettere la sincronizzazione di bit, e per questo motivo sono stati definiti ulteriori codici derivati, come ad esempio

**HDB3** E' utilizzato per trasmettere il segnale PCM a 2 Mbps (vedi § 6.3.1), e l'acronimo significa *High-Density Bipolar-3-zeroes*. Come per AMI, rappresenta gli uni con polarità alternate, ma rimpiazza le sequenze di quattro zeri consecutivi forzando una *violazione* della regola sull'ultimo bit dei quattro, in modo che il ricevitore, rilevando la violazione, è in grado di riportare il bit a zero. Dato però che la presenza della violazione creerebbe la comparsa di una componente continua nel segnale, sono inseriti anche dei bit di *bilanciamento,* per rimuovere quest'ultima. I bit di bilanciamento si collocano al posto del primo dei quattro zeri, e la loro polarità è scelta in modo che la sequenza delle violazioni abbia un polarità alternata; in definitiva, dopo la prima violazione, si usa sempre anche il bit di bilanciamento.

**Codici differenziali** Sono ancora di tipo bipolare, e la forma d'onda non è più legata al solo valore dei bit, ma anche alla loro relazione temporale (vedi anche § 13.6.1). Ciò permette di risolvere l'ambiguità presente qualora si scambino le polarità degli estremi del collegamento.

**NRZI** Deriva dall'NRZ, e la **I** sta per *Inverted*. Ora il livello del segnale permane nello stesso stato per i bit pari ad uno, e cambia stato per i bit pari a zero. L'assenza di valor medio è legata alla statistica che descrive le sequenze di uni e dunque non

[11] La densità spettrale mostrata in figura è relativa all'uso di una $g(t)$ di tipo RZ.

può essere garantita, mentre permangono i problemi legati alla sincronizzazione. La ridotta occupazione spettrale lo rende però interessante.

**Manchester Differenziale** Usa un impulso RZ a piena dinamica come per Manchester, ma la polarità risulta invertita rispetto al bit precedente se il nuovo bit è uno, mentre è mantenuta nel caso arrivi un zero (in corrispondenza delle frecce); pertanto, in presenza degli uni non si verifica transizione al confine tra i periodi di bit. Questa soluzione è utilizzata nel contesto dello standard IEEE 802.5 per LAN *Token Ring*. L'occupazione spettrale è simile a quella osservabile per la codifica Manchester.

### 5.2.2 Segnale dati limitato in banda

Analizziamo i requisiti necessari a realizzare la trasmissione di un segnale dati che impegni una banda esattamente delimitata, evidenziando le condizioni da soddisfare, e descrivendo le conseguenze di alternative di progetto.

#### 5.2.2.1 Requisiti per l'impulso di trasmissione

**Limitazione di banda** Abbiamo già osservato (vedi § 5.1.2) che (sotto opportune condizioni) la densità di potenza del segnale dati (vedi 5.1)) risulta pari a $\mathcal{P}_x(f) = \sigma_A^2 \frac{|G(f)|^2}{T}$, e che la sua trasmissione inalterata è possibile solo disponendo di un canale che presenti una risposta in frequenza di tipo *passa-tutto* (ossia con modulo costante e fase lineare) nella banda di frequenze occupata da $G(f)$.

Nel caso in cui, ad es., si adotti $g(t) = \text{rect}_\tau(t)$, allora $G(f) = \tau \text{sinc}(f\tau)$, con $\tau \leq T_s$; pertanto, il primo passaggio per zero è a frequenza $\frac{1}{\tau} \geq \frac{1}{T_s}$, e la densità di potenza $\mathcal{P}_x(f)$ può essere trascurata solo dopo qualche multiplo di tale valore.

Se invece il canale sopprime contenuti frequenziali non trascurabili dal segnale dati (5.1), allora gli impulsi $g(t)$ *si deformano* (vedi § 5.1.2.2) e non sono più rettangolari; in particolare, possono estendersi per una durata maggiore di $T_s$, causando problemi di *interferenza tra simboli*[12] (ISI).

**Limitazione nel tempo** Il problema della limitazione di banda potrebbe essere risolto adottando, in linea di principio, un impulso elementare di tipo $g(t) = \text{sinc}\left(\frac{t}{T_s}\right)$, che ha trasformata $G(f) = T_s \cdot \text{rect}_{\frac{1}{T_s}}(f)$ strettamente limitata in banda, con frequenza massima $W = \frac{1}{2T_s}$, e che non subisce alterazioni purché il canale abbia un comportamento "passa tutto" in tale ristretto intervallo di frequenze. Notiamo che questa $g(t)$ passa da zero per $t = nT_s$, e pertanto non provoca interferenza tra i simboli collocati agli istanti $nT_s$, come verificabile notando che in tal caso l'espressione (5.1) risulta del tutto simile alla (4.1) relativa alla ricostruzione di una segnale campionato; ora però *non siamo interessati* ai valori del segnale negli istanti intermedi a quelli in cui sono centrati i simboli, mentre desideriamo unicamente recuperare i singoli valori originali, che troviamo *in modo esatto* agli istanti $t = nT_s$.

[12] Come mostrato al § 5.1.2.2, il segnale dati filtrato è costruito con impulsi $g'(t) = g(t) * h(t)$, con una durata pari alla somma delle durate di $g(t)$ e $h(t)$. Pertanto, l'impulso $a_n g(t - nT_s)$ si estende a valori di $t > (n+1)T_s$, e quindi $x((n+1)T_s) = a_{n+1} g(0) + a_n g(T_s)$, introducendo un errore pari a $a_n g(T_s)$, detto appunto *interferenza tra simboli*.

Lo svantaggio di adottare una forma d'onda $g(t)$ limitata in frequenza, è che il suo andamento è illimitato nel tempo, e dunque $g(t)$ può essere realizzata solo in modo approssimato[13].

**Limitazione di precisione** Abbiamo appena mostrato come, adottando una $g(t) = \text{sinc}\left(\frac{t}{T_s}\right)$, si evita l'interferenza tra simboli, purché i campioni vengano prelevati *esattamente* agli istanti $nT_s$ ([14]). Infatti, al di fuori di tali istanti il valore del segnale dipende dal valore delle code degli impulsi $g(t)$ centrati sugli altri simboli.

L'orologio (*clock*) del ricevitore, però, non ha una precisione infinita, e gli istanti di campionamento saranno affetti da errori di fase. Pertanto, è interessante ricercare una soluzione per $g(t)$ che, anche in presenza di errori di precisione nella determinazione degli istanti di campionamento, dia luogo ad errori quanto più ridotti possibile.

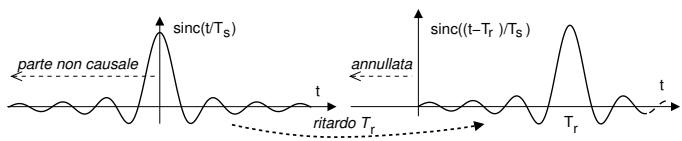
**Riepilogando:** vorremmo soddisfare contemporaneamente le esigenze:

1. Occupare una banda contenuta
2. Ricorrere ad un filtro poco complesso
3. Ridurre la sensibilità agli errori di campionamento

Per i punti 2 e 3, è sufficente adottare $g(t)$ di tipo rettangolare, generando un segnale dati del tipo $x(t) = \sum_k a_k \cdot \text{rect}_\tau(t - kT_s)$, che ha lo svantaggio di occupare un banda infinita, e quindi la sua ricezione *inalterata* è possibile solo per canali ideali.

Prima di esporre una soluzione di compromesso a tutti e 3 i problemi, consideriamo che la trasmissione di $x(t)$ su di un canale non ideale $H(f)$ determina un effetto di cui al § 5.1.2.2 abbiamo tenuto conto considerando di ricevere un segnale dati caratterizzato da un impulso $g'(t) = g(t) * h(t)$. Pertanto, nel caso in cui $H(f)$ sia noto a priori, i risultati che troveremo (validi per $g'(t)$) individueranno in realtà un formatore di impulsi con $G(f) = \frac{G'(f)}{H(f)}$.

[13] Un sistema fisico che debba realizzare una risposta impulsiva $g(t) = \text{sinc}\left(\frac{t}{T_s}\right)$, non può presentare $g(t) \neq 0$ per $t < 0$: questo equivarrebbe infatti ad un sistema in grado di produrre una uscita *prima ancora* che sia applicato un segnale al suo ingresso. Se $g(t)$ ha estensione temporale illimitata, occorre ricorrere ad una versione ritardata e limitata $g'(t) = \begin{cases} g(t - T_R) & con\ t \geq 0 \\ 0 & altrimenti \end{cases}$. Se $T_R \gg T_s$, l'entità dell'approssimazione è accettabile, ed equivale ad un semplice ritardo pari a $T_R$; d'altro canto, quanto maggiore è la durata della risposta impulsiva, tanto più difficile (ossia costosa) risulta la realizzazione del filtro relativo.




[14] 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.

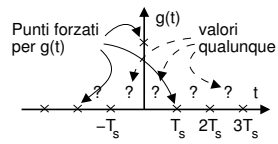
#### 5.2.2.2 Condizioni di Nyquist

Torniamo a riferirci alla (5.1) per osservare che, affinché $x(nT_s)$ dipenda da uno solo degli $\{a_k\}$, deve risultare

$$g[(n-k)T_s] = \begin{cases} 1 & se\ n = k \\ 0 & se\ n \neq k \end{cases} \tag{5.3}$$

e cioè $g(t)$ deve passare da zero in tutti gli istanti multipli di $T_s$, tranne che per $t = 0$ dove deve valere 1; infatti, in tal caso dalla (5.1) si ottiene:

$$x(nT_s) = \sum_k a_k \cdot g((n-k)T_s) = a_n$$




Le condizioni (5.3) prendono il nome di *condizioni di Nyquist per l'assenza di interferenza intersimbolo* (ISI, *Inter Symbol Interference*) nel dominio del *tempo*. Se una forma d'onda $g(t)$ soddisfa tali condizioni, allora viene detta *impulso di Nyquist*([15]).

Dalle condizioni di Nyquist *nel tempo* se ne derivano altre *in frequenza*, mediante i seguenti passaggi. Moltiplicando $g(t)$ per un treno di impulsi $\pi_{T_s}(t) = \sum_k \delta(t - kT_s)$, si ottiene

$$g(t) \cdot \pi_{T_s}(t) = \delta(t)$$

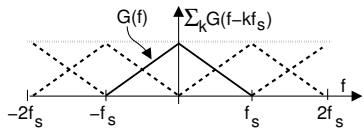
dato che $g(nT_s) = 0$ e $g(0) = 1$. Trasformando (vedi § 3.8.3) si ottiene:

$$1 = G(f) * \frac{1}{T_s} \cdot \Pi_{\frac{1}{T_s}}(f) = G(f) * \frac{1}{T_s} \cdot \sum_k \delta\left(f - k\frac{1}{T_s}\right)$$

Indicando con $f_s = \frac{1}{T_s}$ la frequenza di simbolo, ed eseguendo le convoluzioni, risulta infine

$$\sum_k G(f - kf_s) = T_s \tag{5.4}$$

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, sommata con le sue repliche traslate di multipli di $f_s$, dia luogo ad una costante.




In questo caso si dice che $G(f)$ è una *caratteristica di Nyquist*. Notiamo che, seppure $G(f)$ possa essere qualsiasi, anche non limitata in banda, il nostro interesse è appunto per le $G(f)$ limitate in banda, come quella triangolare dell'esempio a lato.

#### 5.2.2.3 Caratteristica a coseno rialzato

Una famiglia parametrica di caratteristiche di Nyquist limitate in banda, è quella cosiddetta a *coseno rialzato,* che è composta da 2 archi di coseno raccordati da una retta (vedi Fig. 5.3). La banda occupata ha espressione

$$B = \frac{f_s}{2}(1 + \gamma) \tag{5.5}$$

[15] Ad esempio, l'impulso rettangolare è di Nyquist, in quanto $\text{rect}_{T_s}(t) = \begin{cases} 1 & se\ t < \frac{T_s}{2} \\ 0 & se\ t = kT_s \end{cases}$ .

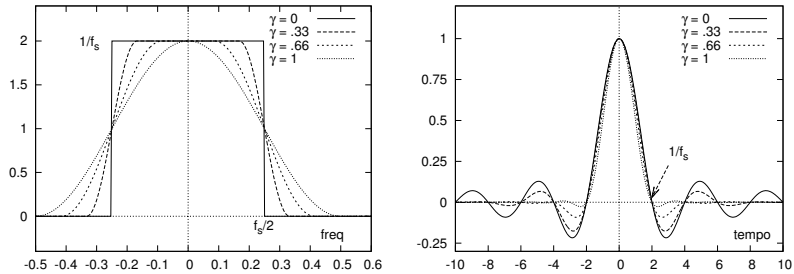



Figura 5.3: Caratteristica a coseno rialzato e impulso di Nyquist per $f_s = 0.5$, al variare di $\gamma$

in cui $\gamma$ è chiamato *coefficiente di roll-off*[16], è compreso tra 0 e 1, e rappresenta un indice di dispersione del ramo di coseno. La banda di $G(f)$ varia quindi da un minimo (per $\gamma = 0$) pari a $B = f_s/2$, con $G(f)$ rettangolare, ad un massimo (per $\gamma = 1$) pari a $B = f_s$, con $G(f)$ che è proprio un periodo di coseno.

Il caso di *banda minima* si ottiene per $\gamma = 0$, che corrisponde a $G(f) = \frac{1}{f_s} rect_{f_s}(f)$, ovvero ad una $g(t) = sinc(f_s t)$, come già discusso a pag. 70 al § sulla limitazione nel tempo. Occupare una banda inferiore a quella minima non è possibile, perchè in tal caso non sarebbero verificate le condizioni di Nyquist in frequenza, in quanto nella $\sum_k G(f - kf_s)$ resterebbero dei "buchi".

Abbiamo già osservato alla nota (13) a pagina 71 come la realizzazione di $G(f)$ *a banda minima* sia complessa, e produca una eccessiva sensibilità agli errori di campionamento. La situazione però migliora decisamente usando $\gamma > 0$, con $\gamma$ via via più grande. Se $\gamma \neq 0$, per $g(t)$ si può ottenere l'espressione

$$g(t) = sinc(tf_s) \cdot \frac{\cos \gamma \pi t f_s}{1 - (2\gamma t f_s)^2} \tag{5.6}$$

a cui corrisponde una forma d'onda simile al $\frac{\sin(x)}{x}$, ma che va a zero molto più rapidamente, come verificabile osservando la parte destra di Fig. 5.3. Pertanto, se $\gamma \to 1$ ogni singola onda $g(t - kT_s)$ estenderà la sua *influenza* ad un numero di impulsi limitrofi molto ridotto rispetto al caso $\gamma = 0$, in quanto le oscillazioni sono molto più smorzate, e dunque il termine di errore di ampiezza in presenza di un errore di istante di campionamento è ridotto, dato che dipende da un minor numero di impulsi limitrofi.

La fig 5.4 mostra l'andamento del segnale dati con $g(t)$ fornito dalla (5.6), calcolata per $\gamma = 0.5$, e per $a_k$ a due valori, pari a 0 e 1. Notiamo che, al difuori degli istanti caratteristici $t = kT_s$ il segnale può assumere valori arbitrari, anche esterni alla dinamica degli $a_k$.

La rappresentazione fornita dal diagramma ad occhio, sempre mostrata in fig 5.4, permette di valutare meglio la precisione di temporizzazione che è necessaria ad evitare ISI, e che è pari a metà della *apertura orizzontale* dell'occhio. Gli ultimi due diagrammi permettono il confronto con i casi ripettivamente a banda minima e massima (per $\gamma = 1$), evidenziando l'influenza del roll-off.

[16] Il termine ROLL-OFF può essere tradotto come "*rotola fuori* ".

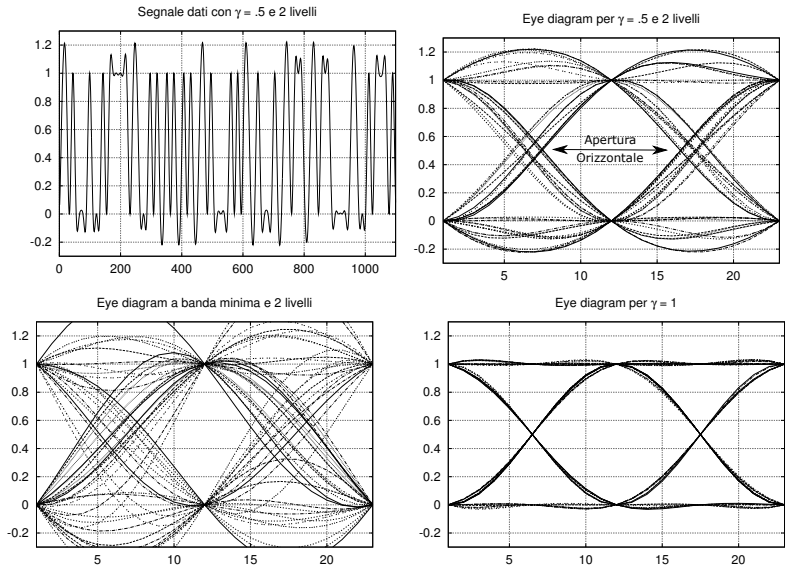
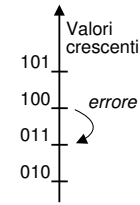



Figura 5.4: Segnale dati ed *eye diagram* per diversi valori di *roll-off*

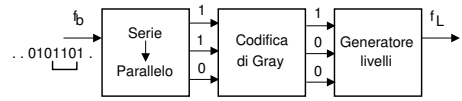
#### 5.2.2.4 Codice di Gray

Si tratta di un accorgimento che consente di ridurre l'effetto degli errori commessi dal decisore (vedi §7.5) in presenza di trasmissione multilivello.

Per fissare le idee, supponiamo che il valore *analogico* dei livelli presenti nel segnale dati sia in corrispondenza lineare con il valore *numerico* dell'uscita del convertitore serie-parallelo (vedi fig. a pagina 67), ovvero che i livelli siano "numerati in binario", e livelli contigui rappresentino configurazioni di bit "in sequenza naturale", come mostrato in figura. Allora, se trasmettiamo ad esempio il livello associato a 100, ed il decisore commette l'errore[17] di ritenere di aver ricevuto il livello contiguo, che rappresenta la sequenza 011, abbiamo tutti e tre i bit sbagliati!




Il *codice di Gray* consiste quindi in una tabella di conversione che sostituisce ai bit uscenti dal convertitore serie-parallelo, una diversa configurazione di bit. Possiamo immaginare l'operazione come quella di un accesso a memoria, in cui la parola originaria costituisce l'indirizzo, per mezzo del quale si individua la parola codificata da trasmettere al suo posto. La conversione è biunivoca (a partire dal codice si risale alla parola originaria), e le parole del codice di Gray hanno la proprietà di rappresentare i livelli di segnale contigui come configurazioni di bit che differiscono solo in una cifra binaria (ossia in un bit).




[17] L'errore è causato dal rumore additivo che, sommandosi al segnale ricevuto (vedi Fig. a pagina 63), ne modifica il valore, fornendo al decisore valori diversi da quelli trasmessi.

Con riferimento alla tabella, osserviamo che (ad esempio) per trasmettere la sequenza 110 di ingresso, si usa il livello numero 100, ossia il quarto (partendo da zero), lo stesso dell'esempio precedente; se il decisore sbaglia e ritiene di aver ricevuto il terzo livello (011, stesso errore precedente), a questo il decodificatore di Gray associa la sequenza 010, che infatti differisce dall'originale per un solo bit (il primo).

| Ingresso binario | Uscita (livello) |
|---|---|
| 100 | 111 |
| 101 | 110 |
| 111 | 101 |
| 110 | 100 |
| 010 | 011 |
| 011 | 010 |
| 001 | 001 |
| 000 | 000 |

Il procedimento illustrato, in presenza di un errore sul simbolo, produce un solo bit errato. Ciò comporta che la probabilità di osservare un bit errato è pari a $P_e^b = P_e^s/M$, con $M$ pari al numero di bit/simbolo. Infatti

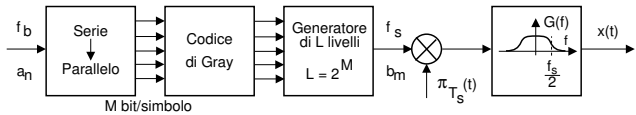
$$P_e^b = \frac{\text{N.Bit errati}}{\text{N.Bit totali}} = \frac{\text{N.Simboli errati}}{\text{M·N.Simboli}} = P_e^s \frac{1}{M}$$

**Riassumendo** La figura che segue mostra la sequenza generale delle operazioni da intraprendere per generare un segnale dati mutilivello, con codifica di Gray, e con caratteristica a coseno rialzato. Notiamo dunque che mentre alla sequenza binaria $a_n$ compete una velocità di $f_b$ bit/secondo, la sequenza multilivello $b_m$ possiede invece un ritmo di

$$f_s = \frac{f_b}{M} = \frac{f_b}{\log_2 L}$$

simboli/secondo, ed il segnale dati risultante $x(t)$ occupa una banda a frequenze positive (vedi eq. 5.5)

$$B = \frac{f_s(1+\gamma)}{2} = \frac{f_b(1+\gamma)}{2\log_2 L}$$




### 5.2.3 Equalizzazione numerica

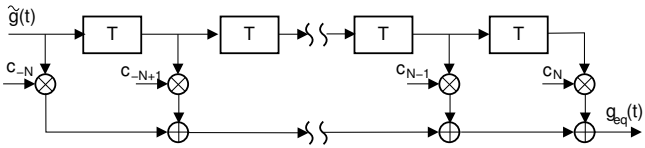
Quando il canale attraversato da un segnale dati presenta una $H(f)$ sconosciuta e non ideale a priori (vedi pag. 333), la *distorsione lineare* da essa introdotta causa la presenza di ISI nel segnali dati in uscita, che appare quindi costituito da impulsi distorti $\tilde{g}(t) = g(t) * h(t)$ anziché $g(t)$. Tale fenomeno può essere *compensato* facendo transitare il segnale ricevuto attraverso un *filtro di equalizzazione* $H_{eq}(f)$ tale che

$$H(f)\,H_{eq}(f) = a\mathrm{e}^{-j2\pi f\tau} \qquad \text{ovvero} \qquad H_{eq}(f) = \frac{a\mathrm{e}^{-j2\pi f\tau}}{H(f)}$$

Se il lato ricevente conosce $H(f)$, può calcolare la $H_{eq}(f)$ teorica, e realizzare un filtro trasversale (vedi § 9.7) che la approssimi, calcolandone i coefficienti mediante uno sviluppo in serie. Il troncamento di tale sviluppo, necessario ad ottenere un $h_{eq}(t)$ di durata finita, comporta un errore residuo, che viene accettato purché piccolo. Ma spesso ciò che è direttamente disponibile al ricevitore *non è* l'$H(f)$ del canale, ma solamente *ciò che ne esce*; inoltre, nel caso delle trasmissioni numeriche siamo interessati ai soli valori di segnale *negli istanti* multipli del periodo di simbolo. Il metodo di equalizzazione presentato appresso, tiene conto di queste due particolarità.

#### 5.2.3.1 Zero forcing equalization

Questa tecnica prevede che gli impulsi ricevuti distorti $\tilde{g}(t)$ attraversino un filtro trasversale, i cui coefficienti sono calcolati in modo da ripristinare in forma approssimata le condizioni di Nyquist (nel dominio del tempo) per l'impulso $g_{eq}(t)$ in uscita dal filtro. Mostriamo come.




La fig. a lato mostra un filtro trasversale con $2N+1$ coefficienti ed un ritardo totale $2NT$. Il messaggio informativo è preceduto dalla trasmissione di un impulso isolato di apprendimento, la cui forma d'onda distorta $\tilde{g}(t)$ ricevuta presenta un picco a $t=0$ ed ISI su entrambi i lati (Fig. 5.5a). Ponendo $\tilde{g}(t)$ in ingresso al filtro $h_{eq}(t)$, all'uscita si ottiene l'impulso

$$g_{eq}(t) = \sum_{n=-N}^{N} c_n \tilde{g}(t - nT - NT)$$

che campionato agli istanti $t_k = kT + NT$ fornisce

$$g_{eq}(t_k) = \sum_{n=-N}^{N} c_n \tilde{g}(kT - nT) = \sum_{n=-N}^{N} c_n \tilde{g}_{k-n} \tag{5.7}$$

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

$$g_{eq}(t_k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, \cdots, \pm N \end{cases}$$

assicurando così l'assenza di ISI *almeno* nei confronti degli $N$ simboli precedenti e successivi. Il valore dei coefficienti $c_n$ in grado di soddisfare queste condizioni si ottengono risolvendo il sistema di $2N+1$ equazioni nelle $2N+1$ incognite, impostato a partire dall'espressione (5.7), ed rappresentato in forma matriciale come

$$\begin{bmatrix} \tilde{g}_0 & \cdots & \tilde{g}_{-2N} \\ \vdots & & \vdots \\ \tilde{g}_{N-1} & \cdots & \tilde{g}_{-N-1} \\ \tilde{g}_N & \cdots & \tilde{g}_{-N} \\ \tilde{g}_{N+1} & \cdots & \tilde{g}_{-N+1} \\ \vdots & & \vdots \\ \tilde{g}_{2N} & \cdots & \tilde{g}_0 \end{bmatrix} \begin{bmatrix} c_{-N} \\ \vdots \\ c_{-1} \\ c_0 \\ c_1 \\ \vdots \\ c_N \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.8}$$

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

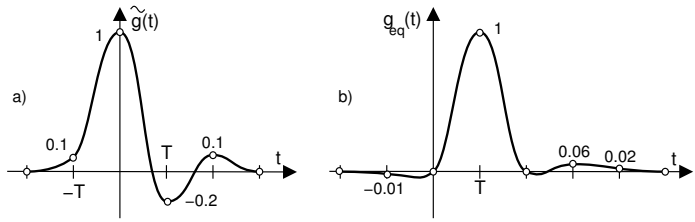



Figura 5.5: Impulso con ISI prima e dopo equalizzazione

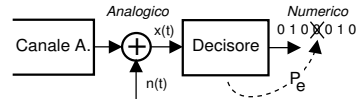
di più campioni ricevuti, eseguita dal filtro di equalizzazione, può avere l'effetto di aumentare la potenza di rumore, ma spesso questo peggioramento di qualità è più che compensato dal miglioramento dell'ISI.

**Esempio** Vogliamo realizzare un equalizzatore del terzo ordine, applicando il principio illustrato all'impulso con ISI mostrato alla fig. 5.5a. Inserendo quei valori di $\tilde{g}_k$ nella matrice dei coefficienti (5.8), otteniamo

$$\begin{bmatrix} 1.0 & 0.1 & 0.0 \\ -0.2 & 1.0 & 0.1 \\ 0.1 & -0.2 & 1.0 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

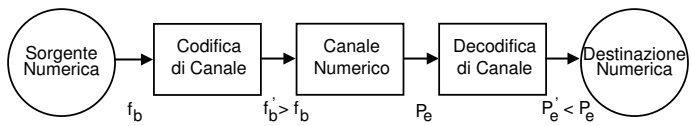
da cui è possibile calcolare $c_{-1} = -0.096 \quad c_0 = 0.96 \quad c_1 = 0.2$. Inserendo ora ad uno ad uno i campioni di $\tilde{g}(t)$ in un filtro trasversale con questi coefficienti, si ottengono i valori di $g_{eq}(t)$ mostrati in fig. 5.5b assieme ad una curva interpolata. Notiamo che sebbene sia stato ottenuto il risultato desiderato di azzerare $g_{eq}(t)$ ai due istanti di simbolo ai lati del picco, risulta ancora essere $g_{eq} \neq 0$ ad istanti più remoti.

## 5.3 Errori di trasmissione



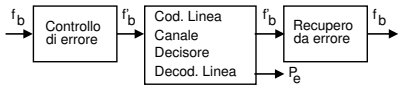

Riportiamo a lato la figura già proposta al primo capitolo (pag. 3) che mostra l'uso di un canale analogico per la trasmissione del segnale dati, e come la presenza di rumore additivo in ingresso al decisore di ricezione provochi *errori* nel flusso di bit ricostruito.

Il calcolo di questa $P_e$ è svolto al § 7.5, mentre alle pagine precedenti abbiamo fatto notare l'esistenza di una ulteriore possibile fonte di errore, quella dovuta all'ISI, legata alla precisione della temporizzazione del decisore, ed alla $g(t)$ complessiva (vedi § 5.1.2.2) per il segnale ricevuto. Sempre al primo capitolo è proposta la figura replicata appresso, che evidenzia come la $P_e$ possa essere ridotta adottando tecniche di *codifica di canale* e *controllo di errore* che sono discusse di seguito.




### 5.3.1 Controllo di errore

Con questo termine si individuano le strategie atte a *proteggere* le informazioni da trasmettere, aumentando il numero effettivo dei bit inviati (che passano da $f_b$ a $f'_b > f_b$




bit/secondo, vedi figura), in modo che i bit aggiunti siano dipendenti dagli altri, permettendo la gestione degli eventuali errori di trasmissione.

**FEC e ARQ** La quantità di bit aggiunti può permettere di *accorgersi* della presenza di errori di trasmissione, oppure di *correggere* detti errori. Se il sistema di trasmissione viene considerato unidirezionale[18], lo scopo del controllo di errore è quello di permettere al ricevitore di correggere gli errori che si sono verificati; questo approccio è anche indicato come *Forward Error Correction* (FEC, o correzione di errore *in avanti*), e si è sviluppato nel contesto funzionale della *codifica di canale* (vedi § 5.3.3).

Se al contrario è presente un canale di comunicazione a ritroso, e non sussistono rigidi vincoli temporali sul massimo ritardo tra trasmissione e ricezione corretta, l'effetto di un errore *rivelato* ma non correggibile immediatamente può essere annullato chiedendo la ritrasmissione del dato errato, dando luogo alle strategie di *Automatic Repeat reQuest*[19] o ARQ (ossia richiesta automatica di ripetizione) (vedi § 5.4), che danno origine alla definizione dei *protocolli a finestra*, evoluti storicamente nel contesto della trasmissione dati.

**Compromesso velocità-distorsione** A pag. 145 sono svolte delle considerazioni su come l'adozione di tecniche di controllo di errore consenta di *contrattare* una migliore qualità del segnale ricevuto, a prezzo di una maggiore velocità di trasmissione (e quindi banda occupata) del segnale, mentre al § 17.2 è svolta la trattazione analitica relativa a questo fenomeno.

#### 5.3.1.1 Errori su parole

I metodi di controllo di errore operano tipicamente su gruppi di $n$ bit, ognuno dei quali può arrivare sbagliato con probabilità $P_e = p$; se gli eventi di errore sono statisticamente indipendenti (§ 7.1.5), la probabilità di *un* bit errato *su* $n$ è pari a $np$, cioè $n$ volte la probabilità di un bit errato[20].

D'altro canto, la probabilità di $i$ bit errati su $n$ è fornita dalla distribuzione Binomiale (vedi § 8.1) $P(i,n) = \binom{n}{i} p^i (1-p)^{n-i}$ che, se $p \ll 1$, può essere appossimata come

$$P(i,n) \approx \frac{n(n-1)\cdots(n-i+1)}{i!} p^i \tag{5.9}$$

[18] O perchè lo è effettivamente, oppure perchè la trasmissione coinvolge informazioni generate in tempo reale, non memorizzate in trasmissione, e *consumate* immediatamente in ricezione (ad es. un segnale numerico ottenuto mediante campionamento e quantizzazione), per le quali non ha senso richiedere la trasmissione.

[19] L'aggettivo *automatic* si riferisce al fatto che spesso la gestione della ritrasmissione avviene a carico di uno strato protocollare di livello *inferiore* a quello che effettivamente consuma il messaggio, che in definitiva neanche si avvede della presenza del meccanismo di ritrasmissione.

[20] La probabilità che solo il primo bit su $n$ sia sbagliato è pari a $p_1 = p \cdot (1-p)^{n-1} \simeq p$ se $p \ll 1$ ed $n$ non è troppo grande; lo stesso risultato si ottiene anche per gli altri $n-1$ casi possibili di un bit sbagliato e gli altri corretti, cosicché la probabilità di un solo *generico* bit sbagliato assomma quelle di tutti i casi.

che è molto inferiore[21] alla probabilità $np$ di *un singolo* errore su $n$. Infine, sempre per $p \ll 1$, risulta che $P(i+1,n) \ll P(i,n)$, e quindi si può considerare la probabilità di ricevere $i$ *o più* bit errati su $n$, praticamente uguale a quella di osservare solo $i$ errori.

All'aumentare di $p$ e di $n$, le approssimazioni non sono più valide; in tal caso però, il sistema di trasmissione è praticamente inusabile, e dunque ci si trova sempre nelle condizioni di $p$ sufficentemente piccolo da permettere le approssimazioni.

L'esposizione prosegue illustrando al § 5.3.2 tre tecniche utilizzate comunemente per la detezione di errore, e formalizzando quindi (al § 5.3.3) il problema della codifica di canale, mediante la descrizione di alcune soluzioni che la realizzano. Infine, al § 5.4 sono descritti i protocolli ARQ adottati nel contesto della trasmissione dati.

### 5.3.2 Detezione di errore

In questo caso ci si limita a mettere il ricevitore in grado di accorgersi della presenza di errori, senza poterli correggere. Le parole errate possono essere scartate, oppure reinviate su richiesta.

#### 5.3.2.1 Parità

Viene comunemente usata nell'ambito della trasmissione asincrona e sincrona orientata al carattere, per rivelare errori sul bit, e consiste nell'aggiungere alla parola da trasmettere un ulteriore bit, in modo che in totale ci sia un numero *pari* di uni[22], applicando quindi una regola di parità *pari* (EVEN). Il caso opposto, ossia l'aggiunta di un bit in modo da rendere *dispari* il numero di uni, prende nome di parità ODD.

In entrambi i casi[23], quando il ricevitore raggruppa i bit pervenuti, esegue un *controllo* detto appunto *di parità*, semplicemente contando il numero di uni, ed accorgendosi così se nella parola si sia verificato un errore (uno zero divenuto uno o viceversa). In tal caso, il ricevitore invierà all'altro estremo del collegamento una richiesta di ritrasmissione del gruppo di bit errati. Se invece si fosse verificato un errore che coinvolge due bit della parola, questo passerebbe inosservato, in quanto la parità prescritta verrebbe mantenuta. Infatti, la *distanza di Hamming* (vedi pag. 83) tra codeword ottenute aggiungendo il bit di parità è pari a due[24].

**Esempio:** indicando la probabilità di errore sul bit con $p$ (es $10^{-3}$) ed applicando la (5.9) si ottiene che la probabilità di $i = 2$ errori in $n$ bit (es $n = 10$) vale $\frac{1}{2} n (n-1) p^2$ (es $4.5 \cdot 10^{-5}$), e rappresenta il *tasso residuo* di errore su parola legato all'uso di un bit di parità; la trasmissione di $n-1$ bit non protetti sarebbe invece stata esposta ad una probabilità $np$ (es $0.9 \cdot 10^{-2}$) di un bit errato su $n$.

[21] Ad esempio, la probabilità di due bit errati può essere approssimata come $\frac{1}{2} n (n-1) p^2$. Infatti, la distribuzione Binomiale fornisce $P(2,n) = \binom{n}{2} p^2 (1-p)^{n-2}$, in cui $\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}$ e, se $p \ll 1$ ed $n$ non è troppo elevato, $(1-p)^{n-2} \simeq 1$. All'aumentare di $p$ e di $n$, l'approssimazione non è più valida, e la probabilità di più di un bit errato può risultare maggiore di quella di un solo bit errato.

[22] 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.

[23] Il ricevitore deve comunque essere al corrente del fatto se la parità sia ODD o EVEN !

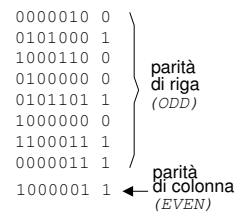
[24] Considerando parole di 3 bit, le codeword (di 4 bit, in cui l'ultimo è una parità pari) risultano: (0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111). E' facile constatare che ognuna di esse differisce da tutte le altre per due bit.

Il calcolo del bit di parità *pari* può essere effettuato svolgendo la *somma modulo due*[25] di tutti i bit che compongono la parola (ovvero complementando il risultato, nel caso di parità *dispari*).

#### 5.3.2.2 Somma di controllo

Quando il messaggio è composto da $M$ diverse parole di $N$ bit, la probabilità che almeno una di queste sia errata aumenta in modo circa proporzionale ad $M$, in base ad un ragionamento del tutto analogo a quello della nota 20.

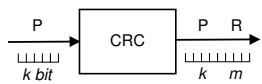
Per aumentare le capacità di rivelazione del controllo a parità applicato sulle singole parole (indicato ora come parità *di riga*, o *trasversale*), si aggiunge al gruppo di $M$ parole una ulteriore parola (detta *somma di controllo*), i cui bit si ottengono applicando il controllo di parità a tutti i bit "omologhi" delle $M$ parole incolonnate, generando così una parità *di colonna* (o *longitudinale*), come esemplificato in figura.




A volte, si preferisce calcolare la somma di controllo mediante una operazione di somma *modulo uno*[26], direttamente realizzabile in software in modo veloce. In tal caso, il ricevitore calcola una nuova somma di controllo lungitudinale, includendo anche la somma di controllo originaria: in assenza di errori, il risultato deve fornire zero.

#### 5.3.2.3 Codici polinomiali e CRC

L'utilizzo di una somma di controllo può produrre risultati scadenti nel caso di distribuzioni temporali dei bit errati particolarmente sfavorevoli, mentre l'uso del "CRC" garantisce prestazioni *più uniformi*. Questo metodo di controllo di errore consiste nell'aggiungere ad una parola $P$ di $k$ bit che si desidera trasmettere, un gruppo $R$ di $m < k$ ulteriori bit *di protezione*, calcolati a partire dai primi $k$, e tali da permettere la detezione di eventuali errori; in questo senso, i codici polinomiali sono classificabili come *codici a blocchi* (§ 5.3.3.1).




L'aggettivo *polinomiale* trae origine dalla associazione tra un numero binario $B$ di $n+1$ bit, indicati con $b_i$, $i = 0, 1, \ldots n$, ed un polinomio[27] $B(x)$ 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 $m$, i cui coefficienti binari identificano una parola $G = g_m g_{m-1} \ldots g_1 g_0$ di $m+1$ bit.

Indicando ora con $P$ la sequenza dei $k$ bit $p_i$ da proteggere, *aggiungiamo* a destra di questi un gruppo di $m$ bit pari a zero, ottenendo una nuova parola $P \cdot 2^m$ lunga

[25] La somma modulo due è equivalente alla operazione di OR esclusivo, viene a volte indicata con il simbolo $\oplus$, e corrisponde alle definizioni: $0 \oplus 0 = 0$, $0 \oplus 1 = 1$, $1 \oplus 0 = 1$, $1 \oplus 1 = 0$.

[26] 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.

[27] 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.

$k+m$ bit, che quindi dividiamo per $G$ (mediante aritmetica modulo due[28]), ottenendo un quoziente $Q$, ed un resto $R$ con al massimo $m$ bit. Pertanto, possiamo scrivere

$$\frac{P \cdot 2^m}{G} = Q \oplus \frac{R}{G}$$

Le sequenze $Q$ ed $R$ costituiscono rispettivamente i coefficienti dei polinomi quoziente $Q(x)$ e resto $R(x)$, ottenibili dalla divisione di $P(x) \cdot 2^m$ per $G(x)$. Gli $m$ bit $R$ del resto sono quindi utilizzati come *parola di protezione*, in modo da esprimere la sequenza $T$ da trasmettere come $T = P \cdot 2^m \oplus R$ di $k+m$ bit, ovvero con i $k$ bit più elevati pari a $P$ e gli $m$ bit in coda pari ad $R$. Il ricevitore effettua anch'esso una divisione, stavolta tra $T$ e $G$, che in assenza di errori produce un *resto nullo*

$$\frac{T}{G} = \frac{P \cdot 2^m \oplus R}{G} = \frac{P \cdot 2^m}{G} \oplus \frac{R}{G} = Q \oplus \frac{R}{G} \oplus \frac{R}{G} = Q$$

in quanto sommare un numero per se stesso in aritmetica modulo due, produce un risultato nullo. Pertanto, se $T/G = Q$ con resto nullo, la parola $P$ è riottenuta semplicemente shiftando $T$ a destra di $m$ posizioni.

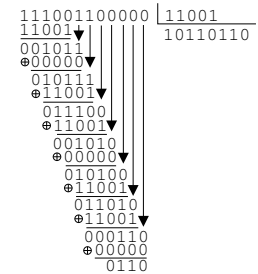
Nel caso invece in cui si siano verificati errori, indichiamo con $E$ la sequenza binaria di errore, di lunghezza $k+m$ bit, ognuno dei quali è pari ad uno se in quella posizione si verifica errore, o zero in caso contrario, in modo da rappresentare il segnale ricevuto $R$ come $R = T \oplus E$. Se $E \neq 0$ la divisione operata al ricevitore ora fornisce

$$\frac{R}{G} = \frac{T \oplus E}{G} = \frac{T}{G} \oplus \frac{E}{G} = Q \oplus \frac{E}{G}$$

e quindi si verifica la presenza di un resto pari ad $E$ e diverso da zero, che indica appunto la presenza di errori, tranne nei casi in cui $E$ risulti perfettamente divisibile per $G$, evento con bassa probabilità se $G$ è scelto oppurtunamente.

Per applicare il metodo, sia il trasmettitore che il ricevitore devono utilizzare lo stesso generatore $G(x)$, per il quale esistono diverse scelte standardizzate[29]. Si può dimostrare che le operazioni discusse permettono di rivelare

[28] Per fissare le idee, consideriamo $k = 8$ bit a da proteggere, pari a $P = 11100110$, $m = 4$ bit di CRC, ed un generatore $G = 11001$. La sequenza $P \cdot 2^m$ risulta pari a 11100110 0000, e la divisione modulo 2 tra $P$ e $G$ fornisce un quozionte $Q = 10110110$ (che viene ignorato) ed un resto $R$ pari a 0110. Pertanto, viene trasmessa la sequenza $T = P \cdot 2^m \oplus R = 11100110\,0110$ con $k+m = 12$ bit.




La divisione modulo 2 si realizza come mostrato nella figura a lato, calcolando un OR-ESCLUSIVO $\oplus$ bit-a-bit tra i bit più significativi del divisore $P$, o del resto parziale, e quelli del dividendo $G$. Per ognuna di queste operazioni, si *abbassano* una o più cifre binarie del divisore, in modo che il resto parziale abbia *la stessa lunghezza* del divisore; in tal caso, si ottiene un bit di risultato di quoziente pari ad uno, mentre se il resto parziale è troppo corto, per ogni bit abbassato in più si ottiene un bit di resto parziale pari a zero. Quando tutti i bit del divisore sono stati usati, l'ultima operazione $\oplus$ fornisce il resto $R$ cercato.

[29] 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 | $G(x) = x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x + 1$ |

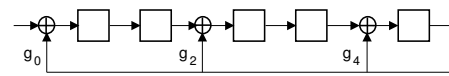
Come discusso, un polinomio di ordine $k$ genera un CRC di $k$ bit; pertanto il CRC-12, che è usato per caratteri a 6 bit, genera 12 bit di CRC, mentre CRC-16 e CRC-CCITT, utilizzati in America ed in Europa rispettivamente per caratteri ad 8 bit, producono 16 bit di CRC. In alcuni standard di trasmissione sincrona punto-punto, è previsto l'uso di CRC-32.

- tutti gli errori singoli;
- tutti gli errori doppi se $G$ possiede almeno tre uni;
- qualsiasi numero dispari di errori, se $G(x)$ contiene il fattore $(x+1)$;
- tutti gli errori a burst che si estendono per $k$ o meno bit;
- la maggior parte degli errori impulsivi più estesi.

**Calcolo del CRC** Gli $m$ bit aggiunti prendono il nome di CYCLIC REDUNDANCY CHECK (CRC o *ridondanza di controllo ciclico*) in virtù del particolare modo in cui questi possono essere calcolati: la parte del procedimento descritto che appare più complessa, ossia la divisione modulo due, è invece quella che ha maggiormente contribuito al successo del metodo, dato che è realizzabile a livello circuitale in modo relativamente semplice.

Si tratta infatti di utilizzare un registro a scorrimento controreazionato (vedi figura seguente), in cui sono immessi ad uno ad uno i $k$ bit da proteggere, seguiti da $m$ zeri consecutivi. Per ogni valore immesso, quelli già presenti *scorrono* a destra nel registro, ed il bit che *trabocca* alimenta gli OR esclusivi presenti nel registro, in corrispondenza degli uni di $G(x)$, tranne che per il bit corrispondente al termine più significativo; l'esempio in figura rappresenta il caso di $G(x) = x^5 + x^4 + x^2 + 1$.

Al termine dell'inserimento di $k+m$ bit, lo stato del registro a scorrimento costituisce proprio il resto $R$, da usare come CRC.



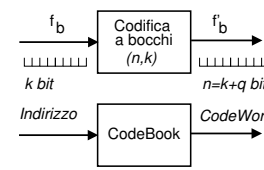

### 5.3.3 Correzione di errore e codifica di canale

Sono ora esposte alcune tecniche, indicate come *codifiche di canale*, utilizzabili anche per la semplice rivelazione di errore, ma più spesso finalizzate al tentativo di rendere il ricevitore in grado di *correggere* gli errori sopraggiunti nella trasmissione. L'argomento viene ripreso al § 17.3.

#### 5.3.3.1 Codici a blocchi

La *codifica a blocchi* opera opera su blocchi di dati disgiunti, e per ogni gruppo di $k$ bit in ingresso, distinto dagli altri gruppi di $k$, determina un gruppo di $q$ bit di protezione, dipendenti dai $k$ in ingresso, e che sono inviati assieme ai $k$ originali per un totale di $n = n + q$ bit.

L'operazione compiuta dal codificatore può essere pensata o realizzata come un accesso a memoria (denominata CODEBOOK = *dizionario*), dove i $k$ bit da codificare rappresentano un *indirizzo* che individua $2^k$ differenti *parole di codice* (CODEWORD), costituite ognuna da $n = k + q$ bit. Un codice siffatto è detto codice $(n, k)$.




**Ridondanza** Una misura delle prestazioni del codificatore di canale è rappresentata dalla *ridondanza* $\rho$, definita come il rapporto tra il numero di bit di protezione e quelli di informazione, e che è espressa in termini percentuali come

$$\rho = \frac{q}{k} \cdot 100\,\%$$

Ad esempio, in una trasmissione con ridondanza del 50%, per ogni due bit di informazione ne viene inserito (da qualche parte) uno di protezione.

**Distanza di Hamming** L'entità della ridondanza introdotta migliora la capacità del codice di correggere uno o più errori, in quanto non tutte le $2^n$ possibili codeword (di $n$ bit) sono utilizzate, ma solamente $2^k$ tra esse: le configurazioni assenti dal codebook possono comunque presentarsi in ricezione, a causa di errori di trasmissione, ma non essendo previste dal codebook, è possibile rilevare l'errore ed eventualmente correggerlo. Le capacità di rivelazione e correzione sono valutate quantitativamente per mezzo del concetto di *distanza di Hamming* $d_H$ tra parole del codebook, che rappresenta il minimo numero di bit diversi tra due parole di codice. Si dimostra infatti, che un codice di canale è in grado di correggere $\frac{d_H-1}{2}$ errori per parola, e di rivelare $d_H - 1$ errori per parola[30]. Un codice a blocchi $(n, k)$ presenta una $d_H \leq n - k + 1$.

**Efficienza** L'efficienza del codice è misurata dal *tasso di codifica* (CODE RATE)

$$R_c = \frac{k}{n} < 1$$

che rappresenta la frazione di bit informativi sul totale di quelli trasmessi, e che consente di scrivere la velocità di uscita dal codificatore come

$$f_b' = \frac{f_b}{R_c} > f_b$$

Ad esempio, in una trasmissione con un tasso di codifica pari a 0.5, il numero di bit uscenti (per unità di tempo) dal codificatore di canale, è il doppio del numero dei bit entranti.

L'argomento dei codici a blocchi è molto vasto[31], e fornisce molteplici soluzioni, la cui trattazione esauriente trascende dallo scopo attuale del testo; gli stessi tre casi di controllo di errore già trattati (parità[32], somma di controllo e CRC) possono essere inquadrati nel contesto dei *codici a blocchi*. Nel seguito, sono fornite alcune definizioni di base, ed un esempio di codice a correzione molto elementare, il codice *a ripetizione*; l'argomento viene ripreso al § 17.3.2, dove è illustrato il codice *di Hamming*.

**Codice a ripetizione** Un esempio di codice a blocchi molto semplice e con proprietà correttive è il codice a ripetizione $n : 1$, che per ogni bit in arrivo ne produce $n$ identici in uscita, in modo che (se gli errori sono indipendenti) il decisore possa correggere

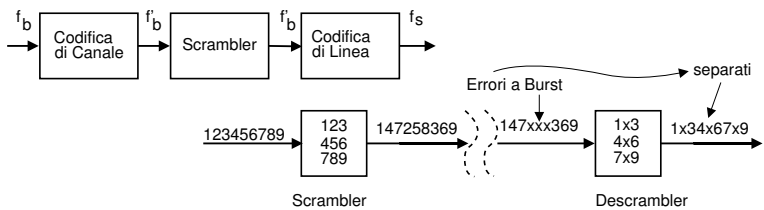
[30] Ad esempio, se $d_H = 3$, la presenza di un solo bit errato fa sì che la sequenza ricevuta abbia un solo bit di differenza rispetto alla codeword trasmessa, ed al minimo due bit di differenza rispetto a tutte le altre codeword, permettendo al ricevitore di correggere l'errore. Se invece sono presenti due errori, la procedura di correzione porterebbe a scegliere una codeword errata. Occorre quindi decidere a priori se utilizzare il codice a fini di correzione oppure di detezione, e nel secondo caso, lo stesso codice con $d_H = 3$ può rivelare fino a due errori per parola.

[31] 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...*

[32] Il concetto di parità può essere il ulteriormente esteso, se ognuno dei $k$ bit aggiunti è calcolato applicando la regola della parità ad un sottoinsieme degli $m$ bit di ingresso, con sottoinsiemi eventualmente sovrapposti. Un codice del genere prende il nome di codice di *Hamming*, descritto al § 17.3.2.

l'errore in base ad una "votazione a maggioranza" (*majority voting*). Ponendo ad esempio $n = 3$, definiamo il codice a ripetizione $3 : 1$, composto da due codeword, 000 ed 111, per le quali risulta $d_H = 3$: pertanto, il codice è in grado di correggere un errore e rivelarne due[33]. L'esercizio a pag. 148 dimostra come, se la probabilità di errore per un bit $P_e$ è molto piccola, l'applicazione della decisione a maggioranza permetta di ridurne il valore a $P_e^{dec} \backsimeq 3P_e^2$. Il codice a ripetizione infine, è l'unico che fornisce esattamente $d_H = n - k + 1$, ed essendo $k = 1$, si ottiene $d_H = n$, a discapito di un tasso di codifica mediocre, e pari a $R_c = 1/n$.

**Errori a burst ed interleaving** Nonostante le capacità di correzione dei codici a blocchi possano sembrare adeguate a ridurre il valore di probabilità di errore, gli errori possono presentarsi in maniera non indipendente, ma concentrati in un breve intervallo di tempo: questo circostanza prende il nome di fenomeno degli *errori a pacchetto* (od a BURST=*scoppio*). In tal caso, si usa ricorrere alla tecnica nota come SCRAMBLING[34] o INTERLEAVING[35], attuabile a patto di accettare un ritardo. Si tratta infatti di modificare l'ordine dei dati inviati, in modo che gli errori che avvengono su bit *vicini* si riflettano in errori su bit... *lontani* e quindi appartenenti a codeword differenti. Ovviamente, occorre prevedere un processo inverso *(descrambling* o *deinterleaving)* all'altro capo del collegamento. E' appena il caso di notare che lo scrambler (similmente al codice di Gray) *non altera* il *numero* dei bit trasmessi.




## 5.4 Protocolli ARQ

Le trasmissioni ARQ prevedono l'esistenza di un canale di ritorno, mediante il quale chiedere la ri-trasmissione delle trame ricevute con errori[36]; 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 efficenza maggiore.

[33] 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.

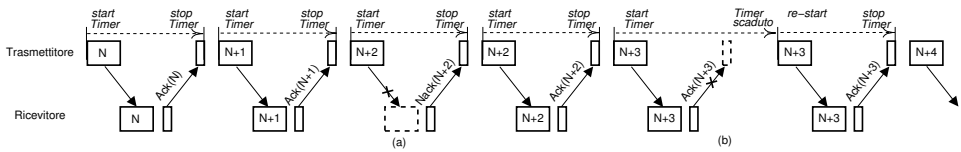
[34] Letteralmente: arrampicamento, ma anche "arruffamento", vedi *scrambled eggs,* le uova strapazzate dell'*english breakfast.*

[35] LEAVE = *foglia, sfogliare, rastrellare,* ed il termine potrebbe essere tradotto come *intercalamento.*

[36] 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.

### 5.4.1 Send and wait

Viene trasmessa una trama alla volta, e si attende un riscontro (*ACKnoledgment*) 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 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 al ricevitore, nel caso (b), di riconoscerela 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 $U = 0.33$;

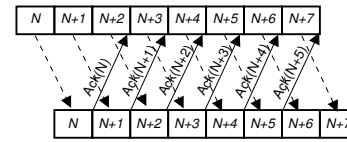
**c)** per le velocità di 1 kbps ed 1 Mbps si ottiene $a = 0.167$ ed $a = 167$ rispettivamente, a cui corrispondono efficienze pari a $U = 0.75$ e $U = 0.003$.

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.

### 5.4.2 Continuous RQ

A differenza del protocollo *send-and-wait*, ora il trasmettitore invia le trame ininterrottamente, senza attendere la ricezione degli ACK. In presenza di trame ricevute correttamente, il ricevitore riscontra positivamente le stesse, consentendo al trasmettitore di liberare i buffer di trasmissione.

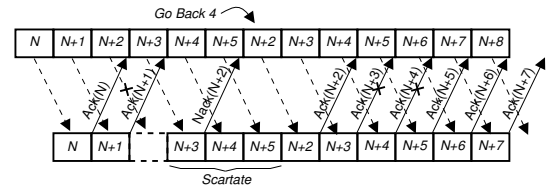
In presenza di trame ricevute con errori, la quantità di memoria tampone utilizzata al ricevitore determina la scelta di due possibili strategie di richiesta di ritrasmissione, denominate *go-back-N* e *selective-repeat*.




#### 5.4.2.1 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[37] 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 $N+i$, finchè 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.

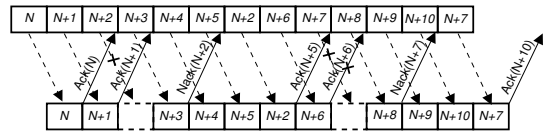



Una variante del metodo, idonea al caso in cui fenomeni di congestione di rete possano determinare la perdita dei NACK, prevede l'uso di un timer al trasmettitore, per re-inviare le trame non riscontrate.

[37] 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.

#### 5.4.2.2 Selective repeat

L'origine di questo nome deriva dal fatto che non è più richiesto al trasmettitore di *tornare indietro* completamente, ma è sufficiente ritrasmettere solamente la trama che ha dato origine ad errore in ricezione, grazie alla capacità del ricevitore di memorizzare temporaneamente più trame, anche se ricevute fuori sequenza.




Come si nota in figura, a seguito della ritrasmissione della trama $N+2$ per cui si è ricevuto il NACK, il trasmettitore continua a trasmettere nuove trame, fino al numero massimo previsto; in assenza di ulteriori ACK, un timer determina la ritrasmissione delle trame non riscontrate.

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 differente da quello naturale. Per questo motivo, anche il ricevitore deve predisporre delle memorie temporanee dove salvare le trame correttamente arrivate, successivamente a quella che invece conteneva errori, e di cui si attende la ritrasmissione.

#### 5.4.2.3 Efficienza

Indicando con $p$ la probabilità di dover chiedere la ritrasmissione di una trama[38] per la quale si sono rilevati errori di trasmissione, il numero totale di trasmissioni necessarie alla sua corretta ricezione (indichiamolo con $m$) è una variabile aleatoria, descritta dalle probabilità $p_M(1) = Pr(m=1) = 1-p$, $p_M(2) = p(1-p)$, $p_M(3) = p^2(1-p)$, $\ldots p_M(m) = p^{m-1}(1-p)$ etc. 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 tenuto conto della relazione espressa alla nota 27 di pag. 172, e del fatto che $\sum_{k=0}^{\infty} k\alpha^{k-1} = \sum_{k=1}^{\infty} k\alpha^{k-1} = \sum_{k=0}^{\infty}(k+1)\alpha^k$. Quindi, per trasmettere una frequenza binaria di $f_b$

[38] Nel caso in cui l'integrità della trama sia protetta da un codice a blocchi $(n,k)$ con $d_H = l+1$, la probabilità che la trama contenga più di $l$ errori e che quindi venga accettata dal ricevitore anche se errata, vale approssimativamente $P(l+1,n) = \binom{n}{l+1}p^i$ (vedi formula (5.9)). Dato che il ricevitore accetta le trame che *non hanno* errori, oppure che hanno più di $l$ errori, la probabilità che venga richiesta una ritrasmissione risulta $p = 1 - P(0,n) - P(l+1,n)$. Considerando ora che un buon codice deve fornire $P(l+1,n) \ll P(0,n)$, si ottiene $p \simeq 1 - P(0,n) = 1-(1-P_e)^n \simeq nP_e$, in cui $P_e$ è la probabilità di errore sul bit (dato che $(1-P_e)^n \simeq 1-nP_e$ se $nP_e \ll 1$).

bps (comprensivi di CRC e *overhead* dei numeri di sequenza), occorre disporre di un canale di capacità $f_b/(1-p)$ bps[39]. Questo risultato approssimato si applica ad un protocollo di tipo *selective repeat*, e trascurando gli errori sul canale a ritroso.

### 5.4.3 Controllo di flusso

Si è illustrato che nei protocoli ARQ il trasmettitore, dopo un pò che non riceve nuovi ACK, cessa a sua volta di inviare trame, dato che esaurisce la memoria temporanea in cui memorizzare le trame già inviate ed in attesa di riscontro. Nel caso in cui il ricevitore non sia in grado di smaltire in tempo i dati ricevuti, può scegliere di sospendere temporaneamente l'invio di riscontri, con il risultato di rallentare la velocità di invio dei dati. Questo meccanismo prende il nome di *controllo di flusso*, per l'evidente analogia idraulica, in cui una conduttura viene ristretta al fine di ridurre il flusso di liquido in transito.

Dato che il ritardo tra la sospensione dell'invio degli ACK, e l'interruzione dell'invio di trame, dipende dalla dimensione delle memorie temporanee, e che questa dimensione incide allo stesso tempo anche sulla efficienza di utilizzo temporale del collegamento in condizioni di ricezione a piena velocità, svolgiamo alcune riflessioni sull'argomento.

#### 5.4.3.1 Round trip time

E' il tempo che intercorre tra l'inizio della trasmissione di una trama, e l'arrivo del relativo ACK. La sua valutazione spesso si avvale della ipotesi di poter trascurare il tempo di trasmissione dell'ACK, e quindi si ottiene $RTT = t_{Tx} + 2t_p$. Se la trasmissione avviene a velocità $f_b$, allora in un tempo pari a $RTT$ possono essere trasmessi $N_{ba} = f_b \cdot RTT$ bit, che possono essere pensati come il numero di *bit in aria*[40]. Se la memoria temporanea del trasmettitore ha dimensioni $W \geq N_{ba}$, allora la trasmissione (senza errori) può avvenire senza soluzione di continuità, impegnando constantemente il collegamento.

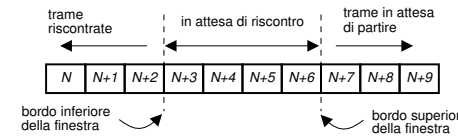
#### 5.4.3.2 Finestra scorrevole

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.*

[39] Dato che $p$ aumenta con $n$ (vedi pag. 78), 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 $n$.

[40] 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.

La condizione di *massima apertura* della finestra identifica la quantità di memoria necessaria al trasmettitore per realizzare un protocollo ARQ, che quindi può essere ri-classificato in questi termini come indicato dalla tabella precedente, dove la colonna *finestra di ricezione* indica anche i requisiti di memoria al lato ricevente[41]. Notiamo infatti che mentre per *send-and-wait* è sufficiente la memoria di una sola trama, per *go-back-N* il trasmettitore deve ricordare fino ad un massimo di $W$ trame in attesa di riscontro, e per *selective repeat* anche il ricevitore ha lo stesso vincolo, allo scopo di riodinare le trame ricevute fuori sequenza.




| | *f. di trasm.* | *f. di ric.* | $N_{Max}$ |
|---|---|---|---|
| send and wait | 1 | 1 | 2 |
| go back N | W | 1 | W+1 |
| selective repeat | W | W | 2W+1 |

#### 5.4.3.3 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 $(0-1)$ perchè, nel caso in cui l'ACK sia corrotto, il ricevitore possa riconoscere la trama ricevuta come duplicata anziché nuova; un ragionamento simile[42] determina la necessità di usare $W+1$ numeri $(0-W)$ nel caso *go-back-N*, e $2W+1$ numeri $(0-2W)$ nel caso *selective repeat*.

L'uso di un numero di bit ridotto per indicare il numero di sequenza, permette di limitare la dimensione dell'*overhead* di trama; ad esempio, con una finestra di dimensione 7, l'uso di go-back-N richiede 8 diversi numeri di sequenza, che quindi possono essere codificati utilizzando 3 bit.

# Appendici

## 5.5 Sincronizzazione dati

Come mostrato al § **??**, il segnale dati deve essere campionato al ricevitore, con cadenza pari alla frequenza di simbolo $f_s$, possibilmente al centro dell'intervallo di simbolo, in modo da contrastare gli effetti della limitazione di banda (vedi fig. 5.1); per questo motivo, occorre che il temporizzatore mostrato a pag. 63 determini gli istanti di cam-

[41] La ricezione di una sequenza di trame corrette, determina l'avanzamento alternato dei due bordi della finestra al ricevitore, che è inizialmente vuota, quindi contiene solo la trama ricevuta (avanza bordo superiore), e quindi è di nuovo svuotata, non appena viene trasmesso l'ACK (ed avanza il bordo inferiore). In presenza di errori, il bordo inferiore non avanza, ma resta fermo sulla trama ricevuta con errori, e di cui si attende la ritrasmissione. Mentre il trasmettitore continua ad inviare trame, il ricevitore le memorizza e fa avanzare il bordo superiore, finché non siano state ricevute tutte quelle trasmissibili senza riscontro, e pari alla dimensione massima della finestra in trasmissione.

[42] 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.

pionamento più idonei, effettuando la *sincronizzazione di simbolo*[43]; le diverse scelte per l'onda elementare $g(t)$ discusse in 5.2.1, determinano differenti gradi di "difficoltà" nel conseguimento della sincronizzazione di simbolo.

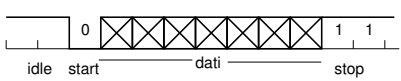
Un differente aspetto della sincronizzazione, riguarda il problema di ricostruire la struttura *sintattica*[44] del segnale binario. Infatti, la sequenza di bit ottenuta al ricevitore, è spesso il risultato della serializzazione (al lato del trasmettitore) di informazioni *a carattere* (o *parola*, o WORD), come nel caso di un *file* di testo, oppure dei campioni di un segnale[45]. Pertanto, il ricevitore deve essere in grado di delimitare, nell'ambito del flusso di bit ricevuti, le unità di *parola*, e di raggruppare le parole in *trame*.

Nel seguito, analizziamo le esigenze e le soluzioni di sincronizzazione che emergono nell'ambito di due diverse tecniche di trasmissione, indicate come *asincrona* e *sincrona*, che si differenziano per il fatto che mentre nella prima le parole sono separate tra loro, nella seconda fluiscono senza interruzione.

### 5.5.1 Trasmissione asincrona

Viene adottata, ad esempio, nel caso di un terminale *stupido*[46] 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*[47].

Quando è pronto un carattere da trasmettere, il segnale viene portato nello stato *zero* (detto SPACE) per la durata di 1 simbolo, che prende il nome di *bit di start*: la transizione *in discesa* viene rilevata dal ricevitore, che si predispone a contare un numero fisso di simboli (7 in figura) basandosi su di un suo orologio indipendente.




Segue poi un bit di *parità* (se prevista, vedi § 5.3.2.1 per la sua definizione) ed uno o due di *stop* (realizzato con MARK), presenti per assicurare una durata minima dello stato di IDLE, prima della trasmissione del carattere successivo.

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[48], che viene *risvegliato* in corrispondenza del bit di start. Tale semplicità operativa produce una inefficienza, in

[43] 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 uno stesso circuito stampato.

[44] 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 spontamenamente 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*).

[45] In appendice 5.6 è riportata la codifica in termini di sequenze binarie dei caratteri stampabili, definita dallo standard ASCII; al § 6.3.1 si mostra la struttura della *trama* PCM, che trasporta i campioni di più sorgenti analogiche campionate.

[46] 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.

[47] In tal caso la linea ".. IS MARKING TIME" (sta marcando il tempo).

[48] Ovviamente, occorre stabilire un accordo a priori a riguardo la velocita della trasmissione.

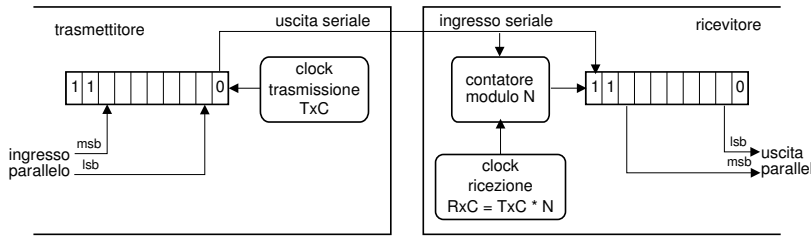



Figura 5.6: Trasmissione asincrona

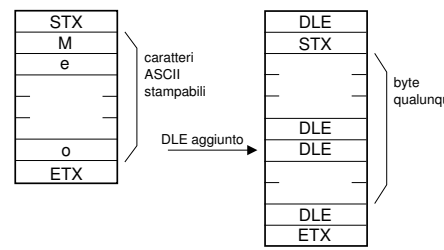
quanto oltre ai dati ed al bit di parità, si introduce anche lo start e lo stop, utili solo ai fini della sincronizzazione ma privi di contenuto informativo.

In fig. 5.6 è mostrato uno schema di funzionamento di trasmettitore e ricevitore, che mostra come le parole entrino in modo parallelo nel trasmettitore, e vengano serializzate alla frequenza `TxC` di trasmissione, assieme ai bit di start e di stop.

**Sincronizzazione di bit e di parola** Il ricevitore dispone di un orologio interno operante ad un ritmo (`RxC`) multiplo di quello di trasmissione, che incrementa un contatore modulo $N$, il quale divide l'orologio, e con il suo azzeramento determina il processo di decisione per un bit, il suo caricamento nel registro di ricezione, e lo shift a destra dei contenuti dello stesso. Alla ricezione del bit di start, il contatore è inizializzato con il valore $N/2$, in modo che il suo azzeramento avvenga a metà del periodo di bit. A seguito dell'azzeramento, il contatore torna a dividere `RxC` per $N$, ed i suoi successivi azzeramenti avvengono sempre a metà del periodo di bit, fino alla ricezione di una intera parola. Al suo termine (segnalato dall'arrivo in prima posizione del bit di start a zero), il registro di ricezione è letto in modo parallelo, recuperando una intera parola.

**Sincronizzazione di trama** I caratteri trasmessi possono far parte di messaggi più estesi, come ad esempio i paragrafi di un *file* di testo.

Per questo motivo, può essere necessario inserire dei caratteri speciali tra quelli trasmessi, con lo scopo di delimitare i gruppi di caratteri che appartengono ad uno stesso messaggio. Se le diverse parole da trasmettere non sono tutte quelle possibili in base alla lunghezza di parola adottata[49], 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. 95), 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

[49] 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.

dati trasmessi[50], causando un *troncamento* prematuro del messaggio. In tal caso, sia STX che ETX vengono fatti precedere da un terzo carattere speciale, il DLE (*Data Link Escape*). Il trasmettitore, dopo aver inserito la coppia DLE-STX iniziale, ispeziona ogni carattere da inviare, e se questo *simula* un DLE, *inserisce* un secondo DLE, attuando una procedura detta CHARACTER (o BYTE) STUFFING. Il ricevitore a sua volta, per ogni DLE ricevuto, controlla se la parola successiva è un ETX, nel qual caso considera terminata la trasmissione; altrimenti, controlla se è un secondo DLE, che è stato inserito dal trasmettitore, e lo rimuove. Altri casi non sono possibili, e se si verificano, rivelano la presenza di un errore di trasmissione.

### 5.5.2 Trasmissione sincrona

La trasmissione dei bit di start e di stop necessaria per effettuare una trasmissione asincrona è fonte di inefficienza, e per questo a velocità più elevate si preferisce non inframmezzare i dati da trasmettere con delimitatori aggiuntivi. Ciò comporta l'esigenza di adottare in ricezione soluzioni apposite per individuare gli istanti di decisione corretti, e quindi conseguire il sincronismo di simbolo. Il sincronismo di parola si basa in generale sull'uso di parole di lunghezza costante, mentre quello di trama prevede due possibili soluzioni, l'una orientata al carattere, e l'altra al bit.

**Sincronizzazione di simbolo** La figura 5.7-a mostra uno schema idoneo ad estrarre la temporizzazione `RxC` dal segnale ricevuto, basata sull'uso di un circuito DPLL (*Digital Phase Lock Loop*), il cui funzionamento richiede la presenza di transizioni nel segnale ricevuto. Analogamente allo schema già analizzato nel caso di trasmissione asincrona, un orologio locale opera ad una frequenza $N$ volte più elevata di quella nominale, e il DPLL (fig. 5.7-b) ne divide l'orologio per $N$, fornendo il segnale `RxC` necessario al decisore per individuare gli istanti posti al centro di un intervallo di simbolo. La divisione per $N$ è realizzata all'interno del DPLL mediante un contatore all'indietro, che al suo azzeramento produce il segnale `RxC` di sincronismo, ed è ri-caricato con la costante $N$. Nel caso in cui si verifichi uno *slittamento di fase* tra il segnale ricevuto e l'orologio locale, questo può essere rilevato osservando che la transizione (quando presente) nel segnale non ricorre nella posizione presunta, ossia a metà del conteggio, ma in anticipo od in ritardo (fig. 5.7-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[51].

Nel caso di una differenza di velocità tra l'orologio di ricezione ed il ritmo di segnalazione, le correzioni avvengono di rado, e sono di entità ridotta. Al contrario, all'inizio di una trasmissione la differenza di fase può essere qualsiasi: per questo motivo, prima dei dati veri e propri, viene trasmessa una sequenza *trailer*, allo scopo di permettere appunto l'acquisizione del sincronismo di simbolo. La durata del trailer

[50] 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.

[51] In termini generali, questo circuito è assimilabile ad un circuito di controllo, in quanto il suo principio di funzionamento si basa sul tentativo di azzeramento di una grandezza di errore. Infatti, la sincronizzazione dell'orologio del campionatore dello stadio ricevente 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, ed il segnale di errore alimenta un circuito di controreazione che mantiene il clock locale *al passo* con quello dei dati in arrivo. Un caso particolare di questo stesso principio è analizzato ai §11.2.1.3 e 11.3.1.1 a proposito del PLL.

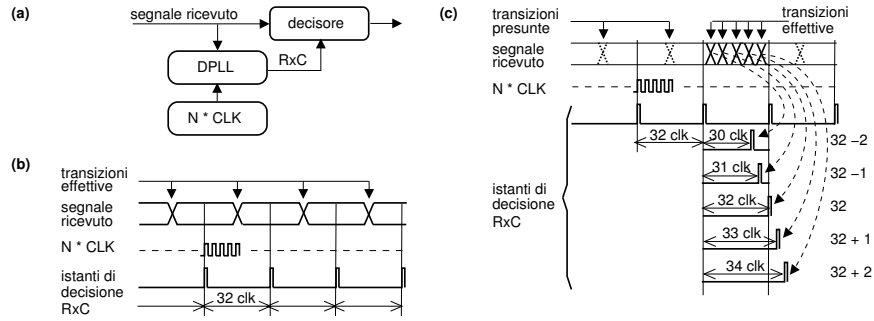
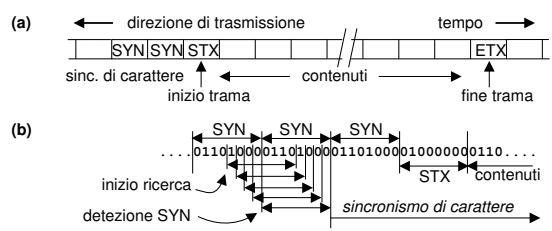



Figura 5.7: Funzionamento del *DPLL*: (a) schema circuitale; (b) condizioni di sincronismo; (c) correzione di fase

dipende dalla velocità di convergenza della procedura, per accelerare la quale sono imposte correzioni di maggiori entità in corrispondenza di errori di fase più elevati.

**Trasmissione orientata al carattere** La trasmissione orientata al carattere è usata principalmente nel caso di contenuti testuali, come per i file ASCII. In assenza dei bit di start e di stop, la sincronizzazione di carattere è ottenuta per mezzo della trasmissione, prima dei dati veri e propri, di una sequenza di caratteri SYN (*Synchronous Idle)*, che permettono sia di conseguire (o mantenere) il sincronismo di bit, che di consentire l'inviduazione dei confini di carattere, e quindi il sincronismo di carattere.

La figura seguente 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, entra in un modalità di ricerca, verificando (fig(b)) se l'allineamento di 8 bit consecutivi corrisponde al carattere SYN, ed in caso negativo, ripete il tentativo bit a bit.

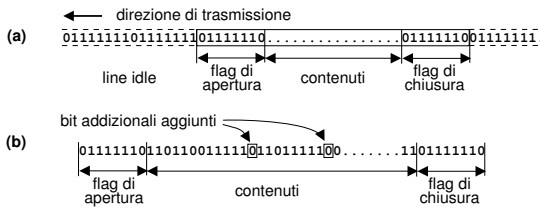



Una volta individuato il SYN, il ricevitore ha conseguito l'allineamento sul carattere, ed inizia ad aspettare il carattere STX, che indica l'inizio della trama, che è terminata da un ETX.
Nel caso in cui la trasmissione contenga caratteri qualunque, e dunque L'ETX posssa essere simulato dai dati, si 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. 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 bytes* 01111110 per indicare sia l'inizio che la fine di una trama.

La figura che segue mostra in (a) un esempio di trama, ed in (b) la soluzione 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. In tal modo, quando il ricevitore osserva un bit pari a 0 preceduto da 5 bit pari ad uno consecutivi, lo rimuove, conseguendo così la *trasparenza dai dati*, e permettendo il corretto rilevamento del flag byte di fine trama. Ovviamente, la procedura di bit stuffing/destuffing è applicata solamente al *contenuto* della trama.




## 5.6 Codifica di carattere

Il codice ASCII (*American Standard Code for Information Inter-change*) è un codice a 7 bit, e molti codici ad 8 bit (come l'ISO 8859-1) si riducono ad ASCII nella loro metà bassa (con il bit più significativo a zero); i primi 32 codici corrispondono a caratteri *non stampabili*, detti codici di controllo, ottenibili su di una tastiera mediante la pressione del tasto CONTROL, e che hanno un significato speciale, come il *carriage return* (CR), il *line feed* (LF), *start of text* (STX), *backspace* (BS), *data link escape* (DLE). La tavola 5.1 mostra i 128 caratteri ASCII. La controparte internazionale dell'ASCII è nota come ISO 646; lo standard è stato pubblicato dallo *United States of America Standards Institute* (USASI) nel 1968.

### 5.6.1 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

Tabella 5.1: codici ASCII

| *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 |

- i restanti CodePoints sono rappresentati mediante quattro byte

# Capitolo 6

# Reti di trasmissione a circuito

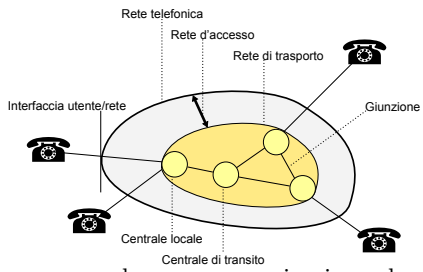
In questo capitolo si espongo i principi di funzionamento della rete pubblica commutata, dalle origini della telefonia analogica, alle tecniche di segnalazione, di multiplazione plesiocrona e sincrona della telefonia numerica, alla commutazione, ed ai sistemi di trasporto.

## 6.1 Introduzione

I casi di trasmissione finora analizzati fanno riferimento a collegamenti *punto-punto*, in cui una unica sorgente di informazione intende comunicare con un unico destinatario. Nella realtà, è assai più frequente il caso in cui i soggetti coinvolti nella comunicazione affidino la stessa ad una *rete* di collegamento, consegnando il messaggio al *nodo* di commutazione a cui hanno accesso. Il messaggio quindi, una volta determinato un percorso di *attraversamento* che coinvolga i nodi della rete più opportuni, giunge al destinatario, grazie anche alla presenza di informazioni addizionali, dette *di segnalazione*.

Le nozioni che seguono fanno esplicito riferimento alle *reti di telefonia*, dette anche a *commutazione di circuito*, allo scopo di fornire una panoramica dei principali aspetti delle stesse. Altri aspetti legati alle *reti di trasmissione dati* saranno illustrati a seguito della trattazione relativa alla teoria del traffico, assieme alle reti a *commutazione di pacchetto*.

### 6.1.1 Elementi della rete telefonica




Con riferimento alla figura, si anticipa che discuteremo prima dei metodi di *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, nel § 6.8 si esaminano i metodi di *commutazione* ed *instradamento* con cui viene individuato il percorso che una comunicazione deve intraprendere tra l'ingresso e l'uscita dalla rete di trasporto.

### 6.1.2 La rete di accesso

E' la parte più rilevante e dispendiosa della rete, e consiste nel *doppino* in rame (pag. 362) che raggiunge la presa telefonica casalinga. All'interfaccia utente/rete sono così resi disponibili i servizi noti nel loro insieme come

- POTS (*Plain Old Telephone Service*) - vedi § 6.9.1;
- ISDN (*Integrated Service and Data Network*) - vedi § 6.9.2;
- ADSL (*Asimmetric Digital Subscriber Line*) - che in realtà usa POTS solo come tramite per raggiungere una *rete IP*- vedi § 6.9.4.

Oltre a questi, sono ormai contemplati nella rete di accesso anche tipi di collegamento diversi dal cavo, come

- accesso ottico - come nel caso FTTH (*Fiber To The Home*)[1], e che permette
  - di interconnettere un insieme più numeroso di collegamenti POTS già multiplati assieme, come nel caso di un grosso centralino aziendale, 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[2] - noto anche come sistema cellulare di seconda generazione[3], usa una rete diversa da PSTN, ma vi si interconnette in modo naturale. Il GSM nasce come standard aperto, favorendone la diffusione mondiale e l'interoperabilità tra gestori (*roaming*), e si sviluppa in forma completamente numerica, sia per la codifica vocale, che per il meccanismo di accesso multiplo al mezzo trasmissivo[4], che adotta una organizzazione in trame; inoltre, ha introdotto la comunicazione i messaggi sms[5];
  - GPRS[6] e UMTS[7] - 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[8] e WiMax[9] - mentre il primo distribuisce l'accesso ADSL su di un'area di estensione casalinga, il secondo ha una copertura di qualche chilometro, e permette collegamenti in mobilità. Entrambe permettono l'interconessione 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

[1] http://en.wikipedia.org/wiki/Fiber_to_the_x
[2] http://it.wikipedia.org/wiki/Global_System_for_Mobile_Communications
[3] La prima generazione si riferisce al sistema analogico TACS http://it.wikipedia.org/wiki/TACS
[4] http://en.wikipedia.org/wiki/Time_division_multiple_access
[5] http://it.wikipedia.org/wiki/SMS
[6] http://it.wikipedia.org/wiki/General_Packet_Radio_Service
[7] http://it.wikipedia.org/wiki/Universal_Mobile_Telecommunications_System
[8] http://it.wikipedia.org/wiki/Wi-Fi
[9] http://it.wikipedia.org/wiki/WiMAX

- CDN (*Circuito Diretto Numerico*) - offre la connettività diretta e continuativa con un altra (ben specifica) interfaccia utente/rete, e pertanto viene a mancare la componente di commutazione;
- VPN (*Virtual Private Network*) - come sopra, con la differenza che in questo caso la connettività è basata su di una comunicazione a pacchetto anziché a circuito.

## 6.2 Multiplazione

In generale, raggruppare assieme più comunicazioni dirette alla medesima destinazione, in modo che condividano uno stesso mezzo trasmissivo, permette di

- tentare di occupare tutta la banda messa a disposizione dal mezzo trasmissivo
- massimizzare la percentuale di utilizzo del mezzo trasmissivo, nel caso di sorgenti non continuamente attive (vedi § 8.3.4)
- semplificare la gestione e la manutenzione dei collegamenti a lunga distanza, essendo questi minori in numero

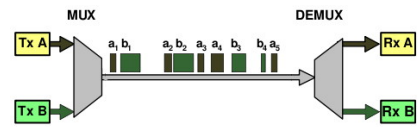
Le tecniche di multiplazione, nella evoluzione storica delle telecomunicazioni, possono operare secondo le diverse modalità di

- *divisione di frequenza* - ogni comunicazione usa una banda di frequenze diversa, come descritto al § 10.1.1, nel contesto dello studio dei segnali modulati;
- *divisione di tempo* - le comunicazioni avvengono in intervalli di tempo disgiunti, mediante segnali numerici, e sono affrontate nel resto di questo capitolo;
- *divisione di codice* - tutte le comunicazioni usano la stessa banda allo stesso tempo, ma ogni diverso destinatario è ancora in grado di distinguere il proprio messaggio, mediante una operazione di correlazione tra segnali ortogonali (vedi § 9.4.4.2). Dato che il segnale trasmesso in questo caso si trova ad occupare una banda più estesa di quella originaria, la tecnica è detta *ad espansione di spettro* ed esposta al § 13.6.5.

### 6.2.1 Multiplazione a divisione di tempo

Per quanto riguarda la multiplazione a divisione di tempo, nella pratica questa è attuata solo a partire da segnali numerici, come sono sempre stati i segnali dati, ed in quel contesto si è sviluppato un approccio basato sull'uso di un *pacchetto dati* (vedi § 8.5.1), attuando uno schema detto

**Multiplazione statistica e commutazione di pacchetto** In questo caso il mezzo trasmissivo non è impegnato in modo esclusivo, ma la trasmissione può avvenire in modo sporadico, ed i dati inviati ad intervalli irregolari. Questo motivo, assieme alla dimensione variabile delle singole comunicazioni, porta a suddividere la comunicazione in unità autonome indicate come pacchetto dati.




La multiplazione dei pacchetti avviene quindi in modo statistico, senza riservare con esattezza risorse a questo o quel tributario: il multiplatore si limita ad inserire i pacchetti ricevuti in apposite code, da cui li preleva per poterli trasmettere in sequenza, attuando una modalità di trasferimento *orientata al ritardo* (vedi § 8.4). La presenza di code comporta

- il determinarsi di un ritardo variabile ed impredicibile
- la possibilità che la coda sia piena, ed il pacchetto in ingresso venga scartato

D'altra parte, ogni pacchetto reca con sé le informazioni necessarie al suo recapito, facilitando il compito dell'instradamento. A seconda dell'adozione di un principio di commutazione di tipo *a circuito virtuale* oppure a *datagramma* (vedi § 8.5.2.2), può essere presente o meno una *fase di setup* precedente l'inizio della comunicazione.

**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 § 8.5.2.1) che determina un paradigma noto come *commutazione di circuito*, per il motivo che ora illustriamo.

Alle origini storiche della telefonia, nell'epoca dei telefoni *a manovella*, con la cornetta appesa al muro, la comunicazione si basava, grazie all'operato di un centralinista umano, sulla creazione di un vero proprio *circuito elettrico* che collegava fisicamente tra loro le terminazioni dei diversi utenti. Nel caso in cui intervengano più centralinisti in cascata, la chiamata risulta instradata attraverso più centralini. Da allora, il termine commutazione di circuito individua il caso in cui

- è necessaria una fase di *setup* precedente alla comunicazione vera e propria, in cui vengono riservate le risorse;
- nella fase di setup si determina anche l'*instradamento* della chiamata nell'ambito della rete, che rimane lo stesso per tutta la durata della medesima;
- le risorse trasmissive restano impegnate in *modo esclusivo* per l'intera durata della conversazione.

Le cose non sono cambiate di molto (da un punto di vista concettuale) con l'avvento della telefonia numerica: in tal caso, più segnali vocali sono campionati e quantizzati in modo sincrono, ed il risultato (numerico) è multiplato in una *trama PCM* (§ 6.3.1), in cui viene riservato un intervallo temporale per ognuno dei flussi tributari.

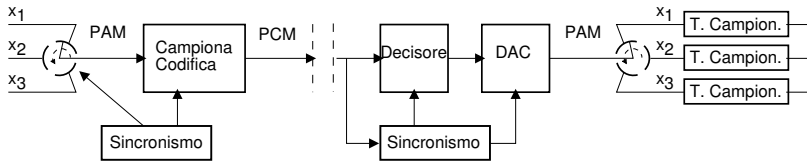
Ad ogni buon conto, si noti che un risultato della teoria del traffico (pag. 174) mostra come l'adozione di una strategia orientata al ritardo migliora notevolmente l'efficienza di utilizzo del mezzo stesso.

## 6.3 Rete plesiocrona

Questo termine si riferisce alla modalità di funzionamento quasi-sincrona adottata dalle centrali telefoniche, almeno finché la rete di trasporto non è divenuta capace di realizzare una modalità di multiplazione sincrona. In entrambi i casi, i segnali vocali sono trasportati in forma numerica, multiplandone i campioni a divisione di tempo in modo deterministico, in accordo ad una organizzazione di trama realizzata presso la centrale di accesso, come descritto di seguito.

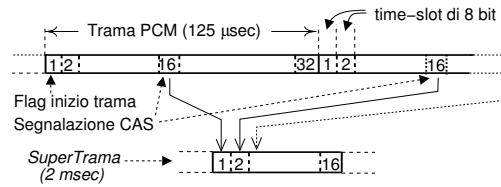
### 6.3.1 Trama PCM

Nella figura seguente sono rappresentati tre segnali *tributari*, campionati a turno alla stessa frequenza di 8 KHz, quantizzati ad 8 bit per campione con quantizzazione logaritmica (vedi § 7.6.1), e trasmessi (8 bit alla volta[10]) a turno su di un unico collegamento, producendo un segnale binario che prende il nome di *PCM* (PULSE CODE MODULATION[11]). In figura è evidenziato inoltre un blocco di sincronismo necessario a ricostruire la corretta sequenza ricevuta, in modo da redistribuire correttamente i campioni ai filtri di restituzione.




La struttura temporale ripetitiva che ospita i campioni dei singoli tributari prende il nome di *trama* (FRAME[12]), ed è composta 32 intervalli detti *time-slot*. Trenta di questi ospitano a turno i bit di un campione proveniente da un numero massimo di 30 tributari[13], mentre i rimanenti due intervalli convogliano le informazioni di segnalazione[14], che indicano lo stato dei singoli collegamenti (il $16^o$ intervallo) e forniscono il sincronismo relativo all'inizio della trama stessa (il primo). La velocità binaria complessiva risulta quindi di 32 intervalli * 8 bit/intervallo * 8000 campioni/secondo = 2048000 bit per secondo; per questo motivo, all'insieme ci si riferisce come alla *trama PCM a 2 Mbit*. La durata della trama deve essere invece la stessa del periodo di campionamento, ossia 1/8000 = 125 $\mu$sec.

Il primo time-slot della trama contiene una configurazione di bit sempre uguale, chiamata FLAG (*bandiera*), che ha lo scopo di indicare ai circuiti di sincronismo l'inizio della trama stessa. I dati di segnalazione contenuti nel $16^o$ intervallo devono essere *diluiti* su più trame, per poter rappresentare tutti i 30 tribu-

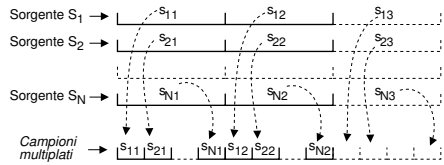



[10] La tecnica di multiplare un *blocco di bit* (in questo caso 8) alla volta prende il nome di *word interleaving*, distinto dal *bit interleaving*, in cui l'alternanza è a livello di bit.

[11] Il segnale PCM ispira il suo nome dal PAM (vedi § 6.9.5) in quanto ora, anzichè trasmettere le *ampiezze* degli impulsi, si inviano i *codici* binari dei livelli di quantizzazione.

[12] frame SIGNIFICA PIÙ PROPRIAMENTE TELAIO, E IN QUESTO CASO HA IL SENSO DI INDIVIDUARE UNA STRUTTURA, DA "RIEMPIRE" CON IL MESSAGGIO INFORMATIVO.

[13] 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).

[14] Vedi anche la sezione 6.3.2.

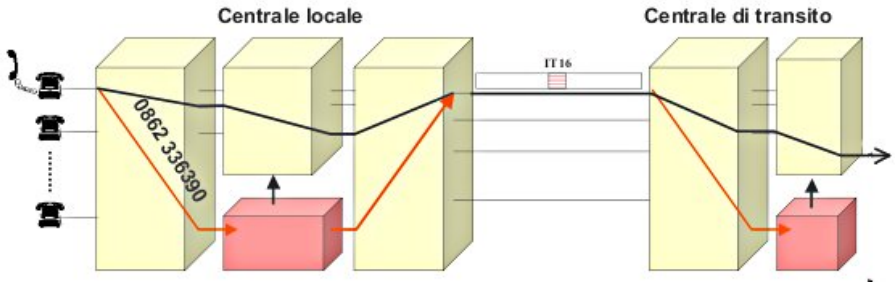
tari[15]. Si è stabilito che occorra prelevare il $16^o$ intervallo di 16 trame successive, per ricostruire una stuttura detta *supertrama* (di 16*8 = 128 bit) che rappresenta le informazioni di tutti i tributari (disponendo così di 4 bit/tributario/supertrama), e che si ripete ogni 16*125 = 2000 $\mu$sec = 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".

### 6.3.2 Messaggi di segnalazione

Come illustrato al § 6.9.1, la rete di accesso è sede di uno scambio di informazioni tra terminale e centrale locale, detta *segnalazione di utente*, e che ha lo scopo di indicare la disponibilità della rete, il numero chiamato, l'attivazione della suoneria, ed i messaggi a ritroso di libero/occupato. Queste informazioni, quando devono essere propagate verso il lato-rete della centrale di accesso, possono essere gestite secondo due diversi approcci.

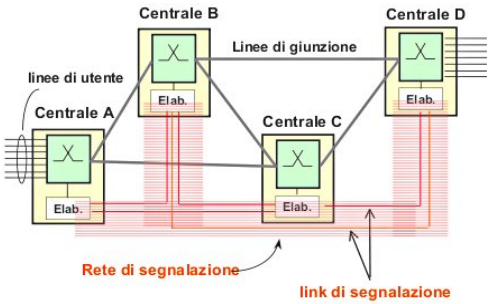
**Segnalazione associata al canale** In questo caso la centrale di accesso inserisce le informazioni di segnalazione relative ad un tributario all'interno della *supertrama* di segnalazione ottenuta collezionando i valori presenti nel $16^o$ time-slot. Questa modalità viene indicata come CAS (*Channel 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. Con la numerizzazione, si è inizialmente scelto di mantenere la segnalazione *associata* al segnale vocale, con la contropartita che quando, nell'attraversare una centrale di transito, una comunicazione è commutata su di una diversa linea di uscita, deve essere commutata anche la segnalazione associata.
La figura seguente mostra come la numerazione venga recepita da un organo di *controllo centrale*, che provvede a impostare il dispositivo di commutazione (§ 6.8), in modo che la comunicazione sia instradata verso la linea di uscita in direzione della destinazione. Quindi, l'informazione di segnalazione viene ri-associata nell'intervallo 16.




[15]Gli 8 bit del $16^o$ intervallo sono infatti insufficienti a codificare lo stato dei 30 tributari che contribuiscono al segnale TDM.

**Segnalazione a canale comune** Il primo passo evolutivo è stato quindi quello di provvedere ad un *canale comune* di segnalazione direttamente collegato agli organi di controllo, su cui poter convogliare la segnalazione relativa a tutte le comunicazioni in transito tra le due centrali.

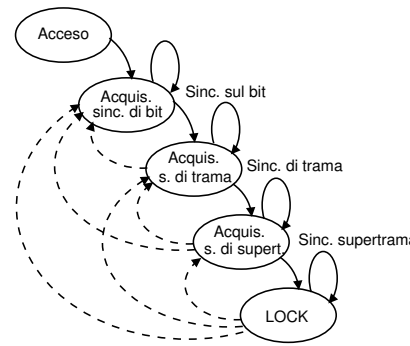

I messaggi di segnalazione, per loro natura, devono essere trasmessi solo quando si verificano degli eventi significativi, e per questo motivo sono ora inviati mediante dei *pacchetti dati*. Il passo successivo è quindi stato quello di realizzare una intera rete a *commutazione di pacchetto*, parallela a quella di transito su cui viaggiano (in modalità a circuito) le conversazioni vocali.




In tal modo, gli organi di controllo delle centrali sono in comunicazione diretta tra loro, secondo la modalità cosiddetta CCS (*Common Channel Signaling*), mediante una rete a pacchetto dedicata alla segnalazione, sulla quale viaggiano i messaggi definiti da un apposito *sistema di segnalazione* (vedi § 6.9.3), e che permette di centralizzare il controllo e la configurazione di tutte le centrali coinvolte nell'instradamento di una stessa comunicazione, rendendo così possibile la disponibilità di servizi come il trasferimento di chiamata, la conversazione a tre, l'avviso di chiamata....

### 6.3.3 Sincronizzazione di centrale

Nella figura a lato sono mostrati i diversi stati attraverso cui deve evolvere il dispositivo di sincronizzazione che opera sui flussi PCM CAS, prima di entrare nello stato di LOCK (*aggancio*) ed iniziare a poter leggere e smistare i contenuti dei diversi time-slot. Occorre infatti acquisire innanzitutto il sincronismo sul bit, sfruttando le caratteristiche del codice di linea utilizzato; quindi si sfrutta la conoscenza della configurazione scelta per il flag di inizio trama, per individuare da dove iniziare a conteggiare gli



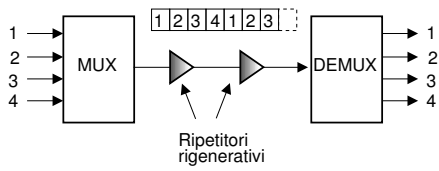
intervalli temporali. Infine, viene individuato l'inizio della supertrama, grazie ad un ulteriore configurazione prefissata, posta all'inizio della stessa. In ogni stato poi, esiste la possibilità (fortunatamente remota) di perdere il sincronismo ed *indietreggiare* (linee tratteggiate) nel diagramma di stato, perdendo le comunicazioni in corso.

### 6.3.4 Multiplazione asincrona e PDH

L'argomento di questo paragrafo non va confuso con la *trasmissione* asincrona (quella START-STOP mostrata al § 5.5.1), e che descrive una modalità di *inviare* informazioni numeriche; qui invece si tratta di *multiplare*, ossia come *mettere assieme* più comunicazioni.

Via via che la rete di trasporto è interconnessa mediante centrali di livello gerarchico superiore, associate ad aree di influenza geografica più estesa (vedi § 6.5), i collegamenti di giunzione trasportano un numero di tributari sempre più elevato, ottenuti raggruppando assieme tutte le conversazioni contemporaneamente dirette verso la stessa destinazione. Considerando allo stesso tempo le problematiche legate al dover svolgere nelle centrali la funzione di commutazione, ci si pone il problema di individuare dei metodi efficienti di raggruppare assieme più tributari, anche a velocità diverse, facendo in modo che l'operazione di inserimento/rimozione di un singolo tributario sia relativamente agevole. Rimandiamo al § 6.4 l'analisi di come avvenga il processo di multiplazione nel caso in cui esista una perfetta sincronizzazione tra gli elementi della rete, e trattiamo nel seguito il caso della rete *plesiocrona*.

Nella trama PCM (§ 6.3.1), tutti i 30 canali sono campionati congiuntamente, e piú flussi a 2 Mbit possono a loro volta essere "messi assieme" in modalità *bit interleaved* (prendendo un bit alla volta da ogni tributario) da appositi dispositivi *multiplatori* (o MULTIPLEXER, o MUX). Il collegamento può prevedere piú dispositivi detti *ripetitori rigenerativi*, che oltre ad amplificare il segnale, lo "puliscono" dal rumore accumulato, decodificando i dati in ingresso per poi generare ex-novo il segnale numerico.
Il problema con questo modo di procedere è che i singoli tributari possono ragionevolmente avere origine da centrali differenti, ognuno con un proprio orologio indipendente, e quindi le velocità possono essere lievemente differenti l'una dall'altra[16], 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[17], nota come *Plesiochronous Digital Hierarcy* (PDH), secondo la quale ad esempio 4 flussi da 2 Mbps sono multiplati in uno da 8 Mb/sec: notiamo che sebbene siano teoricamente sufficienti 8192 Mb/sec, in realtà il Multiplexer ne produce di piú (8448). Questo avviene proprio per permettere

[16] 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.

[17] *Comité Consultif International pour la Telephonie et Telegraphie.* Questo organismo non esiste più. ed ora l'ente di standardizzazione ha nome ITU-T.

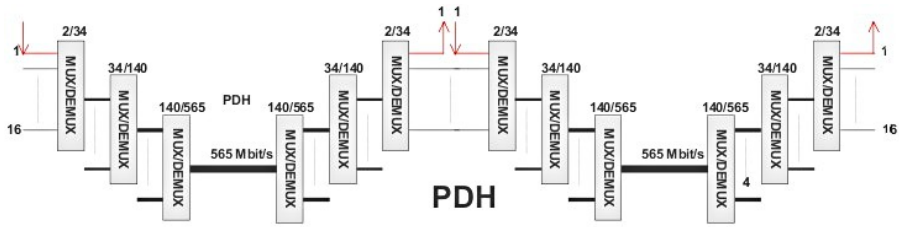
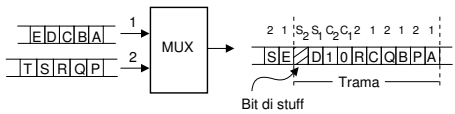



Figura 6.1: Gerarchia di multiplazione PDH e complessità di un ADM

la trasmissione di segnali non necessariamente sincroni, mediante la tecnica del *Bit Stuffing*[18].

#### 6.3.4.1 Bit stuffing

Consideriamo 2 tributari i cui bit vengono inseriti alternativamente in una trama da 4 bit/canale; il secondo risulta lievemente piú lento. I primi 3 + 3 bit vengono trasmessi comunque, mentre il $4^o$ può essere trasmesso o meno, a seconda se i tributari lo abbiano pronto. Per ottenere questo risultato, i bit $C_1$ e $C_2$ (di controllo) valgono 0 oppure 1 a seconda se l'intervallo seguente ($S_1$ e $S_2$) contenga un dato valido oppure sia solo un *bit di stuff*, cioè vuoto, in quanto il tributario corrispondente è piú 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.

#### 6.3.4.2 Add and Drop Multiplexer - ADM

La modalità *bit interleaved* con cui è realizzata la gerarchia PDH è particolarmente problematica qualora di desideri estrarre e/o introdurre un singolo tributario da/in un segnale 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. 6.1) demultiplare l'intero flusso, compresi tutti gli altri tributari, e successivamente ri-multiplare di nuovo il tutto.

Questa caratteristica limita notevolmente la flessibilità nelle configurazioni di rete che si possono ottenere con questa tecnologia, e per i tributari passanti comporta l'aggiunta di un tempo di ritardo addizionale dovuto alle operazioni di demultiplazione e 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.

[18] Da: TO STUFF = *riempire*.

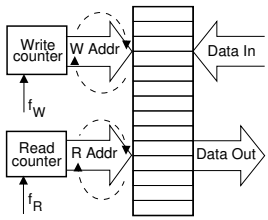
### 6.3.5 Sincronizzazione di rete

Se tutti i nodi della rete operassero alla stessa velocità, non sussisterebbero problemi nella multiplazione di piú tributari. Nel caso in cui la sincronizzazione tra nodi sia completamente affidata ad un orologio di centrale di elevata precisione, si verifica il caso di funzionamento *plesiocrono*, che è quello prescritto per le centrali che interconnettono le reti di due diverse nazioni, o di due diversi operatori di telecomunicazioni. Ma questa non è l'unica soluzione.

Una alternativa è la sincronizzazione *mutua* tra centrali, in cui ognuna di queste emette dati in uscita ad una frequenza pari alla media delle frequenze dei dati in ingresso. A parte fenomeni transitori durante i quali la rete è soggetta ad oscillazioni di velocità, relativi all'inserimento od alla disattivazione di centrali "topologicamente importanti", il metodo funziona ragionevolmente bene. Una seconda soluzione è una sincronizzazione di rete di tipo *gerarchico* in cui le centrali ricevono informazioni di sincronismo da soggetti "piú importanti", come per configurazioni *Master-Slave* in cui il Master è una centrale ad elevata precisione, od un riferimento in comune come ad esempio un satellite in orbita terrestre.

#### 6.3.5.1 Elastic Store

Si tratta di un accorgimento[19] idoneo ad *assorbire* le fluttuazioni della velocità di trasmissione, come ad esempio nel caso della sincronizzazione mutua. Mentre il *bit stuffing* (§ 6.3.4.1) è adottato nella 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 (§ 6.8).




E' realizzato mediante un banco di memoria (di dimensione pari ad una trama), riempito (ciciclicamente) 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[20] appunto a velocità $f_W$, e che torna a puntare all'inizio della memoria una volta raggiunto l'indirizzo più elevato. Un secondo puntatore READ viene utilizzato per leggere la memoria, alla velocità $f_R$ richiesta, e prelevare i dati da inviare in uscita: se $f_R$ e $f_W$ sono differenti, READ e WRITE prima o poi si sovrappongono, causando la perdita o la ripetizione di una intera trama, e nulla piú[21].

## 6.4 Gerarchia digitale sincrona

**Definizione dei livelli della gerarchia** Come anticipato, la *Synchronous Digital Hierarcy* (SDH[22]) è una metodologia di multiplazione che presuppone un funzionamento perfettamente sincrono degli elementi di rete, ed ha solo una variante (nel Nord

[19] Letteralmente: magazzino elastico.

[20] 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.

[21] 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.

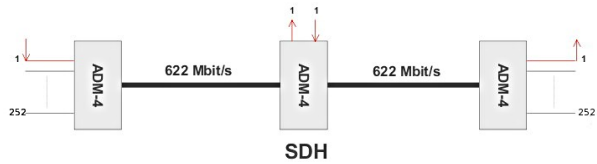
[22] `http://it.wikipedia.org/wiki/Synchronous_Digital_Hierarchy`

| SONET | SDH | payload (kbps) | v. trasm. (kbps) |
|---|---|---|---|
| STS-1 | - | 48.960 | 51.840 |
| STS-3 | STM-1 | 150.336 | 155.520 |
| STS-12 | STM-4 | 601.344 | 622.080 |
| STS-24 | STM-8 | 1,202.688 | 1.244.160 |
| STS-48 | STM-16 | 2,405.376 | 2.488.320 |
| STS-96 | STM-32 | 4,810.752 | 4.976.640 |
| STS-192 | STM-64 | 9,621.504 | 9.953.280 |
| STS-768 | STM-256 | 38,486.016 | 39.813.120 |
| STS-1536 | STM-512 | 76,972.032 | 79.626.120 |

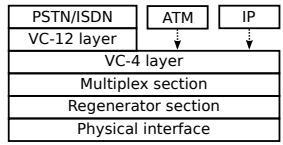
Tabella 6.1: Nomenclatura della gerarchia ottica e relative velocità

America), denominata SONET (*Synchronous Optical Network*), i cui livelli sono siglati STS oppure OC nel caso in cui ci si riferisca al segnale ottico corrispondente, e che interopera abbastanza bene con SDH. La tabella 6.1 elenca le velocità del *payload*[23] e di trasmissione associate ai diversi livelli della gerarchia di multiplazione SDH/SONET.

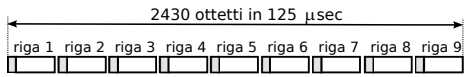
**Multiplexer Add and Drop** La differenza strutturale rispetto a PDH, è che in SDH i tributari usano tutti lo stesso clock, da cui deriva la possibilità di aggiungere e togliere un singolo tributario senza alterare il flusso in cui è immerso, come esemplificato in figura, in cui 252 flussi PDH E1 concorrono a formare un multiplex STM-4.




**Eterogeneità del trasporto**
SDH nasce allo scopo di consentire il trasporto di dati di diversa origine (PCM telefonico, ISDN, pacchetti Ethernet ed IP, celle ATM), come illustrato nella figura a fianco, che rappresenta impilate le diverse elaborazioni che i tributari devono subire per essere immessi nel flusso SDH.

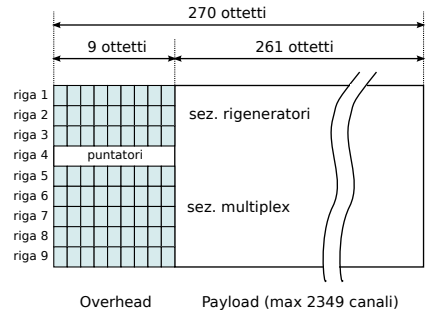



**Struttura di trama**
SDH si basa su di una struttura di trama di durata di 125 $\mu$sec, durante i quali sono trasmessi in modalità *byte interleaved* una sequenza di ottetti provenienti da diversi tributari a 64 kbps che condividono la medesima sorgente di temporizzazione, cosicché ogni tributario può essere inserito o prelevato semplicemente scrivendo o leggendo sempre nello stesso punto (con la stessa fase) un ottetto ogni trama.

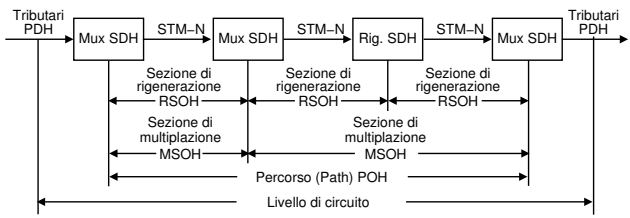


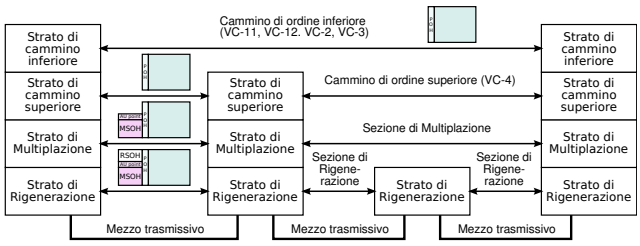

[23] Con il termine *payload* si indica il *carico pagante*, ossia i dati che vengono trasportati

**Synchronous Transport Module STM-1** Il livello più basso della gerarchia è indicato come STM-1 (*Synchronous Transport Module*), che opera ad una velocità di 155.52 Mbit/s, e può trasportare 63 flussi PDH E1, ovvero 2016 canali PCM. STM-1 è caratterizzato infatti da una trama composta da 63 flussi E1 x 32 timeslot/flusso = 2430 ottetti, di cui 81 di segnalazione e 2349 di dati[24], 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 ottetti). 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*. Infatti, il percorso (*Path*) compiuto da un singolo tributario, si snoda 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, è definita una nuova *sezione di rigenerazione*. Per ognuna di queste sezioni, è definito un Overhead (OH) specifico per le operazioni OAM associate.



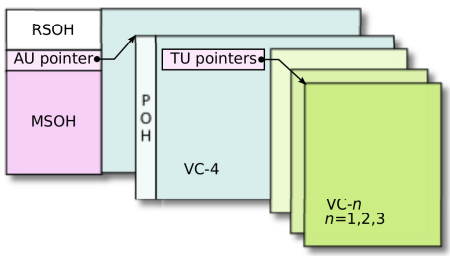




Dato che un ADM è anche rigeneratore, e che i dispositivi di ingresso - uscita del tributario sono anche ADM, si determina la *stratificazione funzionale* per la segnalazione raffigurata a fianco, in cui è evidenziato come l'Overhead associato alle sezioni più esterne venga *impilato* su quello delle sezioni interne. Ma a differenza dell'incapsulamento (pag. 8.5.2.3) proprio

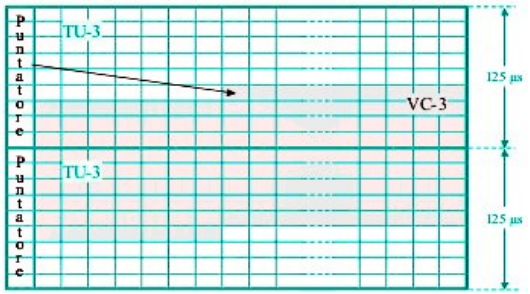
[24] Notiamo che la differenza tra i 2349 ottetti di payload ed i 2016 canali voce fornisce 2349 - 2016 = 333 ottetti, che suddivisi per le nove righe, danno luogo a 37 ottetti per riga *in più*.

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 seguente.

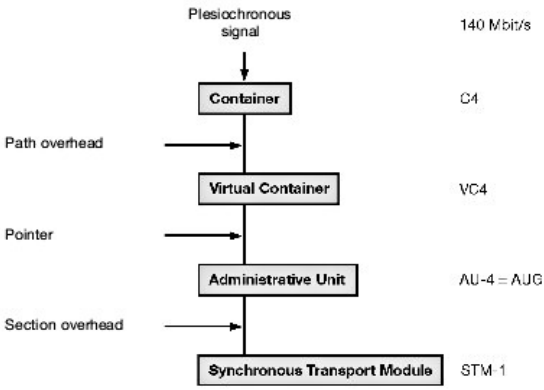
**Il puntatore all'unità amministrativa** Nelle prime tre righe dell'overhead della trama STM-1 trova posto l'RSOH, che viene scritto dai dispositivi di rigenerazione, e quindi letto e ri-scritto ad ogni rigeneratore successivo; in particolare, alla prima riga sono presenti i flag che consentono di acquisire 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 di questo puntatore deriva dalla volontà di ridurre al minimo l'uso di buffer e l'introduzione di ritardi di consegna; pertanto i dati da trasmettere *non* vengono inseriti nella struttura di trama all'inizio della stessa, bensì *al primo ottetto possibile* al momento della disponibilità dei dati stessi. Quindi, è più che normale il caso in cui la AU inizi a metà di una trama, e termini a metà della trama successiva, come illustrato nella figura che a lato. La coppia AU ed AU Pointer prende quindi il nome di *Administrative Unit Group* (AUG).

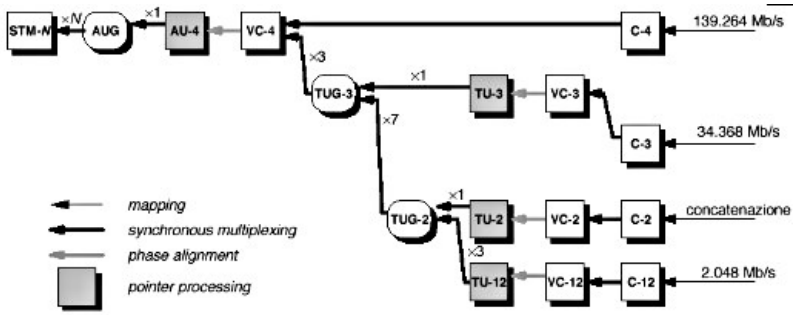



**Virtual Container e Tributary Unit** Il riempimento della AU con i dati da trasmettere, avviene (vedi figura a lato) seguendo una serie di passi successivi, che vedono prima la creazione di una struttura dati detta *Container*, a cui si aggiunge il POH per ottenere un *Virtual Container*, da cui dopo l'aggiunta del puntatore deriva la AU. Notiamo ora che non necessariamente la AU deve essere riempita da un unico tributario; al



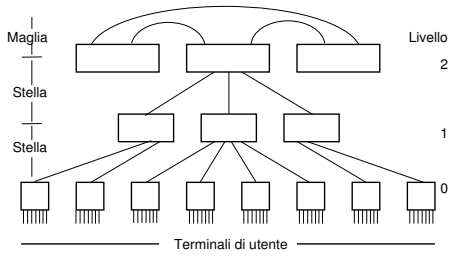
contrario, la multiplazione serve appunto ad ospitarne diversi!! A questo scopo, più VC a bassa velocità possono essere a loro volta multiplati in modalità *byte interleaved*, per produrre una struttura dati intermedia indicata TU (*Tributary Unit*), che a sua volta può essere inserita assieme ad altre TU, all'interno del VC di ordine superiore.

Non approfondiamo oltre questo argomento, che richiede una buona dose di pazienza per essere analizzato a fondo, e ci limitiamo ad inserire un diagramma che mostra le possibilità di combinazione di tributari differenti, in accordo alle specifiche di ETSI.




## 6.5 Topologia di rete

Nella figura seguente è riportata una possibile topologia di rete, a scopo puramente esemplificativo, dato che è un elemento su cui gli operatori godono di massima libertà.




Il nostro esempio si riferisce ad una rete a 2 livelli, in cui i terminali di utente che fanno capo ad una stessa centrale locale di livello zero accedono alla rete per mezzo di un collegamento che multipla le comunicazioni provenienti dalla stessa centrale.

Come vedremo al capitolo 8, il numero di collegamenti contemporaneamente possibili è inferiore al numero di terminali di utente, ed è dimensionato applicando i metodi della teoria del traffico al caso in questione. Le centrali locali sono connesse *a stella* a quelle di livello 1, che quindi realizzano instradamenti relativi alle comunicazioni tra utenti geograficamente vicini, ma connessi a centrali locali diverse. Se, al contrario, due utenti sono connessi alla stessa centrale di livello zero, l'instradamento non esce dalla centrale locale.

Le centrali di livello 1 sono a loro volta connesse a stella a quelle di livello 2, che gestiscono il traffico a livello nazionale; alcune di queste poi, consentono di instradare anche i collegamenti verso altre reti (ad es. di altre nazioni od operatori). Le centrali di $2^o$ 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, la denominazione attribuita alle centrali dei diversi livelli segue lo schema mostrato nella parte sinistra della figura 6.2: i terminali di utente sono attestati presso gli *Stadi di Linea* (SL) tramite la rete di accesso, mentre gli SL

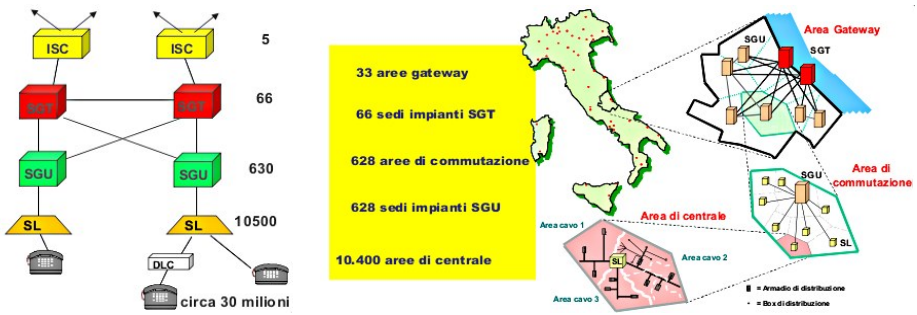



Figura 6.2: Topologia della rete telefonica Italiana

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 Abbruzzo, individuando la ripartizione del territorio, e mostrando come ad un livello inferiore agli stadi di linea, la rete di accesso si dirami ulteriormente attraverso gli armadi ed i box di distribuzione.

## 6.6 Rete in fibra ottica

Nel periodo iniziale le fibre ottiche sono state usate prevalentemente nella rete di trasporto tra centrali di grado gerarchicamente elevato, mentre ora trovano impiego anche nella sezione di accesso. Per ciò che riguarda le modalità di trasmissione ottica, si rimanda al § 15.4; nel seguito illustriamo i dispositivi utilizzati, la topologia risultante, ed i sistemi di protezione.

### 6.6.1 Dispositivi SDH

Come anticipato, la trasmissione SDH si avvale di elementi (vedi Fig. 6.3) che possono essere descritti in termini funzionali secondo la seguente classificazione:

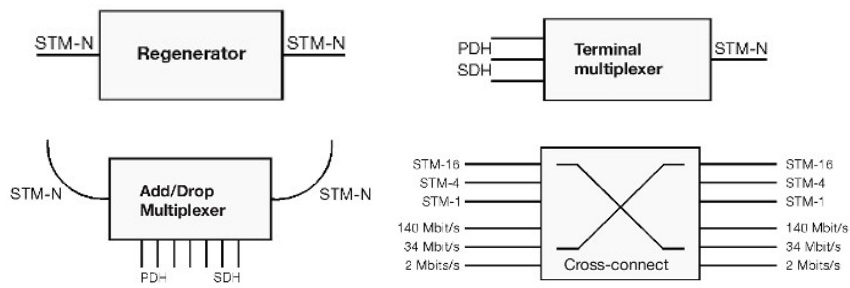



Figura 6.3: Dispositivi SDH

**Rigeneratori** Sono gli elementi di base, che consentono di suddividere su più tratte i collegamenti più lunghi, e che eliminano dal segnale in transito gli effetti del rumore e della dispersione temporale.

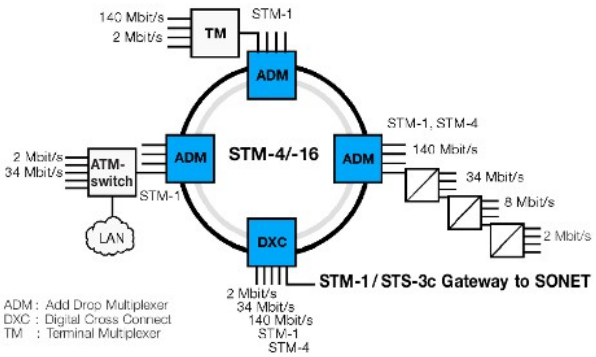
**Multiplatori** Combinano tributari PDH ed SDH, in modo da inserirli in flussi a velocità più elevate.

**Multiplatori Add and Drop** Permettono l'inserimento e l'estrazione di tributari a bassa velocità in/da un flusso in transito, e consentono la creazione di strutture ad anello.

**Digital Cross Connect** A differenza di un ADM, un DXC è interconnesso a più di un flusso SDH, e quindi può inserire un tributario (od un container) prelevato da un flusso entrante, all'interno di un diverso flusso uscente, realizzando così la funzione di commutazione.

### 6.6.2 Topologia ad anello

Le reti in fibra ottica sono quasi sempre realizzate mediante degli *anelli* che congiungono tra loro i nodi di commutazione in forma ciclica. Alcuni di questi nodi (DXC, ovvero *Digital Cross Connect*) sono interconnessi a più di un anello, e svolgono la funzione di commutazione delle comunicazioni che devono essere inoltrate verso gli altri anelli.




#### 6.6.2.1 Rete di trasporto

Al 2002, l'interconnessione dei collegamenti SDH nazionali risultava permessa da struttura su tre livelli riportata in Fig. 6.4.

#### 6.6.2.2 Rete di accesso in fibra

La capacità del trasporto SDH di accettare tributari di tipo Ethernet o IP, facilita la realizzazione di una rete completamente ottica, anche nella sezione di accesso. La fig. 6.5 mostra alcuni casi pratici di accesso in fibra ottica. Iniziando da destra, sono mostrate delle reti Gigabit Ethernet (pag. 197) residenziali, interconnesse mediante *switch di livello 2* ad un PoP (*Point of Presence*), il cui Router si interconnette ad un anello SDH a 655 Mbps, sul quale sono instradati i pacchetti IP diretti verso Internet, per il tramite del PoP primario. In basso a sinistra, sono mostrati accessi a due Megabit, contenenti sia traffico voce che dati, che vengono inseriti in anelli SDH da 155 Mbps: quello al centro inoltra i canali voce verso la PSTN, mentre quello di sinistra si interconnette nel *backbone* IP da 2.5 Gbps.

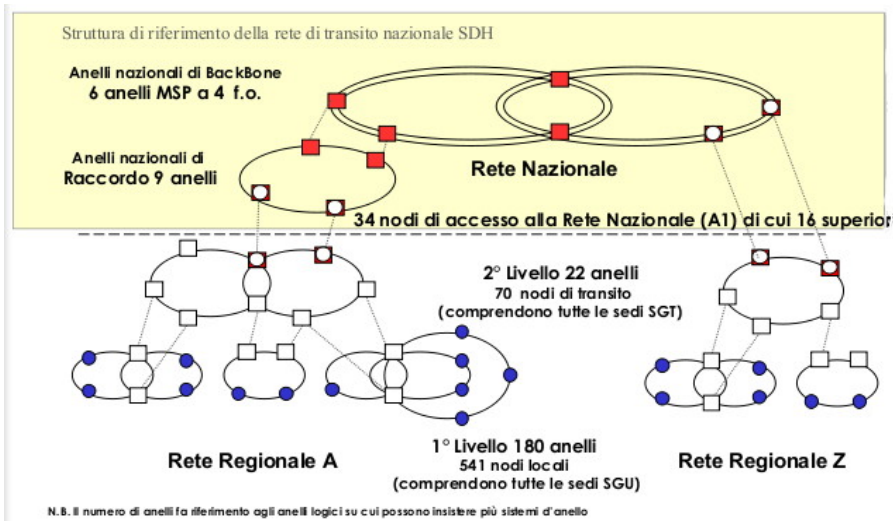



Figura 6.4: Struttura collegamenti SDH nazionali - anno 2002

### 6.6.3 Sistemi di protezione automatica

L'abbondanza di ottetti OAM nella multiplazione SDH permette un monitoraggio costante della qualità del collegameno e di eventuali malfunzionamenti, al punto che gli stessi apparati di commutazione sono ora in grado di svolgere compiti di rimpiazzo

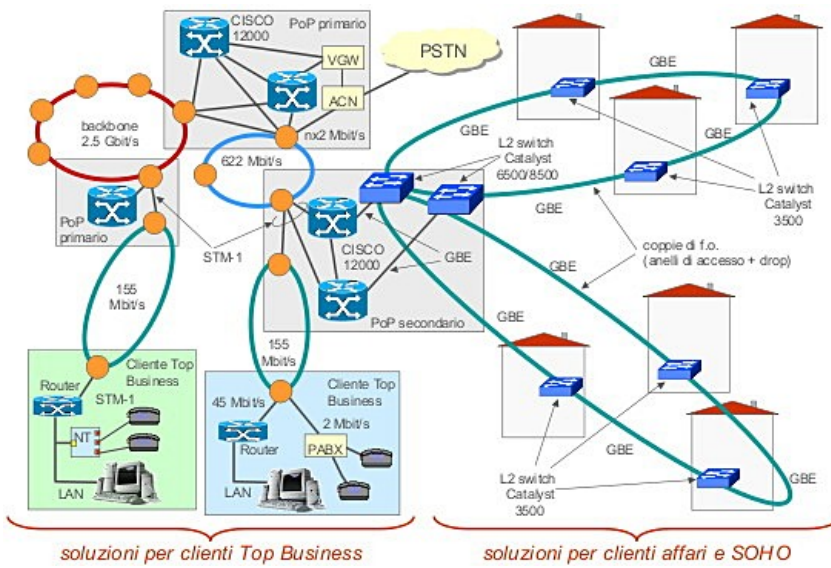



Figura 6.5: Diffusione della fibra ottica nella rete di accesso

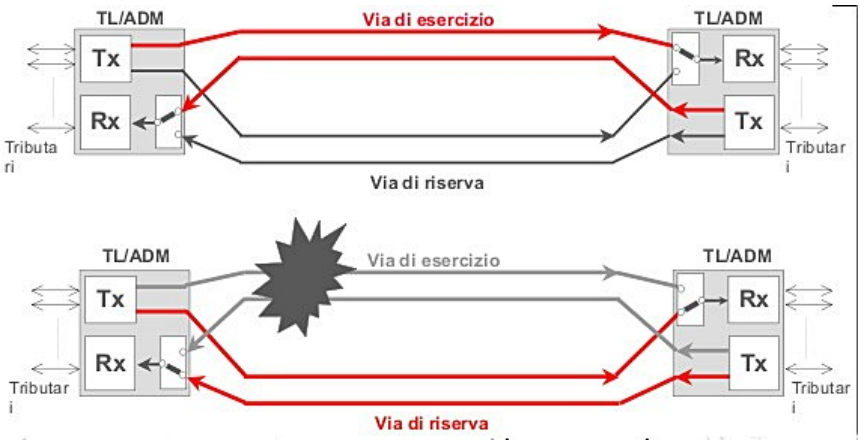



Figura 6.6: Sistema di protezione 1+1

automatico tra la linea andata fuori servizio, ed una riserva presente, come indicato nei seguenti schemi.

**Protezione 1+1** In questo caso, ogni collegamento (vedi fig. 6.6) è provvisto di un collegamento di riserva. Qualora la via di esercizio vada fuori servizio, i terminali di linea che sono posti agli estremi se ne avvedono pressochè immediatamente, e provvedono a commutare la comunicazione sulla via di riserva.

**Collegamento ad anello** Nel caso di collegamenti unidirezionali, la via di ritorno (vedi fig. 6.7) si sviluppa investendo l'altra metà della circonferenza, percorsa nello stesso senso di rotazione. Aggiungendo un secondo anello di riserva, anch'esso unidirezionale ma diretto in senso opposto al primo, la comunicazione può continuare anche nel caso in cui entrambi i collegamenti (generalmente co-locati) tra due nodi vadano fuori servizio.

## 6.7 Instradamento

Per questo argomento sono fornite solo alcune definizioni estremamente sommarie di tre possibili strategie, adottate nel corso della evoluzione delle reti telefoniche:

**END to END** o *Right-through* (da estremo ad estremo): la scelta del percorso è effettuata dalla centrale di origine, ad esempio in base al prefisso od all'inizio del numero, utilizzando delle *tabelle di routing* statiche. E' la modalità dell'inizio della telefonia, in cui i commutatori erano elettromeccanici, ed i collegamenti verso altre centrali erano *cablati*. Ha l'enorme svantaggio che i cambiamenti alla topologia della rete si devono riflettere in cambiamenti di tutte le tabelle - *o dei morsetti!*

**Link-by-link** o *Own-Exchange* (tratta per tratta): ogni centrale decide in autonomia dove instradare (in uscita) le connessioni entranti, in base a sue tabelle dinamiche, ovvero informazioni che giungono dalla rete stessa. Si adatta alle modifiche

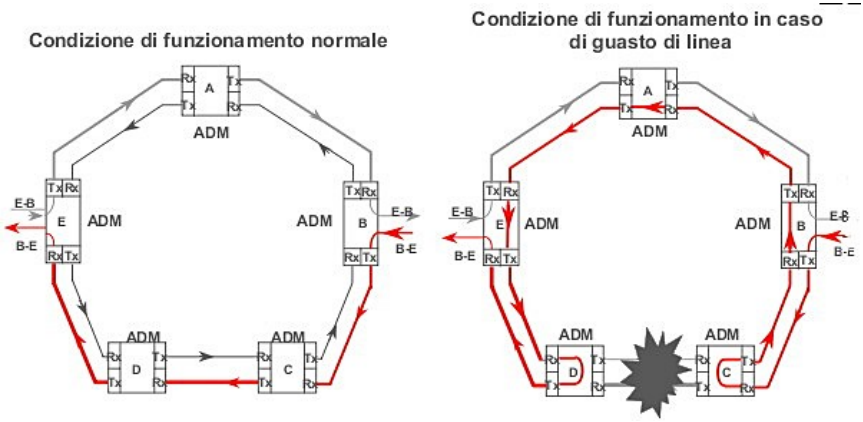



Figura 6.7: Configurazione ridondante ad anello

della topologia ma non è affidabile al 100 %, potendo ad esempio produrre dei *loop* (o circoli viziosi);

**Tramite CCS** (*Common Channel Signaling*, segnalazione a canale comune): le decisioni sull'instradamento sono demandate ad una rete di segnalazione parallela ed indipendente da quella del traffico smaltito, e che collega tutte le centrali ad un unico organo di controllo (il *canale comune*), il quale determina l'instradamento in base alla sua conoscenza dello stato del traffico nella rete, e comunica contemporaneamente a tutte le centrali coinvolte nell'instradamento, come configurare i propri organi di commutazione per realizzare il collegamento richiesto.

## 6.8 Commutazione

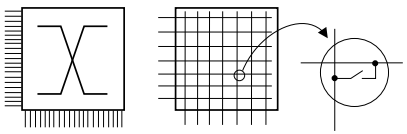
Illustriamo ora l'architettura di dispositivi che consentono la cosiddetta *commutazione di circuito*, ovvero la creazione di un collegamento *stabile* tra due porte del commutatore, con un *impegno permanente* di *risorse fisiche* per tutta la durata del collegamento. Un'altra modalità di commutazione, quella di *pacchetto*, sarà illustrata al Capitolo 8.

### 6.8.1 Reti a divisione di spazio

Sono chiamati cosí gli organi di commutazione che realizzano un collegamento fisico (elettrico) tra uno degli $N$ ingressi ed una delle $M$ uscite. Nel caso in cui $N > M$, la rete è un *concentratore*[25] , mentre se $N < M$ la rete è un espansore; se $N = M$ la rete è quadrata e *non bloccante*.

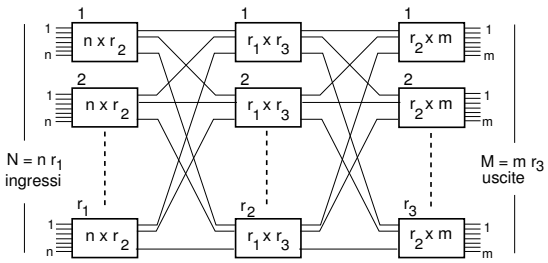
Il commutatore è rappresentato da un blocco con una "X" (in inglese *cross*, od incrocio), e può essere pensata come una matrice binaria in cui ogni elemento (1 o 0) rappresenta lo stato (chiuso od aperto) di un interruttore (realizzato ad esempio mediante un transistor) che collega una linea di ingresso ad una di uscita.

[25] 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*.



Realizzare in questo modo una rete non bloccante prevede l'uso di un numero di interruttori pari ad $N \cdot M$, dei quali solo $\min(N, M)$ sono utilizzati, anche nelle condizioni di massimo carico. Inoltre, nessun interruttore può essere "eliminato" senza precludere irrimediabilmente la possibilità di collegare qualunque ingresso a qualunque uscita. Allo scopo di utilizzare un numero ridotto di interruttori, sia per costruire reti non bloccanti oppure bloccanti con bassa probabilità di blocco, si ricorre alle...

### 6.8.2 Reti multistadio




...di cui in figura è riportato un esempio a 3 stadi, in cui gli $N$ ingressi sono ripartiti su $r_1$ reti piú 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[26]). Una connessione da sinistra a destra ha ora la possibilità di scegliere attraverso quale matrice intermedia passare.

Nel caso di reti quadrate ($N = M$), ponendo $n = m = \sqrt{\frac{N}{2}}$, si ottiene un numero complessivo di incroci pari a $4\left(\sqrt{2}N^{\frac{3}{2}} - N\right)$, che risulta inferiore ad $N^2$ (e dunque vantaggioso rispetto ad un commutatore monostadio) a partire da $N \geq 24$.
Ovviamente, la problematica relativa alle matrici di commutazione è molto articolata, coinvolgendo topologie piú complesse, filosofie di instradamento, e tecniche per la stima delle probabilità di blocco. Tralasciamo ulteriori approfondimenti, per illustrare invece come realizzare dispositivi di commutazione per trasmissioni numeriche *a divisione di tempo*.

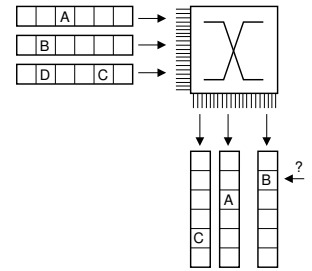
### 6.8.3 Commutazione numerica a divisione di tempo

Consideriamo il caso in cui si debbano commutare le comunicazioni associate ai singoli *time-slot* presenti in diversi flussi[27] numerici organizzati in trame. Avendo a dispo-

[26] E' una condizione *sufficiente* a scongiurare il blocco anche nella condizione *peggiore*. Tale circostanza si verifica quando:
▷ una matrice del primo stadio ($i$) ha $n - 1$ terminazioni occupate
▷ una matrice del terzo stadio ($j$) ha $m - 1$ terminazioni occupate $e$
▷ tali terminazioni non sono connesse tra loro, anzi le connessioni associate impegnano ognuna una diversa matrice intermedia $e$
▷ si richiede la connessione tra le ultime due terminazioni libere di $i$ e $j$
$\Rightarrow$ in totale si impegnano allora $m - 1 + n - 1 + 1 = m + n - 1$ matrici intermedie.

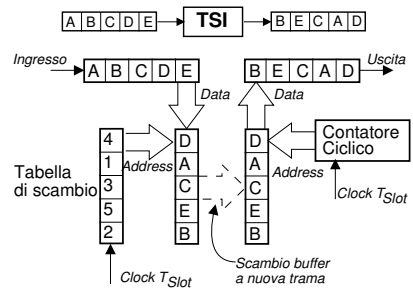
[27] Le comunicazioni presenti in uno stesso flusso, ovvero appartenenti alla stessa trama, condividono la stessa origine/destinazione.

sizione 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; pertanto, non può (ad esempio) inviare le conversazioni $B$ e $D$ sulla stessa linea uscente, in quanto si verifica un conflitto temporale. E' quindi evidente la necessità di introdurre uno stadio di *commutazione temporale*.



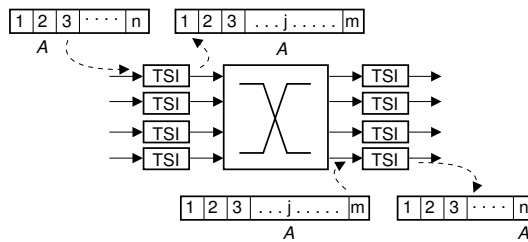

#### 6.8.3.1 Time Slot Interchanger

Questo dispositivo è indicato come TSI *(time slot interchanger)* ed ha la funzione di produrre in uscita una sequenza di dati identica a quella in ingresso, tranne per averne cambiato l'ordine temporale. In figura è mostrato un possibile schema di funzionamento: una trama entrante viene scritta, agli indirizzi ottenuti leggendo sequenzialmente la tabella di scambio, in un buffer di memoria (*es.: entra* E *e lo scrivo al* $4^o$ *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[28], che a sua volta viene letto con ordine sequenziale (partendo dall'alto), per creare la nuova trama in uscita. Ovviamente, è possibile anche la realizzazione opposta, con scrittura sequenziale e lettura secondo il nuovo ordinamento.




#### 6.8.3.2 Commutazione bidimensionale

Cosí come un commutatore spaziale non è sufficiente, anche un TSI "da solo" è di scarsa utilità, non potendo instradare le comunicazioni su vie diverse. Combinando assieme le due funzioni, si giunge a realizzare commutatori sia di tempo che di spazio, come la struttura a 3 stadi in figura, chiamata "TST" perché alterna uno stadio temporale, uno spaziale ed uno temporale.




[28] La tecnica prende il nome di *Double Buffering*.

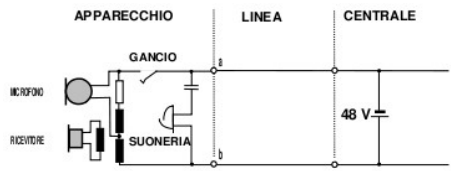
Notiamo subito che, in questo schema, il numero di intervalli temporali in uscita dai TSI di ingresso è $m > n$ ([29]): ciò determina, per lo stadio spaziale, una frequenza di commutazione piú elevata della frequenza dei time-slot in ingresso. Una generica conversazione "A" che occupa il $2^o$ slot del 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 $m = 2n - 1$([30]).

Analizziamo i vantaggi conseguiti dalla commutazione numerica con un semplice esempio. Poniamo di voler commutare con lo schema illustrato 4 flussi PCM (con $n = 30$): i 4 * 30 = 120 canali presenti sono commutati utilizzando solo 4 * 4 = 16 interruttori, contro i 120 * 120 = 14.400 interruttori necessari ad una matrice spaziale monostadio che svolga la commutazione dei 120 canali analogici !

## 6.9 Appendici

### 6.9.1 Plain Old Telephone Service

Il *buon vecchio servizio telefonico* consiste nel collegamento audio, nella banda del canale telefonico, attuato mediante un *terminale di utente* (telefono), e nella *segnalazione* (sempre *di utente*) necessaria ad instaurare il collegamento. L'insieme degli apparati che permette di interconnettere tra loro i telefoni di rete fissa è spesso indicato con l'acronimo PSTN (*Public Switched Telephone Network*), da cui si sono evoluti tutti gli sviluppi successivi delle telecomunicazioni.




Quando la centrale locale deve far squillare il telefono, invia sul doppino una tensione alternata che ne attiva la suoneria. Quando la cornetta dell'apparecchio telefonico viene sollevata[31], nel telefono si chiude un interruttore che determina lo scorrimento di una corrente continua nel *subscriber loop*, indicando la risposta da parte del chiamato. Se viceversa siamo dal lato chiamante, sollevando la cornetta *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*[32], in grado di separare il segnale in ingresso da quello in uscita, in modo da inviare il primo all'altoprlante, 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. 6.8) sono

[29] Ovviamente, $m - n$ intervalli sono lasciati vuoti, in ordine *sparso* tra gli $m$.

[30] Si confronti questo risultato con la condizione di Clos, fornita al § 6.8.2.

[31] In inglese si dice andare OFF-HOOK, con riferimento storico al gancio su cui riporre la cornetta, presente nei primi modelli di telefono.

[32] `http://en.wikipedia.org/wiki/Hybrid_coil`

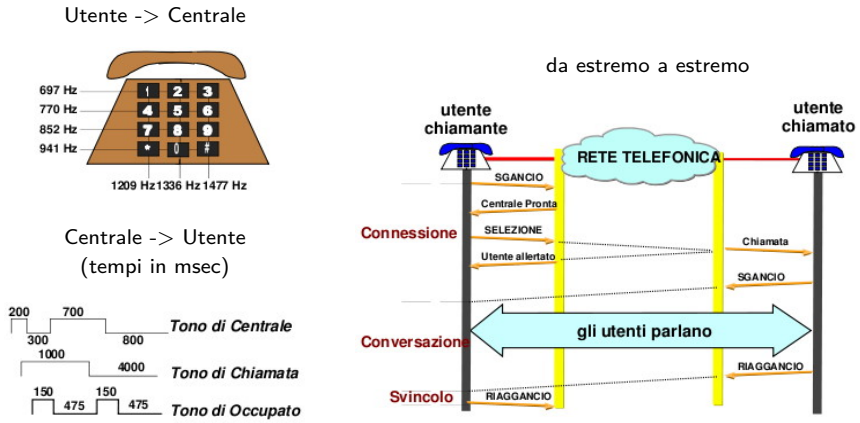



Figura 6.8: Segnalazione di utente

associate *due frequenze* che individuano la cifra (od il simbolo * e #) premuta, come descritto dalla figura. Viceversa, la segnalazione di utente nella direzione centrale -> utente avviene per mezzo di un codice basato su di un tono intermittente a 440 Hz[33], le cui durate sono descritte in basso a sinistra nella figura.

A seguito della ricezione del numero, la centrale *di origine* coinvolge il resto della rete, impegnando risorse della stessa, ed individuando quali nodi attraversare per giungere a destinazione (fase di *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. 6.8, in cui è evidenziato come ogni conversazione è in realtà composta da tre fasi imprescindibili:

- formazione della connessione (*call setup*), in cui sono svolte le funzioni di indirizzamento, e vengono riservate da parte della rete le risorse necessarie alla comunicazione
- mantenimento (*hold*), durante la quale le risorse impegnate sono utilizzate in modo esclusivo dalle parti in conversazione
- svincolo (*release*) in cui le risorse impegnate sono liberate

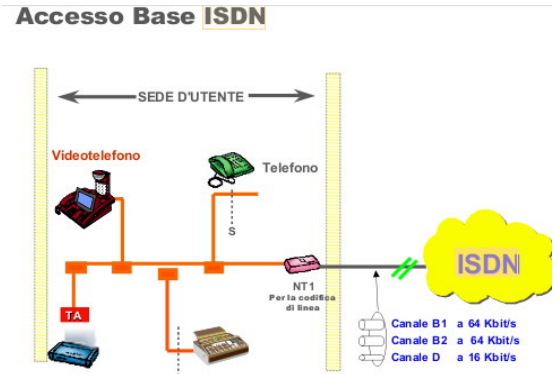
Il passaggio dalla telefonia analogia a quella numerica, in cui il segnale vocale è campionato e quantizzato come PCM, non ha di fatto alterato la presenza di queste tre fasi.

[33] corrispondente al *la* centrale del pianoforte. Ho provato a verificare, e... a me arriva un la bemolle!

### 6.9.2 ISDN

La *Integrated Service Data Network*[34] è una modalità di accesso *numerico* alla rete telefonica, definito da una serie di standard reperibili presso l'ITU[35]. In ISDN la conversione A/D avviene all'interno del terminale di utente, il quale può collegare allo stesso bus ISDN (interfaccia S[36] 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[37], su cui è trasmesso un segnale a quattro livelli noto come 2B1Q[38], 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[39] 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[40], come il protocollo Q.931[41].




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 § 6.3.4.

Dato che l'accesso ISDN preserva il flusso binario inviato sui canali B da estremo a estremo della rete, su quegli stessi canali posso essere inviate anche informazioni niente affatto vocali, ma bensì nativamente numeriche, purché il ricevente condivida le stesse modalità di interpretazione dei bit in arrivo. Struttando tale possibilità, sono ad esempio stati definiti i primi standard di videotelefonia H.320[42].

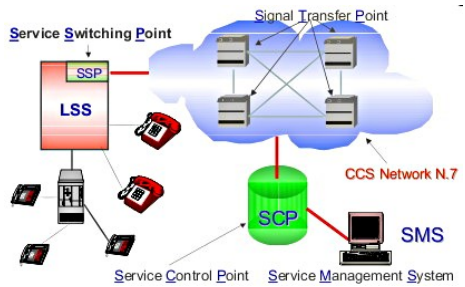
### 6.9.3 Sistema di segnalazione n 7

Il *Signaling System #7* (SS7[43]) è un insieme di protocolli di segnalazione telefonica a canale comune, usato per controllare la maggior parte delle chiamate telefoniche della

[34] http://it.wikipedia.org/wiki/ISDN
[35] http://www.itu.int/rec/T-REC-I/e
[36] http://hea-www.harvard.edu/~fine/ISDN/n-isdn.html
[37] http://www.ralphb.net/ISDN/ifaces.html
[38] http://it.wikipedia.org/wiki/2B1Q
[39] http://telemat.die.unifi.it/book/corso_telematica/lez_100/grp_4.html
[40] http://www.rhyshaden.com/isdn.htm
[41] http://www.javvin.com/protocolQ931.html
[42] http://it.wikipedia.org/wiki/H.320
[43] http://en.wikipedia.org/wiki/Signaling_System_7

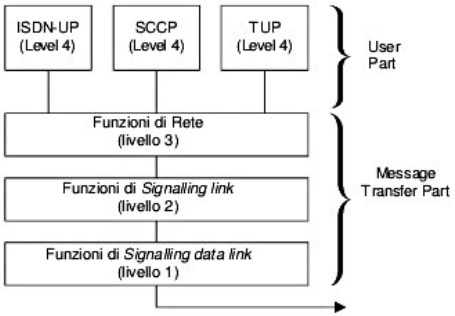
PSTN mondiale, che in questo caso prende il nome di *Intelligent Network* (IN[44]). Oltre ad gestire la fasi di instaurazione e abbattimento della chiamata, permette altri servizi come reindirizzamento, carte prepagate, SMS, numero verde, conferenza, richiamata su occupato...

L'SS7 è descritto dalla serie di raccomandazioni ITU-T Q.700[45], 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



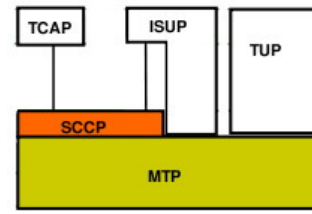

- *Service switching point* (SSP[46]), che termina la segnalazione di utente, ed invia una query all'SCP per determinare come gestire la richiesta di servizio;
- *Signal Transfer Point* (STP[47]), che instrada i messaggi SS7 tra le diverse entità della IN;
- *Service Control Point* (SCP[48]), che interroga un *Service Data Point* (SDP[49]), 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[50]) che opera sopra il *Transaction Capabilities Application Part* (TCAP) della pila protocollare SS7.

Oltre alle entità che prendono parte alla architettura, SS7 è definito anche nei termini della gerarchia protocollare che descrive la stratificazione delle funzioni necessarie allo svolgimento dei servizi richiesti. Il semplice scambio dei messaggi tra le entità è basato su di una rete a commutazione di pacchetto, ed avviene in base alle procedure collettivamente indicate come *Message Transfer Part* (MTP[51]), 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.




[44] http://en.wikipedia.org/wiki/Intelligent_network
[45] http://www.itu.int/rec/T-REC-Q.700/en
[46] http://en.wikipedia.org/wiki/Service_switching_point
[47] http://en.wikipedia.org/wiki/Signal_transfer_point
[48] http://en.wikipedia.org/wiki/Service_Control_Point
[49] http://en.wikipedia.org/wiki/Service_Data_Point
[50] http://en.wikipedia.org/wiki/INAP
[51] http://en.wikipedia.org/wiki/Message_Transfer_Part

Al disopra della MTP possono operare diversi procolli indicati come User Part, come ad esempio il *Signalling Connection Control Part* (SCCP[52]), che arricchisce le funzionalità di rete, offrendo ulteriori capacità di indirizzamento, ed un servizio orientato alla conessione anziché a pacchetto; attraverso SCCP possono operare processi applicativi basati sul *Transaction Capabilities Application Part* (TCAP[53]).



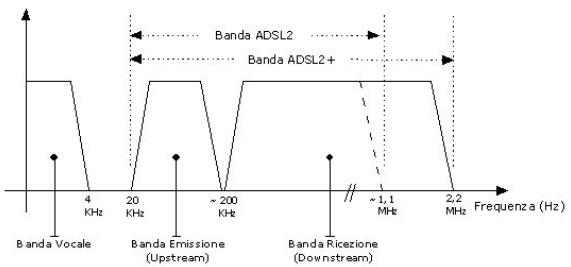

Altri esempi di User Part sono la *Telephone User Part* (TUP[54]) e la *ISDN User Part* (ISUP[55]). TUP è stata la prima UP ad essere definita, e fornisce il supporto all'offerta di servizi PSTN mediante la rete SS7. Attualmente è quasi ovunque rimpiazzato da ISUP, che offre altri servizi, come ad esempio 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*[56].

### 6.9.4 ADSL

L'*Asymmetric Digital Subscriber Loop*[57] è l'insieme di tecnologie trasmissive e di rete per mezzo delle quali viene fornito l'accesso ad Internet *a banda larga* per il tramite del doppino telefonico (*subscriber loop*) in rame, già utilizzato per il normale servizio telefonico POTS.

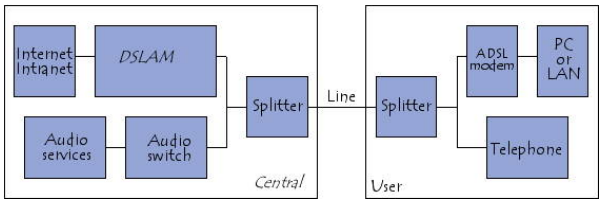
L'uso condiviso del mezzo è reso possibile realizzando la trasmissione numerica ADSL su di una banda di frequenze *più elevevate* di quelle usate da POTS, come mostrato in figura, dove sono rappresentati gli intervalli di frequenze riservati alla telefonia PSTN, ai dati in uscita (*upstream*) ed in ingresso (*downstream*). Le velocità di trasmissione inizialmente sono rispettivamente 1 ed 8 Mbps per i due versi trasmissivi, anche in funzione della lunghezza del collegamento utente - centrale[58]; successivamente la velocità di ricezione arriva rispettivamente a 12 e 20 Mbit/sec per gli standard ADSL2 e ADSL2+.




Le due diverse trasmissioni sono separate su linee differenti[59] inserendo, a valle della presa telefonica casalinga, un doppio filtro passa-alto e passa-basso detto

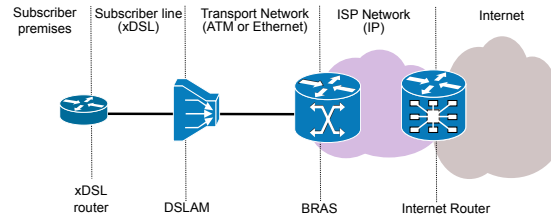
[52] http://en.wikipedia.org/wiki/Signaling_Connection_and_Control_Part
[53] http://en.wikipedia.org/wiki/Transaction_Capabilities_Application_Part
[54] http://en.wikipedia.org/wiki/Telephone_User_Part_(TUP)
[55] http://en.wikipedia.org/wiki/ISDN_User_Part
[56] http://en.wikipedia.org/wiki/SIGTRAN
[57] http://en.wikipedia.org/wiki/Asymmetric_Digital_Subscriber_Line
[58] All'aumentare della lunghezza del collegamento, oltre a ridursi la potenza ricevuta e quindi peggiorare l'SNR, aumenta l'entità della diafonia tra utenti differenti, determinando un ulteriore peggioramento di SNR, che la tecnica di modulazione affronta riducendo la velocità trasmissiva.
[59] Ciò permette di non ascoltare nel telefono il fruscìo della trasmissione ADSL, e di ridurre il rischio che le comunicazioni vocali determino *la caduta* della connessione ADSL.

*splitter*. Un filtro del tutto simile esiste anche dal lato centrale, in modo da inoltrare la componente in banda audio alla centrale POTS, e la componente dati verso un dispositivo DSLAM.




**DSLAM e oltre** Il *Digital Subscriber Loop 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[60] basata su ATM o ETHERNET, che termina il traffico sul *Broadband Remote Access Server* (BRAS)[61] dell'ISP, utilizzato da quest'ultimo anche per terminare il protocollo di strato di collegamento PPP[62], svolgere le funzioni di autenticazione dell'accesso, ed applicare eventuali *policy* a livello IP. Quindi, l'ISP provvede ad interconnettere il traffico del cliente con la rete Internet. Alternativamente, l'ISP può disporre di un *Point of Presence* (POP) nella stessa centrale in cui sono ospitati i DSLAM dei propri clienti[63], che in questo caso producono direttamente traffico IP, inoltrato verso la *core network* dell'ISP usando la sua stessa connettività.




**DMT** Il modem ADSL utilizza a tecnica di modulazione numerica multi-portante detta *Discrete Multi Tone*, in cui il flusso binario viene ripartito su più canali di frequenza contigui, ed il segnale analogico sintetizzato direttamente nel dominio della frequenza mediante il calcolo di una FFT, come previsto dalla tecnica di trasmissione OFDM (vedi § 13.6.4). In questo modo, oltre a semplificare le operazioni di equalizzazione, è possibile variare la velocità di trasmissione in modo indipendente per le diverse portanti, e mantenere buone prestazioni anche nel caso in cui l'SNR vari con la frequenza.

### 6.9.5 TDM mediante modulazione di ampiezza degli impulsi

Al tempo in cui la realizzazione del componente di quantizzazione (vedi § 7.4) presentava discrete difficoltà circuitali, si pensò[64] di sfruttare il teorema del campionamento

[60] Non lasciarsi fuorviare dal ruolo *di trasporto* della rete, che in effetti assolve unicamente un ruolo *di livello due* (o di collegamento), in quanto il punto di uscita non è qualsiasi, ma l'ISP fornitore dell'utente.

[61] `http://en.wikipedia.org/wiki/Broadband_Remote_Access_Server`

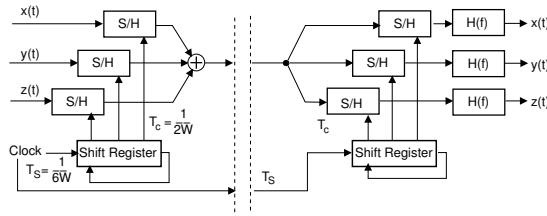
[62] `http://en.wikipedia.org/wiki/Point-to-Point_Protocol`

[63] Come ad es, nel caso dell'Unbundling: `http://it.wikipedia.org/wiki/Unbundling_local_loop`

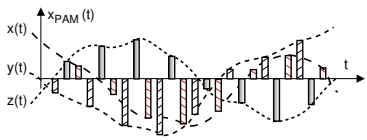
[64] La *pensata* non ebbe molte applicazioni, se non in ambito della commutazione interna ad esempio ad un centralino, a causa della sensibiltà del metodo agli errori di temporizzazione, ed alle caratteristiche del mezzo trasmissivo su cui inviare il segnale PAM.

(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 § 4.1.3 altri segnali simili, ad esempio $y^{\circ}(t)$, $z^{\circ}(t)$ come mostrato in Figura, ognuno campionato a frequenza $2W$, 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 $T_S = \frac{1}{2NW} = \frac{T_c}{N}$, 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 facendone passare gli impulsi campionati a frequenza $2W$ in un filtro di ricostruzione con banda $W$.

## 6.10 Riferimenti

Per questo capitolo un pò particolare, elenco in modo distinto alcune fonti on-line a cui mi sono ispirato, e dalle quali sono state tratte alcune illustrazioni.

- La Rete di Telecomunicazioni `http://net.infocom.uniroma1.it/corsi/impianti/lezioni_new/lez_1.pdf` di *Stefano Paggi* `http://net.infocom.uniroma1.it/corsi/impianti/impianti.htm`
- ISDN `http://www-tlc.deis.unibo.it/Didattica/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 7

# Probabilità, processi ed errori

In questo capitolo sono trattati una serie di argomenti che hanno in comune l'utilizzo dei risultati della teoria delle probabilità, brevemente riassunti nelle prime sezioni. Dopo aver caratterizzato le variabili aleatorie mediante i momenti, e definite le proprietà di ergodicità dei processi, viene valutata l'entità del rumore di quantizzazione, e calcolata la probabilità di errore per una tramissione numerica. Un'ampia appendice è dedicata ad approfondire alcuni aspetti particolari della teoria, come elementi di decisione statistica, la quantizzazione logaritmica, la ricezione ottima, e le distribuzioni di Rayleigh e Rice.

## 7.1 Teoria delle probabilità

Tratta delle caratteristiche *regolari* di fenomeni *irregolari* o *casuali*. Una prima definizione di probabilità è quella fornita dalla teoria frequentistica, la quale asserisce che se, ripetendo $N$ volte un esperimento, si verifica la circostanza $A$ per $n_A$ volte, per essa si osserva una frequenza relativa $n_A/N$, da cui si deriva la probabilità di $A$ come

$$Pr_A = \lim_{N\to\infty} \frac{n_A}{N}$$

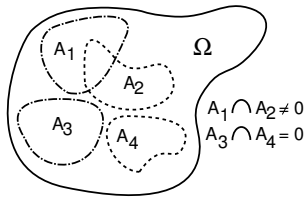
In termini piú astratti, l'insieme di tutte le circostanze possibili può essere pensato come un insieme algebrico, i cui elementi (o punti) sono appunto le diverse circostanze. I punti possono essere raggruppati in sottoinsiemi (eventualmente vuoti o di un solo punto) per i quali valgono le proprietà di unione, intersezione, complemento, inclusione...

I fenomeni fisici sono posti in relazione con i punti degli insiemi suddetti mediante il concetto di *spazio campione* $\Omega$, che è l'unione di tutti i possibili risultati di un *fenomeno aleatorio*. Sottoinsiemi dello spazio campione sono detti *eventi*. L'intero spazio è l'*evento certo*, mentre l'insieme vuoto corrisponde all'evento impossibile $\phi$ (od evento *nullo*). Una *unione* $\bigcup$ di eventi, corrisponde all'evento che si verifica ogni qualvolta se ne verifichi *un suo componente*, mentre *l'intersezione* $\bigcap$ è verificata se *tutti* i componenti lo sono. Esempio: il lancio di un dado genera uno spazio con 6 punti (eventi) disgiunti. Uno spazio campione può avere un numero di punti finito, infinito numerabile, o infinito.

### 7.1.1 Assiomi delle probabilità

Costituiscono le basi su cui sono costruiti i teoremi seguenti, ed affermano che:

- $0 \leq Pr(A) \leq 1$: la probabilità di un evento è compresa tra 0 ed 1;
- $Pr(\Omega) = 1$: la probabilità dell'evento certo è 1;
- Se $Pr(A_i \bigcap A_j) = \phi$ allora $Pr(\bigcup A_i) = \sum Pr(A_i)$: la probabilità dell'unione di eventi *disgiunti* è la somma delle singole probabilità.




### 7.1.2 Teoremi di base

- $Pr(\phi) = 0$: la probabilità dell'evento impossibile è nulla;
- $Pr(A \bigcap B) + Pr\left(A \bigcap \overline{B}\right) = Pr(A)$, e $Pr(B) + Pr\left(\overline{B}\right) = 1$: un evento ed il suo complemento riempiono lo spazio (detto anche teorema delle *probabilità totali*[1]);
- $Pr(A \bigcup B) = Pr(A) + Pr(B) - Pr(A \bigcap B)$: la probabilità dell'evento intersezione si conta una volta sola. Esempio alla nota[2];
- Se $B \subseteq A$ allora $Pr(B) \leq Pr(A)$: quando l'evento $B$ è contenuto in $A$ il verificarsi del primo implica il secondo.

### 7.1.3 Probabilità condizionali

Può avvenire che il verificarsi di un evento *influenzi* il verificarsi o meno di un altro: Si dice allora che lo condiziona, ovvero che l'evento influenzato è *condizionato*. La probabilità che avvenga $A$, noto che $B$ si sia verificato, si scrive $Pr(A/B)$, e si legge probabilità (condizionata) di $A$ *dato* $B$, che è definita[3] come

$$Pr(A/B) = \frac{Pr(A,B)}{Pr(B)}$$

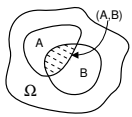
in cui $Pr(A,B) = Pr(A \bigcap B)$ è la probabilità *congiunta* che $A$ e $B$ si verifichino entrambi, ed a patto che $Pr(B) \neq 0$ (altrimenti anche $Pr(A/B)$ è zero!).

---

[1] Utile per scrivere la probabilità di un evento come "1 meno" quella dell'evento complementare.

[2] Lanciando un dado, la probabilità $Pr(pari \bigcup > 2)$ di ottenere un numero pari, *oppure* piú grande di due, è la *somma* delle probabilità dei singoli eventi $Pr(pari) = \frac{3}{6}$ e $Pr(>2) = \frac{4}{6}$, meno quella che si verifichino assieme $Pr(pari \bigcap > 2) = \frac{2}{6}$. Pertanto: $Pr(pari \bigcup > 2) = \frac{3}{6} + \frac{4}{6} - \frac{2}{6} = \frac{5}{6}$.

[3] La relazione può essere verificata ricorrendo al diagramma in figura, ed interpretando $Pr(A/B)$ come il rapporto tra la misura di probabilità dell'evento congiunto, rispetto a quella dell'evento condizionante.



**Esercizio:** Valutare la probabilità condizionata $Pr(A/B)$ che lanciando un dado si ottenga un numero pari (evento $A = (pari)$), condizionatamente all'evento $B$ che il numero sia $>2$. Soluzione alla nota[4].

A partire dalla precedente definizione, si ottiene quella della probabilità *congiunta*: $Pr(A,B) = Pr(A/B)\,Pr(B)$; inoltre, gli eventi condizionante e condizionato possono invertire i rispettivi ruoli, permettendo di scrivere anche: $Pr(A,B) = Pr(B/A)\,Pr(A)$. Eguagliando le due espressioni, si ottiene:

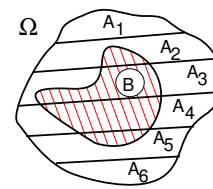
$$Pr(A/B) = \frac{Pr(B/A)\,Pr(A)}{Pr(B)} \quad \text{ed anche} \quad Pr(B/A) = \frac{Pr(A/B)\,Pr(B)}{Pr(A)}$$

Come ultima definizione, rammentiamo che le probabilità $Pr(A)$ e $Pr(B)$ sono indicate come *marginali*.

### 7.1.4 Teorema di Bayes

A volte, non tutti i possibili eventi sono direttamente osservabili: in tal caso la probabilità marginale $Pr(A)$ è indicata come probabilità *a priori*. Qualora l'evento $A$ sia in qualche modo legato ad un secondo evento $B$, che invece possiamo osservare, la probabilità condizionata $Pr(A/B)$ prende il nome di probabilità *a posteriori* perché, a differenza di quella a priori, rappresenta un valore di probabilità valutata *dopo* la conoscenza di $B$.

In generale, però, si conosce solamente $Pr(A)$ e $Pr(B/A)$ (queste ultime sono dette probabilità condizionate *in avanti*), e per calcolare $Pr(A/B)$ occorre conosce anche $Pr(B)$. Quest'ultima quantità si determina *saturando* la probabilità congiunta $Pr(A,B)$ rispetto a tutti gli eventi marginali $A_i$ possibili:




$$Pr(B) = \sum_i Pr(B, A_i) = \sum_i Pr(B/A_i)\,Pr(A_i)$$

a patto che risulti $Pr(A_i, A_j) = 0$ e $\bigcup A_i = \Omega$, e cioè che l'insieme degli $\{A_i\}$ costituisca una partizione dello spazio degli eventi $\Omega$. Tale circostanza è mostrata in figura.

L'ultima relazione ci permette di enunciare il *teorema di Bayes*, che mostra come ottenere le probabilità *a posteriori* a partire da quelle *a priori* e da quelle condizionate *in avanti*:

$$Pr(A_i/B_j) = \frac{Pr(B_j/A_i)\,Pr(A_i)}{\sum_k Pr(B_j/A_k)\,Pr(A_k)}$$

[4]Il risultato è pari alla probabilità $Pr(A,B) = Pr(pari, >2)$ che i due eventi si verifichino contemporaneamente, divisa per la probabilità $P_R(B) = P_R(>2)$ che il numero sia $>2$

Si rifletta sulla circostanza che la probabilità del pari $P_R(A) = \frac{1}{2}$, quella $P_R(B) = \frac{4}{6}$, o quella congiunta di entrambi $P_R(A,B) = \frac{2}{6}$, sono tutte riferite ad un qualunque lancio di dado, mentre $Pr(pari\,/\,>2)$ è relativa ad un numero ridotto di lanci, ossia solo quelli che determinano un risultato $>2$. Pertanto, essendo $Pr(B) \leq 1$, si ottiene $Pr(A/B) \geq Pr(A,B)$; infatti per l'esempio del dado si ottiene $Pr(pari\,/\,>2) = Pr(pari, >2)\,/\,Pr(>2) = \frac{2}{6}/\frac{4}{6} = \frac{1}{2}$, che è maggiore di $Pr(pari, >2) = \frac{1}{3}$ (i valori di probabilità sono ottenuti come rapporto tra il numero di casi favorevoli e quello dei casi possibili).

Si ottiene invece $Pr(A/B) = Pr(A,B)$ solo se $Pr(B) = 1$, ossia se $B$ corrisponde all'unione di tutti gli eventi possibili.

Al §17.2.1 è mostrata l'applicazione del risultato ad un problema di decisione statistica tipico delle telecomunicazioni, relativo alla ricezione binaria. Di seguito, invece, è illustrato un esempio più diretto.

### 7.1.5 Indipendenza statistica

Si verifica quando

$$Pr\left(A/B\right) = Pr\left(A\right)$$

in quanto il verificarsi di $B$ non influenza $A$. Come conseguenza, per due eventi statisticamente indipendenti avviene che

$$Pr\left(A, B\right) = Pr\left(A\right) Pr\left(B\right)$$

#### Esempi

- Un sistema di comunicazione radio è affetto da attenuazioni supplementari causate da pioggia. Indicando con $FS$ l'evento che il sistema vada fuori servizio, e conoscendo le probabilità condizionate $Pr\left(FS/piove\right) = .5$, $Pr\left(FS/non\, piove\right) = .05$ e la probabilità marginale $Pr\left(pioggia\right) = .03$, determinare:

  1. La probabilità di fuori servizio $Pr\left(FS\right)$, indipendentemente dal verificarsi o meno dell'evento piovoso;
  2. La probabilità che stia piovendo, sapendo che il sistema è fuori servizio.

  Risultato ([5]).

- Quale è la probabilità che, lanciando 3 volte un dado, esca 3 volte 1 ? Risultato ([6]).
- Un'urna contiene 2 biglie bianche e 3 nere. Qual è la probabilità che su 2 estrazioni consecutive, escano le 2 biglie bianche ? Risultato ([7]).
- Qual è la probabilità che 2 carte, estratte a caso da un mazzo da bridge da 52, siano K e Q ? Risultato ([8]).

---

[5] La probabilità marginale di fuori servizio si calcola applicando il teorema delle probabilità totali:

$$\begin{aligned} Pr\left(FS\right) &= Pr\left(FS/piove\right) \cdot Pr\left(piove\right) + Pr\left(FS/non\, piove\right) \cdot Pr\left(non\, piove\right) = \\ &= .5 \cdot .03 + .05 \cdot .97 = .0635 = 6.35\% \end{aligned}$$

in quanto $Pr\left(non\, piove\right) = 1 - Pr\left(piove\right) = .97$. Applicando il teorema di Bayes si trova quindi:

$$Pr\left(piove/FS\right) = \frac{Pr\left(FS/piove\right) \cdot Pr\left(piove\right)}{Pr\left(FS\right)} = \frac{.5 \cdot .03}{.0635} = .236 = 23.6\%$$

Si noti come la probabilità *a priori* che piova (3 %) venga rimpiazzata dal suo valore *a posteriori* (23,6 %) grazie alla nuova informazione di cui disponiamo (collegamento fuori servizio). Per una definizione precisa delle probabilità a priori ed a posteriori si veda l'appendice 17.2.1.

[6] E' pari al prodotto delle probabilità marginali, essendo i lanci statisticamente indipendenti, visto che il dado è "senza memoria". Pertanto il risultato è $\left(\frac{1}{6}\right)^3 = \frac{1}{216} \simeq 4.6296 \cdot 10^{-3}$.

[7] Anche qui l'urna è senza memoria; però dopo la prima estrazione le biglie restano in 4! Pertanto ora il prodotto delle probabilità marginali risulta $\frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$.

[8] $Pr\left(K, Q\right) = Pr\left(K\, prima, Q\, seconda\right) + Pr\left(Q\, prima, K\, seconda\right) = Pr\left(K\, prima\right) \cdot Pr\left(Q\, seconda/K\, prima\right) + Pr\left(Q\, prima\right) \cdot Pr\left(K\, seconda/Q\, prima\right) = 2\left(\frac{4}{52}\frac{4}{51}\right) = \frac{8}{663} \cong 1.2 \cdot 10^{-2}$ .

## 7.2 Variabili aleatorie

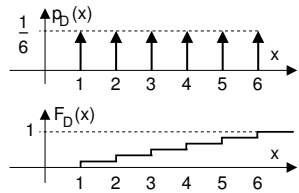
Finora si è parlato di *eventi* in modo astratto, mentre spesso ci si trova ad associare ad ogni punto dello spazio campione un valore numerico: lo spazio campione $\Omega$ diventa allora l'*insieme dei numeri* e prende il nome di *variabile aleatoria*. La realizzazione di un evento corrisponde ora all'assegnazione di un valore (tra i possibili) alla variabile aleatoria; tale valore "prescelto" prende dunque il nome di *realizzazione* della v.a. Distinguiamo inoltre tra variabili aleatorie *discrete* e *continue*, a seconda se la grandezza che descrivono abbia valori numerabili o continui[9]. La caratterizzazione della variabile aleatoria, in termini probabilistici, si ottiene indicando come la "massa di probabilità" si distribuisce sull'insieme di valori che la variabile aleatoria può assumere, per mezzo delle 2 funzioni di variabile aleatoria seguenti.

### 7.2.1 Funzioni di distribuzione e densità e di probabilità

Cosí come la massa di un oggetto *non omogeneo* è distribuita in modo piú o meno denso in regioni differenti del suo volume complessivo, cosí la *densità di probabilità* (o d.d.p.) indica su quali valori della variabile aleatoria si concentra la probabilità. Cosí, ad esempio, la densità della v.a. discreta associata al lancio di un dado può essere scritta:

$$p_D(x) = \sum_{n=1}^{6} \frac{1}{6}\delta(x-n)$$

il cui significato discutiamo subito, con l'aiuto dei due grafici seguenti, in cui $D$ e $x$ indicano rispettivamente la v.a. (il numero che uscirà) ed una sua realizzazione (una delle 6 facce). I 6 impulsi centrati in $x = n$ rappresentano una concentrazione di probabilità nei sei possibili valori; l'area di tali impulsi è proprio pari alla probabilità di ognuno dei sei risultati. E' facile verificare che

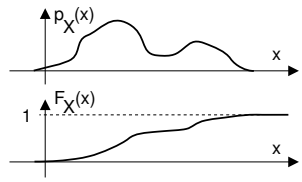



$$\int_{-\infty}^{\infty} p_D(x)\,dx = 1$$

e che risulta

$$\int_a^b p_D(x)\,dx = Pr\{a < D \le b\}$$

ovvero pari alla probabilità che la v.a. $D$ assuma un valore tra $a$ e $b$. In particolare, non potendosi verificare una probabilità negativa, si ha $p_D(x) \ge 0$ con $\forall x$.




Una funzione di v.a. strettamente collegata alla densità è la funzione *distribuzione di probabilità*[10], definita come

$$F_X(x) = \int_{-\infty}^{x} p_X(\xi)\,d\xi = Pr\{X \le x\}$$

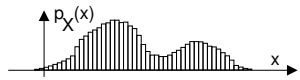
[9] Un esempio classico di v.a. discreta è quello del lancio di un dado, un altro sono i numeri del lotto. Una v.a. continua può essere ad esempio un valore di pressione atmosferica in un luogo, oppure l'attenuazione di una trasmissione radio dovuta a fenomeni atmosferici.

[10] In realtà, l'ordine storico è quello di definire prima $F_X(x)$ come la probabilità che $X$ sia non superiore ad un valore $x$, ovvero $F_X(x) = Pr\{X \le x\}$, e quindi $p_X(x) = \frac{dF_X(x)}{dx}$. Il motivo di tale "priorità" risiede nel fatto che $F_X(x)$ presenta minori "difficoltà analitiche" di definizione (ad esempio presenta solo discontinuità di prima specie, anche con v.a. discrete).

che risulta una funzione non decrescente di $x$, limitata ad un massimo valore di 1, ed il cui andamento mostriamo sotto alla $p_D(x)$, nel caso dell'esempio del lancio del dado[11].

Le definizioni date mantengono validità nel caso di v.a. continua, originando le curve mostrate nei due grafici a lato. Ora è ancora piú evidente la circostanza che $p_X(x)$ è una *densità*, e diviene una probabilità solo quando moltiplicata per un intervallo di $x$([12]).

**Istogramma** Evidenziamo ora la stretta relazione che intercorre tra la densità di probabilità e l'*istogramma*. Quest'ultimo può essere realizzato se si dispone di una serie di realizzazioni della v.a., e si ottiene suddividendo il campo di variabilità della grandezza $X$ in sotto-intervalli, e disegnando rettangoli verticali, ognuno di altezza pari al numero di volte che (nell'ambito del campione statistico a disposizione) $X$ assume un valore in quell'intervallo. Dividendo l'altezza di ogni rettangolo per il numero di osservazioni $N$, si ottiene una approssimazione di $p_X(x)$, via via piú precisa con $N \to \infty$, e con una conseguente riduzione dell'estensione degli intervalli.




### 7.2.2 Medie, momenti e momenti centrati

Indichiamo con $g(x)$ una funzione di variabile aleatoria[13] . Si definisce *valore atteso* (o *media*, *media di insieme*, *media statistica*) di $g(x)$ rispetto alla variabile aleatoria $X$ la quantità:

$$E_X\{g(x)\} = \int_{-\infty}^{\infty} g(x)\, p_X(x)\, dx$$

che corrisponde ad una media (integrale) pesata[14] dei diversi valori $g(x)$, ognuno con peso pari alla probabilità $p_X(x)\, dx$; la notazione $E_X\{.\}$ indica quindi[15] tale operazione di media integrale, assieme alla v.a. $(x)$ rispetto a cui eseguirla.

**Momenti** Nel caso in cui $g(x) = x^n$, il valore atteso prende il nome di *momento di ordine n*, che corrisponde quindi al valore atteso della *n-esima* potenza della v.a., e che si indica come

$$m_X^{(n)} = E\{x^n\} = \int_{-\infty}^{\infty} x^n p_X(x)\, dx$$

[11] Si ricorda che la derivata di un gradino è un impulso di area pari al dislivello. Infatti, sapevamo già che l'integrale di un impulso è una costante - ammesso che l'impulso cada dentro gli estremi di integrazione!

[12] Infatti la probabilità che $X$ cada tra $x_0$ e $x_0 + \Delta x$ vale $\int_{x_0}^{x_0+\Delta x} p_X(x)\, dx \simeq p_X(x_0)\, \Delta x$.

[13] Un esempio di funzione di v.a. potrebbe essere il valore della vincita associata ai 13 in schedina, che dipende dalla v.a. rappresentata dai risultati delle partite, una volta noto il montepremi e le giocate. Infatti, per ogni possibile vettore di risultati, si determina un diverso numero di giocate vincenti, e quindi un diverso modo di suddividere il montepremi. Essendo i risultati improbabili giocati da un ridotto numero di schedine, a queste compete un valore maggiore in caso di vincita, ben superiore al suo *valore atteso*, indicativo invece della vincita media.

[14] Notiamo che se al posto delle probabilità $p_X(x)\, dx$ utilizziamo i valori di un istogramma $Pr(x_i) = \frac{N(x_i < x \leq x_i + \Delta x)}{N} = \frac{N_i}{N}$, l'integrale si trasforma in una sommatoria, il cui sviluppo evidenzia l'equivalenza con una media pesata: $\sum x_i Pr(x_i) = \frac{x_1 N_1 + x_2 N_2 + \ldots + x_n N_n}{N}$.

[15] In effetti, la $E$ simboleggia la parola EXPECTATION, che è il termine inglese usato per indicare il valore atteso.

Nel caso di variabili aleatorie discrete, i momenti sono definiti come $m_X^{(n)} = \sum_i x_i^n p_i$, in cui $p_i = Pr\{x = x_i\}$, pesando quindi le possibili realizzazioni $x_i$ con le rispettive probabilità.

**Media** Verifichiamo subito che $m_X^{(0)} = 1$. Il momento di primo ordine

$$m_X = m_X^{(1)} = \int_{-\infty}^{\infty} x p_X(x)\, dx$$

prende il nome di *media*[16] della v.a. $X$ (a volte denominata *centroide*), mentre con $n = 2$ si ha la *media quadratica* $m_X^{(2)} = \int_{-\infty}^{\infty} x^2 p_X(x)\, dx$.

**Momenti centrati** Nel caso in cui $g(x) = (x - m_X)^n$, il relativo valore atteso è chiamato *momento centrato* di ordine $n$, ed indicato come

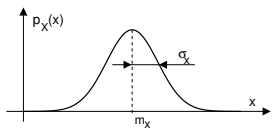
$$\mu_X^{(n)} = E\{(x - m_X)^n\} = \int_{-\infty}^{\infty} (x - m_X)^n p_X(x)\, dx$$

E' immediato constatare che $\mu_X^{(0)} = 1$ e che $\mu_X^{(1)} = 0$.

**Varianza** E' il nome dato al momento centrato del $2^o$ ordine, ed è indicata come

$$\sigma_X^2 = \mu_X^{(2)} = E\left\{(x - m_X)^2\right\} = \int_{-\infty}^{\infty} (x - m_X)^2 p_X(x)\, dx$$

La radice quadrata della varianza, $\sigma_X$, prende il nome di *deviazione standard*. Mentre la media $m_X$ indica dove si colloca il "centro statistico" della densità di probabilità, $\sigma_X$ indica quanto le singole determinazioni della v.a. siano disperse attorno ad $m_x$.

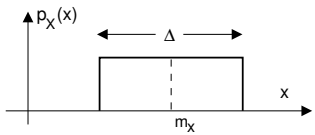



Una relazione notevole che lega i primi due momenti (centrati e non) è ([17]):

$$\sigma_X^2 = m_X^{(2)} - (m_X)^2 \tag{7.1}$$

### 7.2.3 Variabile aleatoria a distribuzione uniforme

Applichiamo le definizioni dei momenti ad un caso pratico: la variabile aleatoria uniforme è caratterizzata da uno stesso valore di probabilità per tutta la gamma di realizzazioni possibili, limitate queste ultime ad un unico intervallo non disgiunto; pertanto, la densità di probabilità è esprimibile mediante una funzione rettangolare:




$$p_X(x) = \frac{1}{\Delta} rect_\Delta(x - m_X)$$

[16] Supponiamo che $X$ rappresenti l'altezza degli individui; l'altezza *media* sarà allora calcolabile proprio come momento del primo ordine.

[17] Infatti risulta

$$\begin{aligned}\sigma_X^2 &= E\{(x - m_X)^2\} = E\{x^2 + (m_X)^2 - 2x m_X\} = E\{x^2\} + (m_X)^2 - 2 m_X E\{x\} = \\ &= m_X^{(2)} + (m_X)^2 - 2(m_X)^2 = m_X^{(2)} - (m_X)^2\end{aligned}$$

Si è preferito usare la notazione $E\{x\}$, più compatta rispetto all'indicazione degli integrali coinvolti; i passaggi svolti si giustificano ricordando la proprietà distributiva degli integrali (appunto), ed osservando che il valore atteso di una costante è la costante stessa.

in cui $\Delta$ rappresenta l'estensione dell'intervallo di esistenza della variabile aleatoria.

E' facile verificare che il parametro $m_X$, che indica l'ascissa a cui è centrato il rettangolo, corrisponde esattamente al momento di primo ordine di $X$. Il calcolo della varianza[18] invece fornisce: $\sigma_X^2 = \frac{\Delta^2}{12}$.

## 7.3 Processi stazionari ed ergodici

Dopo aver descritto come caratterizzare statisticamente singoli valori (denominati variabili aleatorie), occupiamoci del caso in cui si voglia descrivere da un punto di vista probabilistico un intero segnale, la cui reale identità non sia nota a priori[19].

Un segnale siffatto viene detto *membro* (o realizzazione) di un *processo aleatorio*, e può essere indicato come $x(t,\theta)$, che corrisponde ad una descrizione formale che prevede una coppia di insiemi: il primo di questi è l'insieme $\mathcal{T}$ degli istanti temporali (tipicamente un intervallo) su cui sono definiti i membri del processo; il secondo è relativo ad una variabile aleatoria $\Theta$, i cui valori $\theta$ identificano ognuno una particolare realizzazione del processo. Pertanto, una singola realizzazione $\theta = \theta_i$, per cosí dire, *indicizza* il processo, le cui istanze effettive $x(t,\theta_i)$, con $t \in \mathcal{T}$, sono note solo dopo la conoscenza di $\theta_i \in \Theta$ ([20]). Il processo aleatorio è quindi definito come l'insieme dei segnali $\{x(t,\theta)\}$, con $t \in \mathcal{T}$ e $\theta \in \Theta$.

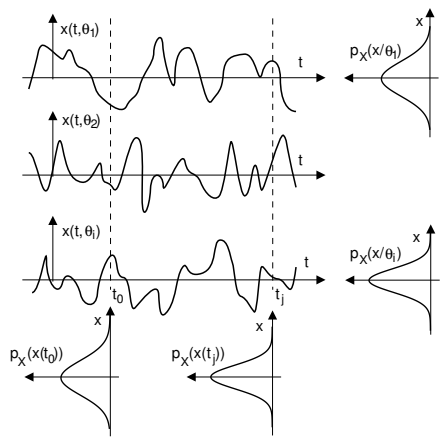



Figura 7.1: Un processo *non* ergodico

Se viceversa fissiamo un particolare istante temporale $t_j$, il valore $x(t_j,\theta)$ è una variabile aleatoria, la cui realizzazione dipende da quella di $\theta \in \Theta$; pertanto, è definita la densità $p_X(x(t_j))$ (indipendente da $\theta$), che possiamo disegnare in corrispondenza dell'istante $t_j$ in cui è prelevato il campione[21]; a tale riguardo, si faccia riferimento alla figura precedente, che mostra le densità di probabilità definite a partire dai membri di un processo.

### 7.3.1 Media di insieme

E' definita come il *valore atteso* di una *potenza n-esima* dei valori del segnale, ossia un suo momento, eseguito rispetto alla variabilità dovuta a $\Theta$, ed è pertanto calcolata

[18] Anzichè calcolare $\sigma_X^2$ per la $p_X(x)$ data, calcoliamo $m_X^{(2)}$ per una v.a. uniforme con $m_X = 0$: in tal caso infatti $m_X^{(2)} = \sigma_X^2$. Si ha: $\sigma_X^2 = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} x^2 \frac{1}{\Delta} dx = \frac{x^3}{3\Delta}\Big|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} = \frac{1}{3\Delta}\left(\frac{\Delta^3}{8} + \frac{\Delta^3}{8}\right) = \frac{1}{3\Delta} 2\frac{\Delta^3}{8} = \frac{\Delta^2}{12}$.

[19] Chiaramente, la maggioranza dei segnali trasmessi da apparati di TLC sono di questo tipo.

[20] Per fissare le idee, conduciamo parallelamente al testo un esempio "reale" in cui il processo aleatorio è costituito da.... la selezione musicale svolta da un dj. L'insieme $\mathcal{T}$ sarà allora costituito dall'orario di apertura delle discoteche (dalle 22 all'alba ?), mentre in $\theta$ faremo ricadere tutte le caratteristiche di variabilità (umore del dj, i dischi che ha in valigia, la discoteca in cui ci troviamo, il giorno della settimana...).

[21] Nell'esempio, $x(t_0,\theta)$ è il valore di pressione sonora rilevabile ad un determinato istante (es. le 23.30) al variare di $\theta$ (qualunque dj, discoteca, giorno...).

come

$$m_X^{(n)}(t_j) = E_\Theta\{x^n(t_j,\theta)\} = \int_{-\infty}^{\infty} x^n(t_j,\theta)\, p_\Theta(\theta)\, d\theta = \int_{-\infty}^{\infty} x^n p_X(x(t_j))\, dx$$

in cui l'ultima eguaglianza evidenzia come una media di insieme dipenda dalla d.d.p. $p_X(x(t_j))$ di $x(t_j,\theta)$ al variare di $\theta \in \Theta$, mostrata in basso in fig. 7.1. Notiamo che in linea di principio, la media di insieme dipende dall'istante $t_j$ in cui è prelevato un valore[22].

### 7.3.2 Medie temporali

In alternativa, possiamo fissare una particolare realizzazione $\theta_i$ di $\Theta$, e quindi identificare un singolo membro $x(t,\theta_i)$, che è ora un segnale certo[23]; per lo stesso, possono quindi essere calcolate le *medie temporali*, indicate con una linea sopra alla quantità di cui si calcola la media $\overline{(.)}$:

$$m_X^{(n)}(\theta_i) = \overline{x^n(t,\theta_i)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^n(t,\theta_i)\, dt$$

In particolare, ritroviamo il *valore medio*

$$m_X(\theta_i) = \overline{x(t,\theta_i)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t,\theta_i)\, dt$$

e la *potenza* [24] (*o media quadratica*)

$$m_X^{(2)}(\theta_i) = \overline{x^2(t,\theta_i)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t,\theta_i)\, dt$$

Notiamo che una generica media temporale:

- non dipende dal tempo;
- è una variabile aleatoria (dipende infatti dalla realizzazione di $\Theta$).

### 7.3.3 Medie temporali calcolate come medie di insieme

L'estrazione da $x(t,\theta_i)$ di un valore ad un istante casuale $t = t_j \in \mathcal{T}$, definisce una ulteriore variabile aleatoria, descritta dalla densità di probabilità (condizionata) $p_X(x/\theta_i)$, che disegnamo a fianco dei singoli membri mostrati in fig. 7.1. Qualora la $p_X(x/\theta_i)$ sia nota, le medie temporali di ordine $n$ possono essere calcolate (per quel membro) come i momenti:

$$m_X^{(n)}(\theta_i) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^n(t,\theta_i)\, dt = \int_{-\infty}^{\infty} x^n p_X(x/\theta_i)\, dx = E_{X/\Theta=\theta_i}\{x^n\}$$

Ciò equivale infatti ad effettuare una media ponderata, in cui ogni possibile valore di $x$ è pesato per la sua probabilità $p_X(x/\theta_i)\, dx$.

[22] Ad esempio, se in tutte le serate il volume aumenta progressivamente nel tempo, la $p_X(x(t_j))$ si *allargherà* per $t_j$ crescenti.

[23] $x(t,\theta_i)$ rappresenta, nel nostro esempio, l'intera selezione musicale (detta *serata*) proposta da un ben preciso dj, in un preciso locale, un giorno ben preciso.

[24] $m_X^{(2)}(\theta_i)$ in questo caso rappresenta la potenza media con cui è suonata la musica nella particolare serata $\theta_i$.

### 7.3.4 Processi stazionari

Qualora $p_X(x(t_j))$ non dipenda da $t_j$, ma risulti $p_X(x(t_j)) = p_X^{\mathcal{T}}(x)$ per qualsiasi $t_j \in \mathcal{T}$, il processo $\{x(t,\theta)\}$ è detto stazionario[25] *in senso stretto*. In tal caso tutte le medie di insieme non dipendono piú dal tempo, ossia $m_X^{(n)}(t) = m_X^{(n)}$ per $\forall t \in \mathcal{T}$, e le $p_X(x(t_j))$ in basso in fig. 7.1 sono tutte uguali.

Se invece solamente le prime due medie di insieme $m_X(t)$ e $m_X^{(2)}(t)$ non dipendono da $t$, il processo $\{x(t,\theta)\}$ è detto stazionario *in media* ed *in media quadratica*, od anche stazionario *in senso lato*[26].

Supponiamo ora di suddividere il membro $x(t,\theta_i)$ in piú intervalli temporali, e di calcolare per ciascuno di essi le medie temporali, limitatamente al relativo intervallo. Nel caso in cui queste risultino uguali tra loro, e di conseguenza uguali alla media temporale $m_X^{(n)}(\theta_i)$, il membro è (individualmente) stazionario[27]. Ovviamente, se tutti i membri sono individualmente stazionari, lo è anche il processo a cui appartengono.

### 7.3.5 Processi stazionari ed ergodici

Questa importante sottoclasse di processi stazionari identifica la circostanza che *ogni membro del processo è statisticamente rappresentativo di tutti gli altri.* Ciò si verifica quando la densità di probabilità (a destra in fig. 7.1) dei valori estratti da un singolo membro $p_X(x/\theta_i)$ è sempre la stessa, indipendentemente dal particolare $\theta_i$, ottenendo in definitiva $p_X(x/\theta_i) = p_X^{\Theta}(x)$ indipendentemente dalla realizzazione e, per la stazionarietà, anche $p_X(x/\theta_i) = p_X^{\mathcal{T}}(x)$, e dunque $p_X^{\Theta}(x) = p_X^{\mathcal{T}}(x) = p_X(x)$. In questo caso le medie temporali $m_X^{(n)}(\theta_i)$, calcolabili come momenti sulla singola realizzazione come illustrato al §7.3.3, sono identiche per tutti i membri[28] $\theta_i$, ed identiche anche alle medie di insieme $m_X^{(n)}(t_j)$ calcolate per un qualunque istante. Enunciamo pertanto la definizione:

> *Un processo stazionario è ergodico se la media temporale calcolata su di una qualunque realizzazione del processo, coincide con la media di insieme relativa ad una variabile aleatoria estratta ad un istante qualsiasi (per la stazionarietà) da una realizzazione qualsiasi (per l'ergodicità).*

**Esempio: la potenza di segnale** Mostriamo come il calcolo della potenza di un membro di un processo ergodico sia equivalente a quello del momento di $2^o$ ordine del processo:

$$\begin{aligned}\mathcal{P}_X(\theta) &= \overline{x^2(\theta)} = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} x^2(\tau,\theta)\,d\tau = \int_{-\infty}^{\infty} x^2 p_X(x/\theta)\,dx = \\ &= \int_{-\infty}^{\infty} x^2 p_X(x)\,dx = m_X^{(2)} = E\{x^2\} = \mathcal{P}_X\end{aligned}$$

[25] *La "serata in discoteca" stazionaria si verifica pertanto se non mutano nel tempo il genere di musica, il volume dell'amplificazione... o meglio se eventuali variazioni in alcune particolari discoteche-realizzazioni sono compensate da variazioni opposte in altrettanti membri del processo.*

[26] In questo caso la $p_X(x(t))$ non è nota, oppure non è stazionaria, ma le maggiori applicazioni della proprietà di stazionarietà dipendono solo da $m_X(t)$ e $m_X^{(2)}(t)$, che possono essere misurati (o per meglio dire *stimati*), e risultare stazionari anche se $p_X(x(t))$ non lo è.

[27] Questo accade se la selezione musicale di una particolare serata si mantiene costante (es. solo raggamuffin) oppure variata ma in modo omogeno (es. senza tre "lenti" di fila).

[28] Volendo pertanto giungere alla definizione di una serata *ergodica* in discoteca, dovremmo eliminare quei casi che, anche se individualmente stazionari, sono decisamente "fuori standard" (tutto metal, solo liscio...).

Questo risultato mostra come sia possibile calcolare la potenza di una realizzazione di un processo, senza conoscerne la forma d'onda.

**Potenza, varianza, media quadratica e valore efficace** In particolare osserviamo che in base alla (7.1) possiamo scrivere

$$\mathcal{P}_X = m_X^{(2)} = \sigma_x^2 + (m_x)^2$$

e per i segnali a media nulla ($m_x = 0$) si ottiene $\mathcal{P}_X = \sigma_x^2$; in tal caso il valore efficace $\sqrt{\mathcal{P}_X}$ coincide con la deviazione standard $\sigma_x$. La radice della potenza è inoltre spesso indicata come *valore RMS* (ROOT MEAN SQUARE), definito come $x_{RMS} = \sqrt{\overline{x^2(t)}}$, ovvero la radice della *media quadratica* (nel tempo). Se il segnale è a media nulla, $x_{RMS}$ coincide quindi con il valore efficace; se $x(t)$ è membro di un processo ergodico a media nulla, $x_{RMS}$ coincide con la deviazione standard.

### 7.3.6 Riassumendo

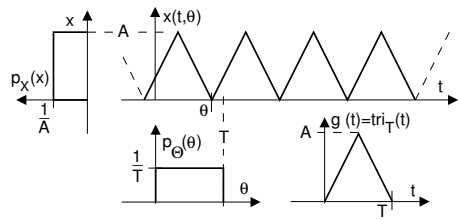
- Se un processo è ergodico, è anche stazionario, ma non il viceversa. Esempio: se $x(t,\theta) = C_\theta$ pari ad una costante (aleatoria), allora è senz'altro stazionario, ma $p_X(x/\theta) = \delta(x - C_\theta)$, e quindi non ergodico.
- Se un processo è ergodico è possibile:
  - calcolare le medie di insieme in forma di medie temporali a partire da una singola realizzazione *oppure*
  - ottenere le medie temporali di una qualunque realizzazione a partire dalle medie di insieme, disponendo della statistica $p_X(x)$.
- Se l'eguaglianza tra medie di insieme e temporali sussiste solo fino ad un determinato ordine e non oltre, il processo *non è* ergodico *in senso stretto*. Per ciò che concerne le Telecomunicazioni, è spesso sufficiente la proprietà di ergodicità *in senso lato*, ovvero limitata al $2^o$ ordine, che garantisce $\overline{x(t)} = E\{x\} = m_x$; $\overline{x^2(t)} = E\{x^2\} = m_x^{(2)}$.

### 7.3.7 Processo ad aleatorietà parametrica

Questo è il nome dato a processi $\{x(t,\theta)\}$ per quali il parametro $\theta$ compare in modo esplicito nella espressione analitica dei segnali membri. Ad esempio, il segnale periodico

$$x(t,\theta) = \sum_{n=-\infty}^{\infty} A \cdot g_T(t - \theta - nT)$$

rappresentato in figura, ha come parametro un ritardo $\theta$, che è una variabile aleatoria che ne rende imprecisata la fase iniziale. Se $\theta$ è (come in figura) una v.a. uniformemente distribuita tra 0 e $T$ (ovvero $p_\Theta(\theta) = \frac{1}{T} rect_T(\theta - \frac{T}{2})$),



allora il processo è stazionario ed ergodico, e la sua densità di probabilità per $g(t) = tri_T(t)$ risulta pari a

$$p_X(x) = \frac{1}{A} rect_A(x - \frac{A}{2})$$

Il valor medio $m_X = E\{x\}$ è pari alla media temporale $\frac{A}{2}$, la varianza è quella della d.d.p.[29] uniforme $\sigma_X^2 = \frac{A^2}{12}$ e la potenza vale

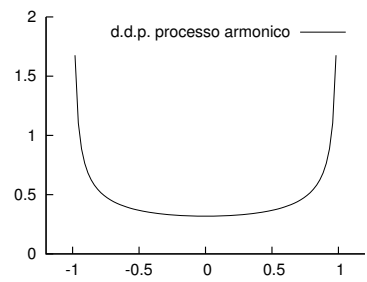
$$\mathcal{P}_X = \sigma_X^2 + m_X^2 = \frac{A^2}{12} + \frac{A^2}{4} = \frac{4A^2}{12} = \frac{A^2}{3}$$

Se la $p_\Theta(\theta)$ fosse stata diversa, il processo avrebbe potuto perdere ergodicità. Se ad esempio $p_\Theta(\theta) = \frac{4}{T} rect_{\frac{T}{4}}(\theta)$, si sarebbe persa la stazionarietà: infatti prendendo ad esempio $-\frac{T}{8} < t < \frac{T}{8}$, *tutte* le realizzazioni avrebbero valori minori del valor medio $\frac{A}{2}$.

**Processo armonico** Si tratta di un processo ad aleatorietà parametrica, i cui membri hanno espressione

$$x(t, \theta) = A\cos(2\pi f_0 t + \theta)$$

dove $\theta$ è una v.a. uniforme con d.d.p. $p_\Theta(\theta) = \frac{1}{2\pi} rect_{2\pi}(\theta)$. In tal caso il processo è stazionario ed ergodico, e si ottiene che una valore estratto a caso da un membro qualsiasi è una v.a. con d.d.p.




$$p_X(x) = \frac{1}{\pi\sqrt{A^2 - x^2}} \tag{7.2}$$

mostrata in figura per $A = 1$.

**Segnale dati** Riprendiamo qui l'espressione (5.1) a cui aggiungiamo un elemento di indeterminazione per quanto riguarda la relazione temporale tra l'origine dei tempi e gli istanti caratteristici, scrivendo

$$x(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT + \theta)$$

con $\theta$ v.a. a distribuzione uniforme tra $\pm\frac{T}{2}$, in modo da rendere il processo ergodico. Mentre il calcolo della sua densità di potenza sarà affrontato al § 9.2.4, qui ci limitiamo ad osservare che, considerando i valori $a_n$ come determinazioni di v.a. indipendenti ed identicamente distribuite, la densità di probabilità di $x(t)$ può euristicamente essere desunta dalla analisi del corrispettivo diagramma ad occhio. Ad esempio, nel caso di segnale a banda infinita e $a_n$ a due livelli equiprobabili (vedi fig. 5.1 a pag. 66) la $p_X(x)$ sarà costituita da due impulsi di area 1/2, mentre nei casi di limitazione in banda e/o adozione di un impulso con caratteristica a coseno rialzato, la stessa assumerà un andamento continuo[30].

[29] d.d.p. è l'abbreviazione comunemente usata per Densità di Probabilità.

[30] In una prossima edizione, potrei calcolare le ddp corrispondenti ai diagrammi ad occhio di fig. 5.4

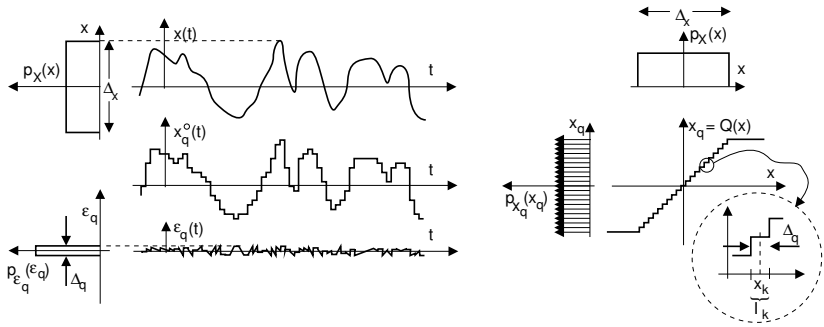
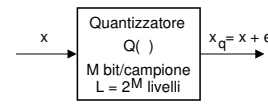



Figura 7.2: Processo di quantizzazione per segnali a distribuzione di ampiezza uniforme

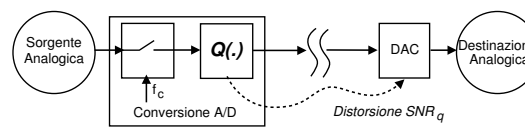
## 7.4 SNR di quantizzazione




Trattiamo ora della questione, lasciata in sospeso, di come scegliere il numero $M$ di bit con cui rappresentare i campioni di un segnale, ovvero della *risoluzione* con cui realizzare il dispositivo già indicato al § 4.1.3 come *quantizzatore*. Tale processo consiste nel rappresentare i valori $x$ in ingresso mediante un insieme finito di $L = 2^M$ valori quantizzati

$$x_q = x + \epsilon_q$$

e rappresentare il valore approssimato del campione mediante la rappresentazione binaria[31] dell'indice $k \in \{1, 2, \cdots, L\}$ che identifica il valore quantizzato $x_q$ più prossimo ad $x$. In tal modo si introduce un errore $\epsilon_q$, il cui effetto si riperquote in fase di restituzione D/A come descritto nella figura che segue.




L'obiettivo è quindi quello di scegliere il numero di livelli $L$ in modo da mantenere il rapporto segnale rumore di quantizzazione $SNR_q$ migliore di un valore desiderato. Dato che l'$SNR_q$ è pari al rapporto tra le potenze del segnale $\mathcal{P}_x$ e del rumore $\mathcal{P}_\epsilon$, procediamo nel determinare queste ultime due, con l'aiuto del grafico mostrato in Fig. 7.2.

Adottiamo l'ipotesi semplificativa che i valori in ingresso al campionatore abbiano origine da un processo ergodico a media nulla, e siano rappresentati da una v.a. con densità di probabilità uniforme $p_X(x) = \frac{1}{\Delta_x} rect_{\Delta_x}(x)$; pertanto la potenza dei campioni, pari alla varianza della v.a., risulterà (vedi § 7.2.3)

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

Lo stesso intervallo di valori di ingresso $\pm\frac{\Delta_x}{2}$ è suddiviso dal quantizzatore in $L-1$ intervalli $I_k$ di eguale ampiezza $\Delta_q = \frac{\Delta_x}{(L-1)}$, centrati sui valori $x_k = k\Delta_q - \frac{\Delta_x}{2}$, con

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

$k = 0, 1, 2, ..., L-1$. Tutti i valori di ingresso $x$, che cadono all'interno di $I_k$, ovvero tali che $x_k - \frac{\Delta_q}{2} \leq x < x_k + \frac{\Delta_q}{2}$, sono codificati con l'intero $k$, rappresentato in binario da una parola di $M = \log_2 L$ bit (scegliendo $L$ come una potenza di 2).

Il componente che, a partire dai valori quantizzati, ricostruisce il segnale $x_q^\circ(t)$ (vedi § 4.1.3) da inviare al filtro di restituzione, tipicamente associa ad ogni intero $k$ il valore centrale $x_q = x_k$ dell'intervallo di quantizzazione, commettendo cosí un errore $\epsilon_q = x_q - x$, di entità limitata entro l'intervallo $\pm\frac{\Delta_q}{2}$. Di nuovo, si suppone che anche $\epsilon_q$ sia una v.a. uniformemente distribuita tra $\pm\frac{\Delta_q}{2}$, ed indipendente[32] da $x_k$; pertanto, la potenza della componente di errore è anche qui pari alla varianza, e cioè

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

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

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

in cui l'ultima approssimazione ha validità nel caso evidente in cui $L \gg 1$. Il risultato mostra che l'$SNR_q$ cresce in modo quadratico con l'aumentare dei livelli, ovvero se $L$ raddoppia $SNR_q$ quadruplica. Ricorrendo alla notazione in decibel[33] per l'SNR, otteniamo il risultato $SNR_q(L)|_{dB} = 10\log_{10} L^2 = 20\log_{10} L$ e, ricordando che $L = 2^M$, si ottiene

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

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

Consideriamo ora cosa accade se il segnale in ingresso $x$ ha una dinamica *minore* di quanto previsto: in tal caso $\sigma_x^2$ si riduce, mentre $\sigma_\epsilon^2 = \frac{1}{12}\left(\frac{\Delta_x}{L-1}\right)^2$ non cambia, e dunque $SNR_q$ peggiora come se avessimo ridotto i livelli. In appendice § 7.6.1 è illustrata la tecnica usata *nella pratica* per mantenere un $SNR_q$ elevato anche con bassi livelli di segnale.

## 7.5 Errori nelle trasmissioni numeriche

Con l'aiuto degli strumenti probabilistici studiati, analizziamo il funzionamento del dispositivo di decisione di una trasmissione numerica dal punto di vista della probabilità di errore. Come descritto al § 5.1.1 ogni ricevitore ha a che fare, oltre che con il segnale effettivamente ricevuto $r(t)$, anche con un segnale $n(t)$ sommato al primo,

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

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

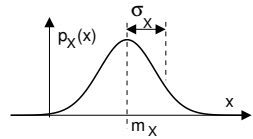
$$SNR_q(dB) = 10\log_{10}\frac{\mathcal{P}_x}{\mathcal{P}_\epsilon} = 10\log_{10}\mathcal{P}_x - 10\log_{10}\mathcal{P}_\epsilon = \mathcal{P}_x\left(dBV^2\right) - \mathcal{P}_\epsilon\left(dBV^2\right)$$

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

la cui forma d'onda effettiva è una realizzazione del processo di *rumore additivo* (vedi cap. 16). Tale processo di rumore è stazionario ed ergodico, e la densità di probabilità del primo ordine che lo descrive è la "famosa" *gaussiana*, che introduciamo subito.

### 7.5.1 Variabile aleatoria gaussiana e funzione $erfc\{.\}$

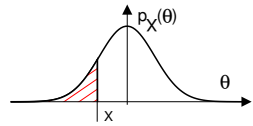
Una variabile aleatoria gaussiana $x$ è descritta da una densità di probabilità di espressione




$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \mathrm{e}^{-\frac{(x-m_x)^2}{2\sigma_x^2}}$$

il cui andamento è mostrato in figura, dove è posto in evidenza come $m_x$ e $\sigma_x$ (media e deviazione standard) siano in relazione la prima con la centratura orizzontale, e la seconda con la dispersione della curva attorno alla media. Oltre che da un punto di vista grafico, i primi due momenti della v.a. descrivono completamente la densità anche dal punto di vista analitico; pertanto, la stima di questi (ad esempio a partire da un buon numero di realizzazioni[34]) è sufficiente per descrivere completamente il fenomeno aleatorio. La v.a. gaussiana descrive bene una moltitudine di fenomeni naturali, ed è dimostrabile analiticamente che la sua densità è caratteristica di grandezze generate dalla somma di un numero molto elevato di cause aleatorie, tutte con la medesima d.d.p.[35] (*teorema centrale del limite*[36]). La funzione di distribuzione di questa v.a. non è calcolabile in forma chiusa; fortunatamente però, stante la necessità di conoscere la probabilità di eventi gaussiani, sono disponibili tabelle e grafici, che riportano il valore numerico dell'integrale che definisce tali valori di probabilità.

Per renderci conto della situazione, proviamo a calcolare la probabilità che $X$ non superi un certo valore $x$, pari per definizione alla funzione di distribuzione, e rappresentata dall'area tratteggiata in figura (per semplicità ci riferiamo ad una gaussiana a media nulla):




$$F_X(x) = Pr\{X \le x\} = \int_{-\infty}^{x} p_X(\theta)\, d\theta = 1 - \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{\theta^2}{2\sigma_x^2}} d\theta$$

Effettuiamo ora un cambio di variabile, ponendo $\frac{\theta^2}{2\sigma_x^2} = \eta^2$, per cui in corrispondenza di $\theta = x$ si ha $\eta = \frac{x}{\sqrt{2}\sigma_x}$, e risulta $d\theta = \sqrt{2}\sigma_x d\eta$. Con queste posizioni, possiamo riscrivere

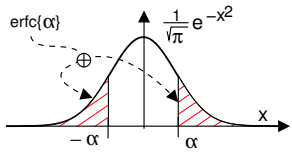
$$\begin{aligned} F_X(x) &= 1 - \int_{\frac{x}{\sqrt{2}\sigma_x}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\eta^2} \sqrt{2}\sigma_x d\eta \\ &= 1 - \int_{\frac{x}{\sqrt{2}\sigma_x}}^{\infty} \frac{1}{\sqrt{\pi}} e^{-\eta^2} d\eta = 1 - \frac{1}{2}\mathrm{erfc}\left\{\frac{x}{\sqrt{2}\sigma_x}\right\} \end{aligned}$$

[34] Disponendo di un insieme $\{x_n\}$ di $N$ realizzazioni di una variabile aleatoria $X$, possiamo effettuare le seguenti stime: $\widehat{m}_x = \frac{1}{N}\sum_{n=1}^{N} x_n$ e $\widehat{m}_x^{(2)} = \frac{1}{N}\sum_{n=1}^{N} x_n^2$, il cui valore tende *asintoticamente* a quello delle rispettive medie di insieme, come $N$ (la dimensione del campione statistico) tende a $\infty$.

[35] Il suo scopritore, K.F. Gauss, denominò la v.a. e la sua ddp come *Normale*, indicando con questo il fatto che il suo uso potesse essere "quotidiano".

[36] `http://it.wikipedia.org/wiki/Teoremi_centrali_del_limite`

Cosa è successo? Semplicemente, abbiamo espresso l'integrale (irrisolvibile in forma chiusa) nei termini della "funzione" $erfc\{.\}$, che rappresenta la probabilità che il valore assoluto di una v.a. gaussiana a media nulla e varianza $\frac{1}{2}$ superi il valore dato come argomento, come mostrato in figura, e pari a




$$\operatorname{erfc}\{\alpha\} = 2\frac{1}{\sqrt{\pi}}\int_{\alpha}^{\infty} e^{-x^2}dx$$

I valori di $erfc$ in funzione del suo argomento sono reperibili sia in forma di tabelle numeriche, sia in forma di diagrammi quotati[37].

In linea generale quindi, volendo calcolare la probabilità che una v.a. gaussiana $X$, con media $m_x$ e varianza $\sigma_x^2$, superi in ampiezza un determinato valore $\overline{x}$, l'unica strada percorribile è quella di utilizzare la funzione $erfc$, avendo cura di porre come argomento il valore di $\overline{x}$ debitamente scalato, per ricondursi ad una gaussiana a media nulla e varianza $\frac{1}{2}$:

$$Pr\{X > \overline{x}\} = \frac{1}{2}erfc\left\{\frac{\overline{x} - m_X}{\sqrt{2}\sigma_x}\right\}$$

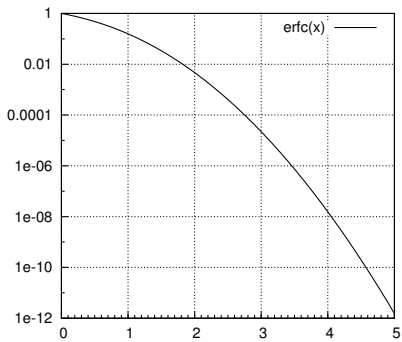
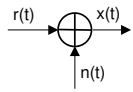



Figura 7.3: Valore di $\operatorname{erfc}\{\alpha\}$ per una gaussiana normalizzata

### 7.5.2 Calcolo della probabilità di errore per simbolo




Come anticipato, il decisore presente al lato ricevente di una trasmissione numerica opera su un segnale $x(t) = r(t) + n(t)$, ovvero il segnale ricevuto $r(t)$ è *corrotto* da un disturbo additivo $n(t)$ realizzazione di un processo stazionario ergodico gaussiano. Nel caso in cui siano presenti più cause di disturbo, anche localizzate in punti diversi del collegamento, si fa in modo (vedi pag. 334) di ricondurle tutte ad un'unica fonte di rumore (equivalente) in ingresso al decisore.

**Rumore limitato in banda** Caratterizziamo quindi il disturbo $n(t)$ come la realizzazione di un processo gaussiano ergodico, a un valor medio nullo, con spettro di densità di potenza *bianco* (ossia costante)

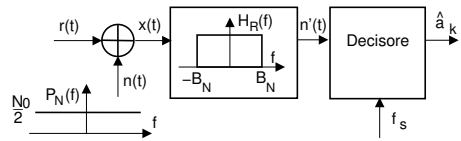
$$\mathcal{P}_N(f) = \frac{N_0}{2}$$

e con una varianza $\sigma_N^2$ pari alla potenza[38] $\mathcal{P}_N$ di una sua realizzazione qualsiasi (per l'ergodicità): tale processo è anche indicato come *Additive White Gaussian Noise* o AWGN.

[37] Il termine $erfc$ sta per *funzione di errore complementare*, e trae origine dai risultati della misura di grandezze fisiche, in cui l'errore di misura, dipendente da cause molteplici, si assume appunto gaussiano.

[38] In realtà $\mathcal{P}_N(f)$ non è costante per qualsiasi valore di $f$ fino ad infinito, ma occupa una banda indeterminata ma limitata: altrimenti, avrebbe una potenza infinita.

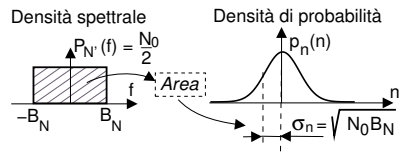
Allo scopo di limitare $\mathcal{P}_N$ alla minima possibile, in ingresso al ricevitore è posto un filtro passa-basso ideale[39] con risposta in frequenza $H_R(f)$ limitata in una banda $\pm B_N$ (detta *banda di rumore*, vedi § 12.1.1), tale da lasciar passare le componenti frequenziali del segnale $r(t)$ per intero, e limitare al tempo stesso la banda (e dunque la potenza) di $\mathcal{P}_N(f)$ al minimo.




Pertanto, la potenza del rumore in uscita da $H_R(f)$ risulta[40] pari a

$$\mathcal{P}_{N'} = \sigma^2_{N'} = \int_{-\infty}^{\infty} \mathcal{P}_N(f)\,|H_R(f)|^2\,df = \int_{-B_N}^{B_N} \frac{N_0}{2}\,df = N_0 B_N$$

e questo valore si riflette (in virtù della ergodicità di $n(t)$) nella dinamica della v.a. di rumore $n$, come esemplificato in figura.




**Soglie di decisione** Proseguiamo ora l'analisi indicando il segnale ricevuto nella forma

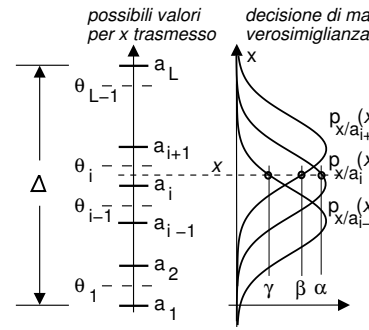
$$r(t) = \sum_k a_k \cdot g(t - kT_s) \tag{7.3}$$

in cui $g(t)$ è una caratteristica di Nyquist, in modo che non insorga di ISI[41], ed ogni *simbolo* $a_k$ è il valore di un elemento di una sequenza aleatoria, pari ad uno tra $L$ possibili valori $a_i$, $i = 1, 2, .., L-1$, distribuiti entro un intervallo con dinamica pari a $a_L - a_1 = \Delta$.

Agli istanti multipli del periodo di simbolo $t = kT_s = k/f_s$, il decisore esamina il valore del segnale $x(t) = r(t) + n(t)$, ed anzichè ritrovare i valori $a_i$ trasmessi, osserva la realizzazione di una variabile aleatoria gaussiana, con media pari al valore $a_i$, e varianza $\sigma^2_N = N_0 B_N$. Chiaramente, il ricevitore non conosce quale valore sia stato trasmesso in quell'istante, ed effettua una decisione *di massima verosimiglianza* o MV (vedi § 17.2.1) confrontando tra loro le densità di probabilità condizionate alle diverse ipotesi $a_i$

$$P_{X/a_i}(x) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left\{-\frac{(x-a_i)^2}{2\sigma^2_N}\right\} \tag{7.4}$$

e scegliendo per l'$\widehat{a_i}$ tale che $P_{X/\widehat{a_i}}(x)$ è la più grande.




Nel caso in cui i valori $a_i$ siano equispaziati, il criterio di massima verosimiglianza equivale (vedi figura) a definire $L-1$ soglie di decisione $\theta_i$, $i = 1, 2, .., L-1$, poste a metà tra i valori $a_i$ ed $a_{i+1}$, e decidere per il valore $a_i$ se il segnale ricevuto $x$

[39] Si veda il § 7.6.2 per la scelta di un diverso filtro di ricezione, individuato applicando le condizioni per la minimizzazione della probabilità di errore.

[40] Per i dettagli relativi al filtraggio di processi, ci si può riferire al § 9.4.3.

[41] Si tratta dell'*Inter Symbol Interference*, definita al § 5.1.2.2

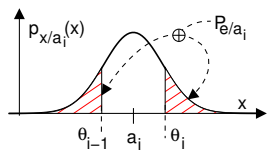
cade all'interno dell'intervallo compreso tra $\theta_{i-1}$ e $\theta_i$ ([42]), dato che ciò corrisponde ad imporre

$$\alpha = P_{X/a_i}(x) > \beta = P_{X/a_{i+1}}(x) > \gamma = P_{X/a_{i-1}}(x)$$

**Probabilità di errore** Si determina considerando che si verifica un errore (ossia il criterio di MV fornisce un risultato diverso dal valore trasmesso) tutte le volte che $x$ oltrepassa una soglia di decisione a causa di un valore particolarmente elevato di rumore, ovvero quando un campione di rumore è (in modulo) più grande di $\alpha = |\theta_i - a_i| = \frac{\Delta}{2(L-1)}$. La probabilità di errore si dice in questo caso *condizionata* alla trasmissione di $a_i$, e vale

$$P_{e/a_i} = 2\int_{\theta_i}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left\{-\frac{(x-a_i)^2}{2\sigma_N^2}\right\} dx = P_\alpha$$

che chiameremo $P_\alpha$. Lo stesso valore $P_\alpha$ è valido per tutti gli indici $i$ compresi tra 2 ed $L-1$, mentre per $a_1$ ed $a_L$ la probabilità di errore è dimezzata perché l'errore si verifica solo su di una soglia di decisione: $P_{e/a_1} = P_{e/a_L} = \frac{1}{2}P_\alpha$.




Applicando il cambiamento di variabile illustrato nella sezione precedente, troviamo che $P_\alpha = erfc\left\{\frac{\theta_i - a_i}{\sqrt{2}\sigma_N}\right\}$; sostituendo a $\theta_i - a_i$ la sua ampiezza espressa in funzione della dinamica di segnale $\Delta$, troviamo

$$P_\alpha = \text{erfc}\left\{\frac{\Delta}{2\sqrt{2}\sigma_N (L-1)}\right\} \tag{7.5}$$

Per arrivare all'espressione della probabilità di errore incondizionata[43] occorre eseguire una operazione di valore atteso rispetto a tutti gli indici $i$, con $i = 1, 2, ..., L$, ovvero pesare le diverse probabilità di errore condizionate per le rispettive probabilità degli eventi condizionanti. Nel caso in cui i valori $a_i$ siano *equiprobabili*, con probabilità $Pr(a_i) = \frac{1}{L}$, si ottiene:

$$P_e = E_{a_i}\left\{P_{e/a_i}\right\} = \sum_{i=1}^{L} Pr(a_i) P_{e/a_i} = \frac{1}{L}\left[(L-2)P_\alpha + 2\frac{1}{2}P_\alpha\right] = \left(1 - \frac{1}{L}\right)P_\alpha$$

in cui si è tenuto conto della diversa probabilità condizionata per i livelli intermedi e per i due agli estremi. Il risultato ottenuto, benchè già idoneo a valutare la $P_e$, può essere ulteriormente elaborato per ottenere espressioni più adatte ai progetti di dimensionamento.

### 7.5.3 Dipendenza di Pe da Eb/No

Sviluppiamo ora la (7.5) in modo da esprimere l'argomento di $erfc\{\}$ in funzione di una grandezza ($E_b/N_0$) che riassume in sé i valori dei *parametri intriseci* del collegamento (potenza ricevuta $\mathcal{P}_R$, densità di rumore $N_0$, frequenza binaria $f_b$). Infatti la grandezza $SNR = \frac{\mathcal{P}_R}{\mathcal{P}_N}$ generalmente indicativa della qualità del segnale ricevuto, in questo caso

[42] Chiaramente, tutti i valori minori di $\theta_1$ provocano la decisione a favore di $a_1$, e quelli maggiori di $\theta_{L-1}$ indicano la probabile trasmissione di $a_L$.

[43] Che non dipende cioè da quale simbolo sia stato trasmesso.

dipenderebbe, oltre che dai parametri suddetti, anche da altre grandezze che invece rappresentano specifiche scelte per dei *parameri di trasmissione*, come $\gamma$ e $L$, ed il cui contributo vorremmo invece mantenere separato. Infatti, la potenza di rumore $\mathcal{P}_N = N_0 B_N$, da cui $SNR$ dipende, a sua volta è funzione dalla banda del filtro di ricezione, che come abbiamo visto è posta pari alla massima frequenza presente in $r(t)$, e quindi (prendendo in considerazione un codice di linea $g(t)$ a coseno rialzato, vedi § 5.2.2.3) pari a

$$B_N = \frac{f_s}{2}(1+\gamma) = \frac{f_b}{2\log_2 L}(1+\gamma)$$

dipendendo quindi anch'essa da $L$ e $\gamma$, oltre che dalla $f_b$. Pertanto, al variare di $L$ e $\gamma$, varia anche $SNR$.

#### 7.5.3.1 Contributo di Eb/No all'SNR

La grandezza $E_b$ rappresenta l'*energia per bit*[44] e la sua definizione

$$E_b = \mathcal{P}_R T_b = \frac{\mathcal{P}_R}{f_b}$$

mostra come essa riassume in sè i parametri di sistema *potenza di segnale* e *velocità binaria*, mentre invece non dipende dai *parametri di trasmissione* $L$ e $\gamma$. Anche $N_0$ costituisce un *parametro di sistema*, rappresentando una grandezza su cui non è possibile intervenire.

Allo scopo quindi di mantenere separati tra loro i diversi elementi che determinano la probabilità di errore, e di porre nella giusta evidenza i parametri di sistema $E_b$ ed $N_0$, esprimiamo dunque le potenze $\mathcal{P}_N$ e $\mathcal{P}_R$ in funzione di $T_b = 1/f_b$:

$$\mathcal{P}_N = N_0 B_N = \frac{N_0(1+\gamma)}{T_b 2\log_2 L} \quad \text{e} \quad \mathcal{P}_R = \frac{E_b}{T_b} \tag{7.6}$$

in modo da ottenere

$$SNR = \frac{\mathcal{P}_R}{\mathcal{P}_N} = \frac{E_b}{T_b}\frac{T_b 2\log_2 L}{N_0(1+\gamma)} = \frac{E_b}{N_0}\frac{2\log_2 L}{(1+\gamma)} \tag{7.7}$$

#### 7.5.3.2 La componente di segnale

Prima di ottenere una espressione di $P_\alpha$ (7.5) in funzione di $^{E_b}/_{N_0}$, occorre esprimere $\mathcal{P}_R$ in funzione dei parametri di trasmissione $L$ e $\gamma$, nonché della dinamica $\Delta$. Si può dimostrare che sotto le ipotesi in cui:

1. si adotti un impulso di Nyquist a coseno rialzato con roll-off $\gamma$;
2. i simboli $a[k]$ siano statisticamente indipendenti ed a valori $a_i$ equiprobabili;
3. tali valori siano distribuiti uniformemente su $L$ livelli, con dinamica $a_L - a_1 = \Delta$;

al § 9.9.4 si ottiene[45]

$$\mathcal{P}_R = \frac{\Delta^2}{12}\frac{L+1}{L-1}\left(1-\frac{\gamma}{4}\right) \tag{7.8}$$

[44] si rifletta sulla circostanza che la potenza è una energia per unità di tempo.

[45] Anche se il risultato sarà dimostrato al § 9.9.4, merita comunque un commento: osserviamo che $\mathcal{P}_R$ diminuisce all'aumentare di $\gamma$ (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".

Se in particolare risulta $L = 2$ e $\gamma = 0$ (come nel caso di trasmissione binaria a banda minima) allora si ha $\mathcal{P}_R = \frac{\Delta^2}{4}$. Ma per essere utilizzata nel prosieguo, la (7.8) deve prima essere invertita, in modo da esprimere $\triangle$ in funzione di $\mathcal{P}_R$:

$$\triangle = \sqrt{12 \frac{L-1}{L+1} \frac{\mathcal{P}_R}{(1-\gamma/4)}} \tag{7.9}$$

#### 7.5.3.3 Espressione della $P_e$ per simbolo

Non resta ora che inserire la (7.9) nella espressione di $P_\alpha$ (eq. 7.5), ricordare che $\sigma_N^2 = \mathcal{P}_N$, e tenere conto della (7.7), in modo da ottenere[46]

$$P_e = \left(1 - \frac{1}{L}\right) \text{erfc} \left\{ \sqrt{\frac{E_b}{N_0} \frac{3 \log_2 L}{(L^2-1)(1+\gamma)\left(1-\frac{\gamma}{4}\right)}} \right\} \tag{7.10}$$

che è graficata alla Fig 7.4, come funzione di $\left.\frac{E_b}{N_0}\right|_{dB}$, per tre condizioni operative. In particolare, notiamo che per $L = 2$ e $\gamma = 0$ si ottiene:

$$P_e = \frac{1}{2} erfc \left\{ \sqrt{\frac{E_b}{N_0}} \right\}$$

Le scelte progettuali ($\gamma$ e $L$) diverse da $L = 2$ e $\gamma = 0$ determinano immancabilmente un peggioramento della $P_e$, ma vengono intraprese per soddisfare esigenze di risparmio di banda (aumentando $L$)[47], e per ridurre i termini di interferenza intersimbolica (aumentando $\gamma$). Due domande riassuntive:

- perché $P_e$ peggiora se aumento i livelli ? Risposta ([48]).

- perché $P_e$ peggiora se aumento $\gamma$ ? Risposta ([49]).

[46] Per chiarezza sviluppiamo i passaggi, piuttosto banali anche se non ovvi:

$$\begin{aligned}
P_e &= \left(1 - \frac{1}{L}\right) P_\alpha = \left(1 - \frac{1}{L}\right) \text{erfc} \left\{ \frac{\Delta}{2\sqrt{2}\sigma_N (L-1)} \right\} \\
&= \left(1 - \frac{1}{L}\right) \text{erfc} \left\{ \sqrt{12 \frac{L-1}{L+1} \frac{\mathcal{P}_R}{(1-\gamma/4)}} \frac{1}{2\sqrt{2\mathcal{P}_N}(L-1)} \right\} \\
&= \left(1 - \frac{1}{L}\right) \text{erfc} \left\{ 2\sqrt{3 \frac{L-1}{L+1} \frac{1}{(1-\gamma/4)}} \frac{1}{2\sqrt{2}} \sqrt{\frac{\mathcal{P}_R}{\mathcal{P}_N}} \frac{1}{(L-1)} \right\} \\
&= \left(1 - \frac{1}{L}\right) \text{erfc} \left\{ \sqrt{\frac{3}{2} \frac{L-1}{L+1} \frac{1}{(L-1)^2} \frac{1}{(1-\gamma/4)} SNR} \right\} \\
&= \left(1 - \frac{1}{L}\right) \text{erfc} \left\{ \sqrt{\frac{3}{2} \frac{1}{L^2-1} \frac{1}{(1-\gamma/4)} \frac{E_b}{N_0} \frac{2\log_2 L}{1+\gamma}} \right\} \\
&= \left(1 - \frac{1}{L}\right) \text{erfc} \left\{ \sqrt{\frac{E_b}{N_0} \frac{3 \log_2 L}{(L^2-1)(1+\gamma)\left(1-\frac{\gamma}{4}\right)}} \right\}
\end{aligned}$$

[47] Aumentando $L$, l'argomento di (7.10) diminuisce, in quanto $\left(L^2 - 1\right)$ cresce più velocemente di $\log_2 L$.

[48] Perché a parità di $\mathcal{P}_R$ gli intervalli di decisione sono più ravvicinati, e le "code" della gaussiana sottendono un'area maggiore.

[49] 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.

**Compromesso banda - potenza** Osservando la fig. (7.4) notiamo che al crescere di $L$, e dunque occupando una banda minore, si può ottenere la stessa $P_e$ solo a patto di aumentare $E_b/N_0$, ovvero (a parità di $f_b$) aumentando la potenza trasmessa: questo è un aspetto di un risultato più generale della teoria dell'Informazione. Claude Shannon ha infatti dimostrato (vedi pag. 425) che *è possibile trasmettere senza errori* (ricorrendo a tecniche di codifica di canale oltremodo sofisticate) purchè la velocità di trasmissione $f_b$ non ecceda la *capacità di canale*, definita come

$$C = B \log_2 \left(1 + \frac{\mathcal{P}_R}{N_0 B}\right)$$

in cui $B$ è la banda del canale, $\mathcal{P}_R$ la potenza ricevuta, e $N_0 B$ la potenza del rumore. Un secondo canale, con $B$ *minore*, dispone di una minore capacità, in quanto $\log_2(.)$ cresce più lentamente di quanto non decresca $B$; pertanto, per mantenere la stessa capacità, è necessario trasmettere con una maggiore potenza di segnale $\mathcal{P}_R$. E' questo il motivo per cui, nel caso in cui vi siano limitazioni di potenza ma non di banda, come ad esempio nelle *comunicazioni satellitari*, conviene occupare la maggior banda possibile, mantenendo $L = 2$, in modo da risparmiare potenza[50].

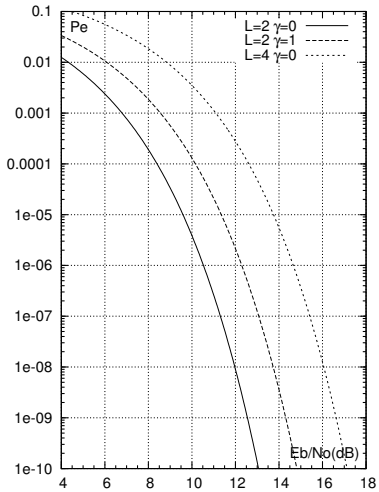



Figura 7.4: Andamento di $P_e$ vs. $E_b/N_0$

**Compromesso velocità - distorsione** Abbiamo appena fatto notare come, riducendo la banda occupata dal segnale dati, occorra aumentare la potenza ricevuta per mantenere la stessa $P_e$. Viceversa, osserviamo anche che l'*aumento* della velocità binaria (e quindi dell'occupazione di banda) dovuta all'adozione di strategie per il controllo di errore (vedi § 5.3 e § 17.2), consente di *ridurre* la probabilità di errore stessa, e dunque migliorare la *fedeltà* del flusso binario, anche a parità di potenza ricevuta.

### 7.5.4 Diagramma ad occhio

Si tratta dello stesso tipo di grafico già descritto a pag. 74, e che ora ci aiuta a valutare in modo visivo la qualità di una trasmissione numerica. In fig. 7.5 sono riportati i grafici per un segnale dati a 4 livelli, in presenza di due diversi livelli di potenza di rumore: notiamo che al peggiorare del rapporto $\frac{E_b}{N_0}$, la zona priva di traiettorie (*l'occhio*) riduce la sua estensione verticale (*tende a chiudersi*). Pertanto, disponendo di un segnale numerico di qualità sconosciuta, questa può essere valutata in modo approssimato, qualora si disponga di un oscilloscopio, esaminando il *grado di apertura dell'occhio*.

[50] Per contro, volendo mantenere immutata la qualità, e dunque la $P_e$, l'*aumento* dell'occupazione di banda prodotta dal controllo di errore (vedi § 5.3), consente di *ridurre* la potenza di segnale!

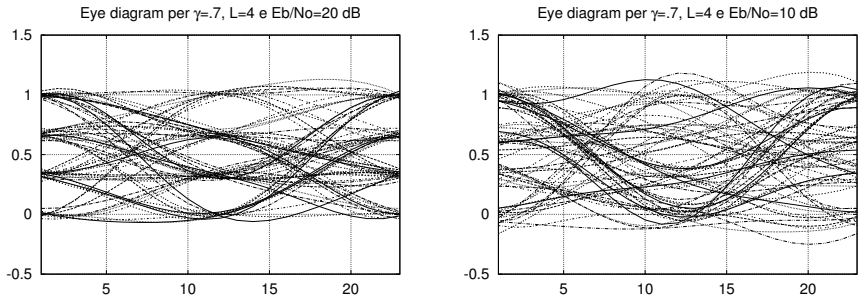



Figura 7.5: Diagramma ad occhio per segnale rumoroso con $E_b/N_0$ pari a 20 e 10 dB, $\gamma = .7$, $L = 4$

### 7.5.5 Uso del codice di Gray e $P_e$ per bit

La probabilità di errore (7.10) identifica l'evento di decidere per la ricezione del *simbolo* $a_i$ quando invece è stato trasmesso $a_{i-1}$ o $a_{i+1}$([51]), mentre ora determiniamo la probabilità che sia errato uno qualunque dei bit ottenibili dopo la *serializzazione* (vedi nota 9 a pag. 66) della codifica associata al simbolo ricevuto. Abbiamo già illustrato al § 5.2.2.4 come esista un modo particolare (il codice di Gray) di assegnare codifiche binarie ai simboli $a_i$, al fine di associare ai simboli relativi a valori (livelli) contigui, parole differenti per un solo bit.

In presenza della codifica di Gray, pertanto, ogni simbolo errato contiene un solo bit errato, e quindi l'evento di errore *sul bit* si verifica quando il simbolo a cui appartiene è errato, ***e*** il bit è quello errato, ovvero: $Pr\{bit\ errato\} = Pr\{simbolo\ errato\} \cdot Pr\{bit\ errato/simbolo\ errato\} = P_e \cdot \frac{1}{\log_2 L}$. Ad esempio, con $L = 256$ livelli, la $P_e$ sul bit si riduce di $\log_2 L = 8$ volte. L'espressione della $P_e$ *per bit* nel caso si adotti una codifica di Gray, diviene quindi:

$$P_e^{bit} = \frac{1}{\log_2 L}\left(1 - \frac{1}{L}\right) erfc\left\{\sqrt{\frac{E_b}{N_0}\frac{3\log_2 L}{(L^2-1)\,(1+\gamma)\left(1-\frac{\gamma}{4}\right)}}\right\} \tag{7.11}$$

Le curve in fig. 7.6 mostrano il valore di $P_e^{bit}$ cosí determinato, per $\gamma = 0$, in funzione di $\frac{E_b}{N_0}$ espresso in dB, per diversi valori di $L$. Valori di $\gamma \neq 0$ determinano un peggioramento[52] di $10\log_{10}(1+\gamma)\left(1-\frac{\gamma}{4}\right)$, che deve essere compensato da un eguale incremento in dB di $\frac{E_b}{N_0}$, per ottenere la stessa $P_e$.

**Dimensionamento di una trasmissione numerica** Una metodologia operativa di progetto può basarsi sull'imporre un determinato valore di $P_e$, a partire dal quale

- in base alle curve di fig. 7.6, si individuano i valori di $E_b/N_0$ (in dB) necessari, per diverse scelte di $L$;
- nota la banda disponibile e la velocità $f_b$, si ottiene il valore di $L$, individuando così la curva appropriata, nell'ipotesi di adottare $\gamma = 0$;
- noto il livello di rumore $N_0$, si determina $E_b$;

[51] La probabilità di un errore legato al *salto* di due o più livelli $\theta$ è così piccola da potersi trascurare.
[52] Di non grande entità: per $\gamma = 1$ il peggioramento risulta di 1.761 dB.

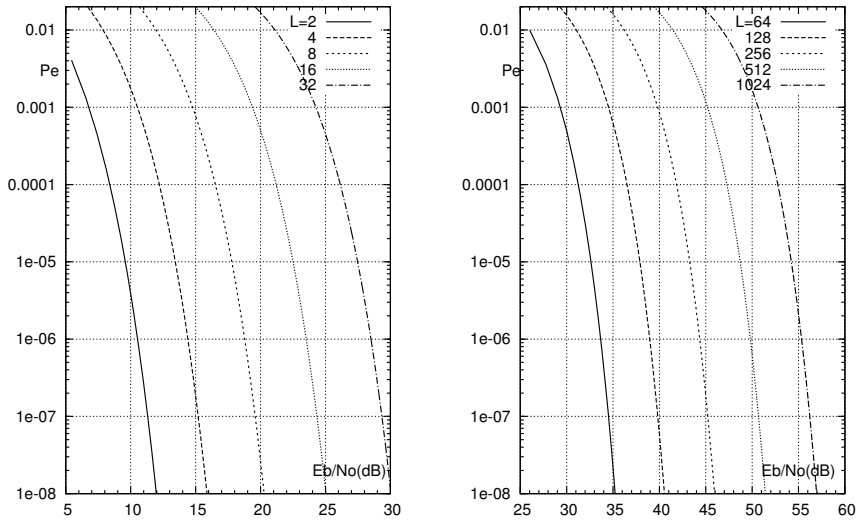



Figura 7.6: Probabilità di errore sul bit per trasmissione multilivello di banda base con codifica di Gray

- note le esigenze di precisione nella temporizzazione, si impone un valore del roll-off $\gamma$, e conseguentemente si aumenta il valore di $E_b$;
- si determina la minima potenza che è necessario ricevere, come $W_{R_{min}} = E_b \cdot f_b$.

**Esercizio** Un sistema di trasmissione basato sul campionamento e sulla trasmissione numerica è rappresentato in figura 7.7. Il canale riportato all'estremità destra è considerato ideale entro una banda $\pm B = \pm 31.5$ KHz, purchè 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 $SNR = \mathcal{P}_x/\mathcal{P}_{z-x} = 10000$. 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 $M$.

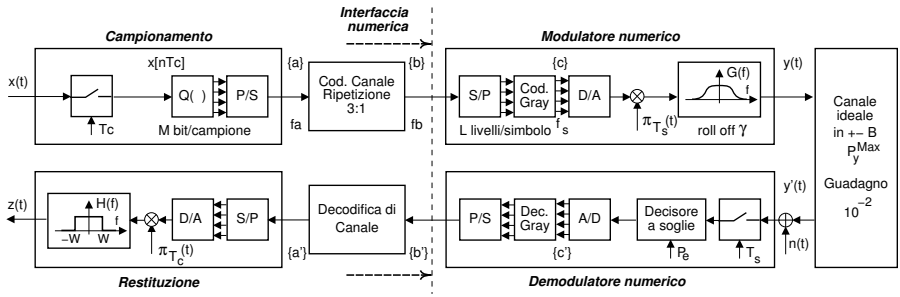



Figura 7.7: Sistema di trasmissione a cui si riferisce l'esercizio

**8)** Determinare la massima banda $W$.

**9)** Se la banda è ridotta a $W' = \frac{1}{2}W$, determinare il nuovo valore di $SNR$ ottenibile.

**Soluzione**

**1)** La banda $B$ occupata dal segnale $y$ vale $B = \frac{f_s}{2}(1+\gamma)$, e quindi deve risultare $f_s = \frac{2B}{1+\gamma} = \frac{2 \cdot 31.5 \cdot 10^3}{1.5} = 42 \cdot 10^3 = 42.000$ baud (*baud = simboli/secondo*).

**2)** Osserviamo che in questo caso la (7.10) 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 $(1+\gamma)\left(1-\frac{\gamma}{4}\right) \backsimeq 1.31$. Inoltre, se $L \gg 1$ (come verificheremo), la (7.10) può essere approssimata come $P_e \simeq erfc\{y\}$, e dunque per $P_e = 10^{-4}$ la figura di pag. 140 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 capitolo seguente, la probabilità di 2 bit errati su 3 è calcolabile dalla distribuzione di Bernoulli, e vale $\binom{3}{2} p_e^2 (1 - p_e) = 3p_e^2 (1 - p_e)$, a cui va sommata la probabilità di 3 bit errati, pari a $p_e^3$. Pertanto $p_e^a = 3p_e^2 (1 - p_e) + p_e^3 = 3p_e^2 - 3p_e^3 + p_e^3 \simeq 3p_e^2$ in cui ovviamente $p_e = P_e^b$, e l'approssimazione è legittima in quanto se $p_e = 10^{-5}$ allora $p_e^2 = 10^{-10}$ e $p_e^3 = 10^{-15}$, trascurabili rispetto a $p_e$. Lo stesso risultato si ottiene osservando che 2 bit errati su 3 hanno probabilità $p_e^2 (1 - p_e)$, e questi possono essere scelti in tre modi diversi ($1^o$ e $2^o$, $1^o$ e $3^o$, $2^o$e $3^o$). 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).

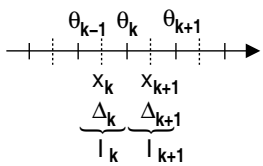
**8)** Come sappiamo, la frequenza di campionamento $f_c = \frac{1}{T_c}$ non può essere inferiore a $2W$; inoltre, la frequenza binaria $f_a$ risulta pari al prodotto dei bit/campione per i campioni a secondo: $f_a = f_c \cdot M$; pertanto $f_c = \frac{f_a}{M} = \frac{140 \cdot 10^3}{7} = 20$ KHz e 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.

## 7.6 Appendici

### 7.6.1 Quantizzazione logaritmica

Abbiamo osservato al termine del § 7.4 che, nel caso in cui il segnale $x$ da quantizzare presenti valori *ridotti* rispetto alla sua dinamica presunta $\Delta_x$, si assiste ad un *peggioramento* di $SNR_q$, in quanto ciò equivale a disporre di un numero di livelli $L = 2^M$ ridotto. Aggiungiamo ora che se il processo $x$ *non ha* distribuzione di ampiezza uniforme come ipotizzato, il risultato $SNR_q \simeq L^2$ non è più valido.

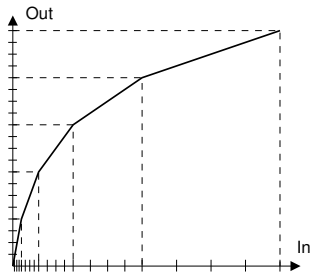



Si può mostrare che, per una generica $p_X(x)$, il quantizzatore *ottimo* (che rende minimo $SNR_q$) *non* suddivide la dinamica $\Delta_x$ in $L-1$ intervalli $I_k$ di *uguale* estensione $\Delta_q = \frac{\Delta_x}{L-1}$, ma utilizza intervalli di estensione $\Delta_q(k)$ *ridotta* nelle regioni per le quali $p_X(x)$ è più grande. Se valori di $x$ più frequenti vengono quantizzati con un errore di potenza $\sigma_q^2(k)$ ridotta rispetto al caso di $\Delta_q$ costante, allora il loro contributo alla potenza totale di errore $\mathcal{P}_\epsilon$ si riduce. Infatti $\mathcal{P}_\epsilon$ si ottiene come valore atteso

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

in cui $p_k = \int_{I_k} p_X(x)\,dx$ è la probabilità che $x \in I_k$. Il modo ottimo di disporre i confini $(\theta_{k-1}, \theta_k)$ che delimitano $I_k$, in modo da rendere minima $\mathcal{P}_\epsilon$, è noto come algoritmo di LLOYD-MAX[53].

Nella quantizzazione del segnale vocale, anche se è arbitrario identificare con esattezza una $p_X(x)$, si verifica strumentalmente che quest'ultima è addensata nelle regioni con valori più piccoli. Per questo motivo, la legge di quantizzazione che si è adottata per ottenere gli 8 bit a campione utilizzati nel PCM segue un andamento logaritmico[54], dimezzando progressivamente la pendenza della caratteristica di ingresso-uscita del quantizzatore.




| *codifica lineare* | *# signif.* | *PCM legge A* |
|---|---|---|
| s0000000wxyza | 5 | s000wxyz |
| s0000001wxyza | 6 | s001wxyz |
| s000001wxyzab | 7 | s010wxyz |
| s00001wxyzabc | 8 | s011wxyz |
| s0001wxyzabcd | 9 | s100wxyz |
| s001wxyzabcde | 10 | s101wxyz |
| s01wxyzabcdef | 11 | s110wxyz |
| s1wxyzabcdefg | 12 | s111wxyz |

La figura mostra un esempio di tale realizzazione (per i soli valori positivi), evidenziando come il risultato possa essere approssimato individuando (a partire dall'origine) regioni di ingresso di ampiezza che man mano raddoppia, e suddividendo la regione in un uguale numero di intervalli equispaziati. La caratteristica non lineare è realizzabile per via completamente numerica: si attua un campionamento con legge lineare, con un numero di livelli molto maggiore del necessario (es con $4096 = 2^{12}$ livelli); i bit più

[53] Il metodo è iterativo, ed inizia suddividendo l'intervallo $\Delta_x$ in modo uniforme. Per ogni iterazione:

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

Le iterazioni si arrestano quando non si riscontrano cambiamenti apprezzabili.

[54] L'andamento esatto della curva segue uno di due standard, denominati legge $\mu$ (per USA e Giappone) e legge $A$ (per gli altri), livemente diverse nella definizione, ma sostanzialmente equivalenti.

significativi del campione individuano la regione di ingresso, ed i bit rimanenti sono *shiftati* a destra, per mantenere costante il numero di intervalli per regione, ottenendo in definitiva una rappresentazione in *virgola mobile* del valore del campione.

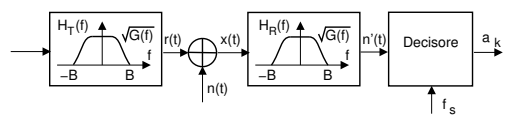
**Esempio** La tabella di destra esemplifica il processo di conversione PCM *legge A*, in cui la codifica lineare è di 13 cifre, e gli 8 bit PCM rappresentano un bit di segno, tre di esponente e quattro di mantissa. Il MSB della codifica lineare rappresenta il segno ($s$), e resta inalterato. Il numero di cifre significative individua in quale degli 8 segmenti cade il valore di ingresso, rappresentato da tre bit nel codice PCM. Infine delle restanti cifre di ingresso, se ne conservano solo le 4 più significative. Ad esempio, il valore 1000000010101, con cinque cifre significative, diviene 1-000-1010 (come risulta dalla prima riga della tabella), mentre 0001100110101 diventa 0-101-1001, come riporta la sesta riga.

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

### 7.6.2 Ricevitore ottimo

Come sarà illustrato al § 9.4.4 in relazione al *filtro adattato*, in presenza di rumore bianco il valore di $SNR$ presente nel punto di decisione *è massimo* se si usa un filtro di ricezione $h_R(t)$ *adattato* alla forma dell'impulso trasmesso $g(t) = h_T(t)$, ovvero per il quale risulti $H_R(f) = G^*(f)$. Nello schema adottato per la figura a pag. 140, il filtro di ricezione possiede invece il solo scopo di limitare la banda del rumore, ed è sempre un passa-basso ideale, indipendente da $g(t)$. Allora, se si adotta una $G(f)$ di Nyquist *non* a banda minima, i campioni di rumore che, sovrapposti a quelli di segnale, danno luogo alle v.a. $x(kT_s)$ gaussiane ma non più indipendenti[55], e 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 le condizioni $H(f) = G^*(f)$ di filtro adattato, decomponendo la caratteristica di Nyquist $G(f)$ in parti uguali tra trasmettitore e ricevitore, e dando quindi luogo allo schema di figura, in cui




$$H_T(f) = H_R(f) = \sqrt{G(f)}$$

[55] Infatti, il segnale $n(t)$ uscente da $H_R(f) = rect_{2B}(f)$ possiede autocorrelazione $\mathcal{R}_N(\tau) = \mathcal{F}^{-1}\{|H(f)|^2\} = 2B\text{sinc}(2B\tau)$ (vedi § 9.2.3), che passa da zero per $\tau = \frac{1}{2B}$. Se si utilizza una $G(f)$ a coseno rialzato con $\gamma > 0$, occorre estendere la banda di ricezione a $B = \frac{f_s}{2}(1+\gamma)$, a cui corrisponde l'incorrelazione tra campioni di rumore prelevati a distanza multipla di $\tau = \frac{1}{2B} = \frac{1}{f_s(1+\gamma)}$. Invece, il rumore è campionato con frequenza pari a quella di simbolo $fs$, 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] Infatti, la dipendenza statistica tra i campioni di rumore, permetterebbe di realizzare un dispositivo *predittore lineare* che, in base alla conoscenza dei precedenti valori di rumore, calcoli una stima del valore corrente la quale, sottratta al valore effettivamente ricevuto, consente di ridurre la varianza della grandezza di osservazione, permettendo una riduzione della probabilità di errore.

Pur non entrando nei dettagli dei metodi di predizione lineare, notiamo che la *correlazione* tra grandezze aleatorie ne determina la *dipendenza*, e che la conoscenza di valori passati consente di ridurre l'incertezza relativa ai nuovi valori. Il valore dei campioni di rumore precedenti, si calcola a partire da quello del simbolo *deciso* senza commettere errore, sottratto al valore del segnale ricevuto in quell'istante.

In tal modo, al decisore giunge esattamente lo stesso segnale di prima[57], mentre la densità di potenza del rumore non è più costante, ma ora vale

$$\mathcal{P}_N(f) = \frac{N_0}{2}|H_R(f)|^2 = \frac{N_0}{2}G(f)$$

Pertanto, i campioni di rumore presi a distanza $T_s$ sono incorrelati, in quanto $\mathcal{R}_N(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_N(f)\}$ è ora un impulso di Nyquist, che passa da zero per $\tau = kT_s$. Notiamo che, essendo $G(f)$ reale pari, la fattorizzazione di $G(f)$ realizza effettivamente la condizione $H_R(f) = H_T^*(f)$ che definisce un filtro adattato.

Per determinare le nuove prestazioni nel caso in cui $G(f)$ sia a coseno rialzato, notiamo che mentre la banda passante di $H_R(f)$ (e dunque del rumore) si è mantenuta pari a $B = \frac{f_s}{2}(1+\gamma)$, la potenza del rumore ora vale[58]

$$\mathcal{P}_N = \int_{-\infty}^{\infty} \frac{N_0}{2}G(f)\,df = \frac{N_0}{2}f_s = \frac{N_0}{2T_b \log_2 L}$$

riducendosi di un fattore $(1+\gamma)$ se confrontata con (7.6), e causando un aumento equivalente per l'$SNR$; lo stesso fattore $(1+\gamma)$ è quindi rimosso anche nella (7.11), portando a

$$P_e^{bit} = \frac{1}{\log_2 L}\left(1 - \frac{1}{L}\right) erfc\left\{\sqrt{\frac{E_b}{N_0}\frac{3\log_2 L}{(L^2-1)\left(1-\frac{\gamma}{4}\right)}}\right\}$$

il valore della probabilità di errore sul bit adottando il ricevitore ottimo ed il codice di Gray. Dato che al massimo $1+\gamma = 2$, questo corrisponde ad un *miglioramento massimo* di 3 dB nel valore di $E_b/N_0$, permettendo di usare ancora le curve di fig. 7.6.

**Conseguenze** L'adozione di un filtro di trasmissione $H_T(f) = \sqrt{G(f)}$ comporta che ora nel segnale trasmesso è presente ISI, che può essere rimossa solo mediante filtraggio dello stesso attraverso il filtro adattato $H_R(f) = \sqrt{G(f)}$.

In figura 7.8 si mostra l'andamento di $g^{\surd}(t) = \mathcal{F}^{-1}\left\{\sqrt{G(f)}\right\}$, posto a confronto con una $g(t)$ a coseno rialzato, per valori di roll-off pari a 0.5 ed 1, ottenuto mediante IFFT della corrispondente risposta in frequenza di modulo unitario nell'origine. Notiamo un aumento sia della durata che della ampiezza delle oscillazioni: questa circostanza determina una maggiore complessità realizzativa del filtro di trasmissione, che deve avere una risposta impulsiva più lunga[59].

Una seconda considerazione può essere svolta a riguardo del caso in cui sia presente un canale di trasmissione con risposta in frequenza $H_C(f)$ non ideale; in tal

[57] Infatti, se $G(f)$ è tutta al trasmettitore, il segnale generato (e ricevuto) ha espressione (7.3) (vedi anche la (5.1)); indicando ora $g^{\surd}(t) = \mathcal{F}^{-1}\left\{\sqrt{G(f)}\right\}$, ed eseguendo un calcolo del tutto analogo a quello svolto in § 5.1.2.2, si ottiene che il segnale ricevuto nel caso di scomposizione di $G(f)$ ha espressione

$$r(t) = h_T(t) * h_R(t) * \sum_k a_k \cdot \delta(t - kT_s) = \sum_k a[k] \cdot g(t - kT_s)$$

in quanto $h_T(t) * h_R(t) = g^{\surd}(t) * g^{\surd}(t) = g(t)$ per la proprietà di prodotto in frequenza.

[58] Il risultato si può ottenere visivamente, a partire dalla $G(f)$ a coseno rialzato mostrata in fig. 5.3 a pag. 73, considerando la risposta di ampiezza nominale pari ad 1, e in base alle sue proprietà di simmetria attorno ad $f_s$: non è nient'altro che l'area di un rettangolo.

[59] Per una analisi degli effetti della limitazione temporale dell'impulso $g^{\surd}(t)$, vedere l'analisi svolta presso `https://engineering.purdue.edu/~ee538/SquareRootRaisedCosine.pdf`.

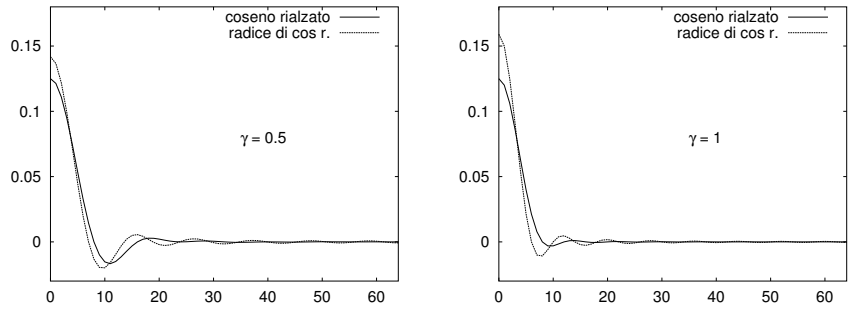



Figura 7.8: Confronto della risposta impulsiva del filtro ottimo e subottimo

caso, occorre realizzare un filtro di trasmissione $H_T(f)$ tale che $H_T(f) H_C(f) = \sqrt{G(f)}$. Spesso l'equalizzazione è invece svolta al lato ricevente, in modo da ottenere $H_{eq}(f) H_C(f) = \sqrt{G(f)}$, ma in tal caso si ottiene una soluzione solamente *sub-ottima*, dato che si perde l'incorrelazione dei campioni di rumore.

### 7.6.3 Funzione caratteristica

E' definita come l'antitrasformata di Fourier di una densità di probabilità, ovvero (equivalentemente) come il valore atteso di $e^{jwz}$:

$$\Phi_z(w) = \mathcal{F}^{-1}\{p_Z(z)\} = E_Z\left\{e^{jwz}\right\} = \int p_Z(z) e^{jwz} dz$$

Osserviamo che, se applicata alla somma di v.a. indipendenti, si ha:

$$\begin{aligned}\Phi_z(w) &= E_Z\left\{e^{jw(x+y)}\right\} = E_Z\left\{e^{jwx}e^{jwy}\right\} = E_X\left\{e^{jwx}\right\} E_Y\left\{e^{jwy}\right\} \\ &= \Phi_x(w)\,\Phi_y(w)\end{aligned}$$

ovvero la funzione caratteristica di una somma di v.a. indipendenti è pari al prodotto delle funzioni caratteristiche.

Effettuando ora l'operazione inversa (trasformata di Fourier della funzione caratteristica della somma) e ricordando che un prodotto in frequenza è una convoluzione nel tempo (e viceversa) si ottiene il risultato $p_Z(z) = \mathcal{F}\{\Phi_z(w)\} = \mathcal{F}\{\Phi_x(w)\,\Phi_y(w)\} = p_X(x) * p_Y(y)$ che ci permette di enunciare:

> *La densità di probabilità della somma di v.a. indipendenti è pari alla convoluzione tra le rispettive densita' di probabiltà marginali.*

La funzione caratteristica ha altri usi... ma non approfondiamo oltre.

### 7.6.4 Trasformazioni di v.a. e cambio di variabili

Quando più v.a. si combinano con leggi diverse dalla somma, il risultato del § precedente non è più sufficiente a fornire una espressione per la d.d.p. risultante. Illustriamo allora il procedimento analitico generale, necessario ad ottenere una espressione per la d.d.p. di generiche funzioni di v.a.

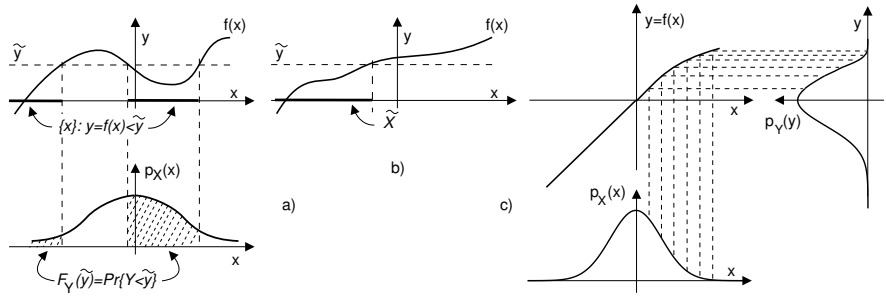



Figura 7.9: Trasformazioni tra variabili aleatorie

#### 7.6.4.1 Caso unidimensionale

Consideriamo una prima v.a. $X$, ed una seconda $Y$ da essa derivata per mezzo della relazione $y = f(x)$, che si applica alle determinazioni di $x$ di $X$. La caratterizzazione statistica di $Y$ in termini della sua d.d.p. $p_Y(y)$ può ottenersi a partire da quella di $X$, nei termini della funzione di distribuzione di $Y$, come:

$$F_Y(y) = Pr\{Y \leq y\} = Pr\left\{X \leq f^{-1}(x)\right\} \tag{7.12}$$

e calcolando poi $p_Y(y) = \frac{dF_Y(y)}{dy}$.

L'espressione (7.12) può dare luogo a risultati più o meno usabili a seconda della natura della trasformazione $f(x)$: in fig. 7.9a) troviamo un esempio in cui i valori $y \leq \tilde{y}$ hanno origine da due diversi intervalli di $X$; in corrispondenza di questi, l'area sottesa dalla $p_X(x)$ individua la probabilità cercata.

Nel caso in cui $f(x)$ sia monotona crescente come in fig. 7.9 b), per ogni valore di $\tilde{y}$ esiste un solo intervallo di $\tilde{X} \subset X$ tale che $y = f\left(x \in \tilde{X}\right) \leq \tilde{y}$, e la (7.12) può essere riscritta come

$$F_Y(y) = Pr\left\{X \leq f^{-1}(x)\right\} = F_X\left(x = f^{-1}(y)\right)$$

che, derivata, permette di giungere alla espressione di calcolo della $p_Y(y)$:

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(x)}{dx}\frac{df^{-1}(y)}{dy} = p_X\left(f^{-1}(y)\right)\frac{df^{-1}(y)}{dy} \tag{7.13}$$

La (7.13) può essere interpretata a parole, osservando che la nuova v.a. $y = f(x)$ possiede una d.d.p. pari a quella di $x$, calcolata con argomento pari alla funzione inversa $x = f^{-1}(y)$, moltiplicata per la derivata di $f^{-1}(y)$. La d.d.p. risultante da una trasformazione di v.a. si presta inoltre ad un processo di costruzione grafica, come esemplificato nella fig. 7.9 c).

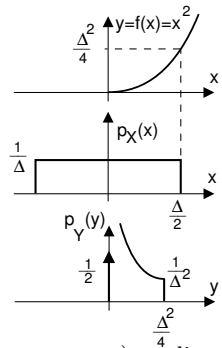
**Esempio** Determinare $p_Y(y)$, qualora risulti $y = f(x) = \begin{cases} 0 & \text{con } x \leq 0 \\ x^2 & \text{con } x > 0 \end{cases}$, nel caso in cui $p_X(x) = \frac{1}{\Delta} rect_\Delta(x)$.

Osserviamo innanzitutto che tutte le determinazioni $x \leq 0$ danno luogo ad un unico valore $y = 0$; pertanto si ottiene $p_Y(0) = \frac{1}{2}\delta(y)$.

Per $0 < y \leq \frac{\Delta^2}{4}$ (corrispondente ad $0 < x \leq \frac{\Delta}{2}$) si applica la teoria svolta, ottenendo $F_Y(y) = Pr\{x \leq \sqrt{y}\} = F_X(\sqrt{y})$, e dunque

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(x)}{dx}\Big|_{x=\sqrt{y}} \frac{d(x=\sqrt{y})}{dy} = \frac{1}{\Delta}\frac{1}{2\sqrt{y}}$$

in cui l'ultima eguaglianza tiene conto che $\frac{dF_X(x)}{dx} = p_X(x)$, che vale $\frac{1}{\Delta}$ per tutti gli $x$ nell'intervallo in considerazione. L'ultima curva mostra la d.d.p risultante per questo esempio.




### 7.6.4.2 Caso multidimensionale

Descriviamo questo caso per mezzo del vettore di v.a. $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, a cui è associata una d.d.p. congiunta $p_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$, e di un secondo vettore aleatorio $\mathbf{Y}$ dipendente dal primo mediante la trasformazione $\mathbf{Y} = \mathbf{G}(\mathbf{X})$, ovvero

$$\begin{cases} y_1 = g_1(x_1, x_2, \ldots, x_n) \\ y_2 = g_2(x_1, x_2, \ldots, x_n) \\ \qquad\vdots \\ y_n = g_n(x_1, x_2, \ldots, x_n) \end{cases}$$

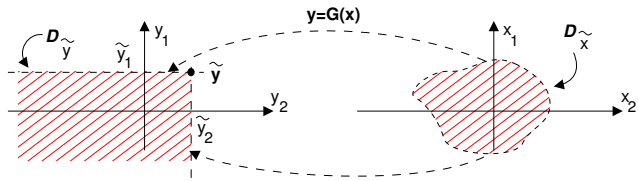
Se esiste la relazione inversa $\mathbf{X} = \mathbf{G}^{-1}(\mathbf{Y})$, composta dall'insieme di funzioni $x_i = g_i^{-1}(y_1, y_2, \ldots, y_n)$ per $i = 1, 2, \cdots, n$, allora per la d.d.p di $\mathbf{Y}$ sussiste[60] un risultato formalmente molto simile a quello valido nel caso monodimensionale, e cioè

$$p_{\mathbf{Y}}(y_1, y_2, \ldots, y_n) = p_{\mathbf{X}}\left(\mathbf{X} = \mathbf{G}^{-1}(\mathbf{Y})\right) \cdot |\mathbf{J}(\mathbf{X}/\mathbf{Y})|$$

in cui $p_{\mathbf{X}}\left(\mathbf{x} = \mathbf{G}^{-1}(\mathbf{y})\right)$ è la d.d.p. di $\mathbf{X}$ calcolata con argomento dipendente da $\mathbf{Y}$, e $|\mathbf{J}(\mathbf{X}/\mathbf{Y})|$ è il *giacobiano* della trasformazione inversa $\mathbf{G}^{-1}$, ossia il determinante della

[60] La dimostrazione segue le medesime linee guida del caso precedente, ed è impostata sulla base della considerazione che la funzione di distribuzione di $\mathbf{Y}$, calcolata in un generico punto $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)$, rappresenta la probabilità che $\mathbf{Y}$ appartenga alla regione (dominio) delimitata dal punto $\tilde{\mathbf{y}}$, indicata con $\mathcal{D}_{\tilde{\mathbf{y}}}$:

$$F_{\mathbf{Y}}(\tilde{\mathbf{y}}) = Pr\{\mathbf{Y} \leq \tilde{\mathbf{y}}\} = Pr\{\mathbf{Y} \in \mathcal{D}_{\tilde{\mathbf{y}}}\}$$




Alla stessa regione $\mathcal{D}_{\tilde{\mathbf{y}}}$, ne corrisponde una diversa $\mathcal{D}_{\tilde{\mathbf{x}}}$ nello spazio $\mathbf{X}$, tale che per ogni valore $\mathbf{x}^{\diamond} \in \mathcal{D}_{\tilde{\mathbf{x}}}$ risulti $\mathbf{y}^{\diamond} = \mathbf{G}(\mathbf{x}^{\diamond}) \in \mathcal{D}_{\tilde{\mathbf{y}}}$. Con queste posizioni, la $F_{\mathbf{Y}}(\tilde{\mathbf{y}}) = Pr\{\mathbf{Y} \in \mathcal{D}_{\tilde{\mathbf{y}}}\}$ si calcola a partire dalla d.d.p. $p_{\mathbf{X}}(\mathbf{x})$, integrata sul dominio $\mathcal{D}_{\tilde{\mathbf{x}}}$:

$$F_{\mathbf{Y}}(\tilde{\mathbf{y}}) = Pr\{\mathbf{X} \in \mathcal{D}_{\tilde{\mathbf{x}}}\} = \int_{\mathcal{D}_{\tilde{\mathbf{x}}}} p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

Infine, osservando che

$$p_{\mathbf{Y}}(y_1, y_2, \ldots, y_n) = \frac{\partial^n F_{\mathbf{Y}}(y_1, y_2, \ldots, y_n)}{\partial y_1 \partial y_2 \cdots \partial y_n}$$

si ottiene il risultato mostrato.

matrice costituita da tutte le sue derivate parziali:

$$\mathbf{J}(\mathbf{X}/\mathbf{Y}) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

**Variabile aleatoria di Rayleigh** Come applicazione della teoria sopra illustrata, consideriamo la relazione che lega il modulo $\rho$ e la fase $\varphi$ di un numero complesso $z = x + jy$, alle sue parti reale ed immaginaria, assieme alle corrispondenti funzioni inverse:

$$\begin{cases} \rho = \sqrt{x^2 + y^2} \\ \varphi = \arctan \frac{y}{x} \end{cases} \qquad \begin{cases} x = \rho \cos \varphi \\ y = \rho \sin \varphi \end{cases} \tag{7.14}$$

Nel caso in cui $x$ ed $y$ siano due v.a. gaussiane indipendenti, a media nulla e uguale varianza $\sigma^2$, la d.d.p. congiunta di $(x, y)$ si ottiene come prodotto delle d.d.p. marginali, e vale

$$p_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{7.15}$$

La $p_{P,\Phi}(\rho, \varphi)$ si ottiene come descritto, valutando[61] i due termini $p_{X,Y}(x(\rho, \varphi), y(\rho, \varphi))$ e $|\mathbf{J}(x, y/\rho, \varphi)|$, ed ottenendo quindi

$$p_{P,\Phi}(\rho, \varphi) = \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \quad \text{con} \quad \begin{cases} 0 & < \rho < & \infty \\ -\pi & < \varphi < & \pi \end{cases}$$

Le d.d.p. marginali $p_P(\rho)$ e $p_\Phi(\varphi)$ si ottengono saturando[62] la d.d.p. congiunta rispetto all'altra variabile, ricavando

$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \text{ con } \rho \geq 0;\ p_\Phi(\varphi) = \frac{1}{2\pi} \text{ con } -\pi < \varphi \leq \pi \tag{7.16}$$

L'espressione di $p_P(\rho)$ in (7.16) prende nome di d.d.p. di RAYLEIGH, graficata appresso, mentre il valor medio e la varianza della v.a. $\rho$ valgono rispettivamente

$$m_P = \sigma\sqrt{\frac{\pi}{2}} \quad \text{e} \quad \sigma_P^2 = \sigma^2\left(2 - \frac{\pi}{2}\right)$$

---

[61] Il calcolo dei due termini si esegue come

$$p_{X,Y}(x(\rho, \varphi), y(\rho, \varphi)) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\rho^2(\cos^2\varphi + \sin^2\varphi)}{2\sigma^2}\right) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

$$|\mathbf{J}(x, y/\rho, \varphi)| = \left| \begin{bmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \varphi} \end{bmatrix} \right| = \left| \begin{bmatrix} \cos\varphi & -\rho\sin\varphi \\ \sin\varphi & \rho\cos\varphi \end{bmatrix} \right| = \rho\left(\cos^2\varphi + \sin^2\varphi\right) = \rho$$

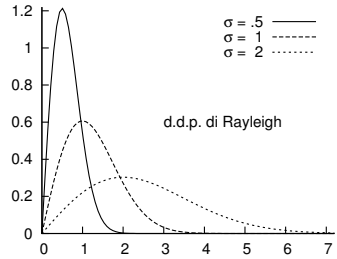
[62] Svolgiamo il calcolo solo per la prima relazione:

$$p_P(\rho) = \int_{-\pi}^{\pi} p_{P,\Phi}(\rho, \varphi)\, d\varphi = \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \cdot \int_{-\pi}^{\pi} d\varphi = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

E' inoltre possibile mostrare che per essa vale la proprietà

$$Pr\{\rho > \gamma\} = \int_{\gamma}^{\infty} p_P(\rho)\, d\rho = \exp\left(-\frac{\gamma^2}{2\sigma^2}\right) \tag{7.17}$$

Quest'ultimo valore può rappresentare la probabilità di *mancare un bersaglio* per una distanza superiore a $\gamma$, nell'ipotesi che gli errori di puntamento orizzontale e verticale siano entrambi gaussiani, indipendenti, a media nulla ed uguale varianza.

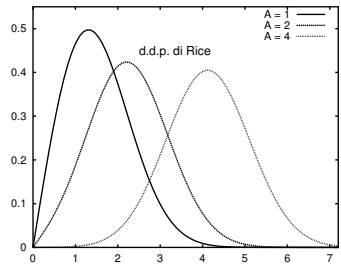



**Variabile aleatoria di Rice** Consideriamo nuovamente la trasformazione (7.14), in cui si considera, al posto di $x$, la v.a. $x'$, somma di una v.a. gaussiana $x$ e di una costante $A$. In tal caso, la d.d.p. $p_P(\rho)$ è detta di RICE, ed ha[63] espressione

$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{\rho A}{\sigma^2}\right) \text{ per } \rho \geq 0 \tag{7.18}$$

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, 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 $z \gg 1$.

Di lato è mostrato l'andamento di $p_P(\rho)$ con $\sigma = 1$ e tre diversi valori di $A$, che possiamo rapportare a quelle del secondo grafico per la d.d.p. di Rayleigh, ottenuto per lo stesso valore di $\sigma$. Notiamo infine che per $A = 0$ si torna al caso di Rayleigh, mentre per valori crescenti di $A$, l'andamento della d.d.p. di Rice approssima sempre più quello di una gaussiana.




### 7.6.5 Detezione di sinusoide nel rumore

Applichiamo i principi della decisione statistica ad una caso "classico", mostrando come la teoria ora svolta si applichi allo schema di demodulazione incoerente mostrato al

[63] Osserviamo innanzitutto che d.d.p. congiunta di partenza si scrive ora come

$$p_{X',Y}(x', y)\, \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x' - A)^2 + y^2}{2\sigma^2}\right)$$

in quanto $x'$ è una v.a. gaussiana con media $A$ e varianza $\sigma^2$. Sostituendo quindi nell'esponente $x' = \rho\cos\varphi$ e $y = \rho\sin\varphi$, si ottiene

$$(x' - A)^2 + y^2 = \rho^2\cos^2\varphi + A^2 - 2\rho A\cos\varphi + \rho^2\sin^2\varphi = \rho^2 + A^2 - 2\rho A\cos\varphi$$

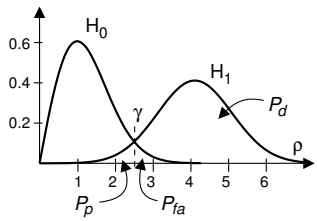
Osservando ora che il giacobiano della trasformazione ha un valore pari a $\rho$ anche in questo caso, otteniamo

$$\begin{aligned} p_{P,\Phi}(\rho, \varphi) &= p_{X',Y}\left(x'(\rho,\varphi), y = y(\rho,\varphi)\right) \left|\mathbf{J}\left(x', y/\rho, \varphi\right)\right| \\ &= \frac{\rho}{2\pi\sigma^2} \exp\left(-\frac{\rho^2 + A^2}{2\sigma^2}\right) \exp\left(\frac{2\rho A\cos\varphi}{2\sigma^2}\right) \end{aligned}$$

La saturazione di questa d.d.p. congiunta, operata eseguendo $p_P(\rho) = \int_{-\pi}^{\pi} p_{P,\Phi}(\rho, \varphi)\, d\varphi$, determina il risultato mostrato.

§ 11.2.1.2. Qualora in ingresso al demodulatore sia presente un rumore di potenza $\sigma^2$, pari a quella delle singole componenti analogiche, ed a questo sia sovrapposta una sinusoide di ampiezza $A$, per il modulo dell'inviluppo complesso $\rho = |\underline{x}(t)|$ valgono esattamente le considerazioni svolte per la v.a. di Rice.

La ricezione del rumore dà infatti luogo (vedi § 12.1.2) a due processi di rumore $x_c(t)$ ed $x_s(t)$ gaussiani ed ergodici, che in assenza di sinusoide hanno media nulla e varianza $\sigma^2$, e quindi la d.d.p. di $\rho$ assume un andamento di Rayleigh, come mostrato alla figura seguente, per il caso indicato con $H_o$, e relativo a $\sigma = 1$. La curva $H_1$ mostra invece la d.d.p. di $\rho$ nel caso di presenza di segnale, con andamento di Rice, per $A = 4$ e $\sigma = 1$.




Il problema nasce qualora il tono sinusoidale non sia presente con continuità, e si desideri operare una decisione relativa all'ipotesi di una sua presenza (indicata con $H_1$) od assenza ($H_0$). Nel caso in cui $A$ e $\sigma$ siano note, il problema è ben posto, e si riduce a determinare una soglia $\gamma$ con cui confrontare $\rho$, e decidere per $H_0$ od $H_1$ nei casi in cui $\rho < \gamma$ e $\rho > \gamma$ rispettivamente. In tal caso, vengono definiti tre possibili eventi, assieme alle rispettive probabilità, che dipendono dal valore assegnato a $\gamma$:

- Probabilità di detezione $P_d = \int_\gamma^\infty H_1(\rho)\, d\rho$
- Probabilità di falso allarme $P_{fa} = \int_\gamma^\infty H_0(\rho)\, d\rho$
- Probabilità di perdita $P_p = \int_0^\gamma H_1(\rho)\, d\rho$

in cui gli ultimi due valori sono riferiti ad eventi di errore.

La nomenclatura adottata è quella tipica dei radar; in tale contesto, può aver senso tentare di spostare $\gamma$ in modo da favorire l'uno o l'altro evento in base a considerazioni strategiche[64]. Nel caso in cui i due errori siano equivalenti, e se le probabilità a priori di $H_0$ ed $H_1$ sono uguali, la probabilità di errore

$$P_e = Pr(H_0)\, Pr(e/H_0) + Pr(H_1)\, Pr(e/H_1) = \frac{1}{2}P_{fa} + \frac{1}{2}P_p \qquad (7.19)$$

risulta minimizzata qualora si adotti una decisione di *massima verosimiglianza*, ponendo $\gamma$ nel punto in cui le due curve si intersecano (come in figura), in modo da preferire l'ipotesi $i$ la cui probabilità a posteriori $Pr(H_i/\rho)$ è più grande.

Una valutazione asintotica delle prestazioni può essere svolta notando che all'aumentare di $\frac{A}{\sigma}$, il valore di $\gamma$ si avvicina (da destra) ad $\frac{A}{2}$; ponendo quindi $\gamma = \frac{A}{2}$ e sostituendo le espressioni (7.16) e (7.18) delle d.d.p. a posteriori in quella (7.19) della $P_e$, otteniamo

$$P_e = \frac{1}{2}\int_{\frac{A}{2}}^\infty \frac{\rho}{\sigma^2}\exp\left(-\frac{\rho^2}{2\sigma^2}\right)d\rho + \frac{1}{2}\int_0^{\frac{A}{2}} \frac{\rho}{\sigma^2}\exp\left(-\frac{\rho^2+A^2}{2\sigma^2}\right) I_0\left(\frac{\rho A}{\sigma^2}\right)d\rho \qquad (7.20)$$

[64] In ambito militare, può aver senso ridurre la probabilità di perdita, a spese di un aumento di quella di falso allarme, tranne nel caso in cui quest'ultimo non provochi conseguenze del tutto irreversibili, e "sbagliate" in caso di errore. Ragionamenti analoghi possono essere svolti in campo medico, in cui si dovrebbe preferire un falso allarme, piuttosto che trascurare l'importanza di un sintomo.

Per ciò che riguarda il primo termine, applicando il risultato (7.17) si trova che il valore $\int_{\frac{A}{2}}^{\infty} \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) d\rho = \exp\left(-\frac{A^2}{8\sigma^2}\right)$. Per il secondo termine, osserviamo che il suo valore è ben più piccolo del primo (si veda la figura tracciata per $A = 4$, o le considerazioni riportato in nota[65]), e quindi può essere trascurato, fornendo in definitiva

$$P_e = \frac{1}{2} \exp\left(-\frac{A^2}{8\sigma^2}\right)$$

per $\frac{A}{\sigma} \gg 1$. Notiamo infine che $\frac{A^2}{2}$ rappresenta la potenza della sinusoide, e che $\sigma^2$ è la potenza del rumore. Pertanto, il risultato trovato ha una immediata interpretazione in termini di $SNR = \frac{A^2/2}{\sigma^2}$:

$$P_e = \frac{1}{2} \exp\left(-\frac{SNR}{4}\right)$$

[65] Come già osservato, la funzione modificata di Bessel può essere approssimata come $I_0\left(\frac{\rho A}{\sigma^2}\right) \sim \frac{\exp\left(\frac{\rho A}{\sigma^2}\right)}{\sqrt{2\pi \frac{\rho A}{\sigma^2}}}$ per $\frac{\rho A}{\sigma^2} \gg 1$, e quindi il secondo termine di (7.20) diviene

$$\begin{aligned} & \frac{1}{2} \int_0^{\frac{A}{2}} \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \exp\left(-\frac{A^2}{2\sigma^2}\right) \exp\left(\frac{\rho A}{\sigma^2}\right) \frac{\sigma}{\sqrt{2\pi\rho A}} d\rho = \\ = \quad & \frac{1}{2} \int_0^{\frac{A}{2}} \sqrt{\frac{\rho}{2\pi\sigma^2 A}} \exp\left(\frac{(\rho - A)^2}{2\sigma^2}\right) d\rho \end{aligned}$$

Notiamo ora che per $\frac{A}{\sigma} \gg 1$ l'integrando è trascurabile tranne che per valori di $\rho$ vicini ad $A$, cosicchè il limite inferiore può essere esteso a $-\infty$, e il termine $\sqrt{\frac{\rho}{2\pi\sigma^2 A}}$ può essere sostituito con $\frac{1}{\sqrt{2\pi}\sigma}$, ottenendo quindi la maggiorazione $\frac{1}{2} \int_{-\infty}^{\frac{A}{2}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{\rho^2 - A^2}{2\sigma^2}\right) d\rho$, che è l'espressione dell'integrale di una gaussiana con media $A^2$, e che può essere espresso come $\frac{1}{2} \cdot \frac{1}{2} erfc\left(\frac{A}{2\sqrt{2}\sigma}\right)$.

Se $z \gg 1$, $\frac{1}{2} erfc(z)$ può essere approssimata come $\frac{1}{2z\sqrt{\pi}} \exp\left(-z^2\right)$, ottenendo

$$\frac{1}{2} \int_{-\infty}^{\frac{A}{2}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{\rho^2 - A^2}{2\sigma^2}\right) d\rho \sim \frac{\sigma}{A\sqrt{2\pi}} \exp\left(-\frac{A^2}{8\sigma^2}\right)$$

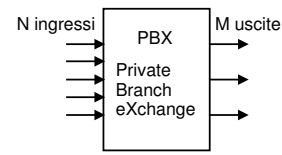
che, essendo per ipotesi $\frac{A}{\sigma} \gg 1$, risulta trascurabile rispetto a $\frac{1}{2} \exp\left(-\frac{A^2}{8\sigma^2}\right)$.

# Capitolo 8

# Traffico, code e reti a pacchetto

Qui trovano spazio argomenti legati alle trasmissioni dati, a carattere sia teorico che estremamente pratico. Gli aspetti teorici si fondano sui concetti di probabilità affrontati al capitolo precedente, e definiscono le relazioni che descrivono la probabilità di blocco nei sistemi di servizio orientati alla perdita, nei casi di popolazione finita ed infinita, e le grandezze tipiche per i sistemi orientati al ritardo, come nel caso di coda infinita e servente unico. Dal punto di vista pratico, invece, sono fornite definizioni e concetti relativi alle reti per trasmissione dati, con particolare riguardo alle reti a pacchetto, nei termini dei diversi modi di funzionamento e delle gerarchie di protocolli adottate, come IP, Ethernet e ATM.

## 8.1 Distribuzione binomiale per popolazione finita




Iniziamo con il chiederci quante linee uscenti $M$ siano necessarie ad un centralino con $N$ interni, in modo che la probabilità di trovare tutte le linee occupate sia inferiore ad un valore massimo, chiamato *grado di servizio*[1]. Per trovare il risultato, calcoliamo prima la probabilità che tutte le linee uscenti siano occupate, assumendo noti $N$ ed $M$.

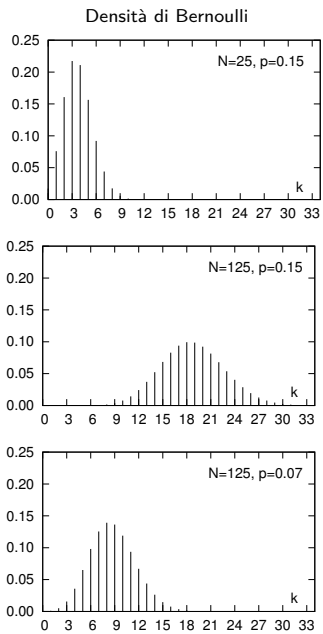
Affrontiamo il problema in termini ancor più generali, chiedendoci quale sia la probabilità $p_B(k)$ che un numero $k$ di persone (su $N$) sia contemporaneamente al telefono. Assumiamo che ognuno degli $N$ interni abbia una probabilità $p$ di telefonare, ossia passi il $p \cdot 100\%$ del suo tempo al telefono, e che le telefonate siano statisticamente indipendenti. Allora, ci saranno in media $Np$ telefoni occupati, e la probabilità che un ben preciso gruppo di $k$ individui telefoni (e $N-k$ no), è pari a $p^k q^{N-k}$ (in cui $q = 1-p$).

Dato che il numero di differenti modi di scegliere $k$ oggetti tra $N$ è pari a $\binom{N}{k} = \frac{N!}{k!(N-k)!} = \frac{N(N-1)\cdots(N-k+1)}{k!}$, allora la probabilità di avere $k$ (qualsiasi) persone al

[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. Nel caso presente, una buona qualità corrisponde a una bassa probabilità di occupato.

telefono è pari a

$$p_B(k) = \binom{N}{k} p^k q^{N-k}$$




Dato che $\sum_{k=0}^{N} p_B(k) = 1$, la funzione $p_B(k)$ è una densità di probabilità di v.a. discreta, nota con il nome di variabile aleatoria di *Bernoulli* [2].

Al variare di $k$, si ottengono tutte le probabilità cercate, rappresentate nella figura a lato nel caso in cui $p = 0.15$ e $N = 25$, oppure $N = 125$; nel secondo caso, si utilizza anche il valore $p = 0.07$, che causa una concentrazione di $p_B(k)$ attorno a valori $k$ inferiori. Inoltre, osserviamo che non si possono avere più di $N$ utenti al telefono.

Per conoscere il numero di linee occorrenti a garantire una probabilità di *congestione* (o di blocco) $P_B$ inferiore ad un massimo, si sommano (partendo *da destra*) i valori di probabilità $p_B(k)$, finché non si supera la probabilità prefissata: allora $M$ sarà pari all'ultimo indice $k$. Infatti in tal modo la probabilità che ci siano più di $M$ interni a voler telefonare è pari a

$$Pr(k > M) = \sum_{k=M+1}^{N} p_B(k) = \sum_{k=M+1}^{N} \binom{N}{k} p^k q^{N-k} < P_B$$

La distribuzione Binomiale è detta anche *delle prove ripetute* poiché può essere usata per calcolare la probabilità di un certo numero di eventi favorevoli, a seguito della ripetizione dello stesso fenomeno aleatorio[3].

Il valore medio della distribuzione Binomiale è $m_B = Np$, e la varianza $\sigma_B^2 = Npq$. Tornando al caso del centralino, il numero medio di linee occupate è $Np$: tale quantità rappresenta il *traffico offerto medio*, che si misura in ERLANG: ad esempio, un traffico medio di 3 Erlang corrisponde ad osservare in media 3 linee occupate.

Il rapporto $\frac{\sigma_B^2}{m_B} = \frac{Npq}{Np} = q < 1$ è un indice di come la variabile aleatoria si distribuisce attorno alla media. Il caso di Bernoulli in cui $\frac{\sigma_B^2}{m_B} < 1$ è rappresentativo di un traffico *dolce*, che deriva dall'ipotesi di popolazione finita, e che si sostanzia nel

[2] La $p_B(k)$ è detta anche *Binomiale*, in quanto i fattori $\binom{N}{k}$ sono quelli della potenza di un binomio $(p+q)^N$, calcolabili anche facendo uso del triangolo di *Pascal* (ma definito prima da *Tartaglia*, e prima ancora da *Hayyām*), mostrato per riferimento di seguito.

```
                1
              1   1
            1   2   1
          1   3   3   1
        1   4   6   4   1
      1   5  10  10   5   1
    1   6  15  20  15   6   1
```

[3] Esempio: si voglia calcolare la probabilità di osservare 3 volte testa, su 10 lanci di una moneta. Questa risulta pari a $p_B(3) = \binom{10}{3} p^3 q^7 = 120 \cdot .5^3 \cdot .5^7 = 0.117$, ovvero una probabilità dell'11,7 %. Come ulteriore esempio, citiamo l'uso della distribuzione Binomiale per calcolare la probabilità di errore complessiva in una trasmissione numerica realizzata mediante un collegamento costituito da $N$ tratte, collegate da ripetitori rigenerativi. In una trasmissione binaria, si ha errore se un numero dispari di tratte causa un errore per lo stesso bit, e cioè $P_e = \sum_{k=1,\ k\ dispari}^{N} \binom{N}{k} p^k q^{N-k}$, in cui $p$ è la prob. di errore sul bit per una tratta; inoltre risulta che se $p \ll 1$ e $Np \ll 1$, allora $P_e \simeq Np$.

fatto che all'aumentare delle linee occupate, diminuisce la probabilità di una nuova chiamata, in quanto diminuiscono le persone *non* al telefono.

**Esercizio** Una linea telefonica risulta occupata per l'80 % del tempo, e le telefonate non durano mai più di 5 minuti. Provandola a chiamare con una cadenza fissa di un tentativo ogni 10 minuti, determinare

1. la probabilità di trovare libero *entro* 3 tentativi
2. la probabilità di trovare libero *almeno* una volta in due ore
3. la probabilità di trovare libero *esattamente* tre volte in due ore

Indichiamo con $p = 0.2$ la probabilità di successo di un singolo tentativo, e con $q = 1 - p = 0.8$ quella di fallimento, identificando così il problema nel contesto delle *prove ripetute*.

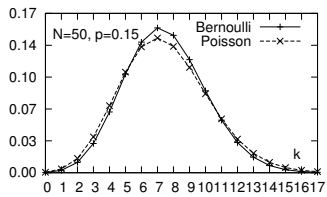
1. Assumendo gli eventi indipendenti, la prob. di trovare libero entro tre tentativi è la somma delle prob. degli eventi favorevoli, ossia subito libero, oppure al secondo, od al terzo tentativo, ovvero $p + p \cdot q + p \cdot q \cdot q = .2 + .2 \cdot .8 + .2 \cdot .2 \cdot .8 = 0.488 = 48.8$ %.
2. In due ore si effettuano $\frac{120}{10} = 12$ tentativi. Conviene in questo caso valutare la probabilità dell'evento complementare $p_0$, quello di fallire tutti i tentativi, pari a $p_B(k)|_{k=0}$, ovvero $p_0 = \binom{12}{0} p^0 q^{12} = \frac{12!}{12!} .8^{12} = 0.0687195$, e quindi la prob. $p_1$ di libero almeno una volta vale $p_1 = 1 - p_0 = 93.12$ %.
3. Trovare libero esattamente tre volte infine ha probabilità $\binom{12}{3} p^3 q^9 = \frac{12 \cdot 11 \cdot 10}{3 \cdot 2} \cdot .2^3 \cdot .8^9 = 0.23$.

## 8.2 Distribuzione di Poisson

Al crescere del numero $N$ di utenti, l'utilizzo della distribuzione Binomiale può risultare disagevole, per via dei fattoriali, e si preferisce trattare il numero di conversazioni attive $k$ come una variabile aleatoria di POISSON, la cui densità di probabilità ha espressione

$$p_P(k) = e^{-\alpha} \frac{\alpha^k}{k!} \tag{8.1}$$

ed è caratterizzata da valor medio e varianza $m_p = \sigma_P^2 = \alpha$.



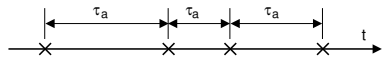

La *Poissoniana* costituisce una buona approssimazione della ddp di Bernoulli, adottando per la prima lo stesso valor medio della seconda $m_P = m_B$, ossia $\alpha = Np$, come mostrato in figura. Ma più in generale, questa densità è impiegata per descrivere la probabilità che si verifichino un numero di eventi *indipendenti e completamente casuali* di cui è noto solo il numero medio $\alpha$([4]).

D'altra parte, al tendere di $N$ ad $\infty$, il modello Bernoulliano precedentemente adottato perde di validità. Infatti, nel caso di una popolazione infinita, il numero di nuove chiamate *non diminuisce* all'aumentare del numero dei collegamenti in corso. In questo caso, gli eventi corrispondenti all'inizio di una nuova chiamata sono invece considerati *indipendenti e completamente casuali*, e descritti unicamente in base ad

[4] Usando il modello Poissoniano pertanto, la probabilità che (ad esempio) si stiano svolgendo *meno* di 4 conversazioni contemporanee è pari a $p_P(0) + p_P(1) + p_P(2) + p_P(3) = e^{-\alpha}\left(1 + \alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{6}\right)$.

una *frequenza media di interarrivo* $\lambda$ che rappresenta la velocità[5] con cui si presentano le nuove chiamate[6].

L'inverso di $\lambda$ rappresenta un tempo, ed esattamente $\overline{\tau}_a = 1/\lambda$ è il *valor medio* della variabile aleatoria $\tau_a$ costituita dall'intervallo di tempo tra l'arrivo di due chiamate.



Tempi di Interarrivo

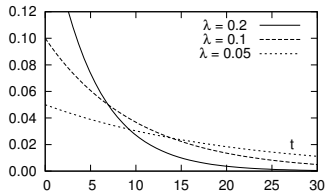
Con queste definizioni, è possibile riferire la v.a. di Poisson ad un intervallo temporale di osservazione $T$, durante il quale si presentano un numero medio $\alpha$ di chiamate[7] pari a $\alpha = \lambda T$. Pertanto, possiamo scrivere la ddp della v.a. Poissoniana come

$$p_P(k)|_T = e^{-\lambda T}\frac{(\lambda T)^k}{k!}$$

che indica la probabilità che in un tempo $T$ si verifichino $k$ eventi (indipendenti e completamente casuali) la cui frequenza media è $\lambda$([8]).

### 8.2.1 Variabile aleatoria esponenziale negativa

La descrizione statistica che la ddp di Poisson fornisce per *il numero* di eventi che si verificano in un (generico) tempo $t$, è strettamente legata al considerare questi come *indipendenti*, *identicamemente distribuiti*, e per i quali *l'intervallo di tempo* tra l'occorrenza degli stessi è una determinazione di variabile aleatoria *completamente casuale*[9], descritta da una densità di probabilità



Densità di v.a. Esponenziale

[5] $\lambda$ viene espresso in *richieste per unità di tempo*.

[6] 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...

[7] Esempio: se da un cavalcavia osserviamo (mediamente) $\lambda = 3$ auto/minuto, nell'arco di $T = 2$ minuti, transiteranno (in media) 3*2 = 6 autovetture.

[8] *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%

[9] 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 quindi di scrivere

$$Pr(t > t_0 + \theta / t > t_0) = Pr(t > \theta)$$

ossia che la probabilità di attendere altri $\theta$ istanti, avendone già attesi $t_0$, non dipende da $t_0$. Per verificare che la ddp esponenziale consente di soddisfare questa condizione, svolgiamo i passaggi, applicando al terzultimo la (8.2):

$$\begin{aligned} Pr(t > t_0 + \theta / t > t_0) &= \frac{Pr(t > t_0 + \theta; t > t_0)}{Pr(t > t_0)} = \frac{Pr(t > t_0 + \theta)}{Pr(t > t_0)} \\ &= \frac{e^{-\lambda(t_0+\theta)}}{e^{-\lambda t_0}} = e^{-\lambda\theta} = Pr(t > \theta) \end{aligned}$$

*esponenziale negativa*[10], espressa analiticamente come

$$p_E(t) = \lambda e^{-\lambda t}$$

valida per $t \geq 0$, e mostrata in figura; tale v.a. è caratterizzata dai momenti $m_E = \frac{1}{\lambda}$ e $\sigma_E^2 = \frac{1}{\lambda^2}$. La probabilità che il tempo di attesa di una v.a. esponenziale superi un determinato valore $t_0$, è allora calcolabile come

$$Pr(t > t_0) = \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_{t_0}^{\infty} = e^{-\lambda t_0} \tag{8.2}$$

e questo risultato ci permette di verificare il legame con la Poissoniana[11].

**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 $\tau_a$, 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 (8.2) 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

$$Pr(t \leq t_0)\big|_{t_0 \to 0} = \lambda t_0 + o(t_0) \tag{8.3}$$

## 8.3 Sistema di servizio orientato alla perdita

Un *sistema di servizio* è una entità in grado di accogliere delle *richieste di servizio*, ovvero eventi che definiscono il cosidetto *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

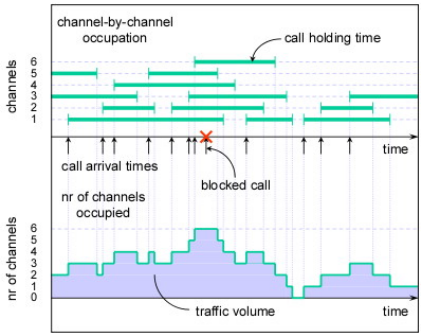
[10] La ddp esponenziale è spesso adottata come un modello approssimato ma di facile applicazione per rappresentare un tempo di attesa, ed applicato ad esempio alla durata di una conversazione telefonica, oppure all'intervallo tra due malfunzionamenti di un apparato.

[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. Allora, se assumiamo che la probabilità di $k$ nascite in un tempo $T$ sia descritta da una v.a. di Poisson, ossia a cui compete una probabilità $p_P(k) = e^{-\lambda T} \frac{(\lambda T)^k}{k!}$, allora la probabilità che durante un tempo $T$ non avvenga nessuna nascita, dovrebbe corrispondere a calcolare $p_P(0)$, ovvero $e^{-\lambda T} \frac{(\lambda T)^0}{0!} = e^{-\lambda T}$, che è esattamente il risultato che fornisce la v.a. esponenziale per la probabilità $Pr(t > T)$ che non vi siano nascite per un tempo $T$.

[12] La dimostrazione della (8.3) si basa sulla considerazione che $Pr(t \leq t_0) = 1 - Pr(t > t_0)$, e sulla espansione in serie di potenze $e^x = 1 + x + {x^2}/{2} + {x^3}/{3!} + \cdots$ che si riduce a $e^x = 1 + x + o(t_0)$ se $x \to 0$. Pertanto, la (8.2) 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*....

successive richieste possono essere poste in coda, individuando così un sistema *orientato al ritardo* (che affrontiamo al § 8.4), oppure rifiutate (vedi la figura a fianco), come avviene per i sistemi *orientati alla perdita*. Scopo della presente sezione sarà pertanto quello di determinare il numero di 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

### 8.3.1 Frequenza di arrivo e di servizio

Mentre il processo di ingresso è descritto in termini della *frequenza media* di arrivo $\lambda$, il tempo medio di occupazione dei serventi (indicato come *processo di servizio*) è descritto nei termini del *tempo medio di servizio* $\tau_S$, 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].

### 8.3.2 Intensità media di traffico

Il rapporto $A_o = \frac{\lambda}{\mu}$ è indicato come *intensità media* del traffico *offerto*[16] e descrive quanti 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).

### 8.3.3 Probabilità di rifiuto

La teoria che porta a determinare la probabilità che una nuova richiesta di servizio non possa essere accolta a causa dell'esaurimento dei serventi, si basa sull'analisi di

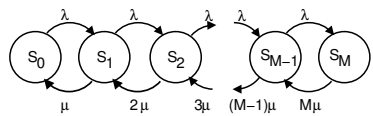
[14] L'ipotesi permette di valutare la probabilità che l'intervallo temporale tra due eventi di ingresso sia superiore a $\theta$, in base alla (8.2), come $e^{-\lambda\theta}$ (ad esempio, la prob. che tra due richieste di connessione in ingresso ad una centrale telefonica passi un tempo almeno pari a $\theta$); allo stesso modo, la probabilità che il servizio abbia una durata maggiore di $\theta$ è pari a $e^{-\mu\theta}$ (ad esempio, la prob. che una telefonata duri più di $\theta$).

[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*.

un cosiddetto *processo di nascita e morte*, che descrive 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.

Descriviamo innanzituto l'evoluzione dello stato del sistema, in cui il numero di *serventi occupati* evolve aumentando o diminuendo di una unità alla volta (come per i processi di nascita e morte), con l'ausilio della figura seguente, dove il generico stato $S_k$ rappresenta la circostanza che $k$ serventi siano occupati, circostanza a cui compete una probabilità $p_k = Pr(S_k)$.




Gli stati del grafo sono collegati da archi etichettati con la frequenza $\lambda$ 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 $\lambda$ e $\mu$ non variano nel tempo, esaurito un transitorio iniziale, il sistema di servizio si troverà in *condizioni stazionarie*, permettendoci di scrivere le *equazioni di equilibrio statistico*

$$\lambda p_k = \mu\,(k+1)\,p_{k+1} \quad \text{con} \quad k = 0, 1, 2, \ldots, M-1 \tag{8.4}$$

che eguagliano la frequenza media con cui il sistema evolve dallo stato $k$ verso $k+1$, alla frequenza media con cui avviene la transizione inversa[18]. La (8.4) 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

$$p_k = \frac{A_o^k}{k!} p_0 \tag{8.5}$$

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

$$p_0 = \left( \sum_{m=0}^{M} \frac{A_o^m}{m!} \right)^{-1} \tag{8.6}$$

[17] Pensiamo ad un ufficio postale visto dall'esterno: la frequenza media $\lambda$ con cui entrano nuove persone non dipende da quanti siano già all'interno, mentre invece la frequenza con la quale escono dipende sia dal tempo medio $1/\mu$ di permanenza allo sportello, che dal numero di sportelli (serventi) $M$ in funzione. La differenza con il caso che stiamo trattando, scaturisce dal fatto che l'ufficio postale è un sistema a coda, e dato che la coda *c'è praticamente sempre* (ossia i serventi sono generalmente tutti occupati) possiamo dire che la frequenza media di uscita è proprio $M\mu$.

[18] E' un pò come se il numero medio di nuove richieste per unità di tempo $\lambda$ si distribuisse, in accordo alle probabilità $p_k$, tra tutti gli stati possibili del sistema: come dire che del totale di $\lambda$, una parte $\lambda p_0$ trovano il sistema vuoto, una parte $\lambda p_1$ 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 $\mu_r$ 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 (8.7)

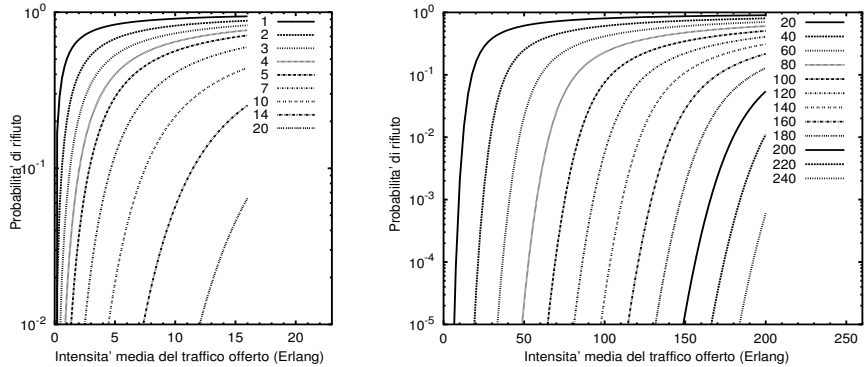



Figura 8.1: Andamento della probabilità di blocco $P_B$ in un sistema orientato alla perdita, al variare di $A_o$, per il numero di serventi indicato sulle curve

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 $M = \infty$, tenendo conto dell'espansione in serie $\sum_{m=0}^{\infty} \frac{A_o^m}{m!} = e^{A_0}$, si ottiene che la (8.6) fornisce appunto $p_0 = e^{-A_0}$, e la (8.5) diviene $p_k = e^{-A_0} \frac{A_o^k}{k!}$, che come riconosciamo immediatamente è proprio la ddp di Poisson (8.1) con valore medio $A_0$[20]. Se invece poniamo $M$ finito, la sommatoria che compare in (8.6) non corrisponde ad una serie nota, e dunque rimane come è, fornendo il risultato

$$p_k = Pr\left(S_k\right) = \frac{\frac{A_o^k}{k!}}{\sum_{m=0}^{M} \frac{A_o^m}{m!}}$$

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 $A_o$:

$$P_B = Pr\left(S_M\right) = p_M = \frac{\frac{A_o^M}{M!}}{\sum_{m=0}^{M} \frac{A_o^m}{m!}} = E_{1,M}\left(A_o\right) \tag{8.7}$$

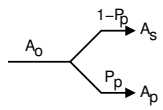
L'andamento di $P_B$ in funzione di $M$ e di $A_o$ è graficato in Fig. 8.1, e mostra come (ad esempio) per una intensità di traffico offerto pari a 40 Erlang, siano necessari più di 50 serventi per mantenere una $P_B$ minore dell'1%, che salgono a più di 60 per una $P_B = 10^{-3}$.

### 8.3.4 Efficienza di giunzione

In presenza di una intensità media di traffico offerto $A_o$, ed una probabilità di perdita

[20] Questo risultato è in perfetto accordo con le la (8.1), quando abbiamo sostituito alla ddp di Bernoulli quella di Poisson, mantenendo inalterato il numero medio di serventi occupati, che ora indichiamo con $A_o$, come definito al § 8.3.2.

$P_p = P_B$, solamente il $(1 - P_p) \cdot 100$ % delle richieste è smaltito, e quindi $A_o$ si ripartisce tra l'intensità media di *traffico smaltito* $A_s = A_o\,(1 - P_p)$, e l'intensità media di *traffico perso* $A_p = A_o P_p$. Possiamo definire un coefficiente di utilizzazione, o efficienza




$$\rho = \frac{A_s}{M} = \frac{A_o}{M}\,(1 - P_p)$$

che rappresenta la percentuale di impegno dei serventi, e di cui la figura 8.2 mostra l'andamento al variare di $A_o$, per una $P_B$ assegnata e pari a $2 \cdot 10^{-3}$, assieme al numero di 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.

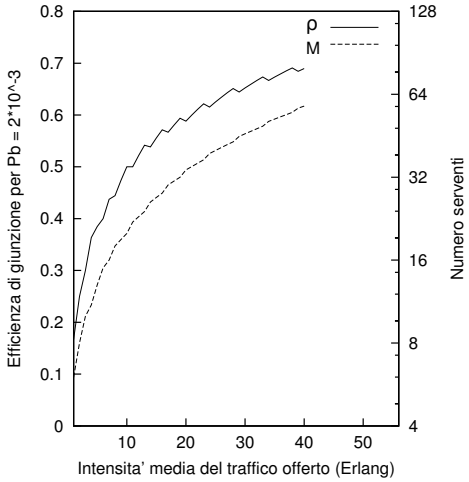



Figura 8.2: Efficienza di giunzione

### 8.3.5 Validità del modello

Le considerazioni esposte si riferiscono ad una ipotesi di traffico completamente casuale con tempi di interarrivo e di servizio 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 $\lambda$ 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.

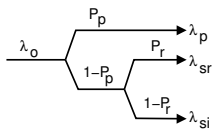
1. Determinare la $f_b$ di una sorgente nelle fasi di attività;
2. determinare il numero massimo di sorgenti contemporaneamente attive;
3. determinare il grado di servizio (Probabilità di rifiuto) ottenibile con il mezzo trasmissivo indicato;
4. indicare la capacità da aggiungere al collegamento per 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. 8.1 si trova una probabiltà 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. 8.1 si osserva che ciò richiede (a parità di $A_o$) almeno 120 (circa) serventi, 20 in più, pari ad una capacità aggiuntiva di $20 \cdot 256 \cdot 10^3 = 5.12$ Mbps.

## 8.4 Sistemi di servizio orientati al ritardo

Mentre i sistemi orientati alla perdita rappresentano il modo di operare delle reti di telecomunicazione a *commutazione di circuito*, in cui ogni connessione impegna in modo esclusivo alcune risorse di rete, che una volta esaurite producono un *rifiuto* della richiesta di connessione, i sistemi *orientati al ritardo* sono rappresentativi di reti a *commutazione di pacchetto*, in cui i messaggi sono suddivisi in unità elementari (detti pacchetti, appunto) la cui ricezione non deve più avvenire in tempo reale, e che condividono le stesse risorse fisiche (degli organi di commutazione e di trasmissione) con i pacchetti di altre comunicazioni. Pertanto, l'invio di un pacchetto può essere *ritardato* se il sistema di servizio è in grado di gestire delle *code di attesa*, in cui accumulare le richieste che eccedono il numero di serventi a disposizione, e da cui prelevare (con ritardo) i pacchetti stessi non appena si rendano disponibili le risorse trasmissive necessarie.




In questo caso, il grafico che mostra la ripartizione dei flussi di richieste si modifica come in figura, dove è evidenziato come la frequenza di richieste $\lambda_o$ si suddivide tra la frequenza delle richieste perse $\lambda_p$, quelle servite con ritardo $\lambda_{sr}$, e quelle servite immediatamente $\lambda_{si}$, 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 $A_s = A_{sr} + A_{si} = (1 - P_p) A_o$ aumenta, a parità di offerta.

### 8.4.1 Risultato di Little

Si tratta di un risultato molto generale, valido per qualsiasi distribuzione dei tempi di interarrivo e di servizio, la cui applicazione può tornare utile nell'analisi, e che recita:

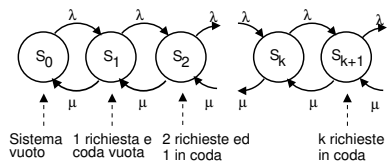
> *Il numero medio $\overline{N}$ di utenti contemporaneamente presenti in un sistema di servizio è pari al prodotto tra frequenza media di smaltimento delle richieste $\lambda_s$ ed il tempo medio di permanenza $\tau_p$ dell'utente nel sistema*

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 $\tau_c$ rappresenta il tempo medio di coda.

### 8.4.2 Sistemi a coda infinita ed a servente unico

Prima di fornire risultati più generali, svolgiamo l'analisi per questo caso particolare, in cui la frequenza di richieste perse $\lambda_p$ è nulla, dato che una coda di lunghezza infinita le accoglie comunque tutte.

Il sistema è descritto da un punto di vista statistico mediante il diagramma di nascita e morte riportato a fianco, in cui ogni stato $S_k$ rappresenta $k$ richieste nel sistema, di cui una sta ricevendo servizio e $k - 1$ sono accodate.




Per procedere nell'analisi, si applica lo stesso principio di equilibrio statistico già adottato a pag. 167, il quale asserisce che, esaurito un periodo transitorio iniziale, la frequenza media delle transizioni tra $S_k$ e $S_{k+1}$ deve eguagliare quella da $S_{k+1}$ ad $S_k$. Indicando con $p_k = Pr(S_k)$ la probabilità che il sistema contenga $k$ richieste, l'equilibrio statistico si traduce nell'insieme di equazioni

$$\lambda_o p_k = \mu p_{k+1} \quad \text{con} \quad k = 0, 1, 2, \ldots, \infty \tag{8.8}$$

Infatti, in base alle stesse considerazioni svolte nella prima parte della nota 18 di pag. 167, $\lambda_o p_k$ è pari alla frequenza media (frazione di $\lambda_o$) 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 $\mu p_{k+1}$ è proprio la frequenza media con cui il sistema passa da $k + 1$ a $k$ richieste accolte.

La relazione (8.8) è di natura ricorsiva, e può esprimersi come

$$p_k = \left(\frac{\lambda_o}{\mu}\right)^k p_0 = A_o^k p_0$$

[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 $\lambda_o (P_p + P_r (1 - P_p))$, ed eguagliando le due espressioni si ottiene $P_p = \frac{P_B - P_r}{1 - P_r}$, che è sempre minore di $P_B$.

Per determinare il valore $p_0 = Pr(S_0)$, uguale alla probabilità che il sistema sia vuoto, 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 $\tau_p$, e scomponibile nella somma $\tau_p = \tau_S + \tau_c$ tra il tempo medio di servizio ed il tempo medio di coda. Possiamo applicare qui il risultato di Little $\overline{N} = \lambda_s \cdot \tau_p$, 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 $\infty$.

---

[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 $A_o = \frac{\lambda_o}{\mu} < 1$; se il servente è unico infatti, una frequenza di arrivo maggiore di quella di servizio preclude ogni speranza di funzionamento, dato che evidentemente il sistema non ha modo di smaltire in tempo le richieste che si presentano.

[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 $\sum_{k=0}^{\infty} k\alpha^k = \alpha \sum_{k=0}^{\infty} k\alpha^{k-1} = \alpha \frac{\partial}{\partial \alpha} \sum_{k=0}^{\infty} \alpha^k = \alpha \frac{\partial}{\partial \alpha} \frac{1}{1-\alpha} = \frac{\alpha}{(1-\alpha)^2}$

[28] Non può essere $\lambda_s > \lambda_o$, perchè si servirebbero più richieste di quante se ne presentano. Se fosse invece $\lambda_s < \lambda_o$, la coda crescerebbe inesorabilmente e sarebbe quindi inutile.

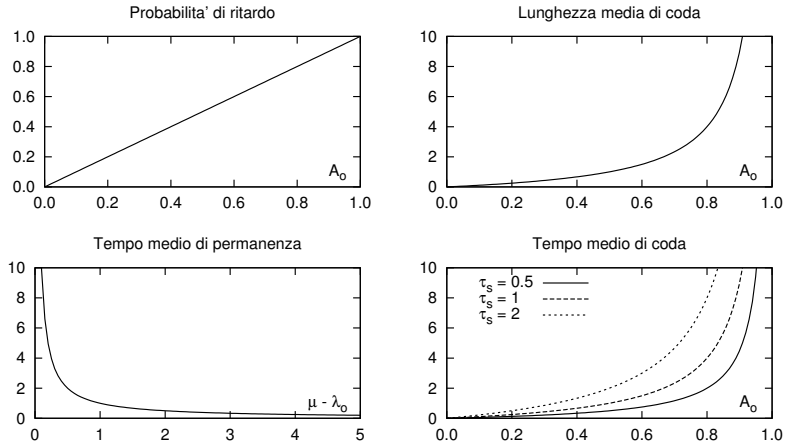



Figura 8.3: Grandezze di interesse per il sistema a coda infinita ed unico servente

**Tempo medio di coda** si calcola come

$$\tau_c = \tau_p - \tau_S = \frac{1}{\mu - \lambda_o} - \frac{1}{\mu} = \frac{\mu - \mu + \lambda_o}{\mu\,(\mu - \lambda_o)} = \frac{A_o}{\mu\,(1 - A_o)} = \frac{1}{\mu}\frac{\rho}{1-\rho} = \tau_S \frac{\rho}{1-\rho}$$

Questo risultato mostra che il tempo medio di coda è legato al tempo medio di servizio e all'efficienza di giunzione, confermando ancora i risultati per $\rho \to \infty$.

La fig. 8.3 mostra l'andamento delle grandezze appena calcolate.

### 8.4.3 Sistemi a coda finita e con più serventi

Riportiamo solo i risultati, validi se entrambi i processi di ingresso e di servizio sono esponenziali con frequenza media $\lambda_o$ e $\mu$, la coda è lunga $L$, i serventi sono $M$ e le sorgenti infinite.

**Probabilità di k richieste nel sistema**

$$p_k\,(A_o) = \begin{cases} \frac{A_o^k}{k!\alpha(A_o)} & 0 \le k \le M \\ \frac{A_o^k}{M^{k-M} M!\alpha(A_o)} & M \le k \le M + L \end{cases}$$

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**

$$P_r = \sum_{k=M}^{M+L} p_k\,(A_o) = p_M\,(A_o)\,\frac{1 - \rho^{L+1}}{1 - \rho} \quad \text{in cui} \quad \rho = \frac{A_o}{M}$$

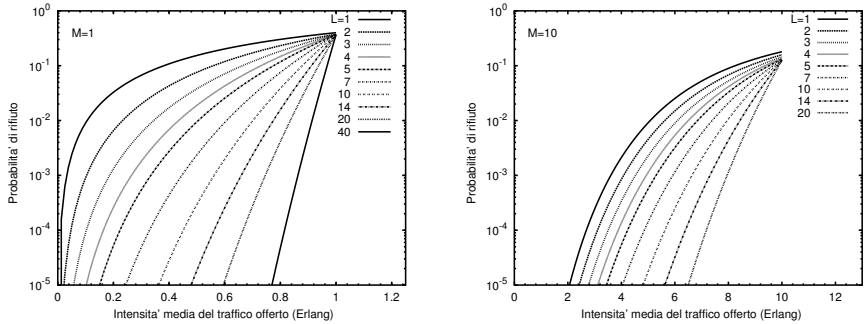



Figura 8.4: Probabilità di perdita per un sistema a coda finita con uno o dieci serventi

**Probabilità di perdita**

$$P_p = p_{M+L}\left(A_o\right) = \frac{A_o^{M+L}}{M^L M! \cdot \alpha\left(A_o\right)}$$

**Tempo medio di coda**

$$\tau_c = \tau_S \frac{P_r - L \cdot P_{M+L}\left(A_o\right)}{M - A_o}$$

La Figura 8.4 descrive la probabilità di perdita per un sistema a servente singolo (a sinistra) e con 10 serventi (a destra), in funzione dell'intensità di traffico offerto e della lunghezza di coda, così come risulta dalla applicazione delle formule riportate. Nel caso di trasmissione di pacchetti di lunghezza fissa, per i quali il tempo di servizio è fisso e *non* a distribuzione esponenziale[29], i risultati ottenuti costituiscono una *stima conservativa* delle prestazioni del sistema (che potranno cioè essere migliori). L'analisi delle curve permette di valutare con esattezza il vantaggio dell'uso di una coda (a spese del tempo di ritardo). Infatti, aumentando il numero di posizioni di coda si mantiene una probabilità di blocco accettabile anche per traffico intenso.

Ad esempio, per $P_b = 1\%$ ed $M = 1$, osserviamo che una coda con $L = 20$ posizioni gestisce un traffico di $A_o = 0.83$ Erlang, contro gli $A_o = 0.11$ Erlang del caso senza coda. Ciò corrisponde ad un aumento dell'efficienza di $\frac{0.83}{0.11} = 7.54$ volte. D'altra parte ora il tempo medio di coda (calcolato in modo conservativo applicando la relazione per coda infinita) è $\tau_c = \tau_S \frac{\rho}{1-\rho} = \frac{0.83}{1-0.83}\tau_S = 4.9\tau_S$, ed è quindi aumentato (rispetto a $\tau_S$) 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 $\tau_a$ tra pacchetti corrispondente ad un traffico di ingresso di 1200 pacchetti/secondo, e l'associata intensità $A_o$;

[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 $\tau_S = \overline{L}_p / f_b$.

[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 velocita' di trasmissione, il prefisso K torna a valere $10^3 = 1000$.

[31] In virtù di quanto esposto alla nota precedente, in questo caso $1M = 10^6 = 1000000$.

**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.

## 8.5 Reti per trasmissione dati

In questa sezione illustriamo le particolarità legate alle *trasmissioni dati*, e come queste possano essere vantaggiosamente sfruttate per conseguire la *maggiore efficienza* che i sistemi di servizio a coda presentano rispetto a quelli orientati alla perdita. Le particolari *modalità* e *funzioni* legate alle trasmissioni dati saranno classificate secondo uno schema che ne consente il confronto in termini di prestazioni e vincoli sulla realizzazione della rete. Infine, verranno formalizzate le esigenze legate alla soluzione dei problemi di trasmissione dati, introducendo i concetti legati alle *architetture protocollari*, assieme ad alcuni esempi reali.

Le trasmissioni dati si prestano bene a comunicazioni in cui siano possibili ritardi temporali variabili, attuando una filosofia di tipo *ad immagazzinamento e rilancio* (STORE AND FORWARD) basata sul suddividere il messaggio in unità informative elementari denominate *pacchetti*, che possono essere inoltrati sulla rete di comunicazione, assieme a quelli prodotti da altre trasmissioni. L'applicazione della stessa metodologia a trasmissioni (ad esempio) vocali non è per nulla semplice, in quanto la presenza di un ritardo variabile per la trasmissione dei pacchetti comporta problemi non trascurabili, a meno di attuare speciali meccanismi di priorità e prenotazione della banda, tuttora oggetto di ricerca.

### 8.5.1 Il pacchetto dati

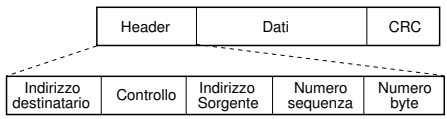
Discutiamo brevemente, in termini generali, i possibili contenuti di un pacchetto dati; il suo formato effettivo dipenderà dal particolare protocollo di trasmissione adottato.

La prima osservazione da fare, è che la suddivisione del messaggio in pacchetti comporta un aumento delle informazioni da trasmettere, in quanto ognuno di questi dovrà

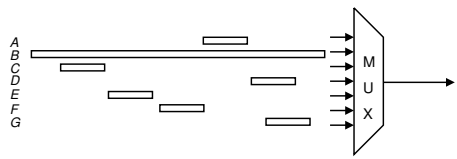
contenere informazioni addizionali per consentire un suo corretto recapito e la sua ricombinazione con gli altri pacchetti dello stesso messaggio. Occorre inoltre affrontare gli ulteriori problemi tipici di una comunicazione dati, ovvero come contrastare gli errori di trasmissione, e come gestire le 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.




Può sembrare vantaggioso mantenere la dimensione dei pacchetti elevata, riducendo così la rilevanza delle informazioni aggiuntive, ma si verificano controindicazioni. Infatti, suddividere messaggi lunghi in pacchetti più piccoli garantisce l'inoltro di (altre) comunicazioni più brevi durante la trasmissione di messaggi lunghi, che altrimenti *bloccherebbero* i sistemi di coda se realizzate con un unico "pacchettone": in figura è mostrato un esempio in cui $B$, presentandosi in ingresso al multiplexer con lieve anticipo rispetto agli altri pacchetti più piccoli, ne impedisce l'inoltro, monopolizzando la linea di uscita per tutta la durata della sua trasmissione.




Infine, all'aumentare della lunghezza di un pacchetto aumenta proporzionalmente la probabilità di uno (o più) bit errati (vedi anche la formula (5.9) a pag. 78 e la discussione al § 5.4.2.3), e dunque l'uso di dimensioni contenute riduce le necessità di ritrasmissione.

### 8.5.2 Modo di trasferimento delle informazioni

È definito in base alla specificazione di 3 caratteristiche che lo contraddistinguono: *lo schema di multiplazione*, il *principio di commutazione* e *l'architettura protocollare*.

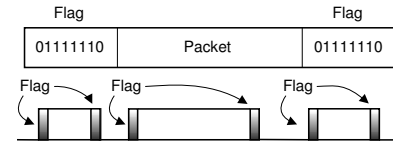
[32] La sigla CRC significa *Cyclic Redundancy Check* (controllo ciclico di ridondanza) ed indica una parola binaria i cui bit sono calcolati in base ad operazioni algebriche (vedi § 5.3.2.3) attuate sui bit di cui il resto del messaggio è composto. Dal lato ricevente sono eseguite le stesse operazioni, ed il risultato confrontato con quello presente nel CRC, in modo da controllare la presenza di errori di trasmissione.

[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).

#### 8.5.2.1 Schema di multiplazione

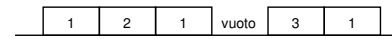
È stato già illustrato uno schema a divisione tempo che prevede l'uso di una *trama* in cui trovano posto diverse comunicazione vocali[34], e che necessita di un funzionamento sincronizzato dei nodi di rete. La trasmissione *a pacchetto* invece non prevede l'uso esclusivo di risorse da parte delle singole comunicazioni, e *non fa uso* di una struttura di trama e pertanto occorrono soluzioni particolari per permettere la *delimitazione* dei pacchetti.

Ad esempio, i protocolli HDLC ed X.25 presentano pacchetti di dimensioni variabili, e fanno uso di un byte di *flag* (vedi pag. 94) costituito dalla sequenza *01111110* in testa ed in coda, per separare tra loro i pacchetti di comunicazioni differenti.

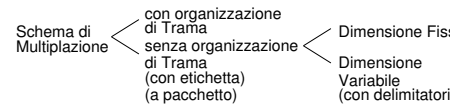



Per evitare che i dati "propri" del pacchetto possano simulare un flag, in trasmissione viene inserito un bit *0* dopo *5* uni di fila, che (se presente) viene rimosso al ricevitore. Se dopo *5* uni c'è ancora un *1*, allora è un flag.

Nel caso in cui il pacchetto invece abbia una *dimensione fissa*[35], ci si trova ad operare in una situazione simile a quella in presenza di trama, tranne che... la trama non c'è, e dunque l'ordine dei pacchetti è qualsiasi, ma viene meno l'esigenza dei flag di delimitazione.




In entrambi i casi (lunghezza di pacchetto fissa o variabile) i nodi della rete non necessitano di operare in sincronismo tra loro; lo schema di multiplazione è quindi detto *a divisione di tempo senza organizzazione di trama, asincrono, con etichetta*. Il termine etichetta indica che ogni pacchetto deve recare con sé le informazioni idonee a ricombinarlo assieme agli altri dello stesso messaggio.

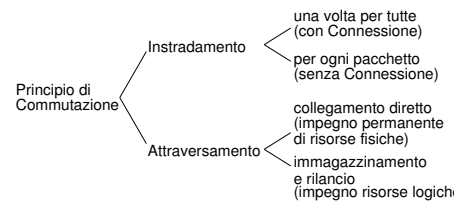



#### 8.5.2.2 Principio di commutazione

È definito in base a come sono realizzate le due funzioni di *instradamento* (come individuare un percorso nella rete) e *attraversamento* (come permettere l'inoltro del messaggio).

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*.



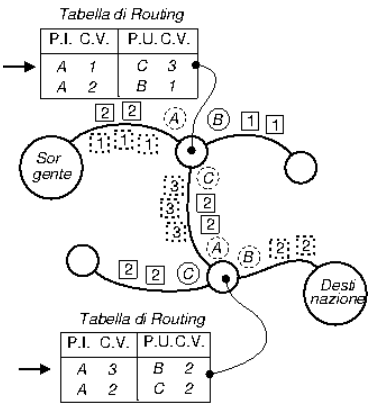

[34] Come nel PCM telefonico, vedi § 6.3.1

[35] Un modo di trasferimento con pacchetti di dimensione fissa è l'ATM (ASYNCHRONOUS TRANSFER MODE che viene descritto sommariamente in appendice.

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[36].

**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 della rete ad essa necessarie, e conseguente ad una *richiesta di connessione* da parte del nodo sorgente. I pacchetti di uno stesso messaggio seguono quindi tutti uno stesso percorso, e l'attraversamento si basa sull'impegno di *risorse logiche*[37] ed avviene per *immagazzinamento e rilancio*. La trasmissione ha luogo dopo aver contrassegnato ogni pacchetto con un *identificativo di connessione* ($IC$) che identifica un *canale virtuale*[38]tra coppie di nodi di rete, e che ne individua 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: una sorgente, a seguito della fase di instradamento, invia i pacchetti con identificativo $IC = 1$ al primo nodo individuato dal ruoting. Consultando la propria tabella, il nodo trova che il canale virtuale *1* sulla *porta di ingresso* (P.I.) $A$ si connette al c.v. *3* sulla *porta di uscita* (P.U.) $C$. Ora i pacchetti escono da $C$ con $IC = 3$ ed una volta giunti al nodo seguente sulla P.I. $A$, escono dalla P.U. $B$ 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[39]. La concatenazione dei canali virtuali attraversati viene infine indicata con il termine *Circuito*

[36]Nel caso del POTS (vedi § 6.9.1) si creava un vero e proprio circuito elettrico (vedi anche pag. 100), 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 § 6.3.1), le risorse allocate cambiano natura (ad esempio consistono anche nell'intervallo temporale assegnato al canale all'interno della trama) ma cionostante vi si continua a far riferimento come ad una rete a *commutazione di circuito*.

[37]Le risorse impegnate sono dette *logiche* in quanto corrispondono ad entità concettuali (i *canali virtuali* descritti nel seguito).

[38]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.

[39]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.

*Virtuale* per similitudine con il caso di commutazione di circuito, con la differenza che ora il percorso individuato è definito solo in termini di tabelle e di etichette, e non di risorse fisiche (tranne che per la memoria della tabella).

Al termine della comunicazione, sul circuito virtuale viene inviato un apposito pacchetto di controllo, che provoca la rimozione del routing dalle tabelle.

**Congestione e controllo di flusso** Durante la fase di instradamento, il percorso nella rete è determinato in base alle condizioni di traffico del momento, ed eventualmente la connessione può essere rifiutata nel caso in cui la memoria di coda nei nodi coinvolti sia quasi esaurita, evento indicato con il termine di *congestione*.

D'altra parte, se alcune sorgenti origine dei Canali Virtuali già assegnati e che si incrociano in uno stesso nodo intermedio, iniziano ad emettere pacchetti a frequenza più elevata del previsto, il nodo intermedio si congestiona (ossia esaurisce la memoria di transito) ed inizia a *perdere pacchetti*, penalizzando tutti i Canali Virtuali che attraversano il nodo.

Per questo motivo, sono indispensabili strategie di *controllo di flusso* che permettano ai nodi di regolare l'emissione delle sorgenti. Il controllo di flusso è attuato anch'esso mediante pacchetti (di controllo), privi del campo di dati, ma contenenti un codice identificativo del comando che rappresentano. Ad esempio, un nodo non invia nuovi pacchetti di un circuito virtuale finché non riceve un *pacchetto di riscontro* relativo ai pacchetti precedenti. D'altra parte, nel caso di una rete congestionata, la perdita di pacchetti causa il mancato invio dei riscontri relativi, e dunque i nodi a monte cessano l'invio di nuovi pacchetti[40]. Dopo un certo periodo di tempo (TIMEOUT) il collegamento è giudicato interrotto e viene generato un pacchetto di *Reset* da inviare sul canale virtuale, e che causa, nei nodi attraversati, il rilascio delle risorse logiche (tabelle) relative al Canale Virtuale.

Discutiamo ora invece di un ulteriore possibile principio di commutazione:

**Commutazione di pacchetto a *datagramma*** Anche in questo caso, *l'attraversamento* dei nodi avviene per *immagazzinamento e rilancio*, mentre la funzione di *instradamento* è svolta in modo distribuito tra i nodi di rete *per ogni pacchetto*, i quali (chiamati ora *datagrammi*) devono necessariamente contenere l'indirizzo completo della destinazione. Infatti, in questo caso manca del tutto la fase iniziale del collegamento, in cui prenotare l'impegno delle risorse (fisiche o logiche) che saranno utilizzate[41]. Semplicemente, non è previsto alcun impegno a priori, ed ogni pacchetto costituisce un collegamento individuale che impegna i nodi di rete solo per la durata del proprio passaggio. L'instradamento avviene mediante tabelle presenti nei nodi, sia di tipo statico che dinamico (nel qual caso tengono conto delle condizioni di carico e di coda dei nodi limitrofi) che indicano le possibili porte di uscita per raggiungere la destinazione scritta sul pacchetto. Quest'ultimo quindi viene fatto uscire *senza nessuna alterazione* dalla porta di uscita.

Uno dei maggiori vantaggi dei datagrammi rispetto ai circuiti virtuali è una migliore resistenza ai guasti e malfunzionamenti: in questo caso infatti, a parte una eventuale

[40] 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), finchè 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*.

[41] Per questo motivo, il collegamento è detto *senza connessione*.

necessità di ritrasmettere i pacchetti persi, il collegamento prosegue attraverso percorsi alternativi; inoltre l'elevato numero di percorsi alternativi, può permettere (in condizioni di carico leggero) di soddisfare brevi richieste di trasmissione a velocità elevate. Allo stesso tempo, in presenza di messaggi molto brevi, l'invio di un singolo datagramma è più che sufficiente, mentre nel caso a circuito virtuale le fasi di instaurazione ed abbattimento sarebbero state un lavoro in più da svolgere (tanto che ad es. l'X.25, che è nato a c.v., prevede anche il funzionamento a datagramma).

**Consegna ordinata e congestione** Uno dei maggiori problemi legati all'uso di datagrammi è che l'ordine di arrivo dei pacchetti può essere diverso da quello di partenza, potendo questi seguire percorsi differenti. Per questo motivo, nei datagrammi è presente un *numero di sequenza* che si incrementa ad ogni pacchetto trasmesso, ed alla destinazione sono predisposti dei *buffer*[42] di memoria nei quali ricostruire l'ordine esatto dei pacchetti.

Nel caso di un pacchetto mancante, il ricevente non sa se questo è semplicemente ritardato oppure è andato perso, rendendo problematico il controllo di flusso. In questo caso si produce un impegno anomalo dei buffer di ingresso, che non possono essere rilasciati perchè 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[43].

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. 188).

Proseguiamo la descrizione delle reti per dati con l'ultima caratteristica di un modo di trasferimento:

#### 8.5.2.3 Architettura protocollare

Definisce la stratificazione delle funzioni di comunicazione, sia per gli apparati terminali che per i nodi di transito, e di come queste interagiscono reciprocamente sia tra nodi diversi, che nell'ambito di uno stesso nodo. Alcune di queste sono già state introdotte, e le citiamo per prime, seguite da quelle più rilevanti illustrate di seguito:

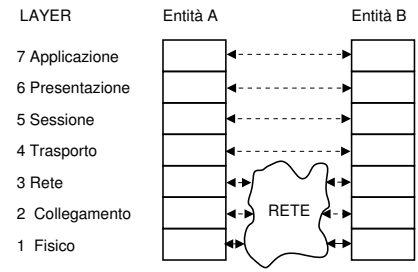
- il *controllo di flusso*, che impedisce la saturazione dei buffer;
- la *consegna ordinata*, per riassemblare messaggi frammentati su più datagrammi;
- la *segmentazione e riassemblaggio*, che definisce le regole per frammentare un messaggio in pacchetti e ricomporli, ad esempio in corrispondenza dei "confini" tra sottoreti con differente lunghezza di pacchetto;
- il *controllo di connessione*, che provvede ad instaurare la connessione, eseguire l'instradamento, impegnare le risorse, supervisionare il controllo di flusso, abbattere la connessione al suo termine;
- il *controllo di errore*, che provvede a riscontrare le unità informative, a rilevare gli errori di trasmissione, a gestire le richieste di trasmissione;

[42] *Il termine* buffer *ha traduzione letterale "respingente, paracolpi, cuscinetto" ed è a volte espresso in italiano dalla locuzione* memoria tampone.

[43] 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.

- l'*incapsulamento*, che aggiunge ai pacchetti di dati da trasmettere le informazioni di protocollo come l'header, gli indirizzi, il controllo di parità...

**Stratificazione iso-osi** Per aiutare nella schematizzazione delle interazioni tra le funzioni illustrate, l'*International Standard Organization* (ISO) ha formalizzato un modello concettuale per sistemi di comunicazione denominato *Open System Interconnection* (OSI)[44], che individua una relazione gerarchica tra i protocolli. In particolare sono definiti sette *strati* o *livelli* (LAYERS) ognuno dei quali raggruppa un insieme di funzioni affini. Gli strati più elevati (4-7) sono indicati anche come *strati di utente*, in quanto legati a funzioni relative ai soli apparati terminali; gli *strati di transito* invece (1-3) riguardano funzioni che devono essere presenti anche nei nodi intermedi.

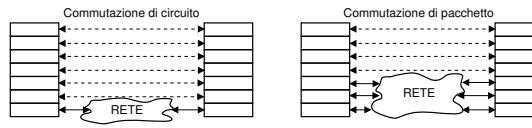



La relazione gerarchica individuata stabilisce tra due strati contigui un rapporto di tipo *utente-servizio*; ovvero lo svolgimento delle funzioni di strato superiore necessita dei servizi offerti dallo strato inferiore. A titolo di esempio, si pensi all'invio di un documento mediante un corriere espresso: ci si affida allora ad uno strato di trasporto che offre all'utente (strato di sessione) un servizio (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.

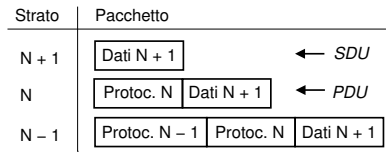
Per terminare l'esempio, facciamo notare come in ogni livello avvengano *due* tipi di colloqui (regolati da altrettanti protocolli): uno è *orizzontale*, detto anche *tra pari* (PEER-TO-PEER), come è ad esempio il contenuto del documento che spediamo, od i rapporti tra corrispondenti locali (che nel caso di un sistema di comunicazione corrisponde allo strato di collegamento, relativo ai procolli tra singole coppie di nodi di rete); il secondo tipo di colloquio avviene invece in forma *verticale*, o *tra utente e servizio*, in quanto per realizzare le funzioni di uno strato *utente* ci si affida ad un *servizio* di comunicazione offerto dallo strato inferiore (che a sua volta può avvalersi dei servizi degli altri strati ancora inferiori)[45].

[44] In virtù dell'intreccio di sigle, il modello di riferimento prende il nome (palindromo) di modello *ISO-OSI*.

[45] 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 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.



**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 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.

## 8.6 Appendici

In questa sezione trova spazio la descrizione di aspetti delle reti per dati come Internet, e le reti IP e ATM, che costituiscono l'immediata applicazione degli argomenti esposti nel capitolo, ossia i sistemi a coda ed a trasmissione di pacchetto, e le architetture protocollari.Viceversa, la realizzazione di sistemi arientati alla perdita ed a commutazione di circuito è discussa al Cap. 6.

Iniziamo subito con il dire che il modello a strati ISO-OSI è una astrazione concettuale utile per individuare raggruppamenti di funzioni, e serve ottimamente come modello per stimolare l'interoperabilità di apparati di diversi costruttori. D'altra parte, alcune delle realizzazioni esistenti (come ad esempio INTERNET) si sono sviluppate precedentemente alla definizione del modello, mentre altre (come ATM) seguono filosofie che solo successivamente sono state incorporate nel modello di riferimento. Pertanto, utilizzeremo le classificazioni ISO-OSI come parametro di riferimento, mediante il quale analizzare le funzioni delle reti reali.

### 8.6.1 La rete Internet

#### 8.6.1.1 Storia

Nel 1964 L. Kleinrock (UCLA) propone un modello di rete non gerarchica e con parti ridondanti, che realizza una modalità di trasferimento senza connessione e senza garanzie di qualità del servizio, rimandando queste ultime ai livelli superiori dell'architettura protocollare. Tale tipologia di servizio è oggi indicata con il termine *best effort*[46]. Nel '69 sono operativi cinque nodi nelle università americane, e nel '72 avviene la prima dimostrazione pubblica di ARPANET, basata su NCP. Nel '73 Kahn e Cerf iniziano a definire TCP, da cui viene successivamente separato l'IP per la convenienza di non dover necessariamente aprire sempre una connessione. Fino all'80, il DoD[47] sovvenziona le università per implementare in ambiente UNIX i protocolli, che nel frattempo si vanno arricchendo di servizi, mentre la trasmissione Ethernet (del 1973) è adottata per realizzare LAN.

Nel 1983 il DoD decreta che tutti i calcolatori connessi a ARPANET adottino i protocolli TCP/IP, e separa la rete in due parti: una civile (ARPANET) ed una militare (MILNET). Negli anni seguenti i finanziamenti dalla *National Science Foundation* permettono lo sviluppo di una rete di trasporto a lunga distanza e di reti regionali, che interconnettono LAN di altre università e di enti di ricerca alla rete ARPANET, alla quale si collegano poi anche le comunità scientifiche non americane.

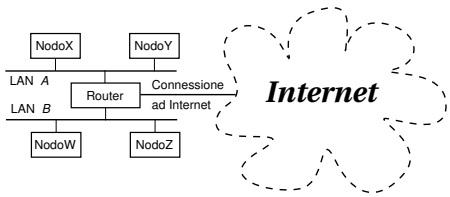
Nel 1990 ARPANET cessa le sue attività, e Barners-Lee (CERN) definisce il WWW, mentre nel '93 Andreessen (NCSA) sviluppa il primo *browser* WWW. Dal 1995 L'NSF non finanzia più la rete di interconnessione, ed il traffico inizia ad essere trasportato da operatori privati.

[46] Migliore sforzo, ossia la rete dà il massimo, senza però garantire nulla.
[47] Department of Defense.

#### 8.6.1.2 Le caratteristiche

La parola *Internet* in realtà è composta da due parole, INTER e NET, in quanto le caratteristiche della rete Internet sono quelle di fondere in una unica architettura una infinità di singole reti locali, potenzialmente disomogenee, e permettere la comunicazione tra i computer delle diverse sottoreti.

Ogni nodo della rete è connesso ad una rete locale (LAN[48]), la quale a sua volta è interconnessa ad Internet mediante dei nodi detti *router*[49] che sono collegati ad una o più LAN e ad Internet, e svolgono la funzione di instradare le comunicazioni verso l'esterno. L'instradamento ha luogo in base ad un *indirizzo IP*[50], che individua i singoli nodi in modo univoco su scala mondiale.




Come anticipato, lo strato di rete (o strato IP) realizza un modo di trasferimento a datagramma e non fornisce garanzie sulla qualità di servizio (QoS, QUALITY O SERVICE) in termini di ritardi, errori e pacchetti persi. La situazione è mitigata dalla strato di trasporto (TCP, TRANSMISSION CONTROL PROTOCOL) che offre ai processi applicativi un servizio a circuito virtuale.

I protocolli di Internet sono realizzati in software e sono pubblici; gli utenti stessi e molte sottoreti private contribuiscono significativamente al trasporto, all'indirizzamento, alla commutazione ed alla notifica delle informazioni. Queste sono alcune ragioni fondamentali per cui Internet *non è di nessuno* ed è un patrimonio dell'umanità.

#### 8.6.1.3 Gli indirizzi

| Strato | Indirizzo |
|---|---|
| Applicazione | *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* |

Iniziamo l'argomento discutendo subito la stratificazione degli indirizzi coinvolti in una comunicazione via Internet. Ogni livello funzionale infatti utilizza le proprie convenzioni di indirizzamento, come illustrato nella tabella a fianco. Se a prima vista questa abbondanza di indirizzi può apparire esagerata, è proprio in questo modo che si realizza l'interoperabilità tra ambienti di rete differenti.

**IP ed Ethernet** I computer connessi ad Internet (detti *nodi*) sono le sorgenti e le destinazioni dell'informazione, e sono inviduati da *un indirizzo IP*, che consiste in un

[48] LOCAL AREA NETWORK, ossia *rete locale*. Con questo termine si indica un collegamento che non si estende oltre (approssimativamente) un edificio.

[49] 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* .

[50] IP = Internet Protocol.

gruppo di 4 byte[51] e che si scrive *x.y.w.z* con ognuna delle 4 variabili pari ad un numero tra 0 e 255.

I nodi sono connessi alla rete mediante una interfaccia a volte indicata come MAC (MEDIA ACCESS CONTROL). Prendendo come esempio[52] i nodi connessi ad una LAN Ethernet, l'interfaccia di rete è individuata a sua volta da un *indirizzo Ethernet* composto da 6 byte. Quest'ultimo è unico in tutto il mondo, ed impresso dal costruttore nella scheda di interfaccia. L'indirizzo Ethernet viene però utilizzato solo nell'ambito della LAN di cui il nodo fa parte, ossia dopo che i pacchetti sono stati instradati dai router, per mezzo dell'indirizzo IP, verso la LAN.

**Sottoreti** Ogni nodo conosce, oltre al proprio indirizzo IP, anche una *maschera di sottorete* composta da una serie di uni seguita da zeri, in numero complessivo di 32 bit, tanti quanti ne sono presenti nell'indirizzo IP. Il termine *maschera* è dovuto all'operazione di AND binario (vedi tabella) operata tra la maschera e gli indirizzi IP, per determinare se questi appartengano alla propria stessa LAN oppure risiedano altrove.

| Indirizzo IP | Maschera Sottorete | Indirizzo sottorete |
|---|---|---|
| 151.100.8.33 | 255.255.255.0 | 151.100.8.0 |

Nel caso in cui la sottorete di un nodo Y verso cui il nodo X deve inviare un pacchetto è la stessa su cui è connesso X, allora questi può individuare l'indirizzo Ethernet del destinatario[53] ed inviargli il pacchetto direttamente. In caso contrario, X invierà il pacchetto al proprio *default gateway* verso Internet.

**Intranet** Alcuni gruppi di indirizzi IP (come quelli 192.168.*w.z* oppure 10.*y.w.z* ) non vengono instradati dai router, e possono essere riutilizzati nelle *reti private* di tutto il mondo per realizzare le cosiddette *reti intranet* operanti con gli stessi protocolli ed applicativi che funzionano via Internet.

**Domain Name Server (dns)** L'utente di una applicazione internet in realtà non è a conoscenza degli indirizzi IP dei diversi nodi, ma li identifica per mezzo di nomi simbolici del tipo nodo.dominio.tld, detti anche *indirizzi Internet*. Il processo di risoluzione che individua l'indirizzo IP associato al nome avviene interrogando un particolare nodo, il DOMAIN NAME SERVER (*servente dei nomi di dominio*). La struttura dei nomi, scandita dai punti, individua una gerarchia di autorità per i diversi campi. Il campo tld è chiamato *dominio di primo livello* (TOP LEVEL DOMAIN[54]), 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

[51] Con 4 byte si indirizzano (in linea di principio) $2^{32} = 4.29 \cdot 10^9$ diversi nodi (più di 4 miliardi). E' tuttora in sviluppo il cosiddetto IPv6, che estenderà l'indirizzo IP a 16 byte, portando la capacità teorica a $3.4 \cdot 10^{38}$ nodi. L'IPv6 prevede inoltre particolari soluzioni di suddivisione dell'indirizzo, allo scopo di coadiuvare le operazioni di *routing*.

[52] 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.

[53] Mostriamo in seguito che questo avviene mediante il protocollo ARP.

[54] 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.

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[55] per conoscerne l'IP. Nella rete sono presenti molti DNS, alcuni dei quali detengono informazioni *autorevoli*[56] riguardo ai nodi di uno o più domini, altri (i DNS *radice*, o ROOT) detengono le informazioni relative a quali DNS siano autorevoli per i domini di primo livello, ed altri fanno da tramite tra i primi due ed i *client* che richiedono una risoluzione di indirizzo. Se il DNS di nodoX *non è autorevole per* nodoY, allora[57] 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 perchè la stessa sequenza di operazioni *non deve* essere effettuata sempre: il DNS utilizzato da nodoX riceve infatti, assieme all'IP di nodoY, anche una informazione detta TIME TO LIVE (TTL o *tempo di vita*) che descrive la scadenza della coppia *nome-IP* ottenuta. Genericamente il TTL è di qualche giorno, e fino alla sua scadenza il DNS *ricorda*[58] la corrispondenza, in modo da fornire la propria copia in corrispondenza delle richieste future, e ridurre sensibilmente il traffico legato alla risoluzione degli indirizzi Internet. L'insieme delle risoluzioni apprese è denominata *cache* del DNS[59].

**Indirizzi TCP** Si è detto che ogni nodo è individuato in Internet mediante il proprio indirizzo IP, ma questo non è sufficiente ad indicare con quale particolare programma (che può essere uno specifico *server* come nel caso del DNS) si vuole entrare in comunicazione. I programmi che sono pronti a ricevere connessioni si pongono *in ascolto* su ben determinate *porte* (o *socket*[60]), identificate da numeri[61], 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,

[55] Il "proprio" DNS viene configurato per l'host in modo fisso, oppure in modo dinamico da alcuni Service Provider raggiungibili per via telefonica, e convenientemente corrisponde ad un nodo situato "vicino" al nodo che lo interroga.

[56] 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.

[57] 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.

[58] 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.

[59] CACHE è un termine generico, che letteralmente si traduce *nascondiglio dei viveri*, e che viene adottato ogni volta si debba indicare una memoria che contiene copie di riserva, o di scorta...

[60] *Socket* è un termine che corrisponde alla... presa per l'energia elettrica casalinga, ed in questo contesto ha il significato di una *presa* a cui si "attacca" il processo che richiede la comunicazione. Per l'esattezza, un *Socket internet* è individuato dal numero di porta TCP e dall'indirizzo IP.

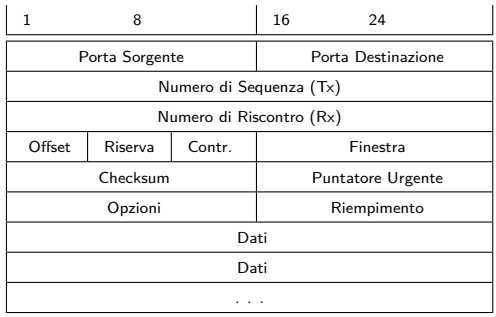
[61] Spesso gli indirizzi che identificano i punti di contatto di servizi specifici vengono indicati come SERVICE ACCESS POINT (SAP), anche per situazioni differenti dal caso specifico delle porte del TCP.

apre con il chiamante una connessione di ritorno, su di un numero di porta diverso, che è stato comunicato dal chiamante al momento della richiesta di connessione, e per il quale il chiamante non ha già aperto altre connessioni differenti.

#### 8.6.1.4 TCP

Discutiamo ora del TCP[62], che offre ai processi applicativi un servizio di trasporto a circuito virtuale, *attaccato* ad una porta[63] di un nodo remoto individuato dall'indirizzo IP. Il suo compito è quello di ricevere dai processi applicativi dei dati, suddividerli in pacchetti, ed inviarli al suo pari che svolge il processo inverso.

**Il pacchetto TCP** La struttura di un pacchetto TCP è mostrata in figura, e comprende una intestazione composta da 6 gruppi (o più) di 4 byte per un minimo di 192 bit, a cui segue un numero variabile di gruppi di 4 byte di Dati, provenienti dagli strati applicativi superiori.




Troviamo subito i numeri delle porte a cui si riferisce la connessione, mentre gli indirizzi IP sono aggiunti dallo strato di rete. I numeri *di Sequenza* e di *Riscontro* servono rispettivamente a numerare i bytes dei pacchetti uscenti, ed a notificare l'altro lato del collegamento del numero di sequenza del prossimo byte che si aspetta di ricevere[64], 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;

[62] TCP = *Transport Control Protocol*.

[63] 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 conessione 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*.

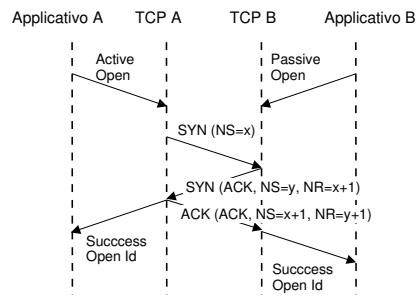
[64] Il numero di sequenza si incrementa ad ogni pacchetto di una quantità pari alla sua dimensione in bytes, ed ha lo scopo di permettere le operazioni di controllo di flusso. Il valore iniziale del numero di sequenza e di riscontro è diverso per ogni connessione, e generato in modo pseudo-casuale da 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, cosicchè il pacchetto duplicato e ritardato risulta fuori sequenza, e non viene accettato.

SYN è pari ad uno solo per il primo pacchetto inviato per richiedere di creare una connessione; FIN indica che la sorgente ha esaurito i dati da trasmettere.

I 16 bit di *Finestra* rappresentano il numero di byte che, a partire dal numero espresso dal *Numero di Riscontro*, chi invia il pacchetto è in grado di ricevere, ed il suo utilizzo sarà meglio illustrato tra breve nel contesto del controllo di flusso. Il *Checksum* serve al ricevente per verificare se si sia verificato un errore, il *Puntatore Urgente* contiene il numero di sequenza dell'ultimo byte di una sequenza di dati urgenti, e le *Opzioni* (di lunghezza variabile) sono presenti solo raramente, ed utilizzate a fini di controllo, ad esempio per variare la dimensione della finestra. Infine, il *Riempimento* conclude l'ultima parola da 32 bit.

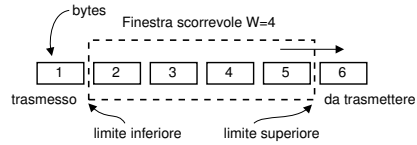
Uno stesso pacchetto TCP può svolgere funzioni di sola segnalazione, o di sola trasmissione dati, od entrambe.

**Apertura e chiusura della connessione** Il TCP offre un servizio di di trasporto a circuito virtuale, e prima di inviare dati, deve effettuare un colloquio iniziale con il nodo remoto di destinazione. In particolare, il colloquio ha lo scopo di accertare la disponibilità del destinatario ad accettare la connessione, e permette alle due parti di scambiarsi i rispettivi numeri di sequenza descritti alla nota 64.




L'estremo che viene "chiamato" riveste il ruolo di *server*, e l'altro di *client*. Dato che anche quest'ultimo deve riscontrare il numero di sequenza fornito dal server, occorrono tre pacchetti per terminare il dialogo, che prende il nome di THREE WAY HANDSHAKE[65]. 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[66]. La chiusura può avvenire per diverse cause: o perchè è terminato il messaggio, segnalato dal bit FIN, o per situazioni anomale, che il TCP indica con il bit RST.

**Protocollo a finestra** Allo scopo di realizzare un controllo di flusso, il TCP prevede l'uso dell'$NR$ inviato dal ricevente per dosare il ritmo con cui trasmettere i propri pacchetti.




La lunghezza di *Finestra* comunicata con il SYN del ricevente, determina la quantità di memoria riservata per i buffer dedicati alla connessione, che viene gestita come una memoria a scorrimento o *finestra scorrevole* (SLIDING WINDOW). Questa memoria è presente per gestire i casi di pacchetti ritardati o fuori sequenza, e contiene i bytes già trasmessi. Il trasmittente (vedi figura) non fa avanzare il limite inferiore finchè non

[65] HANDSHAKE = *stretta di mano.*

[66] Per ciò che riguarda i valori dei numeri di riscontro $NR$, questi sono incrementati di 1, perchè la *finestra* (descritta nel seguito) inizia dai bytes del prossimo pacchetto, a cui competeranno appunto valori di $NS$ incrementati di uno.

riceve un riscontro con $NR$ maggiore di tale limite. In questo modo non occorre attendere il riscontro di tutti i bytes, o di tutti i pacchetti (che devono comunque essere di dimensione inferiore alla finestra), ma ci può avvantaggiare trasmettendo l'intero contenuto della finestra.

Una finestra del tutto analoga è utilizzata dal ricevente, allo scopo di ricomporre l'ordine originario dei pacchetti consegnati disordinatamente dallo strato IP di rete. Non appena il ricevente completa un segmento contiguo al limite inferiore, sposta quest'ultimo in avanti di tanti bytes quanti ne è riuscito a leggere in modo contiguo, ed invia un riscontro con $NR$ pari al più basso numero di byte che ancora non è pervenuto[67].

Nel caso in cui sia settato il bit URG[68], si stanno inviando dati urgenti fuori sequenza, e che non devono rispettare il protocollo a finestra, come ad esempio per recapitare un segnale di interrupt relativo ad una sessione Telnet per terminare una applicazione remota.

**Controllo di errore** Trascorso un certo tempo (detto *timeout*) nell'attesa di un riscontro, il trasmittente ritiene che alcuni pacchetti sono andati persi, e li re-invia[69]. Il valore del *timeout* viene calcolato dinamicamente dal TCP in base alle sue misure di *round-trip delay*[70], ossia del tempo che intercorre in media tra invio di un pacchetto e ricezione del suo riscontro. In questo modo il TCP si adatta alle condizioni di carico della rete ed evita di ri-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 tramissione 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 un dimensione di finestra ridotta, e poi aumentarla nel caso in cui non si verifichino errori, la rete sopporti il traffico, ed i nodi abbiano memoria disponibile.

**Controllo di congestione** Il TCP può usare la sua misura di *round-trip delay* come un indicatore di congestione della rete, e lo scadere di un *timeout* come un segnale del peggioramento della congestione. In tal caso quindi, può essere ridotta la dimensione della finestra di trasmissione, caricando così di meno la rete.

**UDP** Lo *User Datagram Protocol* è ancora un protocollo di trasporto, che opera senza connessione, e sostituisce il TCP per inviare pacchetti isolati, o serie di pac-

[67] 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*.

[68] In tal caso, il campo *Puntatore Urgente* contiene il numero di sequenza del byte che delimita superiormente i dati che devono essere consegnati urgentemente.

[69] 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.

[70] Con licenza poetica: *il ritardo del girotondo*, che qui raffigura un percorso di andata e ritorno senza soste.

chetti la cui ritrasmissione (se perduti) sarebbe inutile. Ad esempio, è utilizzato nella trasmissione di dati in tempo reale, oppure per protocolli di interrogazione e controllo come il DNS.

#### 8.6.1.5 IP

*L'Internet Protocol* costituisce l'ossatura della rete internet, realizzandone i servizi di rete ed interfacciando le diverse sottoreti a cui sono connessi i nodi. Le sue principali funzioni sono pertanto l'indirizzamento, l'instradamento e la variazione della dimensione[71] 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[72], e questi possono essere persi, duplicati o consegnati fuori sequenza.

l'IP riceve dallo strato superiore (il TCP od un altro protocollo) un flusso di byte suddivisi in pacchetti, a cui si aggiunge l'indirizzo IP di destinazione; tale flusso è utilizzato per riempire un proprio buffer di dimensione opportuna, che quando pieno (od al termine del pacchetto ricevuto *dall'alto*) è *incapsulato* aggiungendo una intestazione (*l'header*) che codifica la segnalazione dello strato di rete realizzato dal protocollo IP.

**L'intestazione IP** contiene le informazioni mostrate alla figura successiva.


Il campo *VER* indica quale versione si sta utilizzando, e permette sperimentazioni e mi-glioramenti 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[73] richiesta. L'*identifi-cazione* riporta lo stesso valore per tutti i frammenti di uno stesso datagramma, mentre l'*Offset di frammento* indica la posizione del frammento nel datagramma (con frammenti di dimensione multipla di 8 byte).

Solo 2 dei tre bit di *Flag* sono usati, DF (*Don't Fragment*) per richiedere alla rete di non frammentare il datagramma, e MF (*More Fragments*) per indicare che seguiranno altri frammenti. Il *TTL* (Time To Live) determina la massima permanenza del pacchetto nella rete[74], 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[75].

[71] 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.

[72] 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).

[73] 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 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.

[74] 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.

[75] 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.

Gli *Indirizzi IP* di sorgente e destinazione hanno l'evidente funzione di recapitare correttamente il messaggio, mentre il campo *Opzioni* ha una lunghezza variabile, può essere omesso, e consente ad esempio di richiedere il tracciamento della serie di router attraversati.

**Indirizzamento e Routing** A pagina 185 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[76]. In realtà la gerarchia presenta una ulteriore suddivisione, dettata sia da esigenze amministrative che funzionali.

I bit più significativi dell'indirizzo IP identificano 5 diversi gruppi (o *classi*) di indirizzi, descritti dalla seguente tabella:

| Inizio IP addr | Classe | bit rete/nodo | N. reti | N. nodi per rete |
|---|---|---|---|---|
| 0 | A | 7/24 | 128 | 16 777 216 |
| 10 | B | 14/16 | 16 384 | 65 536 |
| 110 | C | 21/8 | 2 097 152 | 256 |
| 1110 | D | 28 bit di indirizzo multicast per 268 435 456 canali | | |
| 11110 | E | 27 bit per usi futuri e ricerca | | |

Quando una organizzazione decide di essere presente in internet, richiede l'assegnazione di un lotto di indirizzi IP ad apposti organismi, i quali attribuiscono all'organizzazione un gruppo di indirizzi di classe A, B o C in base al numero di nodi che l'organizzazione prevede di mettere in rete. Una rete in classe B ad esempio è individuata da 14 bit (ossia, assieme ai bit di classe, dai primi due bytes dell'indirizzo IP), e quindi esistono $2^{14} = 16384$ diverse reti in classe B, ognuna con una capacità di $2^{16} = 65536$ diversi nodi. Chi è intestatario di un gruppo di indirizzi, provvede ad assegnarli ai singoli nodi della propria sottorete.

**Subnetting e Supernetting** Osserviamo ora che la maschera di sottorete presentata a pag. 185 *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 intestaria della rete in classe B[77]. 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, anzichè un elemento per ogni singola sottorete.

[76] 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.

[77] In questo caso, l'Università di Roma "La Sapienza" è intestataria della rete 151.100.

**Classless Interdomain Routing - cidr** Nella prima metà degli anni '90 apparve evidente che il partizionamento degli indirizzi nelle tre classi A, B e C non era rispondente alle richieste dell'utenza; accadeva infatti che le reti in classe C erano troppo "piccole", mentre quelle in classe B rischiavano di esaurisi a breve, pur essendo sfruttate molto poco[78]. Per questo motivo, è stata rimossa la suddivisione rigida nelle tre classi, e si è sistematicamente applicato il principio del supernetting. In pratica, si è ridefinita la maschera di sottorete, come una sequenza di *uni* allineata a sinistra, permettendo così di definire reti di dimensione pari a una potenza di due qualsiasi. Come risultato, ora una sottorete è identificata da una coppia indirizzo/maschera del tipo (ad es.) 172.192.0.0/12, che rappresenta tutti $2^{20}$ indirizzi che vanno da 172.192.0.0 a 172.207.255.255, che hanno i 12 bit più elevati uguali a 101011001100: questa sequenza prende il nome di *prefisso* della rete. In definitiva quindi, la maschera è espressa come il numero di bit più significativi in comune a tutti i nodi della sottorete.

**Longest Match** Un router decide su che porta instradare un pacchetto IP in base al confronto tra l'indirizzo di destinazione e tutti i prefissi presenti nella tabella di routing, associati ciascuno alla "migliore" porta di uscita verso la sottorete definita dal prefisso. Nel caso in cui si verifichi più di una uguaglianza, si sceglie l'instradamento caratterizzato dal *maggior numero* di bit coincidenti, ossia relativo al prefisso *più lungo*. Infatti, in tal modo viene preferita la direzione *più specifica* verso la destinazione finale. In assenza di uguaglianze invece, il pacchetto è inoltrato in base ad una *default route*, che tipicamente rimanda la decisione ad un router "gerarchicamente più elevato"[79].

**Sistemi Autonomi e Border Gateway** Vi sono router collegati direttamente con le LAN, e configurati per instradare correttamente i pacchetti diretti a destinazioni locali. Vi sono poi router collegati solo ad altri router, che *apprendono* gli instradamenti verso le reti locali mediante appositi *protocolli di routing* che consentono ai router di primo tipo di *pubblicizzare* (ADVERTISE) le reti raggiungibili direttamente, ed ai router del secondo tipo di fare altrettanto nei confronti dei loro pari.

L'insieme di sottoreti (e router, nodi e DNS) gestite da una stessa organizzazione prende il nome di *Autonomous System* (AS), e nel suo ambito sono attivi protocolli di routing detti *Interior Gateway Protocols* (IGP), che distribuiscono le informazioni di raggiungibilità interna. Alcuni router di uno stesso AS svolgono il roulo 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 protocolo 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.

[78] Ad esempio, organizzazioni con poco più di un migliaio di nodi erano costrette a richiedere una intera classe B con capacità di 65536 nodi.

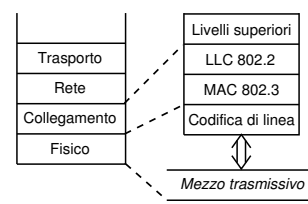
[79] Sebbene la topologia di Internet possa essere qualunque, nella pratica esistono dei *carrier* internazionali che svolgono la funzione di *backbone* (spina dorsale) della rete, interconnettendo tra loro i continenti e le nazioni.

**Multicast** Tornando all'esame della tabella di pag. 191, in cui la classe E costituisce evidentemente una "riserva" di indirizzi per poter effettuare sperimentazioni, la classe D individua invece dei canali *multicast*[80]. Quando un nodo decide di aderire ad un canale multicast, invia un messaggio[81] 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[82] evita infatti di aprire una connessione dedicata per ogni destinatario, ma si suddivide via via nella rete solo quando i destinatari sono raggiungibili da vie diverse.

#### 8.6.1.6 Ethernet

Ci occupiamo qui di un caso particolare di realizzazione dei primi due livelli del modello ISO-OSI. Come anticipato a pag. 185, 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[83]: se infatti questi fossero usati come indirizzi a livello di rete, le tabelle di instradamento dovrebbero essere a conoscenza di *tutti* i nodi esistenti[84]. Puntualizziamo inoltre che un nodo di Internet può essere connesso alla rete anche per via telefonica, o con svariati altri metodi; ci limitiamo qui a descrivere il caso delle LAN Ethernet, peraltro particolarmente diffuso.

Ethernet individua un particolare tipo di pacchetto dati, adottato inizialmente dalla Xerox, adatto ad incapsulare dati provenienti da protocolli diversi. Successivamente, il formato è stato standardizzato dall'IEEE, e per ciò che ci interessa le specifiche sono quelle identificate dalle sigle 802.2 (Logical Link Control, LCC) 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.

[80] Il termine *multicast* è ispirato alle trasmissioni *broadcast* effettuate dalle emittenti radio televisive.

[81] 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.

[82] Data l'impossibilità a stabilire un controllo di flusso con tutti i destinatari, il traffico multicast viaggia all'interno di pacchetti UDP.

[83] E rappresenta quindi ciò che viene detto uno *spazio di indirizzi piatto* (flat address space).

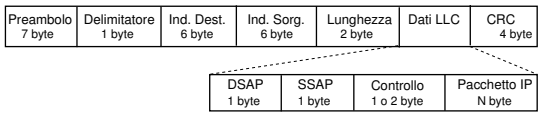
[84] Al contrario, il partizionamento dell'indirizzo IP in rete+nodo permette di utilizzare tabelle di routing di dimensioni gestibili.

**Address Resolution Protocol - arp** Quando un pacchetto IP giunge ad un router, e l'indirizzo IP indica che il destinatario è connesso ad una delle LAN direttamente raggiungibili dal router[85], questo invia su quella LAN un pacchetto *broadcast*, su cui viaggia una richiesta ARP (ADDRESS RESOLUTION PROTOCOL), allo scopo di individuare l'indirizzo Ethernet del nodo a cui è assegnato l'indirizzo IP di destinazione del pacchetto arrivato al router. Se tale nodo è presente ed operativo, riconosce che la richiesta è diretta a lui, ed invia un pacchetto di risposta comunicando il proprio indirizzo Ethernet, che viene memorizzato dal router in una apposita tabella[86].

Operazioni simili sono svolte da ognuno dei nodi della LAN, ogni volta che debbano inviare un pacchetto ad un altro nodo direttamente connesso alla stessa rete locale. Se al contrario l'IP di destinazione non fa parte della stessa LAN, il pacchetto è inviato al *default gateway*.

**Formato di pacchetto** Il pacchetto Ethernet è generato dall'LLC e dal MAC, ognuno dei quali incapsula il pacchetto IP con le proprie informazioni di protocollo.

Nella figura seguente è mostrato il risultato finale delle operazioni. In testa troviamo 7 byte di *preambolo*, necessario a permettere la sincronizzazione dell'orologio del ricevente con quello in trasmissione; dato che la sincronizzazione richiede un tempo non noto a priori, un byte di *flag* segnala l'inizio del pacchetto. Troviamo quindi gli *indirizzi Ethernet* di sorgente e destinazione, due byte che indicano la *lunghezza* della restante parte del pacchetto, e quindi l'incapsulamento dei dati prodotti dall'LLC. In fondo, sono presenti 4 byte che realizzano il *controllo di errore*.




L'LLC inserisce invece, 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*)[87] che pone i campi DSAP, SSAP e controllo pari a `0xAAAA03`, a cui aggiunge altri 5 bytes, dei quali i primi tre sono denominati OUI (*Organizationally Unique Identifier*) che, se posti tutti a zero, stabiliscono che i due byte seguenti (indicati come *protocol ID*) debbano essere interpretati come un codice *Ethertype*[88], lo stesso usato nel formato *Ethernet II* discusso appresso, permettendo quindi di specificare finalmente il protocollo incapsulato.

[85] Ad ogni porta del router è associata una coppia sottorete/maschera (vedi pag. 185) che descrive l'insieme degli indirizzi direttamente connessi alla porta. La verifica di raggiungibilità (o *adiacenza*) è attuata mettendo in AND l'IP di destinazione con le maschere, e confrontando il risultato con quello dell'AND tra le maschere e gli indirizzi delle sottoreti collegate.

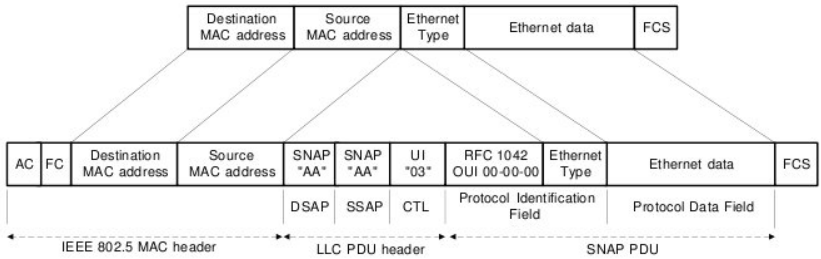
[86] Dato che i nodi possono essere spostati, possono cambiare scheda di rete e possono cambiare indirizzo IP assegnatogli, la corrispondenza IP-Ethernet è tutt'altro che duratura, ed ogni riga della tabella ARP indica anche quando si sia appresa la corrispondenza, in modo da poter stabilire una scadenza, ed effettuare nuovamente la richiesta per verificare se sono intervenuti cambiamenti topologici.

Se il nodo ha cambiato IP, ma non il nome, sarà il TTL del DNS (mantenuto aggiornato per il dominio del nodo) a provocare il rinnovo della richiesta dell'indirizzo.

[87] `http://en.wikipedia.org/wiki/Subnetwork_Access_Protocol`

[88] `http://en.wikipedia.org/wiki/EtherType`

Infine, viene molto frequentemente usato un formato di trama ancora diverso, detto *Ethernet II* o DIX[89], 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.




**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[90] non nullo, e determina un periodo (detto di *contesa*, e che dipende dalla massima lunghezza del cavo) entro il quale un nodo può erroneamente credere che nessun altro stia trasmettendo.

Qualora sia rilevata una contesa, i due nodi smettono di trasmettere, e riprovano solo dopo una attesa di durata casuale.

**Trasmissione** Il segnale relativo al pacchetto Ethernet viene trasmesso adottando una codifica di linea di tipo Manchester differenziale. La configurazione con tutti i nodi collegati su di uno stesso cavo è detta *a bus*, e sono state coniate apposite sigle per identificare il tipo di connessione, come ad esempio 10BASE5 e 10BASE2, relative al

[89] 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

[90] 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 $\mu sec$ perchè un segnale si propaghi da un estremo all'altro. Dato che è permesso di congiungere fino a 5 segmenti di rete per mezzo di ripetitori, e che anch'essi introducono un ritardo, si è stabilito che la minima lunghezza di un pacchetto Ethernet debba essere di 64 byte, che alla velocità di trasmissione di 10 Mbit/sec corrisponde ad una durata di 54.4 $\mu sec$, garantendo così che se si è verificata una collisione, le due parti in causa possano accorgersene.

collegamento di banda base a 10 Mbps, su cavo *tick* e *thin*[91], con estensione massima 500 e 200 metri[92].

#### 8.6.1.7 Fast e Gigabit Ethernet

Mentre si proponeva ATM come una soluzione idonea per quasi tutti gli ambiti, la tecnologia Ethernet ha incrementato la velocità trasmissiva di un fattore pari a mille, e si propone sempre più come soluzione generalizzata.

**Fast Ethernet** Nel 1995 è stato definito lo standard IEEE 802.3u detto *Fast Ethernet*, che eleva la velocità di trasmissione a 100 Mbps ed impiega due diversi cavi UTP[93] per le due direzioni di trasmissione, rendendo eventualmente la comunicazione *full-duplex*[94]. In quest'ambito sono definiti i sistemi 10BaseT e 100BaseT, relativi all'uso del cavo UTP anzichè di un coassiale, e prevedono una topologia *a stella* per la LAN, realizzata utilizzando una unità centrale (detta HUB=*mozzo di ruota*) da cui si dipartono tanti cavi, ognuno che collega un unico nodo. Nel caso di un HUB economico, questo svolge solo le funzioni di ripetitore (ritrasmette tutto su tutte le sue porte) e dunque le collisioni possono ancora verificasi.

**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, cosicchè tutti i pacchetti destinati all'esterno non sono ritrasmessi sugli altri rami della LAN, ed il traffico tra i nodi connessi allo SWITCH non si propaga al resto della LAN.

La lunghezza massima dei collegamenti è ora ridotta a 100 metri, per il motivo che un pacchetto di dimensione minima di 64 byte trasmesso a 100 Mbps, impiega un tempo che è $\frac{1}{10}$ di quello relativo alla velocità di 10 Mbps, e quindi per consentire la detezione di collisione, si è dovuta ridurre di pari misura la massima distanza tra nodi trasmittenti.

**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'efficenza 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 *vituali* (VLAN), che non scambiano traffico, realizzando di fatto molteplici LAN con

[91] TICK = *duro* (grosso), THIN = *sottile*. Ci si riferisce al diametro del cavo.

[92] Le sigle indicano infatti la velocità, se in banda base o meno, e la lunghezza della tratta.

[93] UNSHIELDED TWISTED PAIR (UTP), ossia la coppia ritorta non schermata.

[94] 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.

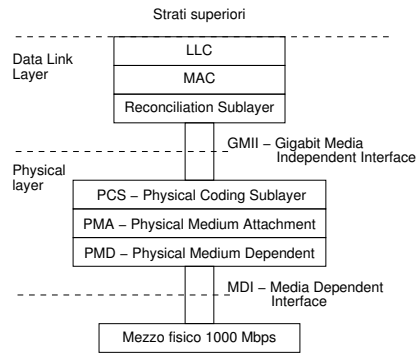
uno stesso cablaggio. Per interconnettere le LAN, occorre attraversare un dispositivo router.

**Gigabit Ethernet** Nel giugno 1998 viene standardizzato l'IEEE 802.3z, che porta ad 1 Gbps la velocità di trasmissione delle trame Ethernet, rimpiazzando lo strato di codifica di linea dell'802.3 con i due strati inferiori dell'ANSI X3T11 *Fiber Channel*. In questo modo, si mantiene la compatibilità con gli strati LLC e MAC di Ethernet, mentre la trasmissione avviene su fibra ottica o su cavo in accordo alla tabella seguente.

| media | distanza | mezzo | sorgente |
|---|---|---|---|
| 1000BASE-SX | 300 m | f.o. multimodo $\phi$ 62.5 $\mu m$ | laser 850 nm |
| | 550 m | f.o. multimodo $\phi$ 50 $\mu m$ | laser 850 nm |
| 1000BASE-LX | 550 m | f.o. multimodo $\phi$ 50 o 62.5 $\mu m$ | laser 1300 nm |
| | 3000 m | f.o. monomodo $\phi$ 9 $\mu m$ | laser 1300 nm |
| 1000BASE-CX | 25 m | cavo STP (*shielded twisted pair*) | |
| 1000BASE-T | 25-100 m | 4 coppie di cavo UTP | |

**Packet bursting** Dato che ora la velocità di trasmissione è 10 volte quella del fast Ethernet, la compatibilità con il MAC CSMA/CD richiederebbe di ridurre la massima lunghezza del collegamento a 10 metri. Al contrario, è stata aumentata la durata minima di una trama portandola a 512 byte, in modo da aumentare la durata della trasmissione e garantire la detezione di collisione. In effetti, il MAC ethernet continua a produrre pacchetti di durata minima 64 byte, e questi sono riempiti (PADDED) fino a 512 byte con una *carrier extension* di simboli speciali. Questa operazione è particolarmente inefficiente se i pacchetti da 64 byte sono frequenti; in tal caso si attua allora il *packet bursting* che, esauriti i 512 byte minimi realizzati come indicato, accoda gli ulteriori pacchetti nello stesso burst trasmissivo, fino ad una lunghezza di 1500 byte.

**Architettura** La figura seguente mostra la pila protocollare per Gigabit Ethernet. La GMII permette di usare lo strato MAC con qualunque strato fisico, ed opera sia in full-duplex che in half-duplex, alle velocità di 10, 100 e 1000 Mbps, mediante due percorsi dati (TX e RX) da 8 bit, più due segnali di strato per indicare presenza di portante e detezione di collisione, che sono mappati dal RS nelle primitive riconosciute dallo strato MAC preesistente.




Lo strato fisico è suddiviso in tre sottolivelli. Il PCS fornisce una interfaccia uniforme al RS per tutti i media. Provvede alla conversione 8B/10B tipica del *Fiber Channel*, che rappresenta gruppi di 8 bit mediante *code group* da 10 bit, alcuni dei quali rappresentano i simboli, ed altri sono codici di controllo, come quelli usati per la *carrier extension*. Il PCS genera inoltre le indicazioni sulla portante e sulla collisione, e gestisce la auto-negoziazione sulla velocità di trasmissione e sulla bidirezionalità del media.

Il PMA provvede alla conversione parallelo-serie e viceversa, mentre il PMD definisce l'MDI, ossia la segnalazione di strato fisico necessaria ai diversi media, così come il tipo di connettore.

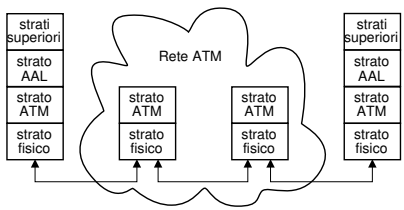
**Ripetitore full-duplex e controllo di flusso** Qualora tutte le porte di un ripetitore siano di tipo full-duplex, non può più verificarsi contesa di accesso al mezzo, e la contesa avviene all'interno del ripetitore, che (non essendo un SWITCH) copia tutte le trame in ingresso (debitamente bufferizzate in apposite code) in tutte le code associate alle porte di uscita. Pertanto, la lunghezza massima dei collegamenti non è più dettata dalla necessità di rilevare collisioni, ma dalle caratteristiche del mezzo trasmissivo. D'altra parte, possono verificarsi situazioni di *flooding* delle code di ingresso; il comitato IEEE 802.3x ha quindi definito un meccanismo di controllo di flusso, che mette in grado i ripetitori (e gli switch) di richiedere ai nodi connessi la sospensione temporanea della trasmissione.

**10 Gigabit Ethernet** Nel 2002 viene definito lo standard IEEE 802.3ae, che stabilisce le modalità operative di un collegamento Ethernet operante solo in full duplex su fibra ottica. Lo standard prevede di interoperare con la trasmissione SONET/SDH.

### 8.6.2 Rete ATM

La sigla ATM sta per *Asynchronous Transfer Mode*, ed identifica una particolare rete progettata per trasportare indifferentemente traffico di diversa natura, sia di tipo dati che real-time[95], che per questo motivo è indicata anche come B-ISDN[96]. Il suo funzionamento si basa sul principio della *commutazione di cella* (CELL SWITCHING), dove per cella si intende un pacchetto di lunghezza fissa di 53 byte. I primi 5 byte delle celle contengono un identificativo di connessione, ed il loro instradamento avviene mediante dei circuiti virtuali. La commutazione delle celle tra i nodi di rete ha luogo in maniera particolarmente efficiente, e questa è una delle caratteristiche più rilevanti dell'ATM.

**Architettura** La rete ATM viene indicata anche come una *Overlay Network*, in quanto operativamente si *sovrappone* ai livelli inferiori di una rete esterna.




Dal canto suo, ATM è strutturata sui tre strati funzionali di adattamento (AAL), di commutazione ATM, e fisico. Mentre i nodi ai bordi della rete devono realizzare tutti e tre gli strati, i nodi interni svolgono solo le funzioni attuate da quelli inferiori. La tabella 8.1 riporta le principali funzioni svolte dai tre strati, e pone in evidenza come in uno stesso strato siano identificabili diverse sotto-funzioni.

**Strato fisico** Il mezzo primario di trasmissione (con cui è in contatto il sotto-strato PM) per ATM è la fibra ottica, in accordo alla struttura di trama dell'SDH/SONET, per la quale sono state standardizzate le velocità di 1.5 e 2 Mbps (DS1/E1), 155 Mbps

[95] Per traffico real-time si intende sia quello telefonico, sia più in generale quello di natura multimediale.

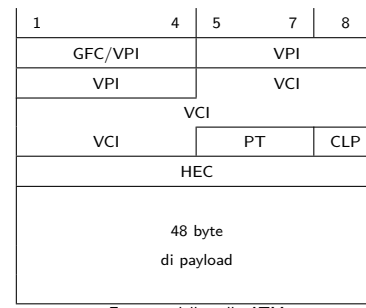
[96] Siamo alla fine degli anni '80, e la definizione *Integrated Service Data Network* (ISDN) si riferisce ad una rete in grado di permettere, oltre al normale trasporto dei dati, anche servizi di rete. La rete ISDN era però limitata ad una velocità massima (presso l'utente) di 2 Mbps, e per questo venne chiamata *narrow-band* ISDN (N-ISDN). A questa, avrebbe fatto seguito la *broad-band* ISDN (B-ISDN) che ha poi dato luogo alla definizione dell'ATM.

| strato | sotto-strato | funzioni |
|---|---|---|
| ATM Adaptation Layer (AAL) | • 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 8.1: Stratificazione delle funzioni in una rete ATM

(OC3) e 622 Mbps (OC12c). La velocità di 155 Mbps è disponibile anche su FIBRE CHANNEL, e su cavo ritorto, mentre la velocità di 100 Mbps è disponibile su FDDI. Infine, sono previste anche velocità di interconnessione di 139, 52, 45, 34 e 25 Mbps.

In funzione del mezzo trasmissivo, può variare la *struttura di trama*[97] (mostrata in figura) in cui vanno inserite le celle. Il quinto byte della intestazione di cella, contiene l'*Header Error Code* (HEC) calcolato sui 4 byte precedenti, che viene usato in ricezione per rivelare due errori e correggerne uno[98]. 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[99]. 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.



Formato della cella ATM

**Strato ATM** Mentre lo strato fisico si occupa di trasmettere e ricevere celle, lo strato ATM si occupa di elaborarle. Nei nodi di frontiera, le celle sono multiplate e demultiplate, mentre dentro la rete, sono commutate tra gli ingressi e le uscite.

Nei primi quattro byte dell'header di cella trova posto *l'etichetta* necessaria a realizzare il trasferimento a circuito virtuale; questa etichetta è suddivisa in due campi, il *Virtual Path Identifier* (VPI) ed il *Virtual Channel Identifier* (VCI)[100].

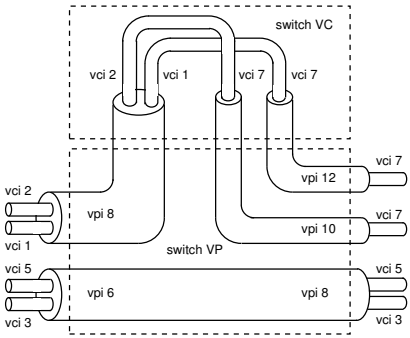
[97] Sono definite due tipi di *interfaccia utente-network* (UNI): quella SDH/SONET, in cui le celle sono inserite nel *payload* della trama SDH, e quella CELL-BASED, che prevede un flusso continuo di celle. Mentre nel primo caso il bit rate lordo comprende *l'overhead* di trama, nel secondo comprende la presenza di celle di tipo *Operation and Maintenance* (OAM).

[98] Nel primo caso la cella viene scartata, mentre nel secondo inoltrata correttamente. La presenza di più di due errori, provoca un errato inoltro della cella.

[99] Le celle IDLE sono riconoscibili in base ad una particolare configurazione dei primi 4 byte dell'header, così come avviene per le celle OAM, nonchè per altri tipi particolari di cella, che trasportano la segnalazione degli strati superiori.

[100] Mentre per VCI sono riservati 16 bit, per VPI si usano 12 bit all'interno della rete, e 8 bit ai

Il motivo della suddivisione risiede nella possibilità di raggruppare logicamente diversi circuiti virtuali che condividono lo stesso percorso nella rete. Nei collegamenti di cui è composto il percorso comune, viene usato uno stesso vpi per tutte le celle, mentre le diverse connessioni su quel percorso sono identificate mediante diversi vci. L'instradamento congiunto delle celle con uguale vpi è effettuato nei nodi (vp switch), che si occupano solo[101] 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 a da zero o più VP switch, ed è delimitata tra due nodi (VC o VP/VC switch[102]) 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)[103] e comprende uno o più VPC, coincidendo spesso[104] con il percorso tra ingresso ed uscita[105] 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)[106]; L'OGGETTO DELLA RICHIESTA PUÒ ESSERE UNA VCC od una VPC, ed in questo secondo caso la VPC verrà usata per tutte le VCC future tra i due nodi.

**Classi di traffico e Qualità del Servizio (QoS)** Nella fase di *setup*, sono attuate delle verifiche dette *Connession Admission Control* (CAC) per assicurarsi che la nuova connessione non degradi le prestazioni di quelle già in corso, nel qual caso la chiamata è rifiutata. La sua accettazione determina invece la stipula tra utente e rete di un *Traffic Contract* a cui la sorgente si deve attenere. Nel corso della trasmissione, i nodi ATM verificano che le caratteristiche del traffico in transito nelle VCC siano conformi al rispettivo contratto, svolgendo un *Usage Parameter Control* (UPC) detto anche *policing*[107]. Prima di proseguire, forniamo però alcune definizioni.

suoi bordi, riservando 4 bit indicati come *Generic Flow Control* (GFC) per regolare il flusso delle sorgenti.

[101] 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 *harwdare* 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.

[102] Nel caso in cui venga invece scambiato solo il VCI, si ottiene uno switch VC puro.

[103] 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.

[104] Può accadere infatti di incontrare uno switch VC puro, in cui è scambiato solo il VCI, ed al quale fanno capo due diverse VCC.

[105] I nodi di ingresso ed uscita sono indicati come *ingress* ed *egress* nella terminologia ATM.

[106] 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.

[107] Letteralmente: POLIZIOTTAMENTO. Il controllo può anche essere effettuato su di una intera VPC.

Come anticipato, ATM si è sviluppata per trasportare diversi tipi di traffico, classificabili come segue, nei termini dei parametri indicati di seguito:

- *Constant Bit Rate* (CBR) identifica il traffico real-time come la voce[108] 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[109].

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 essere ridotta adottando degli *shaper*[110], che riducono la variabilità di ritardo a spese un aumento di CTD;
- *Cell Loss Ratio* (CLR) rappresenta il tasso di scarto di celle del collegamento.

[108] La classe CBR si presta bene a trasportare traffico telefonico PCM. In questo caso, può trasportare solo gli intervalli temporali realmente occupati.

[109] 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.

[110] 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.

Nel caso in cui il policing rilevi che una connessione viola le condizioni contrattuali[111], può intraprendere svariate azioni, e se può, non scarta 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*[112] in caso di congestione in altri nodi. Un ulteriore campo dell'header, il *Paylad Type* (PT), può infine ospitare una *segnalazione in avanti*, che manifesta il fatto che la cella in questione ha subito congestione.

**Indirizzamento** I nodi di una rete ATM sono identificati da un indirizzo di 20 byte, di diverso significato nei casi di reti private o pubbliche, come indicato dal primo byte (AFI). Nel primo caso, detto *formato* NSAP[113], il DCC o l'ICD sono assegnati dall'ISO, e l'indirizzo del nodo è disposto nei 10 byte indicati come *High-Order Domain Specific Part* (HO-DSP). I sei byte dell'*End System Identifier* (ESI) sono forniti dal dispositivo connesso ai bordi della rete, e coincidono con il suo indirizzo *Ethernet*: in tal modo la rete comunica un prefisso che identifica il nodo di ingresso, ed il dispositivo lo associa al proprio ESI per forgiare il proprio indirizzo completo. Infine, il byte SEL può essere usato per 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.

**Strato di adattamento** Come mostrato in tab. 8.1, l'AAL è suddiviso in due componenti, *Segmenting and Reassembly* (SAR) e *Convergence Sublayer* (CS); le funzioni di quest'ultimo sono ulteriormente ripartite tra una *Common Part* (CPCS) ed 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 tabella che segue mostra quattro diverse situazioni.

La classe A è un classico caso CBR, ed in tal caso si adotta un AAL di tipo 1, in cui lo strato CS è assente, ed il SAR utilizza il primo dei 48 byte di cella per inserire informazioni di controllo sull'ordine di consegna, e di ausilio al recupero della temporizzazione di sorgente presso la destinazione.

[111] 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.

[112] 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.

[113] Il formato NSAP si ipira 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 |

classi di servizio della rete ATM

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[114], 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[115], demandando il riconoscimento dell'ultima cella di una stessa CS-PDU ad un bit del campo PT presente nell'header di cella ATM. D'altra parte, la lunghezza della CS-PDU dell'AAL 5 è multipla di 48 byte, aggiungendone un numero appropriato, oltre ai 64 byte di intestazione (ora posta in coda), in cui ora sono presenti anche 8 bit di informazione da utente ad utente.

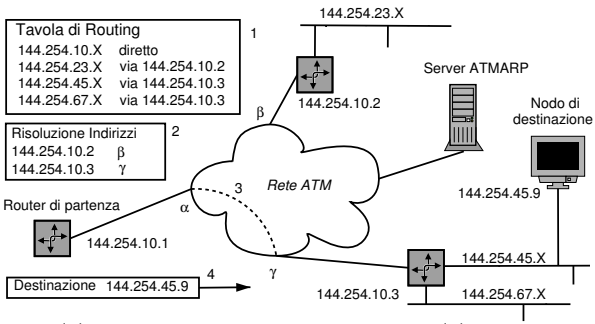
**Ip su ATM classico** Allo stesso tempo in cui di diffonde l'uso di ATM tra gli operatori di TLC, il TCP/IP emerge come *lo* standard comune per l'interconnessione tra elaboratori. Sebbene il TCP/IP si appoggi ad *Ethernet* in area locale, per i collegamenti a lunga distanza[116] l'ATM presenta indubbi vantaggi come la disponibilità di banda su richiesta, la coesistenza con il traffico di tipo diverso, l'elevata efficienza della commutazione, e la possibiltà 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, quando il router di partenza vuole

[114] 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.

[115] 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.

[116] 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 *Metroplitan Area Network* (MAN) per estensione cittadine. Per estensioni ancora maggiori si parla di *reti in area geografica*.




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[117], che risponde comunicando l'indirizzo $\gamma$, il quale è così risolto (2). A questo punto si può instaurare una VCC con $B$ mediante la segnalazione ATM (3), ed effettuare la comunicazione (4). Questa soluzione è nota come VC *multiplexing,* ed i dati sono incapsulati direttamente nella CPCS-PDU di AAL5. In ricezione, l'etichetta VPI/VCI è usata per consegnare il pacchetto al protocollo di strato superiore che ha realizzato la connessione ATM. D'altra parte, questa elaborazione deve avvenire a *diretto contatto* con AAL5, e ciò preclude la possibilità di interlavoro con nodi esterni alla rete ATM.

Nel caso in cui sia antieconomico creare un gran numero di VC, o se si dispone unicamente di un PVC[118], il pacchetto IP viene incapsulato in un header LLC IEEE 802.2 prima di essere consegnato all'AAL5. In tal modo, il router ricevente esamina l'header LLC del pacchetto ricevuto dal nodo ATM di *egress,* per consegnare il pacchetto al protocollo appropriato, realizzando così un *trasporto multiprotocollo* su ATM.

**LANE, NHRP e MPOA** Discutiamo qui brevemente ulteriori possibilità di utilizzo di ATM come trasporto IP, ma a cui verosimilmente sarà preferito l'MPLS.

Mentre l'approccio classico aggiunge un substrato tra IP ed AAL5, per così dire *esterno* alla rete ATM, l'approccio LANE (LAN *Emulation*) ne aggiunge uno *esterno* alla rete IP, che *crede* di avere a che fare con una LAN ethernet. In questo caso anziché una LIS, si definisce una *Emulated* LAN (ELAN), il cui esatto funzionamento prevede diversi passaggi[119].

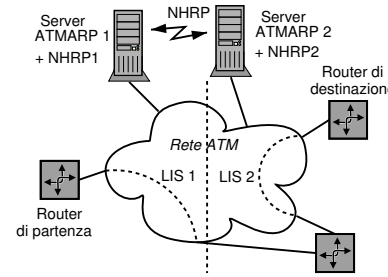
Sia nel caso classico che in quello LANE, se due router IP sono su due LAN (LIS o ELAN) differenti (con prefissi differenti) la comunicazione tra i due deve necessariamente attraversare un terzo router IP, anche se esiste un collegamento diretto tra i primi due, tutto interno alla rete ATM. La situazione è illustrata in figura, per il caso

[117] Tutti i nodi della LIS hanno configurato manualmente l'indirizzo ATM del server ATMARP.

[118] 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 autonomo, eliminando la necessità di sviluppare in proprio un impianto di TLC tra le sedi.

[119] 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.




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 corripondenze *<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 di *passa-parola*[120]*,* per rispondere alle interrogazioni che ricevono. L'applicazione di un meccanismo in parte simile, porta nel caso delle ELAN alla definizione del *Multi Protocol over ATM* (MPOA[121]).

**MPLS** Il *Multi Protocol Label Switching* (MPLS) è una metodo di realizzare una trasmissione a circuito virtuale su reti IP, la cui architettura è descritta nella RFC 3031 dell'IETF, e che verrà esposto meglio in una prossima edizione. Qui illustriamo i legami che MPLS presenta con ATM.

Lo sviluppo di MPLS ha origine dalle iniziative industriali tese a realizzare router internet economici di prestazioni elevate, e capaci di gestire la banda in modo appropriato. Lo IETF ha ricevuto il compito di armonizzare in una architettura standardizzata i diversi approcci, basati sul principio di inoltrare i pacchetti in base ad una etichetta (LABEL) impostata dal primo router della rete, proprio come avviene in ATM. Dato che erano già disponibili i dispositivi hardware per realizzare i nodi di switching ATM, i primi prototipi hanno semplicemente utilizzato tali switch sotto il diretto controllo di un router IP, collegato ad altri simili tramite la rete ATM. L'MPLS è tuttavia più generale, sia verso l'alto (è *multiprotocollo* in quanto si applica oltre che ad IP, a qualunque altro strato di rete) che verso il basso (funziona indifferentemente dall'implementazione dello strato di collegamento, sia ATM, *ethernet* od altro).

La *label* apposta dal primo MPLS *Router* (LSR) dipende dalla destinazione IP del pacchetto; diverse destinazioni possono coincidere con una sola *Forwarding Equivalence Class* (FEC)[122], identificata da una singola *label*. Tutti i pacchetti di una stessa FEC

[120] 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.

[121] 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*.

[122] Nel routing IP tradizionale, una FEC coincide con l'instradamento inviduato dal *longest match*.

sono inoltrati verso il medesimo *next hop*, indicato dalla tabella di routing, indicizzata dalla *label*[123]. Nella stessa tabella, si trova anche la nuova *label* da assegnare al pacchetto, prima di consegnarlo all'LSR seguente. In tutti i LSR successivi, il pacchetto non è riclassificato, ma solo inoltrato verso il *next hop* con una nuova *label* come ordinato dalla tabella di routing. Pertanto, è il primo LSR a decidere tutto il tragitto, ed i pacchetti di una stessa FEC seguono tutti lo stesso *Label Switched Path* (LSP). In tal modo, gli switch possono essere più semplici, si possono stabilire instradamenti diversi per una stessa destinazione[124] 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*[125], e cioè un LSR indica agli LSR dai quali *si aspetta di ricevere* traffico, quale *label* usare in corrispondenza delle FEC per le quali conosce l'instradamento. Dato che la conoscenza di un 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 procolli 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.

[123] 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.

[124] 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.

[125] 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 9

# Densità spettrale e filtraggio

Si descrive l'effetto del passaggio di un segnale, o di un processo, attraverso un sistema fisico, in particolare per quanto riguarda le modifiche spettrali. Per i segnali periodici e di energia siamo già in grado di determinare lo spettro di ampiezza dell'uscita come $Y(f) = H(f) X(f)$, e da questo ottenere lo spettro di densità di potenza o di energia come $|Y(f)|^2$; d'altra parte se $x(t)$ rappresenta una generica realizzazione di un processo, non abbiamo ancora approfondito la teoria che consente di definire il suo spettro di densità potenza $\mathcal{P}_x(f)$: è proprio questo lo scopo della prima parte del capitolo, in cui i concetti statistici già definiti al cap. 7 vengono estesi, giungendo alla definizione di *funzione di autocorrelazione*, la cui trasformata di Fourier è pari appunto alla densità cercata (teorema di WIENER).

Il resto del capitolo procede applicando ad esempi pratici la teoria fin qui sviluppata, presentando alcuni casi tipici di determinazione di uno spettro di densità di potenza, descrivendo possibili architetture dei filtri, fino ad elencare in modo sistematico le alterazioni delle diverse grandezze descrittive dei segnali, in conseguenza del loro passaggio attraverso unità elementari di elaborazione.

Il capitolo termina con una appendice in cui sono riferiti ulteriori risultati, e dimostrate alcune relazioni usate fin qui nel testo.

## 9.1 Correlazione e covarianza

Al § 7.3.5 abbiamo osservato come la caratterizzazione statistica del primo ordine $p_X(x)$ di un processo $\{x(t,\theta)\}$ stazionario ergodico, consenta il calcolo di valor medio $m_x$ e varianza $\sigma_X^2$, nonché della potenza $\mathcal{P}_X = E_X\{x^2\} = \sigma_X^2 + (m_x)^2$, valida per una qualunque realizzazione $\theta$ del processo. Definiamo ora una statistica *del secondo ordine* che permetterà di determinare anche lo spettro di densità di potenza delle realizzazioni del processo.

La statistica di secondo ordine si basa sulla considerazione di 2 istanti $t_1$ e $t_2$, in corrispondenza dei quali estraiamo due variabili aleatorie $x_1 = x(t_1)$, $x_2 = x(t_2)$ a partire da una realizzazione $\overline{\theta}$ di un processo $x(t,\theta)$, come mostrato a sinistra di fig. 9.1. Al variare della realizzazione campionata $\theta \in \Theta$, tutte le coppie di valori estratti sono altrettante determinazioni di una variabile aleatoria *bidimensionale*, descritta da una densità di probabilità *congiunta* $p_{X_1X_2}(x_1x_2;t_1t_2)$, che dipende anche dagli istanti $t_1$ e $t_2$, ed è esemplificata nella parte destra di fig. 9.1, che mostra come questa sottenda

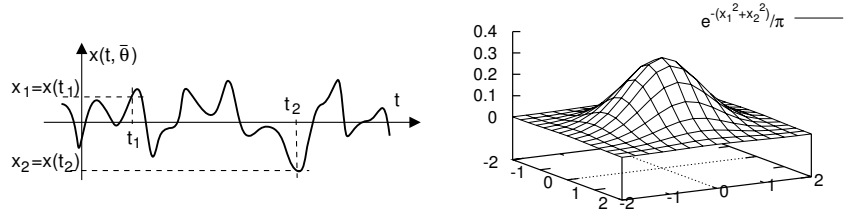



Figura 9.1: Estrazione di due variabili aleatorie da un membro di processo, e loro possible d.d.p. congiunta

un volume unitario, e descriva con il suo andamento le regioni del piano $x_1x_2$ in cui cadono un maggior numero di coppie (ovvero dove la probabilità è più densa).

### 9.1.1 Correlazione

Questa grandezza dipende dalla definizione di un valore atteso, formalmente analogo a quanto già visto per il caso unidimensionale, che prende il nome di *momento misto di ordine* $(i, j)$ e che risulta:

$$m_{X_1X_2}^{(i,j)} = E_{X_1X_2}\left\{x_1^i x_2^j\right\} = \int\int x_1^i x_2^j p_{X_1X_2}\left(x_1x_2; t_1t_2\right) dx_1 dx_2$$

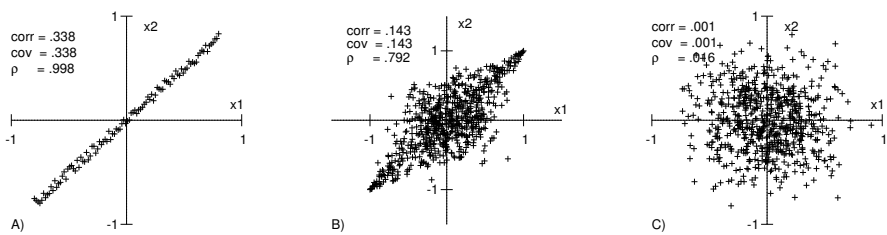
e che nel caso in cui $i = j = 1$ prende il nome di *correlazione* e si indica come

$$\mathcal{R}_X\left(x\left(t_1\right), x\left(t_2\right)\right) = m_{X_1X_2}^{(1,1)}$$

Prima di proseguire, soffermiamoci un istante per meglio comprendere il significato di $m_{X_1X_2}^{(1,1)}$. Il valore di correlazione fra due variabili aleatorie è indicativo del legame che esiste tra le due[1], nel senso di quanto l'una ha un valore che dipende da quello dell'altra, ed ha un valore assoluto tanto più elevato quanto più i valori di $x_1$ e $x_2$ sono legati in modo deterministico[2]: in effetti però, seguendo gli esempi riportati nella nota, non sempre questo accade, e pertanto è opportuno basarsi sull'uso di momenti centrati come descritto al punto successivo.

[1] Intese qui nel senso più generale, per esempio come temperatura e pressione in un punto ben preciso di un cilindro di un motore a scoppio, oppure come pressione e velocità in un circuito idraulico, o pneumatico... astraendo cioè dal caso particolare di due v.a. estratte da una stessa forma d'onda.

[2] Può tornare utile pensare $m_{X_1X_2}^{(1,1)}$ come una media pesata dei possibili valori del prodotto $x_1x_2$; i termini di eguale ampiezza e segno opposto possono elidersi se equiprobabili. Negli esempi che seguono, riportiamo dei *diagrammi di scattering* per sei diversi casi di distribuzione delle coppie di valori $x_1$ e $x_2$, assieme ai valori di correlazione $\mathcal{R}_{x_1x_2}$ (corr), covarianza $\sigma_{x_1x_2}$ (cov), e coefficiente di correlazione $\rho$ (vedi § 9.9.1 per quest'ultimo parametro)



### 9.1.2 Indipendenza statistica e incorrelazione

Nel caso in cui le due v.a. siano *statisticamente indipendenti*, e cioè si possa scrivere $p_{X_1X_2}(x_1, x_2; t_1, t_2) = p(x_1)\, p(x_2)$([3]), l'integrale che definisce la correlazione si fattorizza, e pertanto

$$\begin{aligned} \mathcal{R}_X(x_1, x_2) &= \int\int x_1 x_2 p(x_1)\, p(x_2)\, dx_1 dx_2 = \int x_1 p(x_1)\, dx_1 \cdot \int x_2 p(x_2)\, dx_2 = \\ &= E\{x_1\}\, E\{x_2\} = m_{X_1} m_{X_2} \end{aligned} \tag{9.1}$$

**Covarianza** E' indicata come $\sigma(x_1, x_2)$ e rappresenta il momento misto centrato tra le due v.a., di espressione[4]

$$\begin{aligned} \sigma(x_1, x_2) &= E\{(x_1 - m_{X_1})(x_2 - m_{X_2})\} = \\ &= E\{x_1 x_2\} - E\{x_1 m_{X_2}\} - E\{m_{X_1} x_2\} + E\{m_{X_1} m_{X_2}\} = \\ &= E\{x_1 x_2\} - m_{X_1} m_{X_2} = \mathcal{R}_X(x_1, x_2) - m_{X_1} m_{X_2} \end{aligned} \tag{9.2}$$

**Incorrelazione** Combinando i risultati (9.1) e (9.2) possiamo verificare che

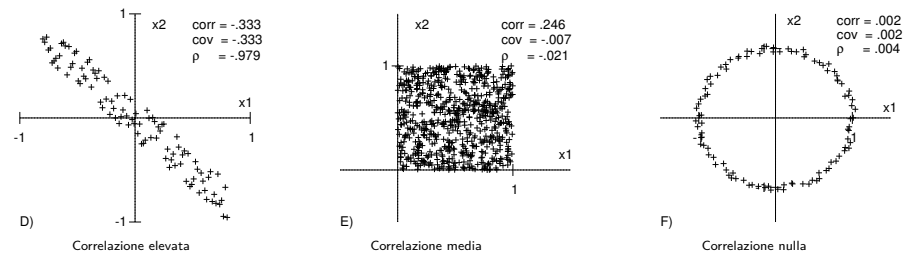
> *Se due variabili aleatorie $x_1$ ed $x_2$ sono statisticamente indipendenti, queste si dicono* INCORRELATE, *in quanto la covarianza $\sigma(x_1 x_2)$ è nulla*[5].

La proprietà esposta ha valore in una sola direzione, in quanto se due v.a. esibiscono $\sigma(x_1, x_2) = 0$ *non è detto* che siano statisticamente indipendenti[6]. L'unico caso in cui l'incorrelazione tra variabili aleatorie *ne implica* l'indipendenza statistica, è quello relativo a v.a. estratte da un processo gaussiano, come mostrato in appendice 9.9.2.

### 9.1.3 Statistiche dei processi

Nel caso in cui il processo da cui si estraggono $x_1$ e $x_2$ sia *stazionario*, si ottiene che

$$\mathcal{R}_X(x(t_1), x(t_2)) = \mathcal{R}_X(x(t_1), x(t_2 = t_1 + \tau)) = \mathcal{R}_X(\tau)$$




In A) e F) le coppie di valori sono legate da una legge pressoché deterministica, mentre in B) e D) c'è più variabilità, ma si nota ancora una certa dipendenza tra le due. Infine nei casi C) ed E), osserviamo due v.a. *statisticamente indipendenti*, dato che $p_{X_1X_2}(x_1, x_2)$ è fattorizzabile come $p_{X_1}(x_1)\, p_{X_2}(x_2)$, e per le quali risulta quindi $\mathcal{R}_X(x_1 x_2) = m_{x_1} m_{x_2}$.

[3] Omettiamo per brevità di indicare la variabile aleatoria a pedice della densità di probabilità.

[4] L'uguaglianza si ottiene ricordando che un valore atteso è in realtà un integrale, ed sfruttando la proprietà distributiva di quest'ultimo.

[5] *Notiamo immediatamente che il termine più corretto sarebbe "incovarianzate"; l'uso (ormai storico e consolidato) dell'espressione* incorrelate *deriva probabilmente dal considerare usualmente grandezze a media nulla, per le quali le due espressioni coincidono.*

[6] Vedi ad esempio il caso F) della nota (2), in cui le variabili aleatorie risultano incorrelate, ma non sono per nulla indipendenti, in quanto l'una dipende strettissimamente dall'altra, dato che le coppie di valori si dispongono su di un cerchio.

e cioè la correlazione *dipende solo dall'intervallo* $\tau = t_2 - t_1$. Infatti se un processo è stazionario, le proprietà statistiche non dipendono da traslazioni temporali.

Nel caso in cui il processo sia anche *ergodico,* allora le medie di insieme hanno lo stesso valore delle corrispondenti medie temporali, e dunque la correlazione (media di insieme) può essere calcolata in base alla sua media temporale equivalente, a partire da una qualunque realizzazione del processo

$$\mathcal{R}_x(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)\, x(t+\tau)\, dt \tag{9.3}$$

Ovviamente è vero anche l'inverso, e cioè: il valor medio del prodotto tra due campioni, estratti (a caso) a distanza $\tau$ a partire da una specifica realizzazione[7], ha un valore che è calcolabile come media di insieme a partire dalla conoscenza della densità di probabilità congiunta $p_{X_1X_2}(x_1, x_2; t_1, t_1+\tau)$.

### 9.1.4 Autocorrelazione e intercorrelazione

La media temporale $\mathcal{R}_x(\tau)$ appena introdotta prende il nome di *integrale di autocorrelazione,* che è definito anche per segnali di energia, come

$$\mathcal{R}_x(\tau) = \int_{-\infty}^{\infty} x^*(t)\, x(t+\tau)\, dt \tag{9.4}$$

in cui l'operatore di coniugato generalizza l'operazione anche al caso di segnali complessi. Simile, ma diversa, è la definizione di integrale di *intercorrelazione*, che esprime lo stesso calcolo, ma generalizzandone il concetto, in quanto relativo a due segnali $x(t)$ ed $y(t)$ estratti da processi diversi:

$$\mathcal{R}_{xy}(\tau) = \int_{-\infty}^{\infty} x^*(t)\, y(t+\tau)\, dt \tag{9.5}$$

che può essere interpretato come *prodotto scalare* o *energia incrociata* (vedi § 3.2) tra $x(t)$ e una copia di $y(t)$ traslata nel tempo. L'*integrale* di autocorrelazione è anche detto *funzione* di autocorrelazione, in quanto il suo argomento è un tempo (l'intervallo tra due campioni) e dunque $\mathcal{R}_x(\tau)$ può essere visto come un segnale (funzione di $\tau$ anzichè di $t$). Nello studio abbiamo già incontrato un integrale (di convoluzione) il cui risultato è una funzione del tempo; la somiglianza tra i due è più profonda di una semplice analogia, in quanto si può scrivere

$$\mathcal{R}_x(\tau) = \int_{-\infty}^{\infty} x^*(t)\, x(t+\tau)\, dt = x^*(-t) * x(t)$$

in cui $*$ è il consueto simbolo di convoluzione[8].

In base a quest'ultima osservazione otteniamo la costruzione grafica mostrata in fig. (9.2), che fornisce il risultato dell'integrale di autocorrelazione, e che è del tutto simile a quella già illustrata per la convoluzione (vedi § 3.5.3), con la differenza che ora non si effettuano ribaltamenti di asse. L'esempio mostrato ne illustra l'applicazione ad un caso noto, per il quale $x(t) = x^*(-t)$, e che fornisce quindi lo stesso risultato di $x(t) * x(t)$.

[7] Si rifletta sulla descrizione ora data a parole dell'operazione di media temporale scritta sopra.

[8] Il risultato ottenuto si basa sul cambio di variabile $\theta = t + \tau$ che permette di scrivere

$$\mathcal{R}_x(\tau) = \int_{-\infty}^{\infty} x^*(\theta - \tau)\, x(\theta)\, d\theta = \int_{-\infty}^{\infty} x^*(-(\tau - \theta))\, x(\theta)\, d\theta = x^*(-t) * x(t)$$

#### 9.1.4.1 Proprietà dell'autocorrelazione

Elenchiamo ora alcuni aspetti *caratteristici* della funzione di autocorrelazione:

**Traslazioni temporali:** se consideriamo i segnali $x(t)$ e $y(t) = x(t+\theta)$, le rispettive autocorrelazioni $\mathcal{R}_x(\tau)$ ed $\mathcal{R}_y(\tau)$ sono identiche[9]. Questo risultato mostra come l'autocorrelazione non tenga conto dell'informazione legata alla fase dei segnali: infatti $x(t)$ e $y(t)$ hanno la stessa densità spettrale, a meno di un contributo di fase lineare, ed hanno uguale autocorrelazione.

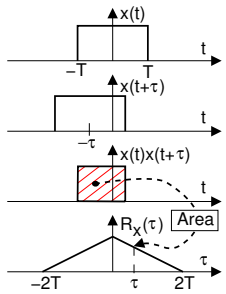



Figura 9.2: Autocorrelazione di due rettangoli

**Durata Limitata:** la funzione di autocorrelazione di un segnale di durata limitata è anch'essa a durata limitata, e di estensione doppia rispetto alla durata del segnale originario.

**Segnali Periodici:** l'autocorrelazione di un segnale periodico di periodo $T$ è anch'essa periodica, con lo stesso periodo. Infatti per $\tau = nT$ il secondo fattore integrando è traslato di un numero intero di periodi.

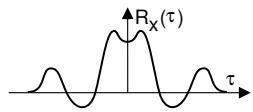
In particolare, le due proprietà seguenti sono da ritenersi *fondamentali* per la possibilità, che come vedremo offrono, di interpretare l'autocorrelazione in senso energetico:

**Massimo nell'origine:** la $\mathcal{R}_x(\tau)$ calcolata in $\tau = 0$ fornisce il valore di $\mathcal{R}_x(\tau)$ più grande di quello ottenibile per qualunque altro valore di $\tau$. In particolare, $\mathcal{R}_x(\tau = 0)$ è uguale alla potenza del segnale $x(t)$, od all'energia se $x(t)$ è di energia, ossia

$$0 \leq \mathcal{R}_x(0) = \begin{cases} \int |x(t)|^2 dt = \mathcal{E}_x > |\mathcal{R}_x(\tau \neq 0)| & \text{se } x(t) \text{ è di energia} \\ \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |x(t)|^2 dt = \mathcal{P}_x \geq |\mathcal{R}_x(\tau \neq 0)| & \text{se } x(t) \text{ è di potenza} \end{cases}$$

e in particolare, se $x(t)$ è periodico, l'ultimo segno $\geq$ è una eguaglianza per $\tau$ multiplo di un periodo.

**Simmetria coniugata:** è possibile verificare[10] che risulta $\mathcal{R}_x(-\tau) = \mathcal{R}_x^*(\tau)$, e ciò consente (vedi § 3.3) di affermare che $\mathcal{F}\{\mathcal{R}_x(\tau)\}$ è reale. Nel caso in cui $x(t)$ è reale, si ottiene $\mathcal{R}_x(-\tau) = \mathcal{R}_x(\tau)$, ovvero l'autocorrelazione di un segnale reale è *reale pari*, alla stregua (come mostreremo ora) della sua trasformata di Fourier.




[9] Infatti otteniamo:

$$\begin{aligned} \mathcal{R}_y(\tau) &= \int_{-\infty}^{\infty} y(t)\, y(t+\tau)\, dt = \int_{-\infty}^{\infty} x(t+\theta)\, x(t+\theta+\tau)\, dt \\ &= \int_{-\infty}^{\infty} x(\alpha)\, x(\alpha+\tau)\, d\alpha = \mathcal{R}_x(\tau) \end{aligned}$$

[10] $\mathcal{R}_x(-\tau) = \int_{-\infty}^{\infty} x^*(t)\, x(t-\tau)\, dt = \int_{-\infty}^{\infty} x^*(\alpha+\tau)\, x(\alpha)\, d\alpha = \mathcal{R}^*{}_x(\tau)$, avendo operato il cambio di variabile $t - \tau = \alpha$, da cui $t = \alpha + \tau$ e $dt = d\alpha$.

## 9.2 Densità spettrale

Come anticipato ad inizio capitolo, mostriamo il metodo con cui determinare lo spettro di densità di potenza nel caso di processi. La cosa decisamente gradevole è che questo stesso strumento è valido anche per gli altri tipi di segnale.

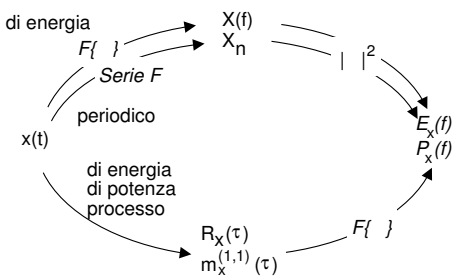
### 9.2.1 Teorema di Wiener[11]

*Lo spettro di densità di potenza $\mathcal{P}_x(f)$ (o di energia $\mathcal{E}_x(f)$) di $x(t)$ è uguale alla trasformata di Fourier della sua funzione di autocorrelazione $\mathcal{F}\{\mathcal{R}_x(\tau)\}$.*

La dimostrazione del teorema è straordinariamente semplice, e per segnali di energia si scrive

$$\begin{aligned}\mathcal{R}_x(\tau) &= \int_{-\infty}^{\infty} x^*(t)\, x(t+\tau)\, dt = \int_{-\infty}^{\infty} X^*(f)\, X(f)\, e^{j2\pi f\tau} df = \\ &= \mathcal{F}^{-1}\{X^*(f)\, X(f)\} = \mathcal{F}^{-1}\{\mathcal{E}_x(f)\}\end{aligned}$$

in cui abbiamo prima applicato il teorema di Parseval, poi la proprietà di traslazione nel tempo, e quindi (vedi § 3.2) espresso $X^*(f)\, X(f)$ come $\mathcal{E}_x(f)$.




Come anticipato, questa proprietà è valida anche per segnali di potenza[12], comprese quindi le realizzazioni di processi. In particolare, se il processo è ergodico, la media di insieme $m_{X_1X_2}^{(1,1)}$ risulta pari alla media temporale espressa dalla (9.3) e calcolata per una qualsiasi realizzazione del processo; pertanto lo spettro di densità di potenza di un processo si ottiene trasformando la funzione di autocorrelazione calcolata come media di insieme, oppure trasformando quella calcolata come media temporale per una delle sue realizzazioni. In virtù del teorema di Wiener, è dunque possibile ottenere $\mathcal{P}_x(f)$ anche per processi e segnali di potenza, oppure fare "la prova del nove" per segnali di energia o periodici.

Applichiamo ora questo metodo di valutazione della densità spettrale di un processo, ad alcuni casi particolari.

---

[11] In realtà le attribuzioni di questo risultato sono molteplici, comprendendo anche *Khinchin*, *Einstein* e *Kolmogorov* - fonte `http://en.wikipedia.org/wiki/Wiener%E2%80%93Khinchin_theorem`

[12] Per segnali di potenza la dimostrazione può essere fornita, senza troppe attenzioni di rigore analitico, partendo dalla stima della densità di potenza ottenuta mediante periodogramma (§ 9.3.1): $\mathcal{P}_{x_T}(f) = \frac{|X_T(f)|^2}{T}$. Osserviamo che $|X_T(f)|^2$ rappresenta lo spettro di densità di energia $\mathcal{E}_{x_T}(f)$ di un segmento di segnale $x_T(t)$ di durata $T$ estratto da $x(t)$, corrispondente alla funzione di autocorrelazione $\mathcal{R}_{x_T}(\tau) = \mathcal{F}^{-1}\{\mathcal{E}_{x_T}(f)\}$ calcolabile per $x_T(t)$ mediante la (9.4). Ma operando il passaggio al limite e la divisione $\lim_{T\to\infty} \frac{1}{T}$ di $\mathcal{R}_{x_T}(\tau)$, si ottiene l'autocorrelazione (9.3) dell'intero segnale, mentre nel dominio della frequenza il limite di $\mathcal{E}_{x_T}(f)$ tende allo spettro di densità di potenza $\mathcal{P}_{x_T}(f)$ del segnale $x(t)$ nella sua interezza.

### 9.2.2 Processo armonico

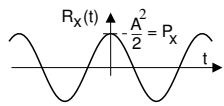
E' definito in base ad una sua generica realizzazione

$$x(t,\theta) = A\cos(2\pi f_0 t + \theta)$$

che, se il parametro $\theta$ è una variabile aleatoria uniformemente distribuita tra $-\pi$ e $\pi$ (ossia $p_\Theta(\theta) = \frac{1}{2\pi} rect_{2\pi}(\theta)$), descrive un processo ergodico, ed in tal caso la d.d.p. $p_X(x) = \frac{1}{\pi\sqrt{A^2-x^2}}$ è graficata a pagina 136.
Sappiamo che una sua realizzazione (ad esempio quella con $\theta = 0$) ha una densità di potenza $\mathcal{P}_x(f) = \frac{A^2}{4}[\delta(f-f_0) + \delta(f+f_0)]$. Possiamo quindi ottenere l'autocorrelazione senza dover svolgere l'integrale:

$$\mathcal{R}_x(t) = \mathcal{F}^{-1}\{\mathcal{P}_x(f)\} = \frac{A^2}{4}\left[e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}\right] = \frac{A^2}{2} cos(2\pi f_0 t)$$

Il risultato ottenuto, ci conferma che l'autocorrelazione di un segnale periodico è periodica; riflettiamo dunque sulla circostanza che anche un seno, od un coseno con qualunque altra fase, avrebbe avuto la stessa $\mathcal{R}_x(t)$. Ciò è d'altra parte evidente, avendo tutti questi segnali uguali densità $\mathcal{P}_x(f)$.



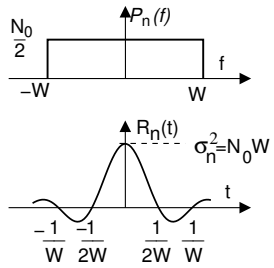

### 9.2.3 Processo gaussiano bianco limitato in banda

Un processo $n(t)$ è chiamato *bianco* qualora sia costante in frequenza, e quindi descritto da una densità di potenza pari a

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

in cui $W$ è l'occupazione di banda a frequenze positive. In tali ipotesi otteniamo

$$\mathcal{R}_n(t) = \mathcal{F}^{-1}\{\mathcal{P}_n(f)\} = N_0 W \text{sinc}(2Wt)$$




da cui possiamo constatare che, con queste posizioni, si ottiene

$$\mathcal{R}_n(0) = \mathcal{P}_n = \int_{-\infty}^{\infty} \mathcal{P}_n(f)\, df = N_0 W = \sigma_n^2$$

in cui l'ultima eguaglianza sussiste in quanto l'assenza di impulsi nell'origine per $\mathcal{P}_n(f)$ corrisponde ad un $n(t)$ a media nulla. Inoltre, dato che $\mathcal{R}_n(1/2W) = 0$, osserviamo che campionando $n(t)$ con periodo $T_c = 1/2W$ si ottengono valori *incorrelati*, e se il processo è gaussiano, anche statisticamente indipendenti. Questo risultato giustifica, almeno da un punto di vista teorico, una ipotesi che viene spesso fatta: quella di trovare sovrapposti ai campioni di segnale, dei campioni di rumore *statisticamente indipendenti*.

All'aumentare di $W$, $\mathcal{R}_n(t)$ tende a zero più rapidamente, cosicchè il rumore si mantiene correlato per un tempo sempre minore, ovvero due campioni estratti ad una stessa distanza $t$ hanno una correlazione sempre minore. Un risultato simile vale anche più in generale, in quanto l'autocorrelazione $\mathcal{R}_x(t)$ di un qualsiasi processo

a media nulla (tranne nel caso periodico, riconducibile ad una combinazione di processi armonici) tende a 0 con $t \to \infty$, ovvero da un certo $t$ in poi la correlazione è trascurabile.

Infine, se immaginiamo il rumore bianco limitato in banda come il risultato del filtraggio di un processo gaussiano a banda infinita (quindi, con $\mathcal{R}_n(t) = \delta(t)$) attraverso il filtro passa basso ideale con $H(f) = rect_{2W}(f)$, ci accorgiamo che la correlazione risultante per i campioni di rumore, è il risultato della memoria introdotta dalla risposta impulsiva $h(t) = 2W\text{sinc}(2Wt)$ sul segnale in transito, dato appunto che l'operazione di convoluzione rende i valori in uscita, una combinazione lineare di quelli passati in ingresso (vedi § 3.5.3).

### 9.2.4 Segnale dati

Abbiamo già descritto al § 5.1.2 un generico segnale numerico come una somma di repliche di una funzione $g(t)$, con ampiezze $a_n$ rappresentative dei valori da trasmettere:

$$x(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT + \theta)$$

La presenza della variabile aleatoria $\theta$ a distribuzione uniforme tra $\pm\frac{T}{2}$ (per cui $p_\Theta(\theta) = \frac{1}{T} rect_T(\theta)$), rende $x(t)$ un processo ergodico.

Si mostrerà in appendice (al § 9.9.3) che, nelle ipotesi in cui le ampiezze $a_n$ siano determinazioni di variabili aleatorie *indipendenti* ed *identicamente distribuite,* a *media nulla* e varianza $\sigma_A^2$([13]), l'autocorrelazione di $x(t)$ vale $\mathcal{R}_x(\tau) = \sigma_A^2 \frac{\mathcal{R}_g(\tau)}{T}$ in cui $\mathcal{R}_g(\tau)$ è l'autocorrelazione di $g(t)$, e dunque

$$\mathcal{P}_x(f) = \sigma_A^2 \frac{\mathcal{E}_g(f)}{T}$$

Osserviamo innanzitutto che è per questa via che al § 5.2.1 si sono caratterizzate le densità di potenza proprie dei codici di linea. Limitandoci a voler interpretare il risultato, notiamo che $\mathcal{E}_g(f)$ è la densità di energia di una singola replica di $g(t)$. La sua ripetizione, con periodo $T$, fornisce una densità di potenza *media* $\frac{\mathcal{E}_g(f)}{T}$. Se ogni replica di $g(t)$ è moltiplicata per una v.a. indipendente a media nulla e varianza (potenza) $\sigma_A^2$, la densità di potenza $\mathcal{P}_x(f)$ aumenta di egual misura (vedi § 9.6.1). Un'ultima avvertenza riguarda il fatto che, se gli $a_n$ non sono indipendenti, il risultato è più complesso (vedi appendice 9.9.3).

## 9.3 Stima spettrale

Il teorema di Wiener ci aiuta qualora si desideri conoscere la densità di potenza per un processo, di cui siamo in grado di stimare o postulare un $m_X^{(1,1)}(\tau) = \mathcal{R}_X(\tau)$. Spesso però si ha a che fare con segnali di cui, pur ricorrendo le ipotesi di appartenenza ad un processo ergodico, si ignorano le statistiche di insieme. Un altro caso tipico è quello di un segnale che, pur se rappresentativo di molti altri, non presenta caratteristiche

[13] Media $m_A$ e varianza $\sigma_A^2$ sono qui riferite ai valori multilivello $a_k$ (con $k = 1, 2, \cdots, L$) che un generico simbolo $a_n$ può assumere, pesati con le rispettive probabilità $p_k$, ossia $m_A = \sum_{k=1}^{L} p_k a_k$ e $\sigma_A^2 = \sum_{k=1}^{L} p_k (a_k - m_A)^2$

spettrali costanti nel tempo, e sono proprio le variazioni di queste ultime ad interessare[14]. In questi casi, tutto ciò che si può fare è di tentare una stima dello spettro di potenza del segnale, a partire da un suo segmento temporale. Esistono al riguardo tecniche differenti, come ad es. quella riportata al § 18.1.2; per ora ci limitiamo ad un caso semplice ma di rilevo teorico.

### 9.3.1 Periodogramma

Definiamo un intervallo $T$ in cui isoliamo un segnale a durata limitata $x_T(t) = x(t,\theta)\, rect_T(t)$ da una realizzazione $x(t,\theta)$. Questo segmento di segnale è di energia, con $\mathcal{E}_{x_T}(f) = |X_T(f)|^2$, e sotto le ipotesi di stazionarietà, da questa si può stimare una densità di potenza

$$\mathcal{P}_{x_T}(f) = \frac{|X_T(f)|^2}{T}$$

che viene detta *periodogramma*[15]. Al tendere di $T$ ad $\infty$, il risultato trovato tende alla densità di potenza $\mathcal{P}_x(f) = \lim_{T\to\infty} \frac{|X_T(f,\theta)|^2}{T}$ della realizzazione $x_T(t,\theta)$ e, se questa appartiene ad un processo ergodico, a quella di un qualunque altro membro.

Nel caso più verosimile in cui $T$ non tende ad infinito, si può mostrare[16] che usando $\mathcal{P}_{x_T}(f)$ come una stima $\widehat{\mathcal{P}}_x(f)$ della vera densità $\mathcal{P}_x(f)$ del processo, si commette un errore espresso come

$$\widehat{\mathcal{P}}_x(f) = \mathcal{P}_x(f) * T(\operatorname{sinc}(fT))^2$$

che è il risultato ottenuto al § 3.9.3 a riguardo del procedimento di *finestratura temporale*, e che mostra come lo stimatore è *polarizzato*[17]. In base allo scopo con cui è condotta la stima spettrale, può essere opportuno adottare al posto della finestra rettangolare un diverso andamento $w_T(t)$, in modo da calcolare $x_T(t,\theta) = x(t,\theta)\cdot w_T(t)$.

[14] Un esempio può essere un segnale sonoro, ad esempio una voce recitante, per il quale vogliamo studiare le caratteristiche spettrali dei diversi suoni della lingua (i fonemi), per confrontarle con quelle di un'altro individuo, o per ridurre la quantità di dati necessaria a trasmettere il segnale in forma numerica, o per realizzare un dispostivo di riconoscimento vocale.

[15] Il termine *periodogramma* trae origine dall'uso che ne fu inizialmente fatto, ossia per scoprire tracce di periodicità all'intero di segnali qualsiasi.

[16] Per una determinata frequenza $f_0$, il valore $\frac{|X_T(f_0)|^2}{T}$ è una variabile aleatoria (dipende infatti da $\theta$), che vorremmo avesse un valore atteso $m_T = E_\theta\{\mathcal{P}_{x_T}(f_0)\}$ pari alla densità del processo ($m_X = \mathcal{P}_x(f_0)$) ed una varianza $\sigma_T^2 = E_\theta\left\{\left(\mathcal{P}_{x_T}(f_0) - m_X\right)^2\right\}$ che diminuisce al crescere di $T$. Per verificare se tali proprietà siano soddisfatte, valutiamo il valore atteso del periodogramma, a partire dalle relazioni fornite dal teorema di Wiener, applicato ad $X_T(f)$, e cioè $|X_T(f)|^2 = \mathcal{E}_{x_T}(f) = \mathcal{F}\{\mathcal{R}_{x_T}(\tau)\}$:

$$\begin{aligned}
E_\theta\{\mathcal{P}_{x_T}(f)\} &= E_\theta\left\{\mathcal{F}\left\{\frac{1}{T}\int_{-\infty}^{\infty} x(t)\, rect_T(t)\, x(t+\tau)\, rect_T(t+\tau)\, dt\right\}\right\} = \\
&= \mathcal{F}\left\{\frac{1}{T}\int_{-\infty}^{\infty} E_\theta\{x(t)\,x(t+\tau)\}\, rect_T(t)\, rect_T(t+\tau)\, dt\right\} = \\
&= \mathcal{F}\left\{\mathcal{R}_x(\tau)\,\frac{1}{T}\int_{-\infty}^{\infty} rect_T(t)\, rect_T(t+\tau)\, dt\right\} = \mathcal{F}\{\mathcal{R}_x(\tau)\cdot \mathrm{tri}_{2T}(\tau)\} = \\
&= \mathcal{P}_x(f) * T(\operatorname{sinc}(fT))^2
\end{aligned}$$

Osserviamo quindi come, all'aumentare di $T$, il nostro stimatore tende al valore vero, dato che $T(\operatorname{sinc}(fT))^2$ tende ad un impulso.

[17] Ci consola verificare che, come commentato alla nota precedente, per $T\to\infty$ la polarizzazione tende a scomparire. Quando il valore atteso di uno stimatore tende al valore vero, si dice che lo stimatore è *non polarizzato* (o *unbiased*); se aumentando la dimensione del campione, la varianza della stima tende a zero, lo stimatore è detto *consistente*.

Il calcolo del periodogramma viene svolto mediante una *Trasformata Discreta di Fourier* (DCT), o meglio ancora con una FFT (vedi § 4.2). In tal caso, l'aumento di $T$ corrisponde all'aumento del numero di campioni utilizzati, ed all'aumento della risoluzione in frequenza ottenibile. In base a queste considerazioni, si può mostrare che la varianza dello stimatore ottenuto per via numerica *non decresce* con $T$[18].

## 9.4 Filtraggio di segnali e processi

Ora che siamo in grado di descrivere da un punto di vista spettrale tutti i segnali, indaghiamo sui valori relativi alla grandezza $y(t) = x(t) * h(t)$ uscente da un filtro con risposta impulsiva $h(t)$. In particolare, ci chiediamo quanto valgano $\mathcal{P}_y$ e $\mathcal{P}_y(f)$, oppure $\mathcal{E}_y$ ed $\mathcal{E}_y(f)$ se $x(t)$ è di energia.

### 9.4.1 Segnali di energia

Sappiamo che per il teorema di Parseval risulta $\mathcal{E}_y(f) = Y(f) Y^*(f)$; dato che $Y(f) = X(f) H(f)$, allora

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

A questo punto, eseguendo l'antitrasformata di Fourier di ambo i membri, si ottiene:

$$\mathcal{R}_y(\tau) = \mathcal{F}^{-1}\{\mathcal{E}_y(f)\} = \mathcal{F}^{-1}\left\{\mathcal{E}_x(f) |H(f)|^2\right\} = \mathcal{R}_x(\tau) * \mathcal{R}_h(\tau)$$

Il risultato ottenuto è posto in evidenza perché è valido anche per i due casi successivi di segnale periodico e di processo, e mostra come l'autocorrelazione dell'uscita di un filtro è pari alla convoluzione tra l'autocorrelazione dell'ingresso e quella della risposta impulsiva.

A corollario di quanto esposto, sussistono le seguenti uguaglianze[19], equivalenti ai fini del calcolo dell'energia totale:

$$\begin{aligned}\mathcal{E}_y &= \int_{-\infty}^{\infty} \mathcal{E}_y(f)\, df = \int_{-\infty}^{\infty} \mathcal{E}_x(f) |H(f)|^2\, df = \int_{-\infty}^{\infty} \mathcal{R}_x(\tau) \mathcal{R}_h(\tau)\, d\tau = \\ &= \int_{-\infty}^{\infty} \mathcal{R}_x(\tau) \mathcal{R}_h^*(\tau)\, d\tau = \mathcal{R}_y(0)\end{aligned}$$

Pertanto è possibile utilizzare tutte queste come relazioni di equivalenza, quando si ha necessità di determinare una grandezza a partire da altre note.

### 9.4.2 Segnali periodici

Se in ingresso ad un filtro è presente un segnale periodico, il segnale in uscita è anch'esso periodico[20], e per esso è valido lo sviluppo in serie di Fourier

$$y(t) = \sum_n Y_n e^{j2\pi nFt}$$

[18] Una possibile alternativa è quella di suddividere l'intervallo di osservazione in diversi sottointervalli, calcolare il periodogramma su ciascuno di essi, e mediare i risultati. In tal modo, all'aumentare dei dati a disposizione, viene mantenuta la stessa risoluzione in frequenza, ma si migliora la varianza della stima spettrale.

[19] La terza uguaglianza sussiste in virtù del teorema di Parseval, mentre la quarta è valida se $\mathcal{R}_H(\tau)$ è reale, ossia $h(t)$ è idealmente realizzabile, proprietà quest'ultima definita al § 9.5.

[20] Questa affermazione è vera, purchè il filtro *non* presenti fenomeni di non linearità (definiti al § 9.5): in tal caso infatti, vale la sovrapposizione degli effetti, e *non possono* prodursi armoniche diverse da quelle in ingresso.

in cui i coefficienti $Y_n$ possono esprimersi in termini dei coefficienti di Fourier dell'ingresso $X_n$ e dei valori della risposta in frequenza (vedi sezione 9.5) come $Y_n = X_n H(nF)$, ovvero in modulo e fase come

$$|Y_n| = |X_n| \, |H(nF)| \, ; \qquad \arg(Y_n) = \arg(X_n) + \arg(H(nF))$$

Per la densità di potenza di un segnale periodico possiamo scrivere $\mathcal{P}_y(f) = \sum_n |Y_n|^2 \delta(f - nF)$, ovvero

$$\mathcal{P}_y(f) = \sum_n |X_n|^2 |H(nF)|^2 \delta(f - nF) = |H(f)|^2 \mathcal{P}_x(f)$$

Di nuovo, antitrasformando si ottiene $\mathcal{R}_y(\tau) = \mathcal{R}_x(\tau) * \mathcal{R}_h(\tau)$.

### 9.4.3 Processi ergodici

Anche in questo caso, si verifica (in appendice 9.9.5) che $m_Y^{(1,1)}(\tau) = m_X^{(1,1)}(\tau) * \mathcal{R}_h(\tau)$, e dunque

$$\mathcal{P}_y(f) = \mathcal{P}_x(f) |H(f)|^2$$

Passiamo quindi a calcolare le altre grandezze rappresentative:

**Media:**

$$\begin{aligned} m_Y &= E\{y(t)\} = E\{x(t) * h(t)\} = E\{x(t)\} * h(t) = \\ &= m_X \int_{-\infty}^{\infty} h(\tau)\, d\tau = m_X H(0) \end{aligned}$$

è pari cioè a quella dell'ingresso, moltiplicata per il guadagno in continua del filtro.

**Potenza:** in linea generale, è sempre vero che $\mathcal{P}_y = \sigma_y^2 + (m_y)^2$; inoltre, valgono le relazioni

$$\begin{aligned} \mathcal{P}_y &= \mathcal{R}_y(0) = \int_{-\infty}^{\infty} \mathcal{P}_y(f)\, df = \int_{-\infty}^{\infty} \mathcal{P}_x(f) |H(f)|^2 \, df = \\ &= \int_{-\infty}^{\infty} \mathcal{R}_x(\tau) \mathcal{R}_h(\tau)\, d\tau \end{aligned}$$

Se ad esempio $x(t)$ è un processo bianco a media nulla e banda finita, con $\mathcal{P}_x(f) = \frac{N_0}{2} \mathrm{rect}_{2B}(f)$ e quindi $\mathcal{R}_x(\tau) = N_0 B \cdot sinc(2Bt)$, si ottiene[21]

$$\sigma_y^2 = \mathcal{P}_y = \frac{N_0}{2} \int_{-B}^{B} |H(f)|^2 \, df = \frac{N_0}{2} \mathcal{R}_h(0)$$

e per la densità di potenza si ha $\mathcal{P}_y(f) = \frac{N_0}{2} |H(f)|^2$: pertanto, il processo in uscita dal filtro *non è più bianco*, ed in questo caso il processo si dice *colorato*. A questo fenomeno corrisponde anche una modifica della funzione di autocorrelazione: questa infatti non è più un *sinc*, ma vale $\mathcal{R}_y(\tau) = N_0 B \cdot \mathcal{R}_h(\tau) * sinc(2Bt)$; mentre prima quindi (per il processo bianco) due suoi valori estratti in istanti multipli di $1/2B$ erano comunque incorrelati, la colorazione introdotta dal filtro ha causato l'insorgenza di una dipendenza statistica tra i valori estratti a tali intervalli[22].

[21] in realtà si fa l'ulteriore ipotesi che la banda passante di $H(f)$ sia minore di $B$

[22] Il motivo di questo risultato può essere meglio compreso ricordando che l'integrale di convoluzione calcola i singoli valori in uscita da un filtro, come dipendenti da tutti gli ingressi passati, ognuno pesato con il valore della risposta impulsiva relativo al ritardo tra ingresso passato ed uscita presente. Pertanto, anche se i singoli valori in ingresso sono incorrelati, quelli di uscita (distanti tra loro per meno della durata della risposta impulsiva) condividono una porzione di storia comune, e quindi i loro valori non sono più indipendenti.

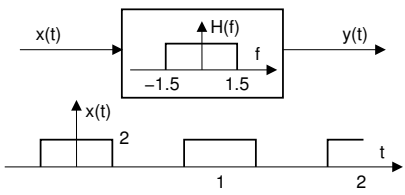
**Densità di probabilità.** A riguardo della $p_Y(y)$ non si può dire nulla di generale, tranne che essa dipende dalla $p_X(x)$ di ingresso e dalle operazioni compiute dal filtro; la sua espressione esatta va però determinata caso per caso. L'unico caso in cui la teoria fornisce una regola certa, è relativo ancora una volta al caso di processi gaussiani: posti in ingresso ad un filtro, producono in uscita un processo anch'esso gaussiano. Questo risultato è una diretta conseguenza della definizione stessa di processo gaussiano, come risultato della somma di infinite cause identicamente distribuite: dato che l'integrale di convoluzione effettivamente esegue una somma tra infiniti valori di ingresso, in linea di principio è lecito affermare che nel transito in un filtro la densità di probabilità dell'uscita *si gaussianizza*.

**Esercizio** Sia dato il filtro in figura, con

$$H(f) = rect_3(f)$$

ed al cui ingresso viene posto il segnale

$$x(t) = 2\sum_{n=-\infty}^{\infty} rect_{\frac{1}{2}}(t-n)$$




Calcolare:

**1)** la potenza in ingresso $\mathcal{P}_x$,
**2)** la potenza in uscita $\mathcal{P}_y$,
**3)** l'espressione di $y(t)$.

**Risposte**

**1)** Calcoliamo la media temporale: $\mathcal{P}_x = \frac{1}{T}\int_{-T/2}^{T/2} x^2(t)\,dt = \frac{1}{T}\int_{-1/4}^{1/4} 2^2(t)\,dt = \frac{4}{2} = 2$;

**2)** Sappiamo che $\mathcal{P}_y = \int_{-\infty}^{\infty}\mathcal{P}_y(f)\,df$, in cui $\mathcal{P}_y(f) = |Y(f)|^2$, ed a sua volta $Y(f) = X(f)H(f)$. Calcoliamo perciò innanzitutto

$$X(f) = F\{2rect_\tau(t) * \sum_{n=-\infty}^{\infty}\delta(t-nT)\} = 2\tau\cdot\mathrm{sinc}(f\tau)\frac{1}{T}\sum_{n=-\infty}^{\infty}\delta\left(f-\frac{n}{T}\right)$$

essendo $\tau = \frac{1}{2}$ e $T = 1$, risulta

$$X(f) = \mathrm{sinc}\left(\frac{f}{2}\right)\sum_{n=-\infty}^{\infty}\delta(f-n) = \sum_{n=-\infty}^{\infty}X_n\delta(f-n)$$

con $X_n = \mathrm{sinc}\left(\frac{n}{2}\right)$. Dunque, dato che gli unici impulsi che cadono entro la risposta in frequenza $H(f)$ sono quelli per $f = -1$, $0$ e $1$, si ha:

$$Y(f) = X(f)H(f) = \sum_{n=-1}^{1}X_nH(n)\delta(f-n)$$

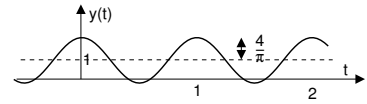
il cui modulo quadro fornisce $\mathcal{P}_y(f) = \sum_{n=-1}^{1}|X_n|^2|H(n)|^2\delta(f-n)$, e pertanto si ottiene

$$\mathcal{P}_y = \int_{-\infty}^{\infty}\mathcal{P}_y(f)\,df = \left(\frac{\sin-\frac{\pi}{2}}{-\frac{\pi}{2}}\right)^2 + 1 + \left(\frac{\sin\frac{\pi}{2}}{\frac{\pi}{2}}\right)^2 = 1 + 2\left(\frac{2}{\pi}\right)^2 = 1.811$$

**3)** Risulta

$$y(t) = \sum_{n=-1}^{1}X_nH(n)e^{j2\pi nt} = 1 + \frac{2}{\pi}\left(e^{j2\pi t} + e^{-j2\pi t}\right) = 1 + \frac{4}{\pi}cos2\pi t$$

Notiamo come il filtro lasci passare solamente la componente continua e la prima armonica.




### 9.4.4 Filtro adattato

Si tratta del filtro da utilizzare da parte di un *detettore di impulso*, ovvero un dispositivo che deve decidere per la presenza o assenza di una forma d'onda nota (a cui è sovrapposto rumore), in modo da rendere minima la probabilità di sbagliare. Supponiamo quindi di trasmettere un segnale $x(t)$, ottenuto facendo transitare un impulso $\delta(t)$ in un filtro con risposta impulsiva (di durata limitata $T$) $h_T(t) = g\left(t - \frac{T}{2}\right)$, e di ricevere lo stesso segnale in presenza di rumore gaussiano a media nulla $n(t)$, con densità spettrale $\mathcal{P}_N(f) = \frac{N_0}{2}$.

Tale ricevitore effettua la decisione per la presenza (ipotesi $H_1$) o l'assenza (ipotesi $H_0$) del segnale $x(t)$ (vedi figura 9.3) dopo che il segnale ricevuto $y(t)$ ha attraversato il filtro di ricezione $h_R(t)$, la cui uscita $z(t)$ è campionata all'istante $t = T$. Il valore $z(T)$ è quindi confrontato con *una soglia* $\lambda$, determinando la decisione per $H_1$ o $H_0$ a seconda se $\lambda$ sia superata o meno. Si commette un errore sia decidendo per $H_1$ in assenza di segnale, sia decidendo per $H_0$ in sua presenza[23]. Si dimostra[24] che la probabilità di errore può essere resa *minima* se $H_R(f)$ è scelto in modo da rendere massimo il rapporto $SNR$ all'istante di decisione, che corrisponde a scegliere[25]

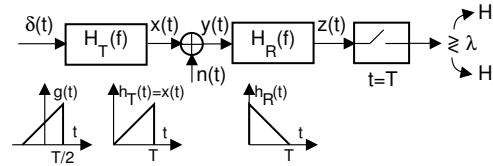



Figura 9.3: Detezione di impulso mediante filtro adattato

$$H_R(f) = H_T^*(f)\,\mathrm{e}^{-j2\pi fT} = X^*(f)\,\mathrm{e}^{-j2\pi fT} = G^*(f)\,\mathrm{e}^{-j2\pi f\frac{T}{2}} \tag{9.6}$$

ovvero

$$h_R(t) = h_T^*(T-t) = x(T-t) = g^*\left(\frac{T}{2} - t\right) \tag{9.7}$$

---

[23] Indicando rispettivamente con $P_{e0}$ e $P_{e1}$ i due tipi di errore, pari a $P_{e0} = \int_{\lambda}^{\infty} p_Z(z/H_0)\,dz$ e $P_{e1} = \int_{-\infty}^{\lambda} p_Z(z/H_1)\,dz$, la probabilità di errore complessiva vale $P_e = P_{e0}P_0 + P_{e1}P_1$, in cui $P_o = Pr(H_0)$ e $P_1 = Pr(H_1)$.

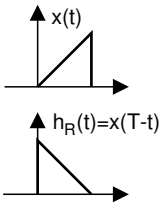
[24] la dimostrazione è rimandata alla nota 31

[25] Avendo definito $h_T(t) = g\left(t - \frac{T}{2}\right)$, risulta che $H_T^*(f) = \left(G(f)\,\mathrm{e}^{-j2\pi f\frac{T}{2}}\right)^* = G^*(f)\,\mathrm{e}^{+j2\pi f\frac{T}{2}}$, e dunque $H_T^*(f)\,\mathrm{e}^{-j2\pi fT} = X^*(f)\,\mathrm{e}^{-j2\pi fT} = G^*(f)\,\mathrm{e}^{-j2\pi f\frac{T}{2}}$.

D'altra parte, potendo scrivere $H_T^*(f)\,\mathrm{e}^{-j2\pi fT} = \left(H_T(f)\,\mathrm{e}^{j2\pi fT}\right)^*$ e ricordando ora la proprietà (3.2) espressa a pag. 34 $\mathcal{F}^{-1}\{X^*(f)\} = x^*(-t)$, otteniamo che

$$h_R(t) = \mathcal{F}^{-1}\left\{H_T^*(f)\,\mathrm{e}^{-j2\pi fT}\right\} = \mathcal{F}^{-1}\left\{\left(H_T(f)\,\mathrm{e}^{j2\pi fT}\right)^*\right\} = h_T^*(\theta+T)\big|_{\theta=-t} = h_T^*(T-t) = x^*(T-t)$$

Infine, essendo $x(t) = g(t - T/2)$ si ottiene anche $h_R(t) = g^*(\theta - T/2)|_{\theta=T-t} = g^*(T/2 - t)$. La fig. 9.3 mostra l'esito di tali operazioni nel caso di una $g(t)$ triangolare.




Con tali scelte, nel caso $H_0$ in cui $x(t)$ è assente risulta $y(t) = n(t)$, e la grandezza di decisione $z(T)$ è una v.a. gaussiana[26] definita come

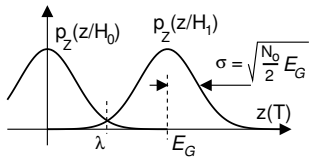
$$\begin{aligned} z(T) &= \int_{-\infty}^{\infty} h_R(\tau)\, y(T-\tau)\, d\tau = \\ &= \int_0^T x(T-\tau)\, n(T-\tau)\, d\tau = \mathcal{R}_{XN}(0) \end{aligned}$$

ossia pari all'intercorrelazione tra $x(t)$ e $n(t)$, calcolata nell'origine, e presenta valore atteso $m_{z(T)}^{H_0}$ nullo[27] e varianza[28] $\sigma_{z(T)}^2 = \frac{N_0}{2}\mathcal{E}_G$ in cui $\mathcal{E}_G$ è l'energia dell'impulso $g(t)$.

Se invece il segnale è presente (ipotesi $H_1$), si ottiene

$$z(T) = \int_0^T x(T-\tau)\left[x(T-\tau) + n(T-\tau)\right] d\tau = \mathcal{R}_X(0) + \mathcal{R}_{XN}(0)$$

producendo in questo caso una grandezza di decisione $z(T)$ con valor medio $m_{z(T)}^{H_1} = \mathcal{E}_G$, mentre per la sua varianza $\sigma_{z(T)}^2$ vale lo stesso risultato precedente[29].




La figura a fianco mostra la d.d.p. di $z(T)$ nelle ipotesi $H_0$ ed $H_1$, che risulta gaussiana come lo è $n(t)$. Notiamo che $m_{z(T)}^{H_1} = \mathcal{E}_G$ *non dipende* dalla particolare $g(t)$ adottata, né dalla sua durata $T$, ma solo dalla sua energia, ed è per questo che il filtro di ricezione è detto *adattato*. Osserviamo quindi che fissando la soglia $\lambda$ al valore $\frac{\mathcal{E}_G}{2}$ si rende minima la probabilità di errore, nel caso in cui le probabilità a priori delle due ipotesi siano uguali, ovvero $P_o = Pr(H_0) = P_1 = Pr(H_1)$.

Il rapporto $\frac{\left(m_{z(T)}^{H_1}\right)^2}{\sigma_{z(T)}^2}$ rappresenta[30] l'$SNR$ all'istante di decisione, ed il suo valore

$$SNR = \frac{\left(\mathcal{E}_G\right)^2}{\frac{N_0}{2}\mathcal{E}_G} = \frac{2\mathcal{E}_G}{N_0} \tag{9.8}$$

---

[26] Ricordiamo che l'uscita di un filtro al cui ingresso è posto un processo gaussiano, è anch'essa gaussiana.

[27] Infatti $m_{z(T)}^{H_o} = E\{\mathcal{R}_{XN}(0)\} = E\left\{\int_0^T x^*(t)\, n(t)\, dt\right\}$, che è pari a zero se $E\{n(t)\} = 0$.

[28] Risulta $\sigma_{z(T)}^2 = E\left\{z^2(T)\right\} = \mathcal{R}_Z(\tau)|_{\tau=0}$. Sappiamo che $\mathcal{R}_Z(\tau) = \mathcal{R}_N(\tau) * \mathcal{R}_{H_R}(\tau) = \frac{N_0}{2}\delta(\tau) * \mathcal{R}_{H_R}(\tau) = \frac{N_0}{2}\mathcal{R}_{H_R}(\tau)$; pertanto

$$\sigma_{z(T)}^2 = \frac{N_0}{2}\mathcal{R}_{H_R}(0) = \frac{N_0}{2}\int_{-\infty}^{\infty} h_R^*(t)\, h_R(t)\, dt = \frac{N_0}{2}\mathcal{E}_{\mathcal{G}}$$

dato che $h_R(t)$ ha la stessa energia di $g(t)$.

[29] Infatti, ora risulta $m_{z(T)}^{H_1} = E\{\mathcal{R}_X(0) + \mathcal{R}_{XN}(0)\}$, in cui il contributo del secondo termine è nullo come già osservato, mentre quello del primo non è aleatorio, e vale $\mathcal{R}_X(0) = \int_0^T x^*(t)\, x(t)\, dt = \mathcal{E}_{\mathcal{G}}$, in quanto il segnale $x(t)$ ha la stessa energia di $g(t)$. Per ciò che riguarda $\sigma_{z(T)}^2$, osserviamo che essendo il filtro di ricezione un operatore lineare, l'uscita si ottiene come sovrapposizione degli effetti delle due cause $x(t)$ ed $n(t)$, e la componente aleatoria dell'uscita è dovuta al solo $n(t)$; pertanto, la sua varianza è la stessa calcolata per il caso $H_0$ di segnale assente.

[30] Il significato fisico del rapporto indicato può essere meglio visualizzato considerandone le radice quadrata, ossia $\frac{m_{z(T)}^{H_1}}{\sigma_{z(T)}}$, che costituisce il rapporto tra l'uscita per $t = T$ in presenza di solo segnale, e la deviazione standard di tale valore introdotta dal rumore. In altre parole, è indicativo della separazione tra le gaussiane riportate in figura. Pertanto, maggiore è questo rapporto, e minore sarà la probabilità di errore.

costituisce *il massimo*[31] che si può ottenere, adottando la (9.7) tra tutte le scelte possibili per il filtro di ricezione $h_{Rx}(t)$, di energia pari a $\mathcal{E}_G$. Notiamo che (9.8) è valida solo in presenza di rumore bianco, mentre se questo è colorato, l'$SNR$ diminuisce, ed il filtro ottimo va determinato in altro modo.

**Rumore colorato** Nel caso in cui $\mathcal{P}_N(f)$ non sia pari ad una costante, la condizione per massimizzare (9.8) non è più la (9.6), bensì deve risultare[32]

$$H_R(f) = \frac{X^*(f)\,\mathrm{e}^{-j2\pi fT}}{\mathcal{P}_N(f)} \tag{9.9}$$

in modo che $H_R(f)$, oltre ad esaltare le frequenze per le quali lo spettro del segnale è maggiore, riesce anche ad attenuare quelle per le quali la potenza di rumore è più grande.

**Assenza di rumore** Se non fosse presente rumore, l'andamento dell'uscita del filtro adattato sarebbe proprio pari alla funzione di autocorrelazione di $g(t)$, che viene campionata in corrispondenza del suo massimo. Notiamo che la $H_R(f)$ *non* presenta modulo costante e fase lineare, dato che lo scopo qui *non è* quello di preservare la forma d'onda in transito, ma di massimizzare l'$SNR$ all'istante di decisione.

#### 9.4.4.1 Segnalazione antipodale

Desiderando distinguere tra due possibili messaggi (ad es, $x_1$ ed $x_2$), e volendo rendere minima la probabilità di errore, la scelta *ottima* consiste nell'adottare $x_2(t) = -x_1(t)$, e di impiegare al ricevitore un filtro adattato ad $x_1(t)$. In tal modo, all'istante di campionamento si avrà (in assenza di rumore) un valore positivo o negativo, a seconda se sia presente $x_1$ od $x_2$, mentre in presenza di rumore si avrà ancora la minima probabilità di errore possibile, ma con un risparmio della potenza trasmessa, dato che

[31] Consideriamo il caso in cui si abbia una $H_R(f) = H(f)$ generica. In presenza di solo segnale, si ottiene

$$|z(T)|^2 = \left| F^{-1}\{Z(f)\}\Big|_{t=T}\right|^2 = \left|\int_{-\infty}^{\infty} H(f)\,X(f)\,\mathrm{e}^{j2\pi fT}df\right|^2$$

Riportiamo ora la diseguaglianza di Schwartz (vedi pag. 27), che afferma la relazione $\left|\int_{-\infty}^{\infty} a(\theta)\,b^*(\theta)\,d\theta\right|^2 \leq \int_{-\infty}^{\infty}|a(\theta)|^2\,d\theta \cdot \int_{-\infty}^{\infty}|b(\theta)|^2\,d\theta$, con l'eguaglianza solo se $a(\theta) = k \cdot b(\theta)$. Se ora facciamo corrispondere $H(f)$ ad $a(\theta)$ e $X(f)\,\mathrm{e}^{j2\pi fT}$ a $b^*(\theta)$, otteniamo che

$$|z(T)|^2 = \left(m_{z(T)/H_1}\right)^2 \leq \int_{-\infty}^{\infty}|H(f)|^2\,df \cdot \int_{-\infty}^{\infty}|X(f)|^2\,df$$

con l'eguaglianza solo se $H(f) = kX^*(f)\,\mathrm{e}^{-j2\pi fT}$, che corrisponde (vedi pag. 27) a $h(t) = kx(T-t)$, ossia se $H(f)$ è *adattata* a $X(f)$. Scegliendo $k = 1$, i due integrali a prodotto hanno lo stesso valore, pari a $\mathcal{E}_G$.

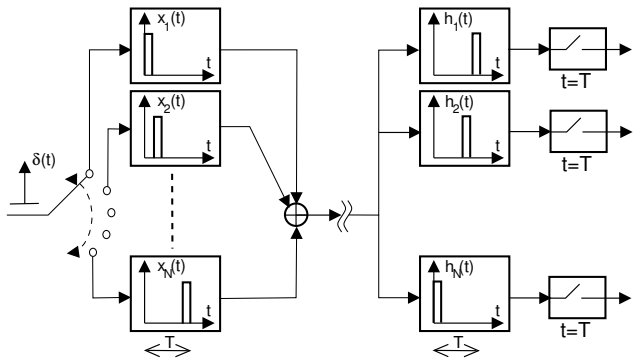
[32] La condizione (9.9) si ottiene anche in questo caso imponendo la minimizzazione di $SNR = \frac{\left(m^{H_1}_{z(T)}\right)^2}{\sigma^2_{z(T)}} = \frac{\left|\int_{-\infty}^{\infty} H(f)X(f)\mathrm{e}^{j2\pi fT}df\right|^2}{\int_{-\infty}^{\infty}|H(f)|^2\mathcal{P}_N(f)df}$ il cui denominatore tiene conto che $\sigma^2_{z(T)} = \int_{-\infty}^{\infty}\mathcal{P}_Z(f)\,df$ è dovuta al solo rumore. Applichiamo ora a $SNR$ la diseguaglianza di Schwartz posta nella forma $\frac{\left|\int_{-\infty}^{\infty} a(\theta)b^*(\theta)d\theta\right|^2}{\int_{-\infty}^{\infty}|a(\theta)|^2 d\theta} \leq \int_{-\infty}^{\infty}|b(\theta)|^2\,d\theta$ e identifichiamo $a(\theta)$ con $H(f)\sqrt{\mathcal{P}_N(f)}$ e $b^*(\theta)$ con $X(f)\mathrm{e}^{j2\pi fT}/\sqrt{\mathcal{P}_N(f)}$. Imponendo di nuovo la condizione $a(\theta) = k \cdot b(\theta)$ con $k = 1$, otteniamo il massimo $SNR$ come $SNR = \int_{-\infty}^{\infty}|b(\theta)|^2\,d\theta = \int_{-\infty}^{\infty}\frac{|X(f)|^2}{\mathcal{P}_N(f)}df$, e quindi scrivendo $a(\theta) = b(\theta)$ ossia $H(f)\sqrt{\mathcal{P}_N(f)} = X^*(f)\mathrm{e}^{-j2\pi fT}/\sqrt{\mathcal{P}_N(f)}$ si ottiene il risultato (9.9).

in questo caso la stessa distanza tra le gaussiane può ottenersi con impulsi di metà energia $\mathcal{E}_G/2$[33].

#### 9.4.4.2 Segnalazione ortogonale

Dovendo trasmettere $N$ diversi messaggi $(x_1, x_2, ..., x_N)$, associamo ad ognuno di essi una forma d'onda $x_i(t)$ tale che $\int x_i(t)\, x_j(t)\, dt = 0$ con $i \neq j$, ovvero in modo che i segnali $x_i(t)$ siano *ortogonali*. In tal caso il ricevitore ottimo è costituito da un banco di filtri, ognuno adattato ad una diversa $x_i(t)$, e in assenza di rumore la ricezione di una delle forme d'onda $x_i(t)$ non produce nessuna uscita sui filtri del banco per $j \neq i$. In presenza di rumore, la decisione su cosa sia stato trasmesso viene presa valutando quale dei filtri presenta il valore massimo in corrispondenza dell'istante di campionamento, realizzando così un *ricevitore a correlazione* (vedi § 13.1.3 a pag. 298).

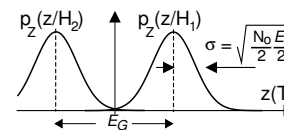
**Esempio** L'impulso $\delta(t)$ entra in uno di filtri mostrati nella figura seguente, le cui risposte impulsive $x_i(t)$ realizzano una famiglia di funzioni ortogonali, dato che le rispettive forme d'onda non si sovrappongono nel tempo. In ricezione, solo uno dei filtri adattati con risposta impulsiva $h_i(t)$ produce una uscita diversa da zero per $t = T$, come verificabile ricordando la costruzione grafica dell'operazione di convoluzione mostrata a pag. 38.




## 9.5 Caratteristiche dei sistemi fisici

Abbiamo già osservato come il legame $\mathcal{T}[x(t)] = y(t)$ definito in 1.7.1 tra ingresso $x(t)$ ed uscita $y(t)$ di un sistema fisico sia definito, in base alla conoscenza della risposta impulsiva, dall'integrale di convoluzione $y(t) = x(t) * h(t)$, che descrive il risultato di una operazione di filtraggio. Alla luce di questo risultato, torniamo ad analizzare le proprietà enunciate al § 1.7.1, assieme ad altre due.

[33] Il senso di questa affermazione può essere meglio apprezzato mediante la figura a fianco, in cui è mostrato il valore di uscita da un filtro adattato per impulsi antipodali di energia $\mathcal{E}_G/2$, e che nelle due ipotesi vale $+\mathcal{E}_G/2$ e $-\mathcal{E}_G/2$, mantenendo così pari ad $\mathcal{E}_G$ la separazione tra i valori in assenza di rumore. Inoltre, dato che la deviazione standard dell'uscita dal f.a. in presenza di rumore dipende dall'energia dell'impulso, ora risulta $\sigma = \frac{1}{\sqrt{2}}\sqrt{\frac{N_0 \mathcal{E}_G}{2}}$, e quindi le gaussiane sono più concentrate, e la probabilità di errore migliore.



**Linearità** Un sistema è lineare se, in presenza di una combinazione lineare di ingressi, l'uscita è la combinazione lineare delle uscite, ossia sussiste la legge di sovrapposizione degli effetti, ovvero

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

Ad esempio, il legame ingresso-uscita descritto dall'integrale di convoluzione è di tipo *lineare*, in virtú della distributività dell'integrale. Al contrario, un operatore basato sulla elevazione a potenza è *non lineare*, come approfondito al § 14.6.

**Memoria** Notiamo che un sistema descritto da una risposta impulsiva $h(t)$ con estensione temporale non nulla, è detto *con memoria,* in quanto i singoli valori di uscita dipendono da tutti i valori di ingresso "raccolti" dalla risposta impulsiva.

**Permanenza** Un sistema è permanente se la risposta ad un ingresso ritardato è anch'essa ritardata. Se la risposta impulsiva non cambia nel tempo, il sistema è permanente.

**Realizzabilità ideale** Un sistema è idealmente realizzabile se $h(t)$ *è reale.*

**Realizzabilità fisica** La proprietà di osservare valori di uscita che dipendono solo dagli ingressi passati, è automaticamente verificata se $h(t) = 0$ con $t < 0$. Abbiamo già osservato (nota 13 a pag. 71) come sistemi non realizzabili fisicamente possano essere approssimati da sistemi realizzabili accettando un ritardo dell'uscita.

**Stabilità** Si può mostrare che la proprietà di stabilità equivale alla condizione $\int |h(t)|\, dt < \infty$, ovvero che $h(t)$ sia un segnale impulsivo. Notiamo che questa circostanza garantisce l'esistenza della sua trasformata $H(f)$.

**Risposta in frequenza** Se un sistema, oltre che stabile, è anche idealmente realizzabile, allora $H(f) = H^*(-f)$, e dunque è sufficiente conoscere la parte a frequenze positive indicata con $H^+(f)$, dato che l'altra metà è ottenibile mediante una operazione di coniugazione. Questo fatto permette di misurare

$$H(f) = M(f)\, e^{j\varphi(f)}$$

(ossia modulo e fase di $H(f)$), che prende il nome di *risposta in frequenza,* utilizzando come ingresso una funzione sinusoidale con ampiezza $A$ e fase $\theta$ note: $x(t) = A\cos(2\pi f_0 t + \theta)$. Il segnale in uscita sarà ancora una cosinusoide[34] e avrà ampiezza $A \cdot M(f_0)$ e fase $\varphi(f_0) + \theta$; pertanto potremo ricavare

$$M(f_0) = \frac{\max\{y(t)\}}{\max\{x(t)\}}, \quad \text{e} \quad \varphi(f_0) = \arg\{y(t)\} - \arg\{x(t)\}$$

[34] Svolgiamo i calcoli nel dominio della frequenza:

$$X(f) = \frac{A}{2}\left(e^{j\theta}\delta(f - f_0) + e^{-j\theta}\delta(f + f_0)\right);$$

$$Y(f) = X(f)\,H(f) = \frac{A}{2}M(f_0)\left(e^{j\theta}e^{j\varphi(f_0)}\delta(f - f_0) + e^{-j\theta}e^{-j\varphi(f_0)}\delta(f + f_0)\right)$$

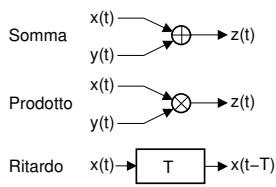
e antitrasformando si ottiene

$$y(t) = A \cdot M(f_0)\cos(2\pi f_0 t + \varphi(f_0) + \theta)$$

Ripetendo il procedimento per diverse $f_0$, possiamo "campionare" $H(f)$.

## 9.6 Unità di elaborazione

Mentre un filtro viene definito, in base alla precedente classificazione, un operatore lineare con memoria, può essere utile considerare altri operatori elementari, che funzionano come "mattoni" di operazioni più complesse. Gli operatori elementari, rappresentati in figura, sono la somma, il prodotto, ed il ritardo.




Forniamo dapprima i risultati della combinazione di processi e segnali certi mediante gli operatori introdotti, e quindi mostriamo alcune semplici realizzazioni di *filtri digitali*, definiti nei termini degli operatori elementari.

### 9.6.1 Prodotto

Nel caso in cui un fattore sia un processo, e l'altro un segnale certo, il risultato (in generale) è un processo *non stazionario*. Infatti ora le medie d'insieme dipendono, istante per istante, dal valore che il segnale certo assume in quell'istante (tranne il caso in cui sia una costante)[35].

Se uno dei due fattori (ad es $y(t) = \tilde{y}$) è una costante, $z(t)$ è un processo della stessa natura di $x(t)$, con media $m_z = m_x \cdot \tilde{y}$, potenza $\mathcal{P}_z = \mathcal{P}_x \cdot \tilde{y}^2$, e autocorrelazione $\mathcal{R}_z(\tau) = \mathcal{R}_x(\tau) \cdot \tilde{y}^2$.

Se i fattori $x(t)$ ed $y(t)$ sono processi statisticamente indipendenti[36], stazionari e congiuntamente[37] ergodici, si ottiene:

**Valor medio**

$$m_z = E\{z(t)\} = E\{x(t)\,y(t)\} = E\{x(t)\}\,E\{y(t)\} = m_x \cdot m_y$$

**Potenza totale**

$$\mathcal{P}_z = E\left\{z^2(t)\right\} = E\left\{x^2(t)\,y^2(t)\right\} = E\left\{x^2(t)\right\}E\left\{y^2(t)\right\} = \mathcal{P}_x \cdot \mathcal{P}_y$$

**Varianza**

$$\sigma_z^2 = E\left\{(z(t) - m_z)^2\right\} = \mathcal{P}_z - (m_z)^2 = \mathcal{P}_x \cdot \mathcal{P}_y - (m_x \cdot m_y)^2$$

[35] Se il segnale certo è periodico, il risultato della moltiplicazione per un processo stazionario dà luogo ad un processo detto *ciclostazionario*, in quanto le statistiche variano nel tempo, ma assumono valori identici con periodicità uguale a quella del segnale certo.

[36] Compreso il caso di processi armonici!

[37] La proprietà di ergodicità congiunta corrisponde a verificare le condizioni ergodiche anche per i momenti misti $m_{XY}^{(1,1)}(x,y)$ relativi a coppie di valori estratti da realizzazioni di due differenti processi.

**Funzione di autocorrelazione**

$$\begin{aligned}\mathcal{R}_z(\tau) &= E\{z(t)\,z(t+\tau)\} = E\{x(t)\,y(t)\,x(t+\tau)\,y(t+\tau)\} = \\ &= E\{x(t)\,x(t+\tau)\}\,E\{y(t)\,y(t+\tau)\} = \mathcal{R}_x(\tau)\cdot\mathcal{R}_y(\tau)\end{aligned}$$

In particolare, notiamo che l'incorrelazione di uno dei due processi, per un certo valore di $\tau$, provoca l'incorrelazione del prodotto.

**Spettro di densità di potenza**

$$\mathcal{P}_z(f) = \mathcal{F}\{\mathcal{R}_z(\tau)\} = \mathcal{F}\{\mathcal{R}_x(\tau)\cdot\mathcal{R}_y(\tau)\} = \mathcal{P}_x(f) * \mathcal{P}_y(f)$$

ossia è pari alla convoluzione tra le densità spettrali dei fattori. Notiamo quindi che la densità di potenza del prodotto presenta una occupazione di banda maggiore di quella dei singoli fattori.

**Densità di probabilità** Si calcola con le regole per il cambiamento di variabile, illustrate al § 7.6.4. Nel caso in cui i due processi siano statisticamente indipendenti, il risultato è

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(\theta)\,p_Y\left(\frac{z}{\theta}\right)\frac{d\theta}{\theta}$$

In Appendice (pag. 232), troviamo l'applicazione di questi risultati al calcolo della densità di potenza di un segnale dati.

### 9.6.2 Somma

Anche in questo caso, se un termine è un processo e l'altro un segnale certo, la somma è (in generale) un segnale non stazionario. Se il segnale certo è costante, si torna al caso stazionario[38]. Procediamo ora nel calcolo delle solite grandezze, con l'ipotesi che $x(t)$ ed $y(t)$ siano processi *statisticamente indipendenti*.

**Valore medio**

$$m_z = E\{x(t) + y(t)\} = E\{x(t)\} + E\{y(t)\} = m_x + m_y$$

**Potenza totale**

$$\begin{aligned}\mathcal{P}_z &= E\left\{(x(t)+y(t))^2\right\} = E\left\{x^2(t)\right\} + E\left\{y^2(t)\right\} + 2E\{x(t)\cdot y(t)\} \\ &= \mathcal{P}_x + \mathcal{P}_y + 2m_x m_y\end{aligned}$$

Pertanto, se almeno uno dei due processi è a media nulla, le potenze dei due processi si sommano.

**Varianza**

$$\begin{aligned}\sigma_z^2 &= E\left\{(z(t)-m_z)^2\right\} = \mathcal{P}_z - (m_z)^2 = \mathcal{P}_x + \mathcal{P}_y + 2m_x m_y - (m_x+m_y)^2 = \\ &= \mathcal{P}_x - (m_x)^2 + \mathcal{P}_y - (m_y)^2 = \sigma_x^2 + \sigma_y^2\end{aligned}$$

[38] Come per il prodotto, se il segnale certo è periodico, la somma si dice *ciclostazionaria* perché la dipendenza temporale non è assoluta, ma periodica.

**Autocorrelazione**

$$\begin{aligned}\mathcal{R}_z(\tau) &= E\{z(t)\,z(t+\tau)\} = E\{x(t)\,x(t+\tau)\} + \\ &\quad E\{y(t)\,y(t+\tau)\} + E\{x(t)\,y(t+\tau)\} + E\{x(t+\tau)\,y(t)\} = \\ &= \mathcal{R}_x(\tau) + \mathcal{R}_y(\tau) + 2m_x m_y\end{aligned}$$

Osserviamo come per $\tau = 0$ si ritrovi il valore della potenza totale.

**Spettro di densità di potenza**

$$\mathcal{P}_z(f) = \mathcal{F}\{\mathcal{R}_z(\tau)\} = \mathcal{P}_x(f) + \mathcal{P}_y(f) + 2m_x m_y \delta(f)$$

**Densità di probabilità** Applicando le regole del cambiamento di variabile (§ 7.6.4), o passando per il calcolo della funzione caratteristica (§ 7.6.3), nel caso di $x(t)$ ed $y(t)$ indipendenti, si ottiene

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(\theta)\, p_Y(z-\theta)\, d\theta = p_X(x) * p_Y(y)$$

La relazione esprime l'importante risultato che la densità di probabilità della somma di variabili aleatorie è la *convoluzione* tra le densità dei termini della somma.

**Esempio** Se $x$ ed $y$ sono ad es. due v.a. a distribuzione uniforme tra $\pm\Delta$, la loro somma ha densità di probabiltà triangolare con base $2\Delta$. Pertanto, nel lancio di 2 dadi il risultato più probabile è 7. Infatti può essere ottenuto come 6+1, 5+2, 4+3, 3+4, 2+5, 1+6, ovvero in 6 modi diversi, ognuno con probabilità $\frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$ e dunque $Pr\{7\} = 6\frac{1}{36} = \frac{1}{6}$.

## 9.7 Filtri digitali

I filtri digitali sono una particolare classe di filtri, per i quali l'integrale di convoluzione si riduce in una sommatoria[39]. La parte sinistra di fig. 9.4 mostra una particolare architettura di filtro digitale denominata *filtro trasversale,* per il quale è facile verificare che la risposta impulsiva ha espressione

$$h(t) = \sum_{n=0}^{N} c_n \delta(t - nT)$$

in cui $N$ è l'ordine del filtro[40]. E' altrettanto facile verificare che la funzione di trasferimento ha espressione

$$H(f) = \mathcal{F}\{h(t)\} = \sum_{n=0}^{N} c_n e^{-j2\pi f n T}$$

[39] Sebbene l'analisi che viene svolta si riferisca ad un *normale* segnale tempo-continuo, i filtri digitali si prestano ad essere realizzati via software (o mediante apposito hardware dedicato), eseguendo le operazioni direttamente su campioni di segnale: in tal caso vengono chiamati anche *filtri numerici*, che non trattiamo, limitandoci a indicare una possibile fonte di approfondimento in `http://www.dspguide.com/ch14/6.htm`).

In tutti i modi, osserviamo che un segnale con banda $< W$ può essere validamente rappresentato (vedi cap. 4) nei termini dei suoi campioni distanziati di $T < 1/2W$. I campioni presi in uscita del filtro trasversale in effetti dipendono solo dai campioni dell'ingresso e pertanto, anche se qui analizzato in termini di segnali continui, il filtro trasversale costituisce allo stempo tempo anche una architettura per filtri numerici operanti su dati campionati.

[40] I coefficienti $c_n$ vengono indicati nei testi inglesi come *taps* (rubinetti) in quanto possono essere pensati "spillare" frazioni del segnale. Per effetto di un processo di trasposizione linguistica, gli stessi coefficienti in italiano vengono a volte indicati discorsivamente come *tappi* (!).

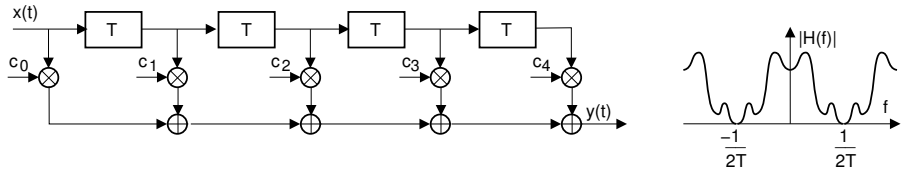



Figura 9.4: Schema simbolico di un filtro trasversale e sua risposta in frequenza

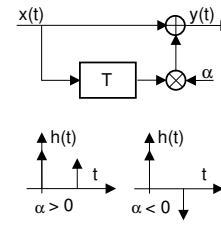
Una tale architettura può essere rappresentativa di un effettivo fenomeno naturale, come ad esempio la presenza di "echi" nel segnale ricevuto. Come rappresentato a destra in fig. 9.4, $H(f)$ risulta periodica (in frequenza) con periodo $F = \frac{1}{T}$, dato che tutti gli esponenziali $e^{-j2\pi fnT}$ lo sono. Pertanto questa architettura può essere adottata per *sintetizzare* una risposta in frequenza desiderata in una certa gamma di frequenze, e "qualsiasi" altrove. Infatti, se il segnale di ingresso $x(t)$ è limitato in una banda (base o traslata) di estensione minore di $\frac{1}{T}$, l'azione filtrante ha luogo appunto nella sola banda del segnale.

Per *sintetizzare* il filtro a partire dall'andamento desiderato di $H(f)$ nella banda di interesse, si calcolano i coefficienti $c_k$ mediante la formula

$$c_k = T \int_{-1/2T}^{1/2T} H(f)\, e^{j2\pi fkT} df$$

che è proprio (a parte un segno) l'espressione dei coefficienti per lo sviluppo in serie di Fourier di un periodo di segnale! Ovviamente, se $H(f)$ è qualsiasi, occorrerebbero un numero infinito di coefficienti $c_k$; usandone un numero inferiore (finito) si produce una approssimazione della $H(f)$ desiderata[41].

### 9.7.1 Filtro trasversale del $1^o$ ordine




E' descritto dalla architettura mostrata a lato, a cui corrisponde una risposta impulsiva

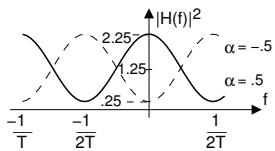
$$h(t) = \delta(t) + \alpha\delta(t - T)$$

[41] Chiaramente il troncamento della serie di coefficienti $c_k$ avverrà in modo simmetrico rispetto a $c_0$, prendendo cioè sia gli indici positivi che quelli negativi. Viceversa, nello schema di filtro trasversale si usano solo coefficienti con indici $\geq 0$. Nel caso in cui l'$H(f)$ da cui partiamo sia reale (e pari), allora i $c_k$ sono una serie reale pari, il che garantisce un filtro idealmente realizzabile, ma la cui $h(t) = \sum_{k=-N/2}^{N/2} c_k \delta(t - kT)$ necessita di una traslazione temporale per risultare anche fisicamente realizzabile. Se invece $H(f)$ ha un andamento qualunque, non si può dire nulla a riguardo di eventuali simmetrie per i coefficienti $c_k$.

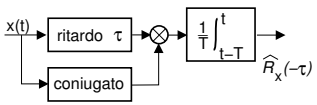
la cui trasformata è $H(f) = 1 + \alpha e^{-j2\pi fT}$([42]), e quindi

$$\begin{aligned} |H(f)|^2 &= (1+\alpha\cos 2\pi fT)^2 + (\alpha \sin 2\pi fT)^2 = \\ &= 1 + 2\alpha\cos 2\pi fT + \alpha^2\left(\cos^2 2\pi fT + \sin^2 2\pi fT\right) = \\ &= 1 + \alpha^2 + 2\alpha\cos 2\pi fT \end{aligned}$$

A lato è riportato l'andamento di $|H(f)|^2$ per due valori di $\alpha = \pm.5$, ed osserviamo che nell'intervallo di frequenze $|f| < \frac{1}{2T}$ può assumere un comportamento passa-alto oppure passa-basso[43], in funzione del segno di $\alpha$. Notiamo inoltre che ponendo $\alpha = -1$ si ottiene un differenziatore, in grado di rimuovere dall'ingresso segnali periodici di periodo $T$.




### 9.7.2 Stima della autocorrelazione di un processo ergodico



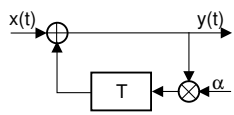

L'architettura riportata in figura *non è* un filtro (c'è un moltiplicatore anzichè un sommatore) ma viene illustrata ora "per similitudine". Si tratta di uno schema idoneo a misurare una stima $\widehat{\mathcal{R}}_x(\tau)$ della funzione di autocorrelazione di una realizzazione $x(t)$ di un processo. Variando il ritardo $\tau$ si ottiene $\widehat{\mathcal{R}}_x(\tau)$ per diversi valori dell'argomento. Dunque sarà poi possibile calcolare $\widehat{\mathcal{P}}_x(f) = \mathcal{F}\left\{\widehat{\mathcal{R}}_x(\tau)\right\}$, che rappresenta una stima dello spettro di densità di potenza di una qualunque realizzazione del processo, in virtú della proprietà di ergodicità.

### 9.7.3 Filtro digitale a risposta impulsiva *infinita* del $1^o$ ordine

In figura è raffigurato un filtro di tipo *ricorsivo,* il cui valore in uscita dipende dalle *uscite precedenti.* La risposta impulsiva corrispondente ha durata infinita, ed è pari a

$$h(t) = \sum_{n=0}^{\infty} \alpha^n \delta(t - nT)$$

La funzione di trasferimento risulta quindi pari a $H(f) = \sum_{n=0}^{\infty} \alpha^n e^{-j2\pi fnT}$ e, ricordando la formula dello sviluppo in serie geometrica $\sum_{n=0}^{\infty} \beta^n = \frac{1}{1-\beta}$, si può scrivere




[42] In questo caso $H(f)$, pur risultando a simmetria coniugata ($H_x(f) = H_x^*(-f)$), è complessa. Pertanto, i coefficienti $c_k$ ottenibili dalla relazione $c_k = T\int_{-1/2T}^{1/2T} H(f)\, e^{j2\pi fkT}\, df$ sono reali, ma non necessariamente pari. Svolgendo i calcoli, si ha: $c_k = T\int_{-1/2T}^{1/2T} \left(1 + \alpha e^{-j2\pi fT}\right) e^{j2\pi fkT}\, df = T\int_{-1/2T}^{1/2T} e^{j2\pi fkT} df + \alpha T \int_{-1/2T}^{1/2T} e^{j2\pi f(k-1)T} df$. Il primo integrale è nullo per $k \neq 0$, mentre il secondo per $k \neq 1$, in quanto le funzioni integrande hanno media nulla sull'intervallo $1/T$; pertanto $c_0 = 1$ e $c_1 = \alpha$, esattamente come è definita la risposta impulsiva.

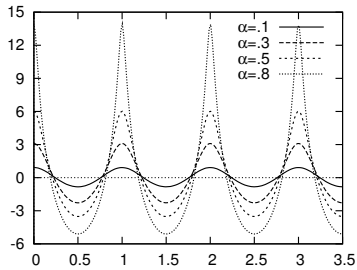
[43] In altre parole, l'andamento ondulatorio di $|H(f)|^2$ rende il filtro idoneo a diversi utilizzi, in funzione dell'andamento in frequenza del segnale di ingresso.

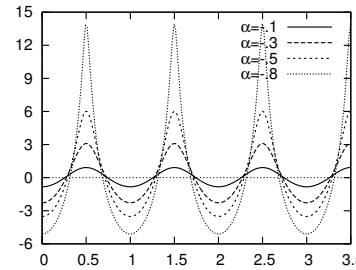
$$H(f) = \sum_{n=0}^{\infty} \left(\alpha e^{-j2\pi fT}\right)^n = \frac{1}{1-\alpha e^{-j2\pi fT}}$$

Notiamo subito che il filtro è stabile purchè $|\alpha| < 1$,
altrimenti si ha una uscita anche senza ingresso, ovvero uscita infinita con ingresso limitato. Per ciò che riguarda $|H(f)|^2$, otteniamo

$$|H(f)|^2 = \frac{1}{(1-\alpha\cos 2\pi fT)^2 + (\alpha \sin 2\pi fT)^2} = \frac{1}{1+\alpha^2 - 2\alpha\cos 2\pi fT}$$

Le curve mostrate appresso rappresentano $10\log_{10}|H(f)|^2$, con $T = 1$ e diversi valori di $\alpha$, positivi a sinistra e negativi a destra. Osserviamo infine che il caso $\alpha = 1$ corrisponde ad avere un integratore perfetto che, ad esempio, produce una rampa in uscita, se in ingresso c'è un gradino.




## 9.8 Filtri analogici

Sono ottenuti mediante componenti elettrici a costanti concentrate come condensatori, induttori e resistori.

Applicando la trasformata di *Laplace* alle equazioni differenziali che descrivono la relazione ingresso-uscita, si ottiene una funzione di trasferimento razionale del tipo

$$H(s) = \frac{\sum_{i=0}^{N} a_i s^i}{\sum_{j=0}^{M} b_j s^j}$$

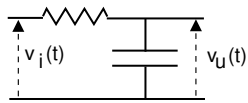
(in cui $N \leq M$), definita su di un piano complesso $s = \sigma + j2\pi f$. Ponendo $s = j2\pi f$ si ottiene la funzione di trasferimento in $f$: $H(f) = H(s = j2\pi f)$. Questo procedimento è valido solo se il filtro è stabile, che nel dominio di Laplace equivale a richiedere che tutti i poli di $H(s)$ siano a sinistra dell'asse immaginario.

### 9.8.1 Filtro analogico ad un polo

Nella figura a fianco è riportato un filtro RC di tipo passa basso, per il quale la relazione tra $v_u(t)$ e $v_i(t)$ è descritta da una risposta impulsiva con espressione

$$h(t) = \frac{1}{RC} e^{-\frac{t}{RC}}$$

L'analisi del circuito mostra che la funzione di trasferimento risulta



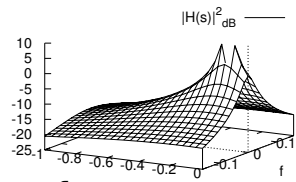

$$H(f) = \mathcal{F}\{h(t)\} = \frac{1/j\omega C}{R + 1/j\omega C} = \frac{1}{1 + j2\pi fRC}$$

ovvero, nel dominio di Laplace

$$H(s) = \frac{1}{1 + sRC}$$

Pertanto, $H(s)$ presenta un polo in $s = -\frac{1}{RC}$ che fa sì che $H(s)|_{s=-\frac{1}{RC}} = \infty$.

A lato, è raffigurato l'andamento di $|H(s)|^2$, espresso in decibel, e con $RC = 8$. Come evidente, $|H(s)|^2$ può essere pensata come una sorta di "cono vulcanico" attorno al polo, le cui falde, quando intersecate dal piano verticale infisso sull'asse $j2\pi f$, individuano la funzione di trasferimento in frequenza $H(f) = H(s = j2\pi f)$. Come si vede dalla figura, $H(f)$ risulta di tipo passa basso, con fianchi tanto più ripidi quanto più il polo è vicino all'origine.

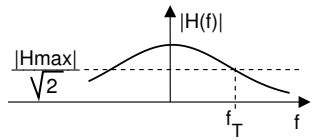



### 9.8.2 Frequenza di taglio

Definiamo *frequenza di taglio* di un filtro la frequenza $f_T$ per la quale

$$|H(f_T)| = \frac{|H_{Max}|}{\sqrt{2}}$$

Nel caso del filtro RC, si ha $|H_{Max}| = H(0) = 1$ e dunque scriviamo



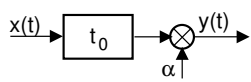

$$|H(f)| = \frac{1}{\sqrt{1 + (2\pi fRC)^2}} = \frac{1}{\sqrt{1 + \left(\frac{f}{f_T}\right)^2}}$$

in cui $f_T = \frac{1}{2\pi RC}$, pari quindi alla frequenza di taglio (infatti $|H(f_T)| = \frac{1}{\sqrt{1+1}} = \frac{1}{\sqrt{2}}$).

Notiamo anche che $|H(f_T)|^2 = \frac{1}{2}$ e dunque $|H(f_T)|^2\big|_{dB}$ = -3 dB; per questo la frequenza di taglio è indicata anche come frequenza a 3 dB.

### 9.8.3 Assenza di distorsioni lineari

Quali proprietà devono essere verificate da un filtro affinché l'uscita non differisca dall'ingresso per più di un fattore di scala ed un ritardo, ovvero si verifichi la proprietà di canale perfetto di pag. 333? La condizione cercata si esprime come $y(t) = \alpha x(t - t_0)$, che corrisponde a $Y(f) = \alpha X(f) e^{-j2\pi f t_0}$, e quindi la risposta in frequanza di tale filtro risulta




$$H(f) = \frac{Y(f)}{X(f)} = \alpha e^{-j2\pi f t_0}$$

e perciò la sua risposta impulsiva è pari a $h(t) = \alpha\delta(t - t_0)$. Pertanto le condizioni poste nel tempo, si riflettono su di una risposta in frequenza con *modulo costante* e *fase lineare,* quantomeno, nella banda del segnale.

## 9.9 Appendici

### 9.9.1 Coefficiente di correlazione

I diagrammi di esempio presentati alla nota (2) a pag. 208 basano la valutazione di quanto una coppia di v.a. $x$ ed $y$ siano correlate, anche sul calcolo del coefficiente di correlazione $\rho_{xy}$, che ha valori compresi tra $+1$ e $-1$, ed è definito come

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

In tal modo, si opera una normalizzazione del valore della covarianza $\sigma_{xy}$, rispetto alle deviazioni standard $\sigma_x$ e $\sigma_y$ delle due v.a., rendendo così il valore di $\rho$ indipendente dalla dinamica dei valori assunti da $x$ ed $y$.

Il coefficiente $\rho$ si presta ad una interessante interpretazione geometrica, una volta identificate (vedi § 2.4.1) la deviazione standard $\sigma_x$ come la norma $\|\overline{x}\|$ di $x$, e la covarianza $\sigma_{xy}$ come il prodotto scalare $(\overline{x}, \overline{y})$ tra $x$ ed $y$[44]. In tale contesto, possiamo indicare due v.a. come *ortogonali* se risulta $\sigma_{xy} = \rho_{xy} = 0$, mentre un valore $\rho_{xy} = \pm 1$ indica che una delle due v.a. è costantemente uguale all'altra, a meno di un coefficiente costante. Notiamo che l'ortogonalità $\rho_{xy} = 0$ esprime unicamente l'assenza di legami di tipo *lineare* tra $x$ ed $y$, come esemplificato dal caso F) della nota (2) a pag. 208.

Per ultima citiamo l'estensione formale del risultato noto come *diseguaglianza di Schwartz* (pag. 27), una volta che al coefficiente di correlazione $\rho_{xy}$ sia stato associato il concetto di coseno tra $x$ ed $y$: tale posizione deriva dall'essere $-1 < \rho_{xy} < 1$, e permette di asserire che $|\sigma_{xy}| \leq \sigma_x \sigma_y$.

### 9.9.2 Gaussiana multidimensionale ed indipendenza statistica

In fondo al § 9.1.2 si è affermato che, *unicamente nel caso di v.a. congiuntamente gaussiane*, il sussistere di incorrelazione tre le stesse ne implica l'indipendenza statistica: mostriamo qui il motivo. Iniziamo con il definire la d.d.p. di una v.a. gaussiana multidimensionale $\boldsymbol{X}$ come

$$p_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \, det \boldsymbol{\Sigma}_x}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{m}_x) \, \boldsymbol{\Sigma}^{-1}{}_x \, (\boldsymbol{x} - \boldsymbol{m}_x)^T \right\} \tag{9.10}$$

in cui $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]$ è un vettore riga che rappresenta le $n$ v.a. marginali, $\boldsymbol{m}_x$ è il vettore dei rispettivi valori medi, e $\boldsymbol{\Sigma}_x$ è la *matrice di covarianza* i cui $n \times n$ elementi risultano pari a $\sigma_{x_i, x_j} = E\left\{ (x_i - m_{x_i}) \left( x_j - m_{x_j} \right) \right\}$. Notiamo che la conoscenza di $\boldsymbol{m}_x$ e $\boldsymbol{\Sigma}_x$ definisce in modo *completo* la densità di probabilità.

Osserviamo ora che nel caso in cui le v.a. siano incorrelate, ossia $\sigma_{x_i, x_j} = 0$ con $i \neq j$, la matrice di covarianza $\boldsymbol{\Sigma}_x$ risulta essere *diagonale*, e così la sua inversa, i

[44] L'analogia non è poi troppo peregrina, considerando che se $x$ è estratta da un processo ergodico a media nulla, la sua varianza $\sigma_x^2$ coincide con la potenza del segnale da cui è estratta, mentre se $x$ ed $y$ sono estratte da segnali congiuntamente ergodici, la covarianza $\sigma_{xy}$ coincide con la funzione di intercorrelazione definita in (9.5) per segnali di energia.

cui elementi risultano in tal caso essere pari a $1/\sigma_{x_i}^2$. Sotto queste ipotesi la (9.10) si esprime come

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_{x_i}}} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(x_i - m_{x_i})^2}{\sigma_{x_i}^2}\right]\right\}$$

e dunque equivale al *prodotto* delle singole d.d.p. marginali[45]

$$p(x_i) = \frac{1}{\sqrt{2\pi}\sigma_{x_i}} \exp\left\{-\frac{1}{2}\frac{(x_i - m_{x_i})^2}{\sigma_{x_i}^2}\right\}$$

Ma dato che questo risultato è proprio la definizione di indipendenza statistica tra le v.a. marginali, abbiamo ottenuto la dimostrazione cercata.

### 9.9.3 Densità spettrale per onda PAM

L'acronimo PAM sta per *Pulse Amplitude Modulation*, e individua una classe di segnali realizzati ripetendo indefinitivamente uno stesso im*pulso* elementare $g(t)$ con periodo $T$, ognuno moltiplicato (o *modulato in ampiezza*) per un diverso coefficiente $a_n$. Entro questa generica definizione rientra sia il segnale dati, che la conversione D/A per segnali campionati, di cui in questa sezione si fornisce una trattazione unificata.

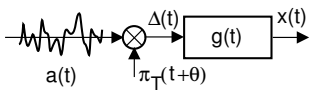
Accingiamoci pertanto a verificare quanto affermato al § 9.2.4, ossia che al segnale dati

$$x(t) = \sum a_n g(t - nT - \theta) \tag{9.11}$$

corrisponde uno spettro di densità di potenza pari a

$$\mathcal{P}_x(f) = \sigma_A^2 \frac{|G(f)|^2}{T} \tag{9.12}$$

Il segnale (9.11) rappresenta un processo stazionario ergodico qualora $\theta$ sia una v.a. aleatoria uniformemente distribuita tra $\pm\frac{T}{2}$, ed i valori $a_n$ realizzazioni di v.a. statisticamente indipendenti, a media nulla ed identicamente distribuite, e varianza $\sigma_A^2 = E\{a_n^2\}$.




Per verificare la (9.12) adottiamo il modello di generazione del segnale (9.11) rappresentato in figura: un processo gaussiano bianco a media nulla $a(t)$ è moltiplicato per un processo impulsivo $\pi_T(t-\theta) = \sum_n \delta(t - nT - \theta)$, dando luogo ad un nuovo processo impulsivo $\Delta(t) = \sum_n a_n \delta(t - nT - \theta)$ in cui i coefficienti $a_n$ sono (per costruzione) statisticamente indipendenti e quindi incorrelati, in modo che $\mathcal{R}_a(\tau) = E\{a_n a_{n+\tau}\} = \begin{cases} \sigma_A^2 & \text{per } \tau = 0 \\ 0 & \text{altrimenti} \end{cases}$, e quindi

$$\mathcal{R}_\Delta(\tau) = \mathcal{R}_\pi(\tau) \cdot \mathcal{R}_a(\tau) = \begin{cases} \sigma_A^2 \cdot \mathcal{R}_\pi(0) & \text{per } \tau = 0 \\ 0 & \text{altrimenti} \end{cases} \tag{9.13}$$

Per quanto riguarda $\mathcal{R}_\pi(\tau)$, osserviamo che

- deve risultare periodica con periodo $T$, come lo è $\pi_T(t)$;

[45] Si verifichi per esercizio che nel caso di una coppia di v.a. congiuntamente gaussiane, a media nulla ed uguale varianza, si ottiene l'espressione (7.15) di pag. 156.

- gli estremi dell'integrale che ne definisce il valore possono limitarsi[46] alla durata di un periodo;
- l'ergodicità consente di eseguire il calcolo su di una particolare realizzazione di $\theta$, ad es. pari a zero

permettendo di scrivere

$$\begin{aligned}\mathcal{R}_\pi(\tau) &= \frac{1}{T}\int_{-T/2}^{T/2}\sum_n \delta(t-nT)\sum_m \delta(t-mT+\tau)\,dt \\ &= \frac{1}{T}\int_{-T/2}^{T/2}\delta(t)\,\delta(t+\tau)\,dt\end{aligned}$$

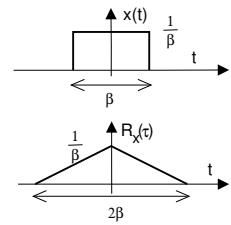
in quanto gli altri termini delle sommatorie cadono al difuori dell'intervallo di integrazione.

Notiamo ora che $\mathcal{R}_\pi(\tau)$ risulta pari a zero se $\tau \neq 0$, mentre per valutare $\mathcal{R}_\pi(0) = \frac{1}{T}\int_{-T/2}^{T/2}[\delta(t)]^2\,dt$ conviene riscrivere $\delta(t)$ in base alla sua definizione, ossia il limite a cui tende una successione di funzioni; ponendo (ad es.) $\delta(t) = \lim_{\beta\to 0}\frac{1}{\beta}\mathrm{rect}_\beta(t)$, si ottiene[47] (per $-\frac{T}{2} \le \tau < \frac{T}{2}$)

$$\begin{aligned}\mathcal{R}_\pi(\tau) &= \lim_{\beta\to 0}\frac{1}{T}\int_{-T/2}^{T/2}\frac{1}{\beta^2}\mathrm{rect}_\beta(t)\,\mathrm{rect}_\beta(t+\tau)\,dt = \\ &= \lim_{\beta\to 0}\frac{1}{T}\frac{1}{\beta}\mathrm{tri}_{2\beta}(\tau) = \frac{1}{T}\delta(\tau)\end{aligned}$$

e quindi in definitiva, dovendo risultare $\mathcal{R}_\pi(\tau)$ periodico, si ottiene

$$\mathcal{R}_\pi(\tau) = \frac{1}{T}\sum_{n=-\infty}^{\infty}\delta(\tau - nT) \qquad (9.14)$$




Sostituendo ora (9.14) in (9.13), otteniamo in definitiva $\mathcal{R}_\Delta(\tau) = \sigma_A^2\frac{\delta(\tau)}{T}$ e $\mathcal{P}_\Delta(f) = \mathcal{F}\{\mathcal{R}_\Delta(\tau)\} = \frac{\sigma_A^2}{T}$. Pertanto, si è dimostrato che l'onda PAM $x(t)$ ha densità di potenza

$$\mathcal{P}_x(f) = \mathcal{P}_\Delta(f)\,|G(f)|^2 = \sigma_A^2\frac{|G(f)|^2}{T} \qquad (9.15)$$

Mediante una diversa trattazione[48] si mostra che, nel caso in cui il valor medio $m_A = E\{a_n\}$ dei valori $a_n$ non sia nullo, risulta

$$\mathcal{P}_x(f) = \sigma_A^2\frac{|G(f)|^2}{T} + \left(\frac{m_A}{T}\right)^2\sum_{n=-\infty}^{\infty}\left|G\left(\frac{n}{T}\right)\right|^2\delta\left(t-\frac{n}{T}\right)$$

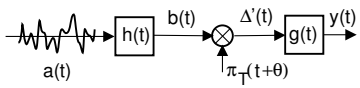
che, rispetto alla (9.15), evidenzia l'aggiunta di un termine *a righe*.

[46] La definizione di autocorrelazione prescrive il calcolo $\mathcal{R}_y(\tau) = \lim_{\alpha\to\infty}\frac{1}{\alpha}\int_{-\alpha/2}^{\alpha/2}y(t)\,y(t+\tau)\,dt$ ma, se $y(t)$ è periodico, il valore dell'integrale su di un intervallo pari ad un periodo fornisce lo stesso valore per qualunque possibile traslazione temporale di $y(t)$

[47] E' facile verificare dalla figura mostrata che sia $\frac{1}{\beta}\mathrm{rect}_\beta(t)$ che $\frac{1}{\beta}\mathrm{tri}_{2\beta}(\tau)$ possiedono area unitaria, e dunque per $\beta \to 0$ convergono ad un impulso.

[48] S. Barbarossa, `http://infocom.uniroma1.it/sergio/PAM.pdf`

Infine, prendiamo in considerazione il caso in cui ai simboli $a_n$ si sostituiscano dei $b_n$ ancora a media nulla, ma *non più indipendenti* tra loro, ovvero la cui autocorrelazione $\mathcal{R}_b(\tau) = E\{b_n b_{n+\tau}\}$ sia diversa da zero per $\tau = nT$. Adottiamo ora il nuovo schema funzionale disegnato appresso per realizzare un modello di ciò che accade.




Mentre $a(t)$ è lo stesso di prima, il suo passaggio attraverso il filtro $h(t)$ introduce una correlazione tra i valori estratti da $b(t)$ mediante moltiplicazione per $\pi_T(t+\theta)$, ottenendo infatti $\mathcal{R}_b(\tau) = \mathcal{R}_H(\tau) * \mathcal{R}_a(\tau)$[49]; indicando ora l'ingresso al filtro $g(t)$ come $\Delta'(t) = \sum b_n \delta(t - nT + \theta)$, si ottiene

$$\mathcal{R}_{\Delta'}(\tau) = \mathcal{R}_\pi(\tau) \cdot \mathcal{R}_b(\tau) = \sigma_A^2 \cdot \mathcal{R}_\pi(\tau) \cdot \mathcal{R}_H(\tau)$$

e dunque[50]

$$\begin{aligned} \mathcal{P}_{\Delta'}(f) &= \mathcal{F}\{\mathcal{R}_{\Delta'}(\tau)\} = \sigma_A^2 \cdot \mathcal{P}_\pi(f) * |H(f)|^2 = \\ &= \frac{\sigma_A^2}{T^2} \cdot \sum_n \delta\left(f - \frac{n}{T}\right) * |H(f)|^2 = \frac{\sigma_A^2}{T^2} * \sum_n \left|H\left(f - \frac{n}{T}\right)\right|^2 \end{aligned}$$

ma dato che $\mathcal{P}_y(f) = |G(f)|^2 \mathcal{P}_{\Delta'}(f)$, in definitiva si ottiene

$$\mathcal{P}_y(f) = \frac{\sigma_A^2}{T^2} |G(f)|^2 \sum_n \left|H\left(f - \frac{n}{T}\right)\right|^2 \tag{9.16}$$

ovvero la densità spettrale dell'onda PAM viene a dipendere da quella dei dati trasportati. Notiamo che esiste un caso particolare in cui ciò non avviene, e si verifica se $\sum_n \left|H\left(f - \frac{n}{T}\right)\right|^2 = cost$, condizione simile a quella (5.4) di Nyquist per l'assenza di ISI. Seguendo gli sviluppi analitici svolti in quel caso, otteniamo che la (9.16) può non dipendere da $H(f)$ nel caso in cui

$$\mathcal{F}^{-1}\left\{\left|H\left(f - \frac{n}{T}\right)\right|^2\right\} = \mathcal{R}_H(\tau) = \begin{cases} 1/T & \text{per } \tau = 0 \\ 0 & \text{per } \tau = nT \\ \forall & \text{altrimenti} \end{cases}$$

che corrisponde alla incorrelazione tra i valori $b$ presi a distanza multipla di $T$.

Osserviamo infine che la (9.16) costituisce una diversa derivazione della (4.2) di pag. 54, nel caso in cui i $b_n$ siano campioni di segnale.

### 9.9.4 Potenza di un segnale dati

Al § 7.5.3.2 si è affermato che la potenza di un segnale dati

$$s(t) = \sum_n a_n g(t - nT)$$

[49] Più precisamente, l'autocorrelazione tra simboli $b_n$ è valutabile solo per distanze temporali multiple del periodo di simbolo, ed in virtù della presenza di $h(t)$, risulta $\mathcal{R}_b((n-m)T) = E\{b_n b_m\} = \sigma_A^2 \mathcal{R}_H((n-m)T)$.

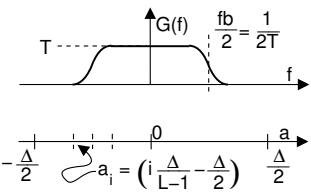
[50] Nell'ultima serie di passaggi si può valutare $\mathcal{P}_\pi(f)$ come $\frac{1}{T^2} \sum_n \delta\left(f - \frac{n}{T}\right)$, in quanto $\Pi(f) = \mathcal{F}\{\pi(t)\} = \mathcal{F}\left\{\sum_n \delta(t - nT)\right\} = \frac{1}{T} \sum_n \delta\left(f - \frac{n}{T}\right)$ e quindi $\mathcal{P}_\pi(f) = |\Pi(f)|^2$, oppure applicando la relazione $\mathcal{P}_\pi(f) = \mathcal{F}\{\mathcal{R}_\pi(\tau)\}$ a partire dalla (9.14).

in cui $g(t)$ è una caratteristica di Nyquist a coseno rialzato con roll-off $\gamma$, e gli $a_i$ sono una sequenza di v.a. discrete, statisticamente indipendenti, a media nulla ed uniformemente distribuite su L livelli in una dinamica $-\frac{\Delta}{2} \leq a_i \leq \frac{\Delta}{2}$, ha valore

$$\mathcal{P}_s = \frac{\Delta^2}{12}\frac{L+1}{L-1}\left(1-\frac{\gamma}{4}\right)$$

Mostriamo ora i passi necessari ad arrivare a questo risultato. Nella precedente appendice, si è mostrato che per lo stesso segnale risulta $\mathcal{P}_s(f) = \sigma_A^2 \frac{|G(f)|^2}{T}$, e dunque

$$\mathcal{P}_s = \int \mathcal{P}_s(f)\,df = \int \sigma_A^2 \frac{|G(f)|^2}{T} df$$




Svolgendo i relativi calcoli, si può mostrare che $\int |G(f)|^2\,df = T\left(1-\frac{\gamma}{4}\right)$, e quindi $\mathcal{P}_s = \sigma_A^2\left(1-\frac{\gamma}{4}\right)$; resta pertanto da calcolare $\sigma_A^2$:

$$\begin{aligned}
&\sigma_A^2 = (a_i \text{ a media nulla}) = E_X\left\{a_i^2\right\} = \textstyle\sum_{i=0}^{L-1} p_{a_i}\cdot a_i^2 = (^{51}) = \\
&= \tfrac{1}{L}\textstyle\sum_{i=0}^{L-1}\left(i\frac{\Delta}{L-1}-\frac{\Delta}{2}\right)^2 = \tfrac{\Delta^2}{L}\textstyle\sum_{i=0}^{L-1}\left(\frac{i^2}{(L-1)^2}+\frac{1}{4}-\frac{i}{L-1}\right) = \\
&= \tfrac{\Delta^2}{L}\left(\tfrac{1}{(L-1)^2}\textstyle\sum_{i=0}^{L-1}(i)^2+\tfrac{L}{4}-\tfrac{1}{L-1}\textstyle\sum_{i=0}^{L-1} i\right) = (^{52}) = \\
&= \tfrac{\Delta^2}{L}\left(\tfrac{L}{4}-\tfrac{1}{L-1}\tfrac{L(L-1)}{2}+\tfrac{1}{(L-1)^2}\tfrac{(L-1)L(2(L-1)+1)}{6}\right) = \Delta^2\left(\tfrac{1}{4}-\tfrac{1}{2}+\tfrac{2L-2+1}{6(L-1)}\right) = \\
&= \Delta^2\tfrac{6L-6-12L+12+8L-8+4}{24(L-1)} = \Delta^2\tfrac{2L+2}{24(L-1)} = \tfrac{\Delta^2}{12}\tfrac{L+1}{L-1}.
\end{aligned}$$

### 9.9.5 Autocorrelazione dell'uscita di un filtro

Al § 9.4.3 si è affermato che, quando un processo attraversa un filtro, il processo di uscita è caratterizzato da $\mathcal{R}_y(\tau) = \mathcal{R}_x(\tau) * \mathcal{R}_h(\tau)$. Mostriamo che è vero.

$$\begin{aligned}
\mathcal{R}_y(t,t+\tau) &= E\{y(t)\,y(t+\tau)\} = \\
&= E\left\{\int h(\alpha)\,x(t-\alpha)\,d\alpha\int h(\beta)\,x(t+\tau-\beta)\,d\beta\right\} = \\
&= \int\int h(\alpha)\,h(\beta)\,E\{x(t-\alpha)\,x(t+\tau-\beta)\}\,d\alpha d\beta = \\
&= \int h(\alpha)\int h(\beta)\,\mathcal{R}_x(\tau+\alpha-\beta)\,d\beta d\alpha = \\
&= \int h(\alpha)\,\mathcal{R}_{xy}(\tau+\alpha)\,d\alpha = \mathcal{R}_{xy}(\tau) * h(-\tau)
\end{aligned}$$

nel cui terz'ultimo passaggio si è assunto che $x(t)$ sia stazionario, e $\mathcal{R}_{xy}(\tau) = \mathcal{R}_x(\tau) * h(\tau)$ è l'intercorrelazione tra $x(t)$ ed $y(t)$[53].

[51] consideriamo valori $a_i$ equiprobabili, ovvero $p(a_i) = 1/L$, e con dinamica bilanciata attorno allo zero, ossia valori compresi tra $-\Delta/2$ e $\Delta/2$

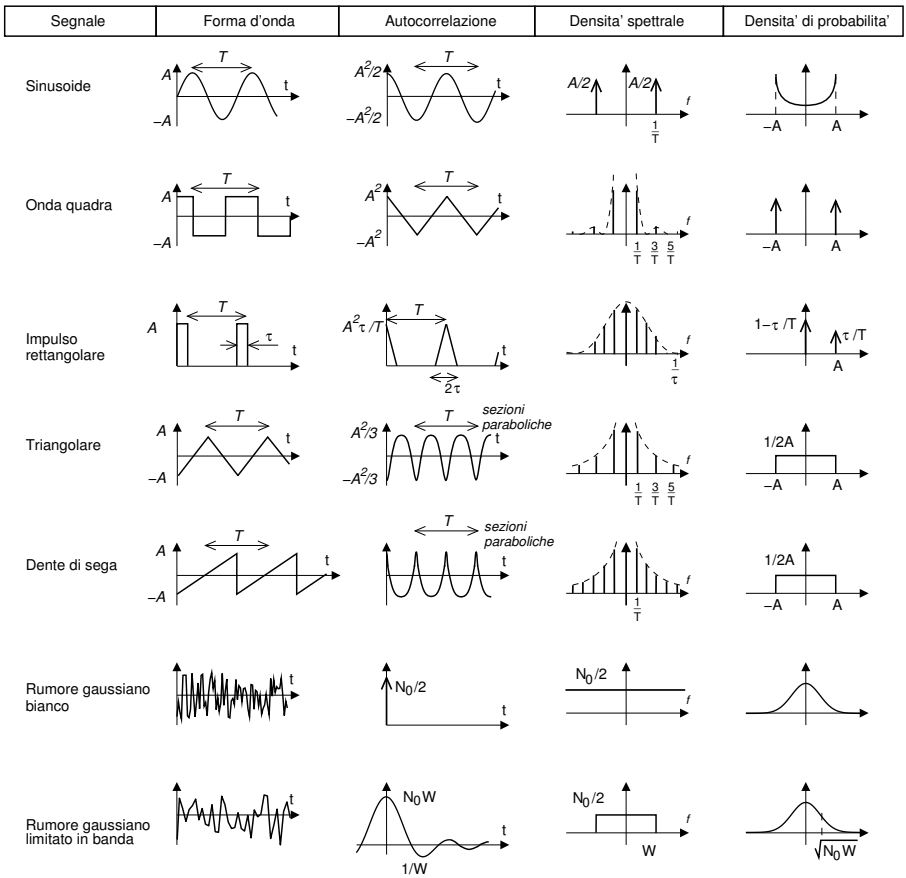
[52] Faccciamo uso delle relazioni $\sum_{n=1}^{N} n = \frac{N(N+1)}{2}$ e $\sum_{n=1}^{N} n^2 = \frac{N(N+1)(2N+1)}{6}$

[53] Infatti, $\mathcal{R}_x(\tau) * h(\tau) = x^*(-\tau) * x(\tau) * h(\tau) = x^*(-\tau) * y(\tau)$, che è appunto la definizione di $\mathcal{R}_{xy}(\tau)$

Scritto in altra forma: $\mathcal{R}_y(\tau) = \mathcal{R}_x(\tau) * h(\tau) * h(-\tau)$, e dunque antitrasformando si ottiene $\mathcal{P}_y(f) = \mathcal{P}_x(f) \cdot H(f) \cdot H^*(f) = \mathcal{P}_x(f) \cdot |H(f)|^2 = \mathcal{F}\{\mathcal{R}_x(\tau) * \mathcal{R}_h(\tau)\}$.

### 9.9.6 Grafici di esempio

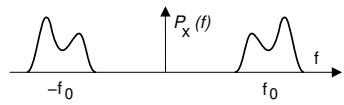
Sono riportati appresso i grafici di forma d'onda, dell'autocorrelazione, della densità spettrale e densità di probabilità, per alcuni segnali tipici.



# Capitolo 10

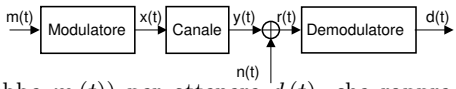
# Segnali modulati

## 10.1 Caratteristiche ed applicazioni




I segnali *modulati*, di cui ci occuperemo ora, sono anche indicati come segnali in *banda traslata*, per la particolarità di esibire una occupazione spettrale limitata ad una banda ristretta attorno ad una frequenza $f_0$ chiamata *portante*.

Questi segnali sono prodotti mediante un *processo di modulazione*, che consiste nell'operare una trasformazione su di un segnale $m(t)$ *modulante* in modo da ottenerne un secondo $x(t)$ *modulato,* il cui contenuto spettrale è ora idoneo alla trasmissione mediante il canale a disposizione.




Il segnale *ricevuto* $r(t)$ (a cui è sovrapposto un processo di rumore $n(t)$) deve quindi essere *demodulato* (operazione che, se applicata ad $x(t)$, restituirebbe $m(t)$) per ottenere $d(t)$, che rappresenta il segnale trasmesso, più eventuali distorsioni $\varepsilon(t)$. Per evidenziare la situazione, scriviamo il segnale demodulato come $d(t) = m(t) + \varepsilon(t)$, in cui $\varepsilon(t) = Dem\{n(t)\} + Dem\{y(t) - x(t)\}$: evidentemente, le distorsioni hanno origine sia dal risultato della demodulazione del rumore in ingresso al demodulatore, sia dagli effetti che la demodulazione ha sulle alterazioni introdotte dal canale sul segnale in transito.

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

1. il canale non permette la trasmissione di frequenze contigue all'origine
2. il canale presenta un comportamento ideale (modulo costante e fase lineare) solo in determinati intervalli di frequenza
3. il canale presenta disturbi additivi solo in determinate regioni di frequenza.

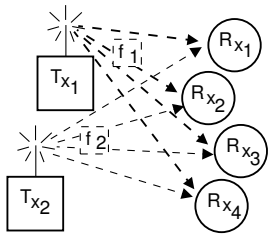
### 10.1.1 Multiplazione a divisione di frequenza - FDM

Come intuibile dal nome, si tratta di una tecnica di trasmissione in cui più comunicazioni contemporanee avvengono condividendo lo stesso mezzo fisico, ed impegnando ognuna una diversa regione di frequenze, per il semplice motivo che se utilizzassero

tutte le medesima banda, costituirebbero termini di interferenza reciproca. Molto spesso tutti i segnali multiplati sono di natura simile, ed ognuno è il risultato di una modulazione operata con una diversa frequenza portante. Portiamo ad esempio tre casi:

#### 10.1.1.1 Collegamenti punto-multipunto

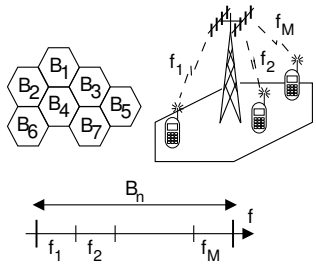
E' una topologia che si verifica ad esempio nel caso di trasmissioni televisive o radiofoniche (dette trasmissioni *broadcast*), in cui ogni emittente (in figura indicata come $T_x$) trasmette a tutti i ricevitori ($R_x$) sintonizzati sulla propria portante (i famosi "canali" della TV), mentre altre emittenti utilizzano contemporaneamente lo stesso mezzo trasmissivo, occupando canali centrati ad altre frequenze.




#### 10.1.1.2 Accesso multiplo

E' la problematica tipica delle comunicazioni mobili, e quindi dei "telefonini". In tal caso, il territorio è suddiviso in *celle*, per ognuna delle quali è definita una regione di radiofrequenze ($B_n$) dedicata alla comunicazione tra i terminali ed un unico ripetitore. All'interno della cella, la banda a disposizione è suddivisa tra più canali, ognuno associato ad una diversa portante ($f_i$), che vengono usate a turno dai terminali che desiderano comunicare[1].

Sotto certi aspetti, questo caso è in qualche modo antitetico rispetto al precedente, e potrebbe essere indicato come collegamento *multipunto-punto*. In realtà la situazione è un pò più complessa, e gli aspetti qualificanti da un punto di vista sistemistico sono i protocolli di rete, necessari per consentire le fasi di richiesta di accesso, la localizzazione dei radiomobili, e la corretta gestione dell'*handover* (il cambio di cella)[2].




[1] Senza entrare nei dettagli, specifichiamo semplicemente che celle limitrofe adottano regioni di frequenza differenti, onde evitare interferenze tra celle; inoltre, nell'ambito di uno stesso canale, è realizzata una struttura di trama, in modo da permettere l'utilizzo dello stesso canale da parte di più terminali contemporaneamente, multiplati a divisione di tempo.

[2] Un minimo di approfondimento però ci sta bene... Aggiungiamo quindi che la scelta del canale su cui comunicare avviene in base alle condizioni di ricezione del singolo radiomobile che, per effetto di cammini multipli del segnale ricevuto, può ricevere meglio certe portanti che non altre.

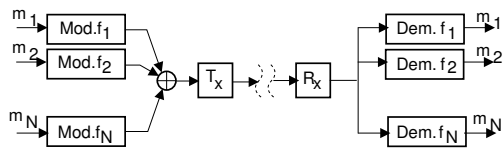
La trasmissione che ha luogo su di una portante, inoltre, può aver origine da più terminali, che si ripartiscono la medesima banda a divisione di tempo, in accordo ad una suddivisione di trama dell'asse dei tempi. Pertanto, dopo che un terminale si è aggiudicato una portante ed un intervallo temporale, la trasmissione (attuata mediante una modulazione numerica) ha luogo solo per brevi periodi, in corrispondenza del time-slot di propria pertinenza.

Dato che i singoli terminali si trovano a distanze diverse dal ripetitore di cella, diversi sono i tempi di propagazione del segnale di sincronismo di trama e di time-slot, e dunque l'intervallo temporale che viene "riempito" da ogni terminale giunge al ripetitore con un ritardo variabile. Per questo motivo, i time-slot della trama sono separati da piccoli periodi di inattività, chiamati *intervalli di guardia*, che garantiscono l'assenza di sovrapposizioni temporali delle trasmissioni originate dai diversi terminali.

#### 10.1.1.3 Collegamenti punto-punto

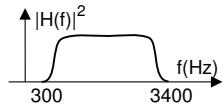
E' una forma di multiplazione più simile a quella già illustrata nel caso della TDM, in cui il collegamento avviene tra due località ben determinate. Un insieme di $N$ segnali $m_i$, $i = 1, 2, ..., N$, transita su di uno stesso mezzo trasmissivo, occupando ognuno una differente banda, centrata su di 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*.

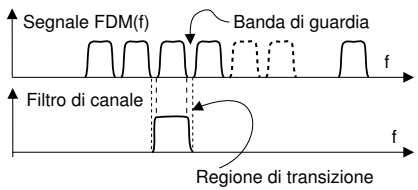



### 10.1.2 Canale telefonico

Le caratteristiche del collegamento offerto dalla comune linea telefonica tengono conto di molteplici aspetti. Uno di questi, forse il principale[3], è la limitazione della banda del canale, per cui la trasmissione è garantita solo in un intervallo di frequenze comprese tra i 300 ed i 3400 Hz, mentre la banda nominale (ovvero l'occupazione di banda in una trasmissione FDM) risulta essere di 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.




Per lungo tempo, il traffico telefonico è stato multiplato su collegamenti FDM punto-punto, con i singoli canali modulati AM-BLU (vedi § 11.1.2), dovendo pertanto rimuovere le componenti frequenziali più basse. Inoltre, la necessità di separare tra loro i canali multiplati FDM mediante i *filtri di canale* che, per essere economicamente realizzabili, devono presentare una regione di transizione di estensione apprezzabile, ha determinato l'esigenza di prevedere tra due canali contigui un intervallo di frequenze detto *banda di guardia* (pari a 900 Hz) che determina la limitazione a 3400 Hz per la massima frequenza di segnale,




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

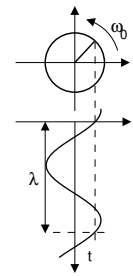
[4] Questo valore massimo nominale determina che la frequenza di campionamento del PCM telefonico è pari a 2*4000 = 8000 campioni al secondo. Utilizzando 8 bit/campione, si ottiene la velocità binaria $f_b = 64000$ campioni/secondo. Velocità inferiori si possono conseguire adottando metodi di codifica di sorgente per il segnale vocale.

[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 § 6.9.1). Nel caso di una linea ISDN, invece, il telefono stesso effettua la conversione numerica, ed i campioni di voce viaggiano nei due sensi (tra utente e centrale) secondo uno schema a divisione di tempo (vedi § 6.9.2).

in modo da ottenere $300 + (4000 - 3400) = 900$ Hz. In caso contrario infatti, all'uscita del filtro di canale si troverebbe anche parte del segnale di un canale contiguo producendo una interferenza tra comunicazioni diverse.

La limitazione in banda di un canale telefonico tra 300 e 3400 Hz è dunque uno dei motivi per i quali (ai tempi della telefonia analogica) la connessione telefonica di un computer ad un fornitore di connettività numerica (ad es. un provider Internet) richiede l'uso di un dispositivo (il modem) che effettui una forma di modulazione sul segnale da trasmettere. Le cose sono notevolmente cambiate nel caso dell'accesso ADSL (vedi § 6.9.4), in cui la connettività numerica inizia direttamente nella centrale del chiamante. Allora, il modem ADSL occupa una banda *disgiunta* da quella del canale telefonico, usando la capacità del doppino che è ad uso esclusivo dell'utente.

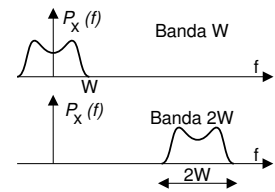
### 10.1.3 Antenne e lunghezza d'onda




La trasmissione di un segnale via onda radio necessita di una antenna di dimensioni comparabili con quelle della lunghezza d'onda. Quest'ultima quantità (indicata con $\lambda$) è pari allo spazio percorso dall'onda in un tempo pari ad un periodo: dato che *spazio* = *velocità* · *tempo*, e considerando che le onde elettromagnetiche si propagano alla velocità della luce ($c = 3 \cdot 10^8$ m/s), si ha $\lambda = c \cdot T = \frac{c}{f}$.

Nel caso di segnali modulati, il valore di $f$ è quello della portante, in quanto in genere il segnale modulato occupa una banda ristretta attorno alla portante. Trasmissioni con frequenze più elevate necessitano di antenne di dimensioni ridotte; 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 ![6]

### 10.1.4 Banda di segnale




La banda occupata da un segnale è la regione di frequenze al di fuori della quale non vi sono componenti energetiche; la sua misura in Hz è indicata come *larghezza di banda*. Per segnali reali l'occupazione di banda è espressa in termini del solo contenuto a frequenze positive; dato che in tal caso lo spettro di potenza è una funzione pari di $f$, la banda totale è doppia. Tale definizione è pertanto non ambigua, ed in accordo alla comune accezione di frequenza (positiva); pertanto, viene spesso indicata come *banda a frequenze positive*[7].

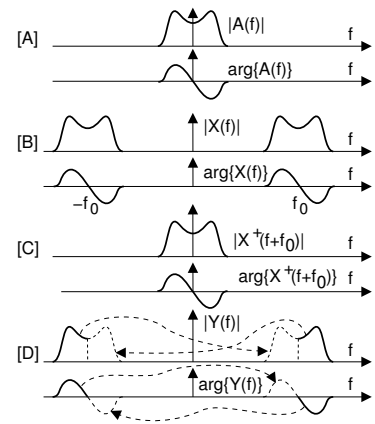
[6] Antenne più corte hanno una efficienza ridotta, ma sono ancora buone. Altrimenti la radio AM (540 - 1600 KHz) avrebbe bisogno di $\frac{3 \cdot 10^8}{1000 \cdot 10^3} = 300$ metri ! Al § 15.5.3 è riportata una tabella dei valori di $\lambda$ per i diversi servizi di TLC.

[7] Se un segnale è strettamente limitato in banda, deve avere durata infinita, e viceversa. E' pratica comune, invece, parlare di limitazione in banda anche per segnali di durata finita. Nel fare questo, si considera un $X(f)$ pari a zero per le frequenze $f$ tali che $|X(f)| < \varepsilon$, ovvero considerare anziché $X(f)$ a banda illimitata, una sua finestra *in frequenza* $X_W(f) = X(f) W(f)$ a banda limitata, la cui antitrasformata $x_W(t)$ è diversa da $x(t)$ (sappiamo infatti che si ha $x_W(t) = x(t) * w(t)$), ma ne costituisce una approssimazione.

### 10.1.5 Trasmissione a banda laterale unica

Consideriamo un segnale $a(t)$ reale e limitato in banda, con $A(f) = A^*(-f)$ (grafico [A]), a simmetria coniugata. In virtù delle proprietà di simmetria coniugata per segnali reali, la conoscenza del solo contenuto a frequenze positive $f > 0$, ovvero di $A^+(f) = A(f)\, rect_W\left(f - \frac{W}{2}\right)$, è sufficiente a definire $a(t)$ in modo completo. Se definiamo $x(t) = a(t)\cos\omega_0 t$, anch'esso reale, otteniamo che $X(f)$ [B], oltre ad essere a simmetria coniugata rispetto all'origine, ha simmetria coniugata anche rispetto ad $f_0$: $X^+(f+f_0) = \left\{X^+(-f+f_0)\right\}^*$ [C].

Questo risultato mostra come sia teoricamente possibile (con una fotocopiatrice ed un paio di forbici!) ottenere il segnale $X(f)$ a partire da un $Y(f)$ [D], con $Y(f)$ ottenuta da $X(f)$ eliminandone metà banda. 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 l'esattezza del procedimento illustrato, che ci consente di ricevere per intero $X(f)$ trasmettendone solo metà (cioè $Y(f)$), osserviamo che anche $Y(f)$ è a simmetria coniugata rispetto a zero (ossia $Y(f) = Y^*(-f)$), e quindi la sua antitrasformata $y(t)$ è reale, e dunque può essere realmente trasmesso.

A parte il "dettaglio" di come ricostruire "veramente" $X(f)$ a partire da $Y(f)$, ci chiediamo: esiste una formula per ottenere $y(t)$ in modo *diretto* a partire da $a(t)$? La risposta è positiva; per provarla occorre però affrontare alcune pagine di teoria, che illustrano un metodo di rappresentazione (nel dominio del tempo) per segnali modulati.

## 10.2 Rappresentazione dei segnali modulati

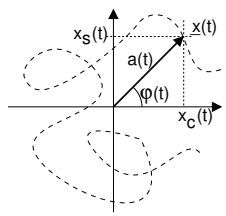
Questa sezione è dedicata alla teoria che permette di scrivere un qualunque segnale $x(t)$ nella forma

$$x(t) = x_c(t)\cos 2\pi f_0 t - x_s(t)\sin 2\pi f_0 t \tag{10.1}$$

che assume una particolare rilevanza nel caso in cui $x(t)$ sia un segnale modulato attorno ad $f_0$, perchè allora $x_c(t)$ e $x_s(t)$ sono segnali limitati in banda con banda contigua all'origine, e le alterazioni prodotte sul segnale modulato, compresa l'estrazione del messaggio modulante $m(t)$, possono essere descritte mediante operazioni condotte su $x_c(t)$ ed $x_s(t)$.

### 10.2.1 Inviluppo complesso

Introduciamo l'argomento ricordando (vedi § 2.1.3) come un segnale $x(t) = a\cos(\omega_0 t + \varphi)$ può essere rappresentato per mezzo del fasore $\underline{x} = ae^{j\varphi}$, mediante la relazione $x(t) = \Re\left\{\underline{x}e^{j\omega_0 t}\right\}$. Estendiamo ora il concetto, definendo l'*inviluppo complesso* $\underline{x}(t)$ come un fasore per il quale il modulo $a$ e la



fase $\varphi$ siano funzioni del tempo

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

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

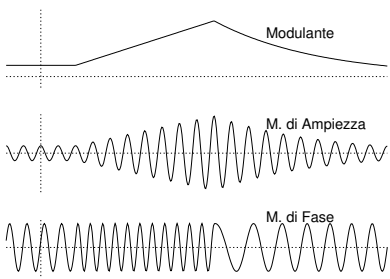
$$x(t) = \Re\left\{\underline{x}(t)\, e^{j\omega_0 t}\right\} = a(t)\cos(\omega_0 t + \varphi(t)) \tag{10.2}$$

corrispondente ad imprimere al piano dell'inviluppo complesso una rotazione antioraria a velocità angolare $\omega_0$. D'altra parte, indicando con $x_c(t) = a(t)\cos\varphi(t)$ ed $x_s(t) = a(t)\sin\varphi(t)$ la parte reale ed immaginaria dell'inviluppo complesso $\underline{x}(t)$[8], la (10.2) è equivalente alla (10.1), dato che[9]

$$\begin{aligned} x(t) &= \Re\left\{\underline{x}(t)\, e^{j\omega_0 t}\right\} = a(t)\cos(\omega_0 t + \varphi(t)) = \\ &= a(t)\left[\cos\omega_0 t\cos\varphi(t) - \sin\omega_0 t\sin\varphi(t)\right] \\ &= x_c(t)\cos 2\pi f_0 t - x_s(t)\sin 2\pi f_0 t \end{aligned}$$

### 10.2.2 Modulazione di ampiezza e/o angolare

L'inviluppo complesso è un potente strumento che permette di descrivere il processo di modulazione in modo semplice ed omogeneo. Ad esempio, il caso (già noto) di traslazione in frequenza del segnale $a(t)$ mediante moltiplicazione per un coseno, corrisponde ad un inviluppo complesso $\underline{x}(t) = a(t)$ a fase nulla: ad esso si dà il nome di *modulazione di ampiezza*[10] per (l'evidente) ragione che l'ampiezza del coseno varia in funzione del segnale $a(t)$; la frequenza $f_0 = \frac{\omega_0}{2\pi}$ prende il nome di *frequenza portante*. Se al contrario consideriamo un inviluppo complesso con modulo costante $\underline{x}(t) = e^{j\varphi(t)}$, l'andamento della fase $\varphi(t)$ imprime alla portante non modulata un diverso tipo di modulazione, detto *modulazione di fase*[11] o *angolare* in quanto il segnale modulante ($\varphi(t)$ in questo caso) altera l'argomento del coseno.




Prima di proseguire, riflettiamo sul risultato mostrato alla figura a lato, in cui si considera un segnale modulante $m(t)$ prima costante, poi a rampa lineare, e quindi decrescente. Ponendo $\underline{x}(t) = m(t)$ si ottiene una portante modulata in ampiezza, mentre con $\underline{x}(t) = e^{jm(t)}$ la portante modulata 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 $2\pi f_0 t + \alpha t = 2\pi\left(f_0 + \frac{\alpha}{2\pi}\right)t$.

[8] $x_c(t)$ e $x_s(t)$ si ottengono a partire dalla rappresentazione polare $\underline{x}(t) = a(t)\, e^{j\varphi(t)}$ di $\underline{x}(t)$, semplicemente sviluppando la stessa come $\underline{x}(t) = a(t)\, e^{j\varphi(t)} = a(t)\cos\varphi(t) + ja(t)\sin\varphi(t) = x_c(t) + jx_s(t)$

[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*).

Per meglio descrivere il caso di modulazione angolare, definiamo una *fase istantanea*

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

ed una *frequenza istantanea*

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

In questi termini, la modulazione angolare viene distinta in *modulazione di fase* propriamente detta quando

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

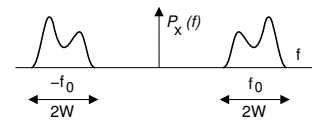
mentre viene detta *modulazione di frequenza* quando

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

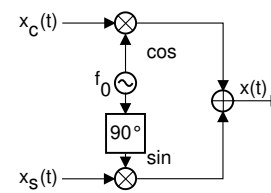
in quanto in questo caso è la frequenza istantanea a dipendere direttamente dal segnale modulante: $f_i(t) = f_0 + k_f m(t)$.

### 10.2.3 Componenti analogiche di bassa frequenza

Abbiamo mostrato come un generico segnale modulato $x(t)$ possa essere rappresentato per mezzo di un diverso segnale chiamato *inviluppo complesso* $\underline{x}(t) = a(t)\, e^{j\varphi(t)}$, le cui parti reale ed immaginaria $x_c(t) = a(t)\cos\varphi(t)$ ed $x_s(t) = a(t)\sin\varphi(t)$, che prendono il nome di *componenti analogiche di bassa frequenza* di $x(t)$ per un motivo presto chiaro, permettono di ri-scrivere $x(t)$ come $x(t) = x_c(t)\cos 2\pi f_0 t - x_s(t)\sin 2\pi f_0 t$.




Mentre il risultato ottenuto è valido per un qualunque segnale, esso riveste una importanza particolare nel caso in cui $x(t)$ sia di tipo limitato in banda con banda $2W$ centrata attorno ad $f_0$, con $W < f_0$: in tal caso infatti, sia $x_c(t)$ che $x_s(t)$ risultano *limitate in banda tra* $\pm W$ *e contigue all'origine.* Che sia vero anche il viceversa, può essere verificato in modo intuitivo, partendo da $x_c(t)$ e $x_s(t)$ limitate in banda, e moltiplicandole per coseno e seno.




L'ultima osservazione ci mostra una via per *sintetizzare* un segnale modulato in ampiezza, od angolarmente, od entrambe le cose, mediante il semplice schema circuitale disegnato a fianco, che si basa sulla conoscenza delle componenti analogiche di bassa frequenza, che a loro volta sono ottenibili a partire da $a(t)$ e $\varphi(t)$. Restano comunque (per ora) aperti i seguenti problemi:

- Noto $x(t)$, come ottenere $x_c(t)$ ed $x_s(t)$ ?
- Noto lo spettro di densità di potenza $\mathcal{P}_x(f)$, che dire di $\mathcal{P}_{\underline{x}}(f)$ ?

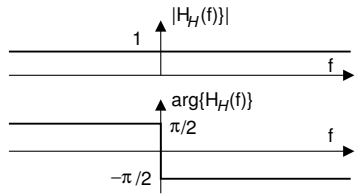
Mentre la prima domanda può trovare una risposta di tipo operativo nella *demodulazione omodina* (vedi § 10.3.3), dal punto di vista formale occorre prima definire il *segnale analitico* (vedi § 10.2.5), e prima ancora, definire il...

### 10.2.4 Filtro di Hilbert

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

$$H_{\mathcal{H}}(f) = -j \cdot \operatorname{sgn}(f) \tag{10.3}$$

ed il cui andamento è rappresentato nella figura a lato, che permette di evidenziare l'andamento costante del modulo $|H_{\mathcal{H}}(f)| = 1$, e quello discontinuo della fase $\angle H_{\mathcal{H}}(f)$, che passa da $\frac{\pi}{2}$ per $f < 0$ a $-\frac{\pi}{2}$ per $f > 0$.




Il risultato del passaggio di un segnale $x(t)$ attraverso[12] il filtro di Hilbert è un un secondo segnale, identificato come la *trasformata di Hilbert* del primo, indicato come $\widehat{x}(t) = \mathcal{H}\{x(t)\}$, ed il cui andamento in frequenza ha espressione

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

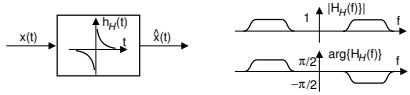
ossia differisce da $X(f)$ per uno sfasamento di $\mp\frac{\pi}{2}$ per frequenze rispettivamente positive o negative.

#### 10.2.4.1 Estrazione delle c.a. di b.f.

Senza soffermarci ora sulle proprietà[13] della trasformata di Hilbert, introduciamo direttamente il risultato[14]

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

[12] L'antitrasformata di Fourier di $H_{\mathcal{H}}(f)$ è calcolata al § 10.5.1, e fornisce l'espressione della risposta impulsiva del filtro di Hilbert $h_{\mathcal{H}}(t) = \mathcal{F}^{-1}\{H_{\mathcal{H}}(f)\} = \frac{1}{\pi t}$, permettendo di scrivere la trasformata di Hilbert nella forma di un integrale di convoluzione: $\widehat{x}(t) = \mathcal{H}\{x(t)\} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau = x(t) * \frac{1}{\pi t}$.




La realizzazione di un filtro di Hilbert con una risposta in frequenza esattamente descritta dalla (10.3) può risultare molto ardua, a causa della brusca transizione della fase in corrispondenza di $f = 0$. In realtà, il filtro di Hilbert si usa principalmente per segnali modulati, che non presentano componenti spettrali a frequenze prossime allo zero. Pertanto, lo stesso scopo può essere svolto da un diverso filtro $\widehat{H}_{\mathcal{H}}(f)$, 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.

[13] Accenniamo brevemente alle principali proprietà della trasformata di Hilbert:

- $\mathcal{H}\{x(t) = x_0\} = 0$: una costante ha trasformata di Hilbert nulla, e la trasformata di Hilbert è definita a meno di una costante. Il valore medio di $x(t)$ non si ripercuote su $\widehat{x}(t)$;
- $\mathcal{H}\{\mathcal{H}\{x(t)\}\} = \widehat{\widehat{x}}(t) = -x(t)$: infatti una rotazione di fase pari a $\pi$ radianti corrisponde 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.

[14] Il risultato è valido solamente se $x_c(t)$ ed $x_s(t)$ sono limitate in banda $\pm W$ con $W < f_0$, come mostrato in appendice 10.5.2.

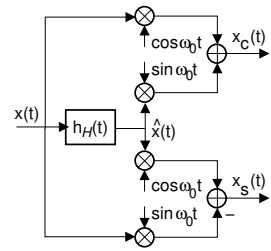
che, assieme all'espressione (10.1) di $x(t)$ in funzione di $x_c(t)$ ed $x_s(t)$, permette di esprimere la trasformata di Hilbert di un segnale modulato, e quindi di impostare un sistema di due equazioni nelle due incognite $x_c(t)$ e $x_s(t)$:

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

Il sistema può essere risolto[15], permettendo in definitiva di esprimere le componenti analogiche di bassa frequenza in termini di $x(t)$ e di $\widehat{x}(t)$:

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

Pertanto, le componenti analogiche di bassa frequenza possono essere estratte direttamente da $x(t)$, utilizzando un filtro di Hilbert per ottenere $\widehat{x}(t)$, e combinando i due segnali per mezzo di oscillatori in quadratura, in accordo allo schema circuitale mostrato nella figura a fianco.




Infine, $x_c(t)$ e $x_s(t)$ permettono di risalire alle modulazioni di ampiezza ed angolare:

$$\begin{cases} a(t) = |\underline{x}(t)| = \sqrt{x_c^2(t) + x_s^2(t)} \\ \varphi(t) = \angle\underline{x}(t) = \arctan\frac{x_s(t)}{x_c(t)} \end{cases}$$

Prima di procedere a calcolare $\mathcal{P}_{\underline{x}}(f)$, occorre introdurre l'ulteriore concetto di *segnale analitico*.

### 10.2.5 Segnale analitico

Il segnale analitico associato ad $x(t)$ corrisponde al suo *contenuto a frequenze positive* $x^+(t)$, introdotto al § 10.1.5; si può mostrare che $x^+(t)$ è esprimibile nei termini di $x(t)$ e $\widehat{x}(t)$, secondo l'espressione[16]:

$$x^+(t) = \frac{1}{2}\left(x(t) + j\widehat{x}(t)\right) \tag{10.4}$$

Molto utile è anche la relazione che lega il segnale analitico all'inviluppo complesso:

$$x^+(t) = \frac{1}{2}\underline{x}(t)\, e^{j\omega_0 t} \tag{10.5}$$

[15] Potremmo notare come la matrice dei coefficienti costituisca una rotazione di assi (vedi as es. il § 14.5.4.1), rotazione che "ruota" letteralmente a velocità angolare $\omega_0$. Tale rotazione stabilisce che le coppie di segnali $(x_c(t),\, x_s(t))$ e $(x(t),\, \widehat{x}(t))$ rappresentano comunque l'evoluzione dell'inviluppo complesso $\underline{x}(t) = a(t)\, e^{j\varphi(t)}$: mentre $x_c(t)$ e $x_s(t)$ lo rappresentano su due assi ad esso solidali, $x(t)$ e $\widehat{x}(t)$ sono definiti su assi ruotanti che tengono conto della frequenza portante.

[16] L'eguaglianza si dimostra valutandola nel domino nella frequenza, ricordando la definizione di filtro di Hilbert, in quanto risulta:

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

infatti, a frequenze negative il prodotto $j \cdot j = -1$ costituisce uno sfasamento di $\pi$ radianti per tutte le frequenze, provocando l'elisione tra $X(f)$ e -$X(f)$ per tutti i valori $f < 0$.

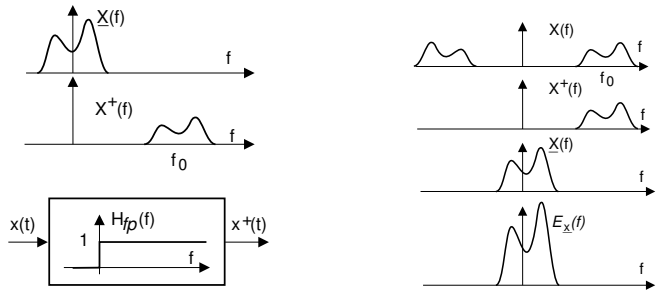



Figura 10.1: Derivazione di $x^{+}(t)$ mediante filtraggio e densità spettrale di $\underline{x}(t)$

che si ottiene tenendo conto dalla (10.4), come illustrato alla nota[17]. Effettivamente, l'ultima relazione rappresenta il contenuto a frequenze positive di $x(t)$, a patto che $\underline{x}(t)$ sia di banda base con frequenza massima $W < f_0$; in tal caso infatti, trasformando la (10.5), si ottiene

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

che giace completamente nel semipiano $f > 0$.

Alternativamente alla (10.4), si può ottenere $x^{+}(t)$ senza utilizzare $\widehat{x}(t)$, pensandolo come il risultato del passaggio di $x(t)$ attraverso un filtro $H_{fp}(f)$[18] (vedi fig. 10.1) con funzione di trasferimento a gradino unitario:

$$x^{+}(t) = x(t) * h_{fp}(t)$$

### 10.2.6 Densità spettrale di segnali passa-banda

Invertendo la (10.5), otteniamo ora $\underline{x}(t) = 2x^{+}(t)\, e^{-j\omega_0 t}$, che trasformata, ci consente di valutare l'espressione di $\underline{X}(f)$:

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

che ci consente di osservare, con riferimento alla fig. 10.1, come in linea di principio $\underline{X}(f)$ non goda di simmetria rispetto allo zero, come prevedibile per $\underline{x}(t)$ complesso. Ricordando ora che $\mathcal{E}_x(f) = |X(f)|^2$, otteniamo

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

Un risultato del tutto simile può essere ottenuto[19] per segnali di potenza, ovvero

$$\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{x^{+}}(f + f_0) \tag{10.6}$$

[17] Sviluppando il secondo membro di (10.5) si ottiene:

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

che corrisponde al secondo membro di (10.4), e quindi a $x^{+}(t)$.

[18] Il pedice $_{fp}$ stà per *frequenze positive*.

[19] La (10.6) può essere motivata seguendo le stesse linee guida indicate alla nota 12 a pag. 212.

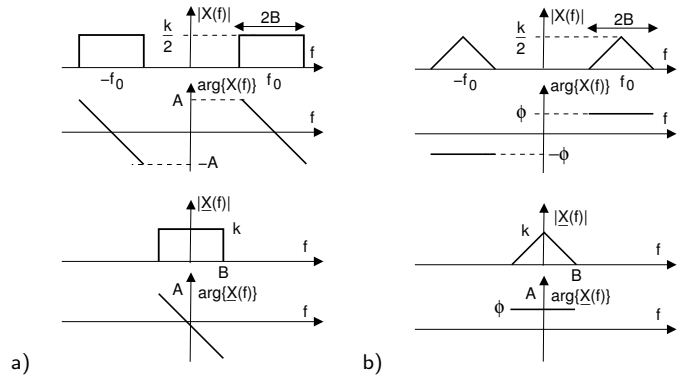



Figura 10.2: Densità spettrali utilizzate negli esempi

Pertanto, la densità di potenza di $\underline{x}(t)$ si ottiene da quella a frequenze positive di $x(t)$, traslata nell'origine e moltiplicata per 4. Infine, la discussione riportata al § 10.4 mostra come un risultato del tutto simile sia valido anche nel caso in cui $x(t)$ è membro di un processo ergodico.

### 10.2.7 Esempi

- Sia dato il segnale $x(t)$ la cui trasformata $X(f)$ è riportata nel lato sinistro in alto di Fig. 10.2. Quali sono le sue componenti analogiche di bassa frequenza, espresse nel dominio della frequenza?
- Notiamo che $\left|X^+(f)\right| = \frac{k}{2}\mathrm{rect}_{2B}(f - f_0)$, e dunque

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

  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[20] $\begin{cases} X_c(f) = k\mathrm{rect}_{2B}(f)\, e^{-j2\pi\frac{A}{2\pi B}f} \\ X_s(f) = 0 \end{cases}$, ed effettuando l'antitrasformata di $X_c(f)$ si ottiene

  $$x_c(t) = 2kB\mathrm{sinc}\left[2B\left(t - \frac{A}{2\pi B}\right)\right]$$

  in cui la traslazione nel tempo è dovuta alla fase lineare presente in $\underline{X}(f)$.
- Lo stesso problema precedente, ma applicato al segnale b), la cui trasformata $X(f)$ è mostrata al lato destro in alto di Fig. 10.2.
- Eseguendo di nuovo le operazioni di traslazione si ottiene l'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

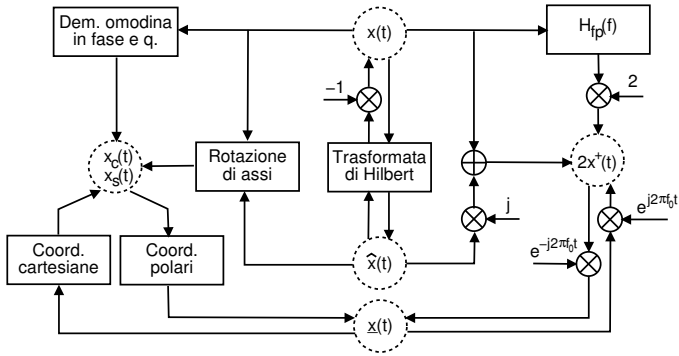
  $$\begin{cases} x_c(t) = kB\left(\frac{\sin\pi Bt}{\pi Bt}\right)^2\cos\phi \\ x_s(t) = kB\left(\frac{\sin\pi Bt}{\pi Bt}\right)^2\sin\phi \end{cases} \Rightarrow \begin{cases} X_c(f) = k\left(1 - \frac{|f|}{B}\right)\cos\phi \\ X_s(f) = k\left(1 - \frac{|f|}{B}\right)\sin\phi \end{cases}$$

[20] Approfittiamo dell'occasione per notare che, pur potendo scrivere $\underline{X}(f) = X_c(f) + jX_s(f)$, non è assolutamente lecito dire che $\Re\{\underline{X}(f)\} = X_c(f)$ e $\Im\{\underline{X}(f)\} = X_s(f)$; infatti sia $X_c(f)$ che $X_s(f)$ possono a loro volta essere complessi (mentre $x_c(t)$ e $x_s(t)$ sono necessariamente reali).

con $|f| < B$.

### 10.2.8 Schema delle trasformazioni

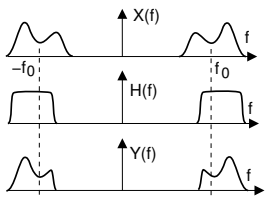
La figura che segue riassume le relazioni esistenti tra le grandezze $x(t)$, $\hat{x}(t)$, $\underline{x}(t)$, $x^+(t)$, $x_c(t)$ e $x_s(t)$.




## 10.3 Transito dei segnali modulati nei sistemi fisici

Trattiamo ora delle trasformazioni subite da un segnale modulato nel passaggio attraverso un sistema fisico, in funzione degli inviluppi complessi del segnale $\underline{x}(t)$ e della risposta impulsiva $\underline{h}(t)$. L'ovvia utilità di questa forma di descrizione è che ci consente di svolgere i calcoli maneggiando segnali di banda base.

### 10.3.1 Filtraggio




Poniamoci nelle condizioni generali di un segnale $x(t)$ che attraversa un filtro $h(t)$, che per entrambi valgano le condizioni di limitazione in banda, e che si adotti una stessa $f_0$ di riferimento. Anche il segnale in uscita $y(t)$ è dello stesso tipo, e può essere mostrato[21] che il suo inviluppo complesso vale:

$$\underline{y}(t) = \frac{1}{2}\underline{x}(t) * \underline{h}(t)$$

[21] 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

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

in cui $h_{fp}(t)$ è la risposta impulsiva del filtro necessario ad estrarre il segnale analitico. Non resta ora che mostrare lo sviluppo per il risultato anticipato:

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

da cui è facile ottenere l'espressione di $y_c(t)$ ed $y_s(t)$ in funzione delle c.a. di b.f. di $x(t)$ e di quelle del filtro:

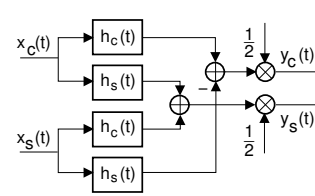
$$\begin{aligned}\underline{y} &= \frac{1}{2}\underline{x} * \underline{h} = \frac{1}{2}[x_c + jx_s] * [h_c + jh_s] = \\ &= \frac{1}{2}[x_c * h_c - x_s * h_s] + j\frac{1}{2}[x_s * h_c + x_c * h_s]\end{aligned}$$

Dunque, per le componenti reale e immaginaria del segnale modulato, dopo il filtraggio, sussistono le relazioni

$$\begin{cases} y_c(t) = \frac{1}{2}[x_c(t) * h_c(t) - x_s(t) * h_s(t)] \\ y_s(t) = \frac{1}{2}[x_s(t) * h_c(t) + x_c(t) * h_s(t)] \end{cases} \tag{10.7}$$

e lo schema riportato appresso raffigura un circuito equivalente alle (10.7), ossia operante su $x_c(t)$ e $x_s(t)$, e che determina lo stesso risultato.

#### 10.3.1.1 Intermodulazione tra componenti analogiche di bassa frequenza




Osservando il risultato (10.7) notiamo che sia $y_c(t)$ che $y_s(t)$ dipendono in generale da entrambe le componenti $x_c(t)$ e $x_s(t)$: questo fenomeno prende il nome di *intermodulazione* tra componenti analogiche di bassa frequenza, ed è fonte di una distorsione *ineliminabile* in banda base. Infatti, le informazioni contenute in $x_c(t)$ ed $x_s(t)$ sono ora mescolate in modo tale che, anche disponendo sia di $y_c(t)$ che di $y_s(t)$, non possono essere separate. Gli unici casi in cui ciò *non* si verifica sono relativi all'evenienza che $\underline{x}(t)$ oppure $\underline{h}(t)$ presentino *una sola* delle due C.A. di B.F., ossia almeno uno dei due sia solo reale o solo immaginario.

#### 10.3.1.2 Equalizzazione di banda base

Nel caso in cui *non avvenga* il fenomeno di intermodulazione suddetto, come ad esempio se $\underline{h}(t) = h_c(t)$, e quindi risulti $\begin{cases} y_c(t) = \frac{1}{2}x_c(t) * h_c(t) \\ y_s(t) = \frac{1}{2}x_s(t) * h_c(t) \end{cases}$, allora $y_c(t)$ e $y_s(t)$ sono affette unicamente da distorsione lineare (§ 14.5), e quindi le componenti *trasmesse* $x_c(t)$ e $x_s(t)$ possono essere ri-ottenute a partire da quelle *ricevute* $y_c(t)$ e $y_s(t)$ mediante un procedimento di *equalizzazione*, che consiste nel loro passaggio attraverso un filtro con risposta impulsiva $g_{eq}(t)$, la cui risposta in frequenza risulta pari a $G_{eq}(f) = {}^{ae^{j2\pi f\tau}}/_{H_c(f)}$, permettendo di ottenere $\begin{cases} x_c(t) = 2y_c(t) * g_{eq}(t) \\ x_s(t) = 2y_s(t) * g_{eq}(t) \end{cases}$. Infatti in tal caso la risposta in frequenza complessiva, risultato del passaggio del segnale modulato prima nel canale, e quindi nell'equalizzatore, risulta essere il prodotto delle due risposte in frequenza $H_c(f)\,G_{eq}(f)$, e quindi pari a quella di un canale perfetto $H(f) = ae^{j2\pi f\tau}$ (§ 14).

Se al contrario sono presenti entrambe $h_c(t)$ e $h_s(t)$, o entrambe $x_c(t)$ e $x_s(t)$, allora non è più possibile eseguire il processo di equalizzazione operando sulle c.a. di b.f., mentre è teoricamente possibile operare direttamente sul segnale modulato, a patto di non incontrare ostacoli teconologici.

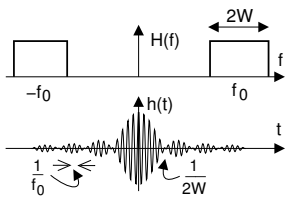
### 10.3.2 Condizioni per inviluppo complesso reale

Data la rilevanza dei segnali con inviluppo complesso ad una sola componente, determiniamo quali condizioni si debbano verificare per dar luogo ad una simile circostanza, iniziando da un esempio.

#### 10.3.2.1 Filtro passa banda ideale

E' descritto da una risposta in frequenza $H(f)$ nulla ovunque, tranne che negli intervalli di frequenze $f_0 - W \leq |f| \leq f_0 + W$ dove ha valore unitario. Pertanto, risulta $H(f) = rect_{2W}(f - f_0) + rect_{2W}(f + f_0)$, da cui si ottiene facilmente

$$\begin{aligned} h(t) &= \mathcal{F}^{-1}\{H(f)\} = 2W\text{sinc}(2Wt)\left(e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}\right) = \\ &= 4W\text{sinc}(2Wt)\cos 2\pi f_0 t \end{aligned}$$




D'altra parte, l'andamento di $H(f)$ è quello tipico dei segnali modulati, e quindi per $h(t)$ vale la sua rappresentazione in termini di inviluppo complesso $\underline{h}(t) = h_c(t) + jh_s(t)$, per cui possiamo scrivere

$$h(t) = h_c(t)\cos 2\pi f_0 t - h_s(t)\sin 2\pi f_0 t$$

Confrontando questa espressione con quella trovata prima, si osserva che deve necessariamente risultare $h_s(t) = 0$ ed $h_c(t) = 4W\frac{\sin 2\pi Wt}{2\pi Wt}$, per cui $\underline{h}(t)$ è reale.

#### 10.3.2.2 Simmetria coniugata attorno ad $f_0$

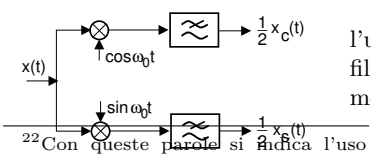
Il filtro passa banda ideale presenta $\underline{h}(t)$ reale, in quanto $\underline{H}(f) = 2H^+(f + f_0) = 2rect_{2W}(f)$ esibisce simmetria coniugata attorno all'origine. E' proprio questa la condizione cercata, che ci permette di enunciare

> *Un segnale modulato $x(t)$ possiede un inviluppo complesso $\underline{x}(t)$ reale, se lo spettro di quest'ultimo $\underline{X}(f)$ ha simmetria coniugata attorno all'origine $\underline{X}(f) = \underline{X}^*(-f)$, ovvero il segnale analitico $X^+(f)$ ha simmetria coniugata attorno ad $f_0$: $X^+(f_0 + \phi) = \left[X^+(f_0 - \phi)\right]^*$ $(|\phi| < W)$.*

In altre parole, $\underline{x}(t) = x_c(t)$ se $X^+(f)$ ha modulo pari e fase dispari rispetto ad $f_0$.

### 10.3.3 Estrazione delle componenti analogiche di bassa frequenza

Lo schema adottato al § 10.2.4 per ottenere le C.A. di B.F., basato sul filtro di Hilbert, non è l'unico.

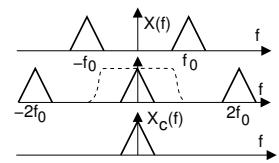



Un secondo metodo molto pratico prevede ancora l'uso di portanti *sincrone* (od *omodina*)[22], e di due filtri passa basso anzichè del filtro di Hilbert, come mostrato a lato. Il suo funzionamento è basato sul

[22] Con queste parole si indica l'uso in ricezione della stessa identica portante usata per la trasmissione, senza errori né di fase né di frequenza, e le cui modalità sono indicate al § 11.2.1.

fatto che, considerando $x(t)$ espresso in termini delle sue C.A. di B.F., si ha[23]:

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




I termini in cui compaiono $\cos 2\omega_0 t$ e $\sin 2\omega_0 t$ rappresentano componenti di segnale centrate attorno a $2f_0$, ed il filtro passa basso[24] (la cui $H(f)$ è tratteggiata in figura) provvede ad eliminarle: la banda del filtro deve quindi essere maggiore di $W$ ma inferiore a $2f_0 - W$. Pertanto, il filtro non deve necessariamente essere ideale, e se $f_0 \gg W$ non sussistono particolari problemi realizzativi. Procedendo in maniera simile, per il ramo in quadratura si ottiene:

$$x(t)\sin\omega_0 t = \frac{1}{2}x_s(t) - \frac{1}{2}x_c(t)\cos 2\omega_0 t - \frac{1}{2}x_s(t)\sin 2\omega_0 t$$

e dunque, anche in questo caso, il filtro passa-basso rimuove le componenti a frequenza doppia.

Se i filtri non sono ideali, ma hanno ad esempio una fase lineare, saranno equivalenti ad un ritardo; se presentano distorsioni più severe (modulo non costante o fase non lineare), allora introducono distorsioni aggiuntive; per ridurre al minimo gli effetti di queste ultime, si tenta almeno di realizzare i due filtri quanto più identici tra loro.

## 10.4 Rappresentazione dei processi in banda traslata

Finora abbiamo trattato i casi di segnali di energia e di potenza; per ottenere una rappresentazione adeguata anche dei processi, occorre ancora un pò di teoria. Il lettore impaziente, o timoroso di perdersi tra i calcoli (che sono effettivamente intricati), può saltare direttamente alle conclusioni (§ 10.4.1), che sono le uniche che ci serviranno per il resto del testo. Altrimenti, armiamoci di pazienza e iniziamo.

Siamo ora interessati ad ottenere, una volta noto uno spettro di densità di potenza $\mathcal{P}_x(f)$ limitato in banda attorno ad $f_0$, delle rappresentazioni utili per gli spettri di densità di potenza delle componenti analogiche di bassa frequenza, ovvero espressioni per le loro funzioni di autocorrelazione. Infatti, come abbiamo visto, il passaggio di un segnale in banda traslata attraverso un filtro può essere scomposto in 4 filtraggi in banda base: pertanto la rappresentazione delle C.A. di B.F. è sufficiente per ottenere tutte le altre grandezze di interesse.

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

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

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

[24] Il simbolo $\boxed{\approx}$ rappresenta un filtro passa-basso, poichè viene cancellata l'ondina superiore. Nello stesso stile, sono a volte indicati un passa-alto $\boxed{\approx}$ ed un passa-banda $\boxed{\approx}$.

Pertanto, non si ottengono direttamente le C.A. di B.F. del processo, come invece accade per segnali di energia di cui è noto $\underline{X}(f)$. D'altra parte, è innegabile che una realizzazione di $x(t)$ sia limitata in banda centrata a $f_0$, e che quindi per essa debba esistere la rappresentazione $x(t) = x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t$; data la natura aleatoria di $x(t)$, gli stessi $x_c(t)$ ed $x_s(t)$ sono realizzazioni di processi. Questi ultimi in generale non sono indipendenti tra loro, in quanto la loro combinazione deve produrre un $x(t)$ che appartiene al processo originario. Si pensi ad esempio al segnale $x(t) = x_c(t)\cos\omega_0 t$, in cui $x_c(t)$ è stazionario ed ergodico: come già osservato al § 9.6.1, $x(t)$ è solamente ciclostazionario.

Come prima cosa, proviamo a calcolare la funzione di autocorrelazione dell'inviluppo complesso di una generica realizzazione:

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

e queste 4 quantità sono calcolate in appendice 10.5.3. Il risultato finale dei calcoli, nel caso in cui $x_c(t)$ e $x_s(t)$ siano stazionari ed ergodici, fornisce le espressioni

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

in cui $\widehat{\mathcal{R}}_x(\tau) = \mathcal{H}\{\mathcal{R}_x(\tau)\}$ è la trasformata di Hilbert di $\mathcal{R}_x(\tau)$. Osserviamo quindi come risulti $\mathcal{R}_{\underline{x}}(\tau) = 2\,[\mathcal{R}_{x_c}(\tau) + j\mathcal{R}_{x_c x_s}(\tau)]$, e pertanto $\mathcal{P}_{\underline{x}}(f) = 2\,[\mathcal{P}_{x_c}(f) + j\mathcal{P}_{x_c x_s}(f)]$.

- Da quest'ultima espressione, sembrerebbe che $\mathcal{P}_{\underline{x}}(f)$ possa assumere valori complessi, perdendo il senso fisico di potenza: ma non è così. Osserviamo infatti che $\mathcal{R}_{x_c x_s}(\tau)$ è un segnale reale dispari[25]: pertanto $\mathcal{P}_{x_c x_s}(f) = \mathcal{F}\{\mathcal{R}_{x_c x_s}(\tau)\}$ è completamente immaginario, e dunque $\mathcal{P}_{\underline{x}}(f)$ è reale.
- Se risultasse $\mathcal{R}_{x_c x_s}(\tau) = 0$ per ogni $\tau$ allora $\mathcal{P}_{x_c x_s}(f) = 0$ e $\mathcal{P}_{\underline{x}}(f) = 2\mathcal{P}_{x_c}(f)$ sarebbe reale pari; la presenza di $\mathcal{P}_{x_c x_s}(f)$ lo può invece rendere asimmetrico, permettendo di ottenere ancora $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_x^+(f + f_0)$ (non dimostrato),
- Corollario del punto precedente è che, se $\mathcal{P}_x(f)$ è simmetrico rispetto ad $f_0$, allora $\mathcal{R}_{x_c x_s}(\tau) = 0$ e le due C.A. di B.F. sono incorrelate; ***se*** inoltre queste sono congiuntamente gaussiane, *allora risultano anche statisticamente indipendenti.*
- Dato che $\mathcal{R}_{x_c x_s}(\tau) = -\mathcal{R}_{x_s x_c}(\tau)$ sono dispari, deve risultare che $\mathcal{R}_{x_c x_s}(0) = -\mathcal{R}_{x_s x_c}(0) = 0$; se i processi sono a media nulla, allora la potenza è pari al valore dell'autocorrelazione per $\tau = 0$, e quindi

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

  In definitiva, le componenti analogiche di bassa frequenza hanno entrambe potenza metà di quella dell'inviluppo complesso.

[25] Infatti $\mathcal{R}_{x_c x_s}(\tau) = \widehat{\mathcal{R}}_x(\tau)\cos\omega_0\tau - \mathcal{R}_x(\tau)\sin\omega_0\tau$, in cui $\mathcal{R}_x(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_x(f)\}$ è pari e $\sin\omega_0 t$ è dispari, mentre $\widehat{\mathcal{R}}_x(\tau)$ è dispari (non è stato dimostrato, ma vale per le trasformate di Hilber di segnali pari) e $\cos\omega_0\tau$ è pari. Inoltre, essendo $x_c(t)$ ed $x_s(t)$ reali, $\mathcal{R}_{x_c x_s}(\tau)$ è reale.

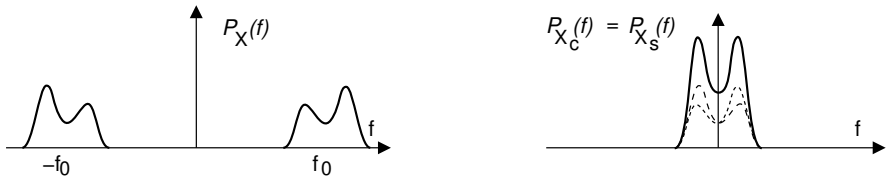



Figura 10.3: Segnale modulato e densità di potenza delle componenti analogiche di b.f.

- $x_c(t)$ e $x_s(t)$ hanno (ciascuna) potenza pari a quella di $x(t)$, ovvero $\mathcal{P}_{x_c} = \mathcal{P}_{x_s} = \mathcal{P}_x$; infatti, ricordando che $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_x^+(f+f_0)$, si ottiene $\mathcal{P}_{\underline{x}} = 4\mathcal{P}_x^+$. Dovendo chiaramente risultare $\mathcal{P}_x = \mathcal{P}_x^+ + \mathcal{P}_x^-$, si ottiene

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

- E' possibile mostrare che, volendo esprimere l'autocorrelazione di $x(t)$ in termini delle sue C.A. di B.F. $\mathcal{R}_x(\tau) = \mathcal{R}_c(\tau)\cos\omega_0\tau - \mathcal{R}_s(\tau)\sin\omega_0\tau$, risulta

$$\begin{cases} \mathcal{R}_c(\tau) = \mathcal{R}_{x_c}(\tau) \\ \mathcal{R}_s(\tau) = -\mathcal{R}_{x_c x_s}(\tau) \end{cases}$$

  da cui è possibile mostrare che $\mathcal{R}_x(\tau) = \mathcal{R}_{\widehat{x}}(\tau)$.
- Volendo valutare $\mathcal{P}_{x_c}(f)$, questo risulta identico a $\mathcal{P}_{x_s}(f)$, in quanto (come già visto) $\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x(\tau)\cos\omega_0\tau + \widehat{\mathcal{R}}_x(\tau)\sin\omega_0\tau$; applicando ora la formula di Eulero per seno e coseno si ottiene

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

  infatti i termini tra parentesi quadre corrispondono alla definizione di componenti a frequenze positive e negative ottenute tramite trasformata di Hilbert.

### 10.4.1 Conclusioni

Abbiamo mostrato che $\mathcal{R}_{x_c}(\tau) = \mathcal{R}_{x_s}(\tau) = \mathcal{R}_x^-(\tau)\,e^{j\omega_0\tau} + \mathcal{R}_x^+(\tau)\,e^{-j\omega_0\tau}$. Risulta quindi:

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

e dunque lo spettro di densità di potenza delle componenti analogiche di un processo si ottiene traslando e sovrapponendo (vedi fig. 10.3) le componenti a frequenze positive e negative del $\mathcal{P}_x(f)$ di partenza.

### 10.4.2 Processo gaussiano bianco limitato in banda

Se $x(t)$ è gaussiano stazionario ergodico e bianco, con $\mathcal{P}_x(f) = \frac{N_0}{2}$ limitato in banda $\pm W$ attorno ad $f_0$ ed a media nulla, allora (vedi fig. 10.4):

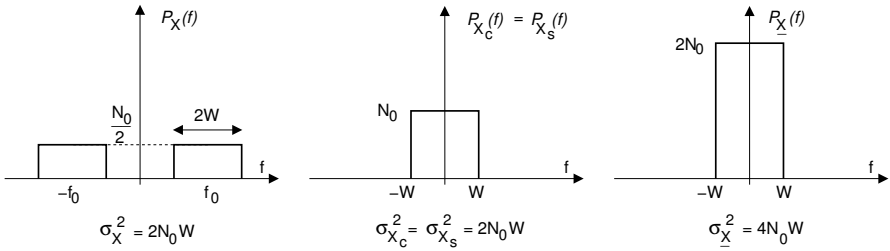



Figura 10.4: Densità di potenza per rumore passabanda

- $x_c(t)$ e $x_s(t)$ sono congiuntamente gaussiane, stazionarie, ergodiche e indipendenti, con $\mathcal{P}_{x_c} = \mathcal{P}_{x_s} = \mathcal{P}_x = 2N_0W$, e pari alle varianze $\sigma_x^2$, $\sigma_{x_c}^2$ e $\sigma_{x_s}^2$. Le rispettive densità di potenza valgono:

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

- L'inviluppo complesso ha potenza doppia:

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

## 10.5 Appendici

### 10.5.1 Risposta impulsiva del filtro di Hilbert

Al § 10.2.4 si è affermato che $h_{\mathcal{H}}(t) = \frac{1}{\pi t}$, ed ora passiamo a dimostrarlo. Innanzitutto osserviamo che $H_{\mathcal{H}}(f)$ può essere scritta in termini di due gradini in frequenza $g(f)$, con quello relativo alle frequenze negative, di tipo anticausale:

$$H_{\mathcal{H}}(f) = -j \cdot g(f) + j \cdot g(-f)$$

Per calcolare l'antitrasformata di Fourier di un gradino in frequenza, utilizziamo il risultato noto per la trasformata di un gradino nel tempo (vedi § 3.9.4) $G(f) = \mathcal{F}\{g(t)\} = \frac{1}{2}\left(\delta(f) - \frac{j}{\pi f}\right)$, a cui applichiamo la proprietà di dualità (vedi pag. 31) che asserisce che se $G(f) = \mathcal{F}\{g(t)\}$, allora $\mathcal{F}^{-1}\{g(f)\} = G(-t)$ , per ottenere

$$\mathcal{F}^{-1}\{g(f)\} = \frac{1}{2}\left(\delta(-t) - \frac{j}{\pi(-t)}\right) = \frac{1}{2}\left(\delta(t) + \frac{j}{\pi t}\right) \tag{10.8}$$

Per scrivere l'antitrasformata del gradino anticausale in frequenza $g(-f)$, occorre tenere conto della proprietà del cambiamento di scala (vedi pag. § 34), che asserisce che $\mathcal{F}^{-1}\{X(-f)\} = x(-t)$, e che applicata alla (10.8), consente di ottenere

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

Ora possiamo quindi scrivere

$$\begin{aligned} h_{\mathcal{H}}(t) &= \mathcal{F}^{-1}\{H_{\mathcal{H}}(f)\} = -j\cdot\frac{1}{2}\left(\delta(t)+\frac{j}{\pi t}\right)+j\frac{1}{2}\left(\delta(t)-\frac{j}{\pi t}\right) \\ &= -j\frac{1}{2}\delta(t)-j^2\frac{1}{2\pi t}+j\frac{1}{2}\delta(t)-j^2\frac{1}{2\pi t} \\ &= \frac{1}{\pi t} \end{aligned}$$

### 10.5.2 Trasformata di Hilbert di un segnale modulato

Limitiamoci a dimostrare che

$$\mathcal{H}\{x_c(t)\cos 2\pi f_0 t\} = x_c(t)\sin 2\pi f_0 t \tag{10.9}$$

Iniziamo con il considerare che dopo aver $\mathcal{F}$-trasformato la (10.9), possiamo evidenziarne le componenti a frequenza positiva e negativa $X_c(f-f_0)$ e $X_c(f+f_0)$

$$x_c(t)\cos 2\pi f_0 t = \frac{x_c(t)}{2}\left(e^{j2\pi f_0 t}+e^{-j2\pi f_0 t}\right) \overset{\mathcal{F}}{\Rightarrow} \frac{1}{2}\left[X_c(f-f_0)+X_c(f+f_0)\right] \tag{10.10}$$

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 (10.9)

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

risultato che, anche se non ancora nella forma anticipata, poteva comunque essere ottenuto anche direttamente a partire dal secondo membro di (10.10), invocando subito la limitazione ad una semibanda di $x_c(t)\,e^{\pm j2\pi f_0 t}$. Per ottenere la (10.9) è ora sufficiente moltiplicare e dividere per $j = e^{j\frac{\pi}{2}}$, ossia

$$\begin{aligned} \frac{x_c(t)}{2}\left(e^{j2\pi f_0 t}e^{-j\frac{\pi}{2}}+e^{-j2\pi f_0 t}e^{j\frac{\pi}{2}}\right)\cdot\frac{e^{j\frac{\pi}{2}}}{e^{j\frac{\pi}{2}}} &= \frac{x_c(t)}{2j}\left(e^{j2\pi f_0 t}+e^{-j2\pi f_0 t}e^{j\pi}\right) = \\ = \frac{x_c(t)}{2j}\left(e^{j2\pi f_0 t}-e^{-j2\pi f_0 t}\right) &= x_c(t)\sin 2\pi f_0 t \end{aligned}$$

in quanto $e^{j\pi} = -1$.

### 10.5.3 Autocorrelazione di processi passa-banda

Svolgiamo qui il calcolo relativo al valore di $\mathcal{R}_{x_c}(\tau)$, $\mathcal{R}_{x_s}(\tau)$, $\mathcal{R}_{x_c x_s}(\tau)$ e $\mathcal{R}_{x_s x_c}(\tau)$. Ricordando che $x_c(t) = x(t)\cos\omega_0 t + \widehat{x}(t)\sin\omega_0 t$, iniziamo a svolgere i calcoli per $\mathcal{R}_{x_c}(\tau)$:

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

Valutiamo quindi i quattro valori attesi singolarmente, 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. 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 calcolati, e quindi possiamo scrivere direttamente lo sviluppo dei calcoli, in cui si applicano nuovamente le identità trigonometriche note:

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

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.

I calcoli relativi al valore di $\mathcal{R}_{x_s x_c}(\tau)$ sono del tutto simili, ed il loro svolgimento porta al risultato $\mathcal{R}_{x_s x_c}(\tau) = -\mathcal{R}_{x_c x_s}(\tau)$.

# Capitolo 11

# Modulazione per segnali analogici

Trattiamo qui delle tecniche comunemente usate per imprimere su una portante l'informazione di un messaggio modulante di natura analogica[1], mostrando allo stesso tempo le caratteristiche spettrali del segnale ottenuto. Sono quindi discussi i metodi piú idonei per realizzare la funzione di demodulazione, e gli accorgimenti che influenzano il risultato finale.

## 11.1 Modulazione di ampiezza - *AM*

Si è illustrato come un segnale modulato possa essere rappresentato nei termini delle sue componenti analogiche di bassa frequenza: $x(t) = x_c(t)\cos\omega_0 t - x_s(t)\sin\omega_0 t$. Nel caso in cui $x_c$ e $x_s$ siano segnali indipendenti, la trasmissione congiunta di entrambi sulla medesima portante costituisce un segnale *QAM* (QUADRATURE AMPLITUDE MODULATION). Nei casi piú tipici invece, i segnali $x_c$ ed $x_s$ non sono qualsiasi, ed in base alla loro definizione sono distinte le seguenti classi di segnali modulati in ampiezza:

- *Banda Laterale Doppia*: la componente $x_s(t)$ è nulla, cosicchè $\mathcal{P}_x(f)$ è simmetrico rispetto ad $f_0$. Si tratta del caso a noi già noto, ed è indicato dagli acronimi *BLD* o *DSB* (DOUBLE SIDE BAND).
- *Banda Laterale Unica*: sono presenti sia $x_c(t)$ che $x_s(t)$, e risulta $x_s(t) = \widehat{x}_c(t)$. Questo fa sí che (come vedremo) la densità $\mathcal{P}_x(f)$ del segnale modulato giaccia tutta all'*esterno* (od all'interno) di $\pm f_0$ (*BLU* o *SSB* - SINGLE SIDE BAND)
- *Banda Laterale Ridotta*: è una via di mezzo tra i due casi precedenti, e cioè $\mathcal{P}_x(f)$ non è simmetrica rispetto ad $f_0$, ma comunque giace da ambo i lati (*BLR* o *VSB* - VESTIGIAL SIDE BAND[2] ).

Per completare la classificazione, per ognuna delle possibilità precedenti può verificarsi uno tra tre sottocasi, che si riferiscono alla presenza o meno, in $\mathcal{P}_x(f)$, di una concentrazione di potenza (ossia di un impulso) a frequenza $f_0$, corrispondente alla trasmissione di potenza non associata al messaggio $m(t)$, ma solamente alla portante, e quindi priva di contenuto informativo ai fini della trasmissione. I tre sottocasi citati sono indicati come:

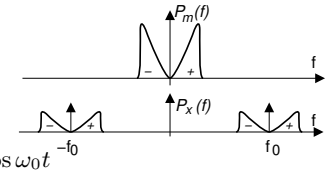
[1] Per i segnali numerici si usano tecniche peculiari, esposte al capitolo 13.

[2] Come sarà più chiaro nel seguito, l'acronimo *VRB* simboleggia che, anziché sopprimere completamente una delle due bande laterali, se ne mantengono *delle vestigia*.

- Portante Soppressa (*PS* o *SC* - SUPPRESSED CARRIER);
- Portante Intera (*PI* o *LC* - LARGE CARRIER);
- Portante Parzialmente Soppressa (*PPS*).

### 11.1.1 Banda laterale doppia - *BLD*

Questo è il caso a cui ci si riconduce in presenza di una sola componente analogica di bassa frequenza, che per convenzione è posta pari a $x_c(t)$[3]. La dipendenza di $x_c(t)$ da $m(t)$ è posta nella forma generale $x_c(t) = a_p + k_a m(t)$, e quindi




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

L'inviluppo complesso pertanto risulta $\underline{x}(t) = a_p + k_a m(t)$ da cui

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

e quindi, dato che $\mathcal{P}_x(f) = \mathcal{P}_{x^+}(f) + \mathcal{P}_{x^-}(f)$ e che, in base alla (10.6) risulta $\mathcal{P}_{x^+}(f) = \frac{1}{4}\mathcal{P}_{\underline{x}}(f - f_0)$ e $\mathcal{P}_{x^-}(f) = \frac{1}{4}\mathcal{P}_{\underline{x}}(f + f_0)$, si ottiene una densità di potenza per il segnale modulato pari a

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

La potenza totale di $x(t)$ risulta pertanto $\mathcal{P}_x = \frac{a_p^2}{2} + \frac{k_a^2}{2}\mathcal{P}_m$, mentre il suo andamento in frequenza è quello riportato in figura, dove si è posto $k_a = 1$.

#### 11.1.1.1 Portante soppressa - *PS*

Esaminando l'ultima espressione trovata per $\mathcal{P}_x$, è evidente come $\frac{a_p^2}{2}$ sia pari alla potenza della portante non modulata (concentrata per metà ad $f_0$ e per metà a $-f_0$), che quindi svanisce per $a_p = 0$, dando luogo in quest'ultima circostanza al sottocaso di *portante soppressa*.

La demodulazione di questo segnale si effettua in modo coerente (§ 11.2.1), dopo aver ricostruito la portante per quadratura (§ 11.4.2), oppure mediante demodulatore ad inviluppo (§ 11.2.2), dopo aver elaborato la portante ricostruita come spiegato al § 11.1.1.3.

#### 11.1.1.2 Portante intera - *PI*

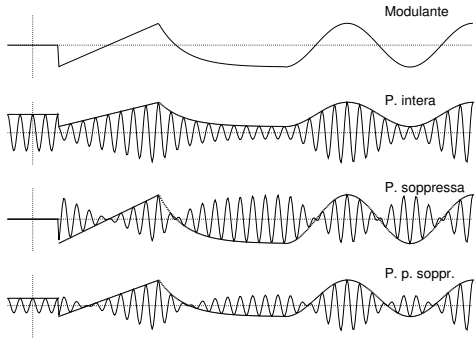
Nel caso in cui $a_p \neq 0$, si può scegliere che risulti sempre $x_c(t) \geq 0$, e quindi $a_p \geq k_a \cdot \max\{|m(t)|\}$, da cui risulta che deve essere

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

[3] Considerando che la portante di modulazione può avere una fase iniziale arbitraria, e che con una traslazione temporale ci si può sempre ricondurre ad usare una funzione $\cos \omega_0 t$, la convenzione posta tratta il caso di un segnale modulato $x(t) = a(t) \cos(\omega_o t + \varphi)$ generico, con $\varphi$ costante.

Queste ultime sono le condizioni che caratterizzano il caso di *portante intera*. Il rapporto $\left(\frac{a_p}{k_a}\right)^2$ rappresenta la massima *potenza istantanea*[4] per $m(t)$, e consente di dimensionare i parametri $a_p$ e $k_a$ in modo da realizzare le condizioni richieste[5].

Quindi, mentre nel caso di portante intera $x_c(t)$ non inverte mai il segno, nel caso di portante soppressa $x_c(t)$ ha invece media nulla (se $m(t)$ ha media nulla) e la portante cambia frequentemente segno, cosicchè per $f = f_0$ non compaiono impulsi in $\mathcal{P}_x(f)$.




La ragione principale dell'utilizzo della portante intera è che in tal caso il processo di decodifica non richiede la conoscenza di $f_0$, e può svolgersi facendo uso di un semplice demodulatore di inviluppo, descritto in 11.2.2.

#### 11.1.1.3 Portante parzialmente soppressa - *PPS*

Se $a_p$ è inferiore al valore necessario per avere la portante intera, ma non è nullo, si ottiene il caso della portante *parzialmente soppressa*, che ci permette di risparmiare potenza (vedi §11.1.1.4). Il residuo di portante presente, può essere usato per rigenerarla "al ricevitore " mediante un PLL (§11.2.1.3), e sommarla al segnale ricevuto, ri-producendo cosí il termine $a_p \cos \omega_0 t$. In tal modo, ci si riconduce al caso *PI*, e si può effettuare la demodulazione di inviluppo (§11.2.2).

#### 11.1.1.4 Efficienza di *PI-PPS*

Nell'espressione 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, notiamo che solo $\mathcal{P}_u = \frac{k_a^2}{2}\mathcal{P}_m$ è relativa al segnale utile, mentre $\frac{a_p^2}{2}$ viene spesa sulla portante, che non trasporta informazione. Pertanto, si definisce una efficienza energetica

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

che rappresenta la frazione di potenza trasmessa, utile ai fini della ricostruzione del messaggio[6].

---

[4] Si definisce come potenza istantanea (o *di picco*) di $m(t)$, il segnale $\mathcal{P}_{M_I}(t) = m^2(t)$, per cui $\mathcal{P}_M = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} \mathcal{P}_{M_I}(t)dt$.

[5] 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 § 7.2.3) per questo caso risulta $\sigma_M^2 = \frac{\Delta^2}{12}$, mentre ad esempio se $m(t) = a\sin 2\pi f_M t$ 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...

[6] Ad esempio, se $m(t) = \sin 2\pi f_M t$ si ha $\mathcal{P}_M = 1/2$ e, nel caso di portante intera, deve risultare $a_p = k_a$ e dunque $\eta = \frac{1}{1+2} = 0.33$. Ovvero solo 1/3 della potenza trasmessa è utile al ricevitore!

### 11.1.2 Banda laterale unica - *BLU*

Come abbiamo visto, la modulazione *BLD* determina una occupazione di banda per $x(t)$ doppia di quella di $m(t)$. Per impegnare invece una banda pari a quella di $m(t)$, il segnale modulato deve dipendere da entrambe le componenti analogiche $x_c(t)$ ed $x_s(t)$, che devono risultare: $x_c(t) = m(t)$ e $x_s(t) = \widehat{m}(t)$. Infatti in tal modo si ottiene:

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

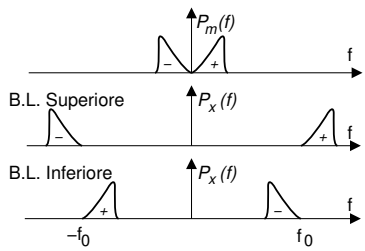
Ricordando ora che $\frac{1}{2}\left[m(t) \pm j\widehat{m}(t)\right] = m^{\pm}(t)$ (vedieq. (10.4)) è proprio il contenuto a frequenze positive (negative), allora se $x(t)$ è di energia, effettuando la trasformata di Fourier di ambo i membri si ottiene

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

e quindi il segnale modulato *AM-BLU* è formato dai contenuti a frequenze positive e negative di $m(t)$, traslati ai lati della portante $f_0$.

Qualora si consideri invece $m(t)$ un processo, si può dimostrare (passando dalla trasformata di $\mathcal{R}_x(\tau)$) un risultato del tutto analogo, ovvero

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




Nel caso descritto abbiamo considerato soppressa la portante, ed il segnale modulato (considerato nel dominio della frequenza) risulta "esterno" ad $f_0$: questa circostanza è indicata con il termine di *banda laterale superiore*. Il caso opposto (*banda laterale inferiore*) si ottiene cambiando segno a $x_s(t)$. Scriviamo dunque

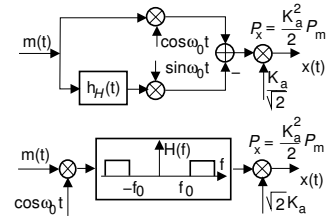
$$x_{BLU}(t) = \frac{k_a}{\sqrt{2}} m(t)\cos\omega_0 t \mp \frac{k_a}{\sqrt{2}}\widehat{m}(t)\sin\omega_0 t$$

con $-$ e $+$ rispettivamente per ottenere un segnale *BLU* con banda superiore o inferiore. Con le costanti indicate, il segnale modulato *BLU* ha una potenza $\mathcal{P}_x = 2\cdot\left(\frac{k_a^2}{2}\cdot\mathcal{P}_M\cdot\frac{1}{2}\right) = \frac{k_a^2}{2}\mathcal{P}_m$ (vedi §11.4.1), eguale a quella di un segnale *AM-BLD* in cui $x_c(t) = k_a m(t)$ e $x_s(t) = 0$.

I vantaggi di un tale metodo di modulazione sono subito evidenti: consente infatti di risparmiare banda, permettendo la trasmissione di piú messaggi in divisione di frequenza (FDM).

#### 11.1.2.1 Generazione di segnali *BLU*

Un segnale *BLU* può essere generato in due diversi modi. Il primo 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. E' subito evidente come si possano presentare problemi se $m(t)$ ha contenuti energetici prossimi a frequenza zero, che rendono assai stringenti le specifiche per approssimare il filtro di Hilbert.

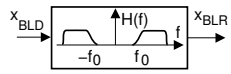



Un problema simile si presenta anche con il secondo metodo di generazione del segnale *BLU*, in cui si genera un segnale *BLD,* che viene filtrato in modo da eliminare una delle bande laterali; la necessità di trasmettere frequenze di $m\,(t)$ prossime allo zero, complica infatti la realizzazione dei filtri.

La trasmissione FDM di segnali *BLU* è stata lungamente usata per i ponti radio telefonici (vedi § 10.1.1.3). Pertanto, la limitazione sulle minime frequenze telefoniche a non meno di 300 Hz sono motivate anche dalla necessità di effettuare modulazioni BLU.

### 11.1.3 Banda laterale ridotta - *BLR*

Si può verificare il caso in cui non si possa assolutamente fare a meno di componenti di segnale a frequenza molto bassa, come avviene, ad esempio, nel segnale televisivo[7] (vedi appendice 11.4.4). Si ricorre allora alla modulazione a Banda Laterale Ridotta (*BLR*), che viene generata inviando il segnale modulato BLD attraverso uno specifico filtro, che presenta una transizione tra la banda passante e quella attenuata, piú dolce di quella di un passa-banda ideale, e che si estende oltre $f_0$.




### 11.1.4 Potenza di un segnale AM

Segue uno schema delle espressioni del segnale modulato per i diversi tipi di modulazione AM, ed i vincoli sui parametri legati all'ottenimento di una potenza totale $\mathcal{P}_x$.

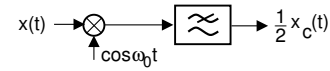
| | Segnale modulato $x(t)$ | Potenza $\mathcal{P}_x$ | $k_a$ per $\mathcal{P}_x$ dato |
|---|---|---|---|
| **BLD-PS** | $k_a m(t)\cos(\omega_0 t)$ | $\frac{k_a^2}{2}\mathcal{P}_m$ | $\sqrt{\frac{2\mathcal{P}_x}{\mathcal{P}_m}}$ |
| **BLU-PS** | $\frac{k_a}{\sqrt{2}}m(t)\cos(\omega_0 t) - \frac{k_a}{\sqrt{2}}\widehat{m}(t)\sin(\omega_0 t)$ | $\frac{k_a^2}{2}\mathcal{P}_m$ | $\sqrt{\frac{2\mathcal{P}_x}{\mathcal{P}_m}}$ |
| **BLD-PI** | $[a_p + k_a m(t)]\cos(\omega_0 t)$ con $a_p \geq k_a \cdot \max\{\lvert m(t)\rvert\}$ | $\frac{a_p^2}{2} + \frac{k_a^2}{2}\mathcal{P}_m$ | $\sqrt{\frac{2\mathcal{P}_x - a_p^2}{\mathcal{P}_m}}$ |

## 11.2 Demodulazione di ampiezza

Un segnale AM può essere demodulato mediante diverse tecniche, ossia *omodina*, *inviluppo*, *in fase e quadratura*, *eterodina*; ognuna di esse ha il suo campo di applicazione, ed i suoi pregi e difetti.

[7] Nel caso ad esempio di ampie zone di immagine a luminosità costante, ed in lento movimento, il segnale è praticamente costante.

### 11.2.1 Demodulazione coerente o omodina




Si tratta del circuito già noto (vedi § 10.3.3) di estrazione della componente in fase $x_c(t)$ mediante moltiplicazione[8] di $x(t)$ per una portante di demodulazione $\cos\omega_0 t$, e rimozione delle componenti a frequenza $2f_0$ mediante un filtro passa-basso, come mostrato dalla figura a lato.

La portante generata localmente deve presentare la stessa fase e frequenza di quella in arrivo, e per questo lo schema viene indicato anche con il nome di demodulazione *omodina* o *sincrona*. Può essere realizzata qualora la portante sia ri-generata al ricevitore mediante un circuito PLL (§ 11.2.1.3) oppure un quadratore (§ 11.4.2). Il metodo è applicabile a tutti i tipi di modulazione di ampiezza, in quanto in tutti la componente in fase dipende da $m(t)$; nella pratica, nei casi di *BLD-PI* e quelli ad esso riconducibili, viene invece adottato il demodulatore di inviluppo (§ 11.2.2).

#### 11.2.1.1 Errori di fase e frequenza

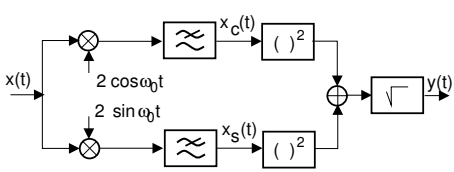
Esaminiamo ora cosa accade nel caso in cui la portante di demodulazione $\cos[2\pi(f_o + \Delta f)t + \theta]$ presenti un errore di fase $\theta$ e/o di frequenza $\Delta f$. Il risultato di una demodulazione coerente risulta[9]:

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

Pertanto, mentre il termine a frequenza (circa) doppia viene eliminato come al solito da un filtro apposito, il segnale utile è affetto da una distorsione pari a :

- per errori di fase: si ottiene $\frac{1}{2}x_c(t)\cos\theta$ che... può annullarsi per $\theta = \pm\frac{\pi}{2}$!
- per errori di frequenza: si ottiene $\frac{1}{2}x_c(t)\cos[\Delta\omega t]$ e dunque il segnale demodulato, oltre ad invertire periodicamente polarità, presenta una notevole oscillazione di ampiezza che, ad esempio, nel caso di segnale audio può rendere il risultato inintelligibile già con $\Delta f$ pari a pochi Hertz.

#### 11.2.1.2 Demodulazione in fase e quadratura




Se il demodulatore dispone anche del ramo in quadratura (quello con il seno) possiamo, in presenza di errore di fase, scegliere ad esempio quale dei due rami è meno attenuato, ed ovviare al problema. Lo stesso schema può essere utile in fase di ricerca della regione di frequenza

[8] In appendice 11.4.3 sono illustrate due tecniche di realizzazione del moltiplicatore.
[9] Si applichi $\cos\alpha\cos\beta = \frac{1}{2}[\cos(\alpha+\beta) + \cos(\alpha-\beta)]$.

in cui è presente un segnale[10], oppure qualora si desideri solo verificare la presenza o meno di un segnale ad una determinata frequenza, come nel caso del radar[11].

Nel caso in cui il segnale ricevuto presenti una fase $\theta$ incognita

$$x(t) = m(t)\cos(\omega_0 t + \theta)$$

l'inviluppo complesso risulta

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

e quindi $x_c(t) = m(t)\cos\theta$ e $x_s(t) = m(t)\sin\theta$.

I due rami del demodulatore in fase e quadratura estraggono proprio $x_c(t)$ ed $x_s(t)$, e dunque il segnale $y(t)$ risulta pari a:

$$y(t) = \sqrt{x_c^2(t) + x_s^2(t)} = |m(t)|\sqrt{\cos^2\theta + \sin^2\theta} = |m(t)|$$

Pertanto, nonostante l'ignoranza della fase $\theta$, siamo ancora in grado di individuare la *presenza* di un segnale modulante. L'operazione di modulo impedisce l'uso dello schema per demodulare generici segnali BLD-PS (mentre il caso PI sarebbe perfettamente demodulabile, ma in tal caso è piú che sufficiente un demodulatore di inviluppo (§ 11.2.2)). Al § 7.6.5 sono esposti alcuni risultati relativi alla probabilità di detezione per questo demodulatore, nel caso in cui l'ingresso sia costituito da rumore, più una eventuale sinusoide.

#### 11.2.1.3 Phase Locked Loop - *PLL*

Trattiamo qui del problema di generare una portante di demodulazione *coerente* (in fase) con quella della portante del segnale ricevuto. Una soluzione molto usata adotta un circuito controreazionato noto come *anello ad aggancio di fase*, e che basa il suo funzionamento su di un dispositivo chiamato *oscillatore controllato in tensione* (*VCO*, VOLTAGE CONTROLLED OSCILLATOR), il quale genera una sinusoide

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

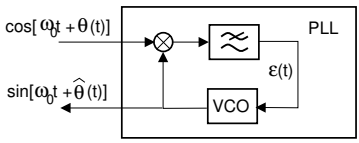
la cui fase varia nel tempo con l'integrale del segnale di ingresso[12].

[10] La ricerca dell'emittente, che può essere banalmente l'azione di sintonizzare la propria radio sul programma preferito, può richiedere interventi automatici, qualora si tratti ad esempio di dover compensare le variazioni di frequenza dovute al movimento reciproco di trasmettitore e ricevitore (*effetto doppler*), come per il caso delle comunicazioni con mezzi mobili.

In questi casi, prima della comunicazione vera e propria, è necessario prevedere una fase di *acquisizione della portante*, svolta ad esempio mediante un circuito del tipo di quello che stiamo discutendo, in cui vengono provate diverse portanti di demodulazione, finché non si produce un segnale in uscita.

[11] Un radar trasmette ad elevata potenza per periodi molto brevi, e stima la presenza di oggetti basandosi sul ritardo con cui il segnale, riflesso da questi, torna indietro. Per questo, il ritardo di fase rappresenta proprio la grandezza che fornirà l'informazione relativa alla distanza, e può essere qualsiasi. Prima di iniziare a stimare tale informazione, è essenziale per il sistema accertarsi che *ci sia* un segnale da stimare.

[12] Se quest'ultimo ad esempio è costante ($x(t) = \Delta$), allora si avrà: $y(t) = \sin(2\pi f_0 t + 2\pi\Delta t) = \sin[2\pi(f_0 + \Delta)t]$, ovvero la frequenza si è alterata di una quantità pari a $\Delta$. Il lettore più attento avrà riconosciuto che il VCO realizza il processo di *modulazione di frequenza*.



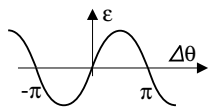

Supponiamo allora di disporre di un segnale AM in cui sia presente un residuo di portante, come ad esempio nel caso di *portante parzialmente soppressa*: in tal caso la portante di demodulazione può essere ottenuta mediante il circuito in figura, che rappresenta appunto un PLL. Indicando con $\widehat{\theta}(t) = 2\pi k_f \int_{-\infty}^{t} x(\tau)\, d\tau$ la fase già integrata dal VCO fino all'istante $t$, all'uscita del moltiplicatore è presente un segnale[13]:

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

Il termine centrato a frequenza doppia ($2\omega_0$) viene eliminato dal filtro passa basso, e dunque rimane

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

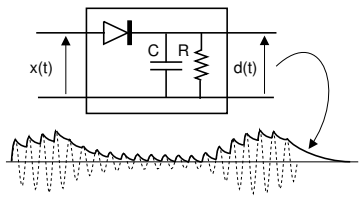
in cui $\Delta\theta(t)$ rappresenta l'errore di fase, e $\varepsilon(t)$ è la grandezza in ingresso al VCO.




Pensiamo ora di avere una $\theta(t)$ di ingresso costante: se risulta $\Delta\theta = 0$, si ha che anche $\varepsilon = 0$, ed il VCO non altera la fase (esatta) della portante generata. Se invece $\Delta\theta \gtrless 0$ e $|\Delta\theta| < \pi$[14], allora $\varepsilon \gtrless 0$, e dunque il VCO è portato ad aumentare (diminuire) la fase della propria portante riducendo di conseguenza l'errore di fase[15].

Nel caso in cui, invece, la fase entrante $\theta(t)$ vari nel tempo, allora il PLL insegue tali variazioni tanto piú da vicino quanto piú è elevato il coefficiente di proporzionalità $k_f$ tra $\widehat{\theta}(t)$ e l'integrale di $\varepsilon(t)$ $k_f$[16].

### 11.2.2 Demodulatore di inviluppo




Si tratta del semplice circuito non lineare riportato in figura[17]. Durante i periodi in cui il segnale in ingresso $x(t)$ è positivo rispetto alla tensione $d(t)$ accumulata dal condensatore, quest'ultimo si carica, inseguendo l'andamento dell'ingresso. Quando diviene $x(t) < d(t)$, il condensatore si scarica sulla resistenza con una costante di tempo $\tau = RC$, abbastanza grande rispetto ad $\frac{1}{f_0}$, e

[13] Si utilizzi $\cos\alpha \sin\beta = \frac{1}{2}[\sin(\alpha+\beta) + \sin(\alpha-\beta)]$.

[14] La grandezza di controllo $\varepsilon(t) \propto \sin(\Delta\theta)$ 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 $\varepsilon$ è ripettivamente *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 $2\pi$, corrispondenti a condizioni di stabilità.

[15] Notiamo che un moltiplicatore, seguito da un filtro passabasso, esegue il calcolo dell'intercorrelazione tra gli ingressi del moltiplicatore (vedi § 9.7.2), che nel nostro caso è una sinusoide.

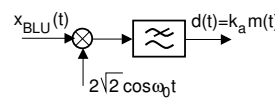
[16] In particolare, le realizzazioni pratiche del PLL dipendono fortemente dalla banda e dall'ordine del filtro di loop, in quanto è quest'ultimo che limita la velocità di variazione di $\varepsilon(t)$ e l'estensione dell'intervallo di aggancio. Lo studio teorico delle prestazioni si basa sull'uso della trasformata di Laplace e sulla linearizzazione di $\sin(\Delta\theta) \simeq \Delta\theta$, in quanto così il PLL può essere studiato come un sistema di controllo "linearizzato". Questa soluzione è brevemente illustrata al § 11.3.1.1.

[17] 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).

tale da permettere la ricostruzione dell'andamento di $x_c(t)$. Le oscillazioni a frequenza $f_0$ (e sue armoniche) infatti possono essere rimosse da un successivo filtro passa-basso, mentre la costante $a_p$ è rimossa mediante un passa alto[18].

La semplicità del circuito è tale da farlo usare nel maggior numero di casi possibili, anche se il suo uso prevalente è per la demodulazione di segnali a *portante intera*. D'altra parte, la presenza di altri segnali modulati, oltre a quello desiderato, rendono obbligatoria l'adozione di ulteriori elaborazioni, come discusso nel § 11.2.3 relativo alla demodulazione eterodina.

#### 11.2.2.1 Segnali a banda laterale unica e *ridotta*




Nel caso di segnali $BLU$, il segnale modulante può essere riottenuto a partire da $x(t)$ utilizzando un demodulatore omodina, in quanto la componente in fase $x_c(t)$ dell'inviluppo complesso è proprio pari al messaggio modulante $m(t)$.

In questo caso, occorre prestare particolare attenzione ad eventuali errori di frequenza e di fase ($\Delta f$ e $\theta$) della portante di demodulazione perché, essendo presenti entrambe le componenti $x_c(t)$ ed $x_s(t)$, in uscita al demodulatore si ottiene (per il caso di banda laterale superiore):

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

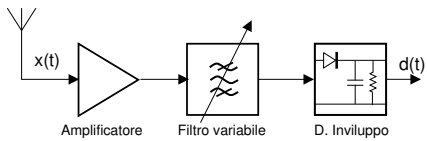
Pertanto si nota come la modulazione $BLU$ sia piú sensibile di quella $BLD$ agli errori della portante di demodulazione, in quanto ora un semplice errore di fase $\theta$ produce non solo un affievolimento, ma una vera intermodulazione tra $m(t)$ e $\widehat{m}(t)$.

Per i segnali $BLU$ a portante intera è possibile anche usare un demodulatore di inviluppo: in tal caso è necessario che l'ampiezza della portante sia sufficiente a non far invertire $x(t)$ neanche per i picchi di segnale piú ampi. L'analisi di questa esigenza determina una efficienza inferiore a quella del caso $BLD$.

Anche nel caso $BLR$, è possibile ricorrere ad un demodulatore di tipo omodina, purchè il filtro $H(f)$ usato in trasmissione per rimuovere parte di una banda laterale presenti alcune condizioni di simmetria attorno a $f_0$[19].

### 11.2.3 Demodulatore eterodina

Con questo nome si indica l'uso di una frequenza di demodulazione *differente* da quella della portante, particolarmente idonea (ma non solo) alla ricezione di una tra diverse trasmissioni operate a frequenze vicine tra loro, come nel caso della diffusione broacast (vedi § 10.1.1.1).

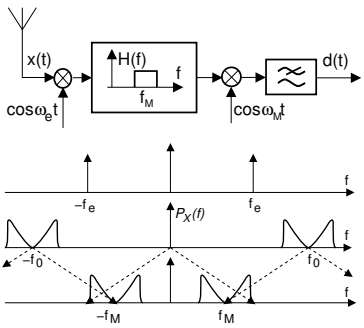



Nel caso in cui si voglia usare un demodulatore di inviluppo (a patto che le emittenti trasmettano a portante intera) occorrerebbe *sintonizzare* la stazione con un filtro passa

[18] Mentre la frequenza di taglio superiore del filtro complessivo può assumere un qualunque valore tra $f_0$ e la massima frequenza di $m(t)$, la frequenza inferiore dev'essere minore di quella minima del segnale. Perciò il metodo non è adatto, nel caso in cui $m(t)$ abbia componenti energetiche prossime a frequenza zero.

[19] 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.

banda variabile (rappresentato in figura da una freccia), la cui realizzazione a radio frequenza può presentare difficoltà non trascurabili[20]. D'altro canto, l'adozione di un demodulatore omodina, pur se elimina il problema del filtro variabile e della sua banda frazionaria, introduce quello di dover generare una portante di demodulazione sincrona a quella della trasmissione desiderata, essendo la realizzazione di oscillatori variabili e di precisione, via via più problematica all'aumentare delle frequenze in gioco.

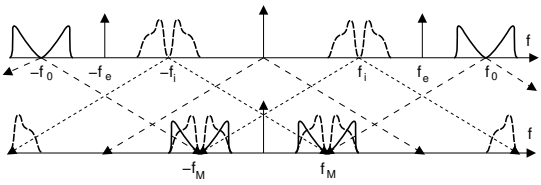



Si ricorre allora ad un diverso schema che potremmo definire *in due passi*: volendo sintonizzare l'emittente con portante $f_0$, il segnale ricevuto viene innanzitutto moltiplicato per una portante *eterodina* $f_e \neq f_0$, ed in particolare $f_e = f_0 - f_M$, ottenendo l'effetto di centrare la frequenza zero dei segnali in ingresso, alle frequenze $\pm f_e$. Allo stesso tempo, la componente a frequenze positive $\mathcal{P}_{x^+}(f)$ centrata in $f_0$ si trasla in $f_0 \pm f_e$, cosí come la componente a frequenze negative $\mathcal{P}_{x^-}(f)$ centrata in $-f_p$ si riloca in $-f_0 \pm f_e$.

Dato che $f_0 - f_e = f_M$, il risultato ottenuto è immesso in un filtro *passa banda*, centrato proprio sulla cosiddetta *media frequenza* (MF) $f_M$ prefissata, in modo che alla sua uscita è ora presente solo l'emittente desiderata, ossia quella per la quale risulta $f_0 = f_M + f_e$, e che potrà essere demodulata da un demodulatore *fisso* di precisione. Per sintonizzare una diversa emittente a frequenza $f_0'$, è sufficiente porre $f_e' = f_0' - f_M$, mentre il resto del ricevitore (con i suoi amplificatori e filtri) opera su di un segnale centrato sempre alla stessa media frequenza $f_M$, indipendentemente dall'emittente.

#### 11.2.3.1 Frequenze immagine

L'uso di un ricevitore eterodina necessita della presenza di un ulteriore filtro in ingresso al ricevitore, tale da impedire che si verifichi il problema seguente. Accade infatti che il filtro a media frequenza si ritrova nella propria banda, oltre alle emittenti centrate a $\pm f_0 = \pm(f_e + f_M)$, anche le emittenti alle portanti $\pm f_i = \pm(f_e - f_M)$, per le quali 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 $f_e$ provoca la traslazione entro la MF sia della stazione desiderata centrata in $f_0 = f_e + f_M$, che della sua immagine a distanza $2f_M$, centrata in $f_i = f_e - f_M$. Pertanto, a monte del mixer va anteposto un filtro, che elimini dal segnale di ingresso le frequenze immagine, ovvero, una volta nota la gamma di frequenze che si vuol sintonizzare, tutte quelle a distanza $2f_M$ dalla banda di interesse.

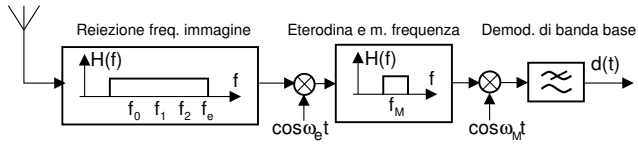
[20] Le difficoltà nascono sia dall'esigenza di *accordare* il filtro attorno alla frequenza portante desiderata, sia dalla necessità di attenuare sufficientemente le trasmissioni che avvengono su frequenze limitrofe, determinando la necessità di realizzare un filtro con regione di transizione molto ripida, problema che può divenire insormontabile se il rapporto tra banda del segnale e portante (la cosiddetta *banda frazionaria*) è particolarmente ridotto.

**Esempio** La maggior parte dei ricevitori di trasmissioni AM Broadcast (540-1600 KHz) utilizza un ricevitore detto *supereterodina*, con $f_e > f_0$ anziché $f_e < f_0$ come discusso sopra[21], e $f_M = 455$ KHz. Volendo ad esempio sintonizzare una stazione con $f_p = 600$ KHz, occorre una $f_e = f_M + f_p = 1055$ KHz. Ma allo stesso tempo anche l'emittente a portante $f_i = f_e + f_M = 1510$ KHz viene traslata nella stessa banda del filtro MF. Dunque, prima del mixer occorre un filtro che elimini le stazioni centrate su portanti $f_i > f_e$.

La scelta $f_M = 455KHz$, inferiore alla minima frequenza di 510 KHz, permette di utilizzare una regione di frequenze libera da altre trasmissioni, che altrimenti potrebbero essere amplificate dagli stadi ad alto guadagno posti dopo il filtro MF. La scelta di $f_e > f_0$ permette di posizionare il *filtro di reiezione* delle frequenze immagine al disotto della $f_e$, rendendo più semplice la realizzazione del filtro.

#### 11.2.3.2 Supereterodina

La figura seguente mostra lo schema finale del ricevitore con $f_e > f_0$.




Elenchiamo di seguito i vantaggi conseguiti:

- la sintonia avviene mediante la variazione di $f_e$, ed il resto non cambia;
- la distanza tra $f_0$ ed $f_M$ scongiura il rischio di instabilità, che potrebbe verificarsi se parte del segnale uscente dal filtro di media frequenza, amplificato, fosse ricaptato dallo stadio di ingresso, mentre ora invece l'amplificazione può aver luogo proprio nello stadio a media frequenza;
- la scelta di $f_M$ è una opportunità di progetto, che consente di realizzare il filtraggio della emittente desiderata a frequenza sufficientemente bassa da non porre grossi problemi;
- lo stadio di eterodina può essere ulteriormente ripartito in due conversioni di frequenza successive (vedi ad es. la fig. 15.7 a pag. 389), di cui la seconda conversione opera la sintonia, mentre la prima ha il solo scopo di translare le frequenze in gioco in un intervallo più basso, più idoneo ad esempio alla sua trasmissione via cavo.

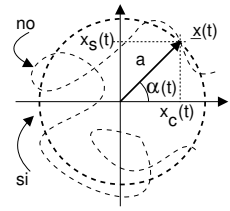
## 11.3 Modulazione angolare

In questo tipo di modulazione, l'informazione contenuta nel messaggio $x(t)$ è impressa sulla portante modificandone la fase: $x(t) = a\cos(\omega_0 t + \alpha(t))$. Si ottiene allora un inviluppo complesso

$$\begin{aligned} \underline{x}(t) &= 2x^+(t)\,\mathrm{e}^{-j\omega_0 t} = 2\cdot a\frac{1}{2}\mathrm{e}^{j\omega_0 t}\mathrm{e}^{j\alpha(t)}\cdot\mathrm{e}^{-j\omega_0 t} = a\mathrm{e}^{j\alpha(t)} \\ &= a\left[\cos\alpha(t) + j\sin\alpha(t)\right] = x_c(t) + jx_s(t) \end{aligned}$$

[21] Scegliere $f_e < f_0$ oppure $f_e > f_0$ praticamente ha l'effetto di scambiare i ruoli tra la portante desiderata, e la sua immagine. In particolare, il risultato non cambia se la modulazione è a banda laterale doppia, ma dato che nel secondo caso, $X^+(f)$ si rialloca sull'asse delle frequenze negative (ed il contrario per $X^-(f)$), per la modulazione BLU o BLR questo fenomeno di inversione delle bande laterali deve essere tenuto in conto nel demodulatore omodina finale.

Notiamo subito che, a differenza della AM, il modulo di $\underline{x}(t)$ è rigorosamente costante e pari ad $a$, mentre la fase $\alpha(t)$ varia continuamente. Si è già mostrato come sia possibile definire 2 diversi tipi di legame tra messaggio $m(t)$ e fase dell'inviluppo complesso $\alpha(t)$, indicati con PM (modulazione *di fase*) ed FM (modulazione di *frequenza*), e riassunti nella tabella che segue.




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

In particolare, le due modulazioni sono espresse anche nei termini della *frequenza istantanea*, che è definita come la derivata della *fase istantanea* $\psi(t) = 2\pi f_0 t + \alpha(t)$:

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

Le due alternative (PM e FM) sono analizzate assieme, in quanto reciprocamente intercambiabili qualora si effettui

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

Illustriamo subito alcune particolarità della modulazione angolare, prima di applicarci al problema della ricezione, ed alla determinazione della densità di potenza del segnale modulato.

**Non linearità** La caratteristica *fondamentale* della modulazione angolare è che il segnale modulato dipende da $m(t)$ in modo fortemente *non lineare,* e pertanto lo spettro di densità di potenza $\mathcal{P}_x(f)$ del segnale modulato non può essere calcolato con le tecniche tradizionali. Infatti, l'inviluppo complesso di un segnale modulato angolarmente può essere espresso[22] come:

$$\underline{x}(t) = a\mathrm{e}^{j\alpha(t)} = a\left[1 + j\alpha(t) - \frac{\alpha^2(t)}{2} - j\frac{\alpha^3(t)}{3!} + ...\right] \tag{11.1}$$

da cui risulta evidente che, anche se $\mathcal{P}_\alpha(f)$ è esprimibile a partire da $\mathcal{P}_M(f)$, nulla può essere detto in generale per $\mathcal{P}_{\underline{x}}(f)$ (e dunque per $\mathcal{P}_x(f) = \frac{1}{4}\mathcal{P}_{\underline{x}}(f - f_0) + \frac{1}{4}\mathcal{P}_{\underline{x}}(f + f_0)$). La presenza delle potenze dell'angolo $\alpha(t)$ infatti, impedisce l'applicabilità del principio di sovrapposizione degli effetti, ovvero, anche se sono noti i risultati della modulazione per due diversi messaggi $x_1(t) = FM\{m_1(t)\}$, $x_2(t) = FM\{m_2(t)\}$, il risultato ottenibile modulando la loro somma, non è quello della somma dei risultati individuali: $FM\{m_1(t) + m_2(t)\} \neq FM\{m_1(t)\} + FM\{m_2(t)\}$.

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

**Ampiezza costante** La circostanza che $\underline{x}(t) = ae^{j\alpha(t)}$ presenti un modulo costante pari ad $a$, indipendentemente dalle ampiezze del segnale modulante, è particolarmente utile, qualora per $m(t)$ siano da aspettarsi forti variazioni di dinamica. Questo è proprio il caso del segnale FDM (pag. 239), utilizzato per trasmettere piú canali telefonici[23]. In questo caso infatti, 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 con una diversa portante) può variare di molto: allora, il segnale complessivo $y(t)$ viene applicato all'ingresso di un modulatore FM e trasmesso come tale.

**Generazione di un segnale a modulazione angolare** Come anticipato, per effettuare una modulazione PM $x(t) = k_\phi m(t)$ si può usare un modulatore FM, in cui $\alpha(t) = 2\pi k_f \int_{-\infty}^{t} m'(\tau)d\tau$, ponendo $m'(t) = \frac{1}{2\pi}\frac{k_\phi}{k_f}\frac{d}{dt}m(t)$. Pertanto, consideriamo nel seguito solo le operazioni di modulazione/demodulazione FM. Un metodo *diretto* di generare un segnale FM è quello di utilizzare un VCO (già introdotto al § 11.2.1.3), ossia un oscillatore controllato in tensione, che produce il segnale $x(t) = a\sin\left(\omega_0 t + 2\pi k_f \int_{-\infty}^{t} m(\tau)d\tau\right)$, e dunque realizza proprio la funzione desiderata. Un secondo metodo verrà illustrato per un caso particolare in appendice § 11.4.5. Infine, è sempre valido il modulatore in fase e quadratura, in cui si pone $x_c(t) = \cos\alpha(t)$ e $x_s(t) = \sin\alpha(t)$.

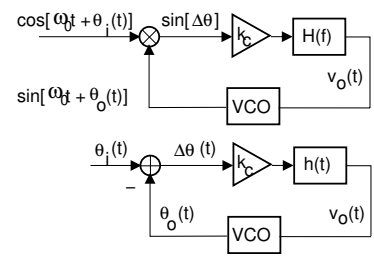
### 11.3.1 Ricezione di un segnale a modulazione angolare

In linea di principio, una volta ottenute $x_c(t)$ ed $x_s(t)$ del segnale modulato (ad esempio mediante un demodulatore in fase e quadratura) è sempre valida la relazione $\alpha(t) = \arctan\frac{x_s(t)}{x_c(t)}$. D'altra parte, tale soluzione si presta esclusivamente a realizzazioni digitali, in quanto è difficile realizzare un dispositivo che presenti esattamente la relazione non lineare di tipo arcotangente. Illustriamo allora i due metodi piú comunemente usati:

#### 11.3.1.1 Ricevitore a PLL

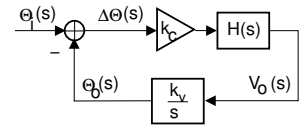
Al § 11.2.1.3 si è già mostrato l'uso del circuito PLL per l'aggancio della fase della portante di modulazione. Lo stesso schema può essere usato per inseguire l'andamento temporale della fase di una portante modulata angolarmente, ottenendo in tal modo l'informazione desiderata.

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




[23] Un altro caso di multiplex FDM è quello del downlink di un trasponder DVB-S, introdotto al § 15.5.2

L'analisi di Laplace fornisce allora il risultato




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

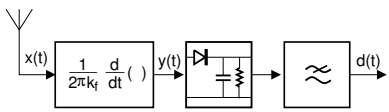
che, antitrasformato, permette di esprimere $\theta_o(t)$
(fase del VCO) come una versione filtrata della fase della portante modulata $\theta_i(t)$, da parte della funzione di trasferimento ad anello chiuso

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

Inoltre, dato che il VCO produce $\theta_o(t) = k_v \int_{-\infty}^{t} v_o(\tau) d\tau$, si riconosce subito che l'uscita $v_o(t)$ del filtro di loop $H(s)$ corrisponde alla ricostruzione del messaggio modulante $m(t)$ nel caso di modulazione FM. Pertanto, l'uscita del filtro di loop del PLL realizza la demodulazione di frequenza.

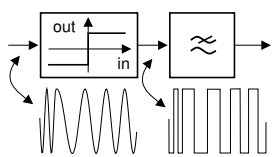
#### 11.3.1.2 Ricevitore a discriminatore

E' realizzato mediante il circuito di figura, in cui il derivatore effettua una conversione FM-AM, di cui viene demodulato l'inviluppo di ampiezza. Infatti, la grandezza $y(t)$ risulta pari a




$$\begin{aligned} y(t) &= \frac{1}{2\pi k_f} \frac{d}{dt} a \cos\left(\omega_0 t + 2\pi k_f \int_{-\infty}^{t} m(\tau) d\tau\right) = \\ &= \frac{1}{2\pi k_f} \left(2\pi f_0 + 2\pi k_f m(t)\right) a \sin\left(\omega_0 t + 2\pi k_f \int_{-\infty}^{t} m(\tau) d\tau\right) \end{aligned}$$

e corrisponde dunque ad un segnale modulato sia angolarmente che in ampiezza, di ampiezza $a(t) = a\left(\frac{f_0}{k_f} + m(t)\right)$. Siamo dunque in presenza di una modulazione di ampiezza *BLD-PI* ( § 11.1.1.2) e quindi, con una scelta opportuna[24] di $\frac{f_0}{k_f}$, il messaggio $m(t)$ può essere estratto mediante un demodulatore d'inviluppo ( § 11.2.2).




Il risultato ottenuto è valido purchè $x(t)$ sia privo esso stesso di variazioni di ampiezza: per questo, spesso il derivatore è preceduto da un blocco *squadratore*, che produce una versione, appunto, "squadrata" del segnale ricevuto e quindi priva di modulazioni di ampiezza. Essendo lo squadratore fortemente non lineare, in uscita saranno presenti, oltre al segnale originario, anche componenti centrate a frequenze multiple di quella della portante, che vengono rimosse dal filtro passa basso a valle dello squadratore.

[24] Per utilizzare il demodulatore inviluppo, deve risultare sempre $\frac{f_0}{k_f} + m(t) > 0$, e dunque occorre scegliere $\frac{f_0}{k_f} > \max_t \{|m(t)|\}$.

### 11.3.2 Densità spettrale di segnali modulati angolarmente

Riprendiamo la relazione (11.1) che esprime l'inviluppo complesso di un segnale modulato angolarmente nei termini di una serie di potenze:

$$\underline{x}(t) = ae^{j\alpha(t)} = a\sum_{n=0}^{\infty}\frac{[j\alpha(t)]^n}{n!} \tag{11.2}$$

Innanzitutto osserviamo che la potenza totale di $\underline{x}(t)$ vale sempre $\mathcal{P}_{\underline{x}} = a^2$, indipendentemente da $\alpha(t)$, e dunque $\mathcal{P}_x = \frac{a^2}{2}$. Per ciò che riguarda $\mathcal{P}_{\underline{x}}(f)$, in linea di principio non si potrebbe neanche affermare che $\underline{x}(t)$ sia limitato in banda, vista la presenza delle potenze di qualunque ordine di $\alpha(t)$. D'altro canto, la presenza dei fattoriali a denominatore fa sí che la serie possa essere troncata ad un certo ordine $\nu < \infty$. Se poniamo ora $\alpha(t) = k_\phi m(t)$, osserviamo che quanto piú $|k_\phi m(t)|$ è piccolo, tanto prima può essere troncata, con errori trascurabili. In particolare, se $\alpha(t)$ si mantiene sempre *molto piccolo*, la (11.2) può essere troncata al primo termine ($n = 1$), dando luogo ad un comportamento praticamente lineare.

Se invece $\alpha(t)$ assume valori *molto elevati*, e quindi (11.2) 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 $f_0$ completamente dipendenti dalle ampiezze di $m(t)$, e quindi l'andamento della densità di potenza $\mathcal{P}_x(f)$ risulta strettamente dipendente da quello della densità di probabilità di $p_M(m)$ che descrive le ampiezze di $m(t)$.

Per valori intermedi della dinamica di $\alpha(t)$, invece, la $\mathcal{P}_x(f)$ risultante sarà una via di mezzo tra i due casi estremi discussi, che pertanto possono essere pensati come casi limite tra cui porre la densità di potenza effettiva.

Come anticipato, la natura non lineare della modulazione angolare rende necessario studiare ogni caso individualmente; pertanto la determinazione di $\mathcal{P}_x(f)$ viene svolta per due casi particolari, considerando per questi le due possibilità estreme di $\alpha(t)$ molto piccolo o molto grande, ed i risultati estrapolati per approssimare altre situazioni; i due casi esaminati sono:

- $m(t)$ sinusoidale e
- $m(t)$ membro di un processo stazionario ergodico.

#### 11.3.2.1 Segnale modulante sinusoidale

Ponendo $m(t) = \cos(2\pi wt)$, si ottiene che la fase modulante $\alpha(t)$ e la frequenza istantanea $f_i(t)$ per i due casi PM ed FM, relativi al segnale $x(t) = a\cos(2\pi f_0 t + \alpha(t))$, risultano:

| | $\alpha(t)$ | $f_i(t)$ | $\Delta\alpha$ | $\Delta f$ |
|---|---|---|---|---|
| **PM** | $k_\phi \cos(2\pi wt)$ | $f_0 + wk_\phi \sin(2\pi wt)$ | $k_\phi$ | $wk_\phi$ |
| **FM** | $2\pi k_f \int_{-\infty}^{t} m(\tau)d\tau = \frac{k_f}{w}\sin(2\pi wt)$ | $f_0 + k_f \cos(2\pi wt)$ | $\frac{k_f}{w}$ | $k_f$ |

in cui si è anche indicata la massima deviazione di fase $\Delta\alpha = \max\{|\alpha(t)|\}$ e di frequenza $\Delta f = \max\{|f_i(t) - f_0|\}$. Notiamo subito che, in entrambi i casi, sia la fase $\alpha(t)$ che la frequenza istantanea $f_i(t)$ variano sinusoidalmente con periodo $\frac{1}{w}$; nel caso PM l'entità di $\Delta f$ aumenta con $w$, mentre nell'FM la $\Delta\alpha$ diminuisce con $w$. Nel

seguito si farà riferimento all'*indice di modulazione angolare* $\beta$, corrispondente alla massima escursione della fase $\Delta\alpha$, che risulta:

$$\beta = \begin{cases} k_\phi & \text{(PM)} \\ \frac{k_f}{w} & \text{(FM)} \end{cases}$$

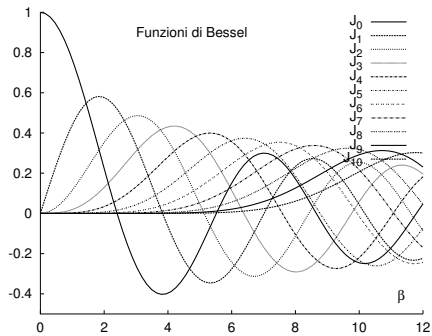
Con questa convenzione, possiamo trattare congiuntamente entrambi i casi PM ed FM riscrivendo l'inviluppo complesso come [25]

$$\underline{x}(t) = a\mathrm{e}^{j\beta\sin(2\pi wt)}$$

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

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

ovvero sono pari[26] alle *funzioni di Bessel* del primo tipo, ordine $n$ ed argomento $\beta$. Queste hanno l'andamento mostrato in figura, assieme alle proprietà che le caratterizzano:




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

e quindi i valori di $X_n$ si ottengono tracciando una linea verticale nel diagramma di figura in corrispondenza del valore $\beta$ adottato, e individuando il valore di ciascuna $\mathcal{J}_n$ per quel $\beta$. Osserviamo ora che l'ultima proprietà mostra come, in presenza di un valore di $\beta$ elevato, le funzioni di Bessel di ordine $n > \beta$ siano praticamente nulle: è quindi lecito in tal caso limitare lo sviluppo in serie di Fourier di $\underline{x}(t)$ ai primi $\beta$ termini (positivi e negativi), ovvero: $\underline{x}_{FM}(t) \simeq a\sum_{n=-\beta}^{\beta} \mathcal{J}_n(\beta)\mathrm{e}^{j2\pi nwt}$. Pertanto, il segnale modulato $x(t) = \Re\left\{\underline{x}(t)\mathrm{e}^{j\omega_0 t}\right\}$ risulta:

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

e quindi lo spettro di densità di potenza di $x(t)$ ha espressione

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

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

[26] Le *funzioni di Bessel* del primo tipo, ordine $n$ ed argomento $\beta$ sono definite come $\mathcal{J}_n(\beta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{j(\beta\sin x - nx)}dx$.

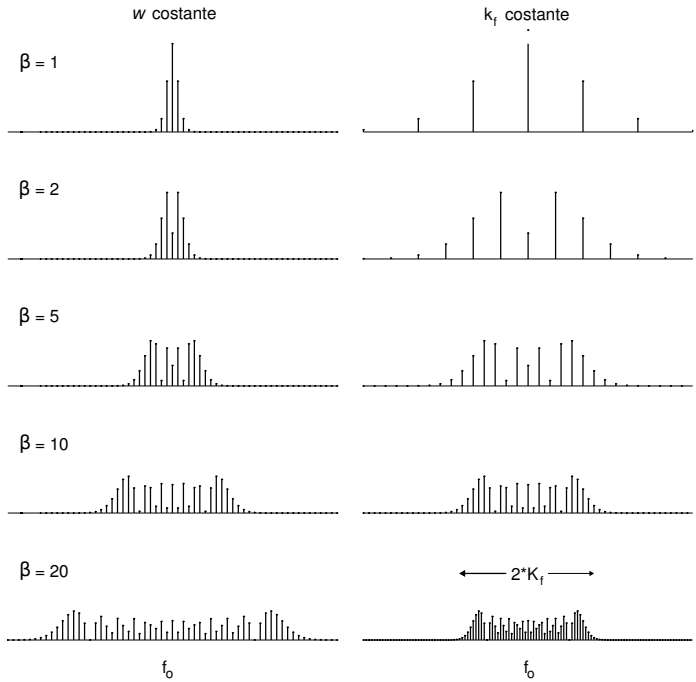



Figura 11.1: Spettro di ampiezza per segnale FM a modulazione sinusoidale

ed è formato da impulsi centrati a frequenze $f = \pm f_0 \pm nw$ ([27]).

La fig. 11.1 mette a confronto $|X(f)|$ per $f$ vicino ad $f_0$, ovvero mostra $\left|X^+(f)\right| = \frac{a}{2}\sum_{n=-\beta}^{\beta}|\mathcal{J}_n(\beta)|\,\delta(f - f_0 - nw)$, calcolato per diversi valori di $\beta$, mantenendo fisso $w$ oppure $k_f$ (a sinistra e destra rispettivamente), e ci aiuta a comprendere i ragionamenti che seguono.

**Modulazione a basso indice** In questo caso $\beta \ll 1$, e tale da rendere trascurabili le funzioni di Bessel $\mathcal{J}_n(\beta)$ con $n > 1$. Allora, $x(t)$ occupa una banda pari a $2w$, in modo del tutto simile all'AM-BLD.

**Modulazione ad alto indice** In tal caso sono presenti piú funzioni di Bessel, ed il comportamento non lineare tende a legare $\mathcal{P}_x(f)$ ai valori assunti da $f_i(t) = k_f m(t)$, realizzando una conversione ampiezza $\rightarrow$ frequenza.

In Fig. 11.1 sono evidenziati gli effetti dell'aumento di $\beta = \frac{k_f}{w}$ nelle due circostanze:

1. Si mantiene $w$ fisso, aumentando $k_f$. Il numero di righe spettrali a frequenza $f_0 \pm nw$ aumenta, occupando una banda crescente, e per $\beta$ molto grande si verifica che $\mathcal{J}_n(\beta) \simeq 0$ per $n > \beta$. Pertanto, la banda occupata tende a $B = 2\beta w = 2\frac{k_f}{w} \cdot w = 2k_f$;

[27] Integrando l'espressione di $\mathcal{P}_x(f)$, e ricordando che $\sum_{n=-\infty}^{+\infty}\mathcal{J}_n^2(\beta) = 1$, si ottiene ancora un risultato già noto, e cioè che la potenza totale del segnale modulato risulta pari a quella della portante non modulata, e pari a $\mathcal{P}_x = \frac{a^2}{2}$, indipendentemente da $\beta$.

2. Si mantiene $k_f$ fisso, diminuendo $w$. La banda occupata tende a ridursi, mentre le nuove righe spettrali a frequenza $f_0 \pm nw$ si infittiscono. Per $\beta \to \infty$, la spaziatura $w$ tra le righe spettrali tende ad annullarsi, producendo una $\mathcal{P}_x(f)$ praticamente continua, e con una banda $B = 2k_f$, ossia pari alla massima deviazione di frequenza istantanea $\Delta f$.

Notiamo che in entrambi i casi, all'aumentare di $\beta$ la $\mathcal{P}_x(f)$ tende ad assumere la densità di ampiezza tipica del processo armonico, descritta dalla (7.2) a pag. 136.

**Regola di Carson** Come mostrato, nei due casi a basso ed alto indice, la banda occupata da $x(t)$ varia tra $2w$ e $2k_f$ rispettivamente. Nei casi intermedi, è pratica comune ricorrere all'espressione

$$B_C \simeq 2(k_f + w) = 2w(\beta + 1) \tag{11.3}$$

nota come *Regola di Carson*[28], che tiene conto di entrambi i fattori che concorrono alla determinazione della banda, e che fornisce i valori esatti sia per $\beta \ll 1$, che per $\beta \to \infty$, in entrambi i casi in cui $k_f \to \infty$ o $w \to 0$.

Sebbene la determinazione approssimata della banda mediante la *Regola di Carson* sia stata ottenuta nel caso di $m(t) = \cos(2\pi wt)$, la stessa espressione è spesso adottata come una buona approssimazione anche per segnali non sinusoidali, ma limitati in banda tra $-W$ e $W$, e contraddistinti da una $\Delta f = k_f \cdot \max\{|m(t)|\}$. In tal caso, la regola di Carson si applica ponendo ora $B_C \simeq 2W(\beta + 1)$ con $\beta = \frac{\Delta f}{W}$. Per un approfondimento della questione, si veda l'appendice 11.4.6 e la sottosezione 11.3.3.1.

A prima vista, l'estensione del risultato per $m(t) = \cos(2\pi wt)$ al caso qualunque appare piú che ragionevole; il comportamento non lineare della modulazione angolare impedisce però una sua verifica analitica. D'altra parte, i risultati sperimentali mostrano che l'approssimazione fornita dalla banda di Carson può effettivamente costituire una stima plausibile della banda occupata per segnali modulanti qualsiasi.

In base alla regola di Carson, notiamo ora che la banda occupata dal segnale modulato può risultare $\beta+1$ volte piú estesa di quella ottenibile mediante modulazione AM. Nonostante questo aumento di banda possa apparire un fatto negativo, vedremo nel capitolo 12 che ciò consente un migliore SNR dopo la demodulazione rispetto al caso AM. Al contrario, se $\beta \ll 1$, il comportamento si avvicina molto a quello lineare (vedi appendice 11.4.5).

### 11.3.3 Densità spettrale FM con processo aleatorio modulante

Riprendiamo il ragionamento iniziato in 11.3.2, relativo all'influenza di $p_M(m)$ su $\mathcal{P}_x(f)$. Considerando che la frequenza istantanea ha espressione $f_i = f_0 + k_f m(t)$, la frazione di potenza tra $f_1$ ed $f_2$ sarà pari alla frazione di tempo che il segnale $m(t)$ si trova tra $m_1 = \frac{f_1 - f_0}{k_f} \leq m(t) \leq m_2 = \frac{f_2 - f_0}{k_f}$. Nel caso in cui $m(t)$ sia sinusoidale, con fase iniziale aleatoria a distribuzione uniforme, $m(t)$ è una realizzazione di un processo armonico, e la frazione di tempo su indicata equivale alla $Prob\{m_1 \leq m(t) \leq m_2\}$. Pertanto le righe spettrali, addensandosi, tendono a disporsi in accordo all'andamento della densità $p_M(m)$[29].

[28] J.R. Carson fu uno dei primi ad investigare l'FM negli anni '20.

[29] In particolare, per $\beta \to \infty$ risulterà $\mathcal{P}_{\underline{x}}(f) = \frac{a^2}{1-\left(f/k_f\right)^2}$, che è l'andamento a cui tendono (per $\beta \to \infty$) i grafici in basso di fig. 11.1.

Il risultato a cui siamo pervenuti nel caso di modulante sinusoidale è generale, e pertanto si può affermare che qualora si generi un segnale FM ad alto indice, a partire da un processo con densità di probabilità nota, lo spettro di densità di potenza del segnale modulato acquisisce l'andamento proprio della densità di probabilità del processo modulante, indipendentemente dal suo spettro di densità di potenza.

La conclusione riportata si mantiene valida purchè $\beta \gg 1$; nel caso contrario, sono validi i ragionamenti sviluppati alla sezione 11.3.3.2.

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

Notiamo subito che la frequenza istantanea $f_i$ rimane limitata tra $f_0 - \frac{\Delta_M}{2}k_f$ e $f_0 + \frac{\Delta_M}{2}k_f$. Inoltre, la potenza totale deve risultare ancora pari a $\frac{a^2}{2}$. Pertanto si ottiene[30]:

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

### 11.3.3.1 Indice di modulazione per processi

Ai fini dell'applicazione della regola di Carson, si è posto l'indice di modulazione $\beta = \frac{\Delta f}{W}$, con $\Delta f = k_f \max\{|m(t)|\}$. Nel caso di processi, può accadere che $m(t)$ non sia limitata in ampiezza, come ad esempio nel caso gaussiano, rendendo problematica la quantificazione di $\beta$. Per risolvere la questione, l'indice di modulazione $\beta$ è ridefinito ancora una volta, e nel caso in cui $m(t)$ sia un generico processo si pone

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

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[31], e $\sigma_\alpha = k_\phi\sqrt{\mathcal{P}_m}$ è la deviazione standard della fase modulante[32]. L'applicazione della regola di Carson con il nuovo valore di $\beta'$, fornisce per la banda un risultato che non indica piú la banda *totale* occupata, ma individua una *banda efficace* entro cui $\mathcal{P}_x(f)$ è in larga parte (ma non completamente) contenuta (vedi anche 11.4.6).

Nel caso in cui *non* risulti $\beta \gg 1$, lo spettro di potenza del segnale modulato FM torna a dipendere da quello del segnale modulante, e si ricade nella trattazione che segue.

---

[30] 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 11.3.3.1) si sarebbe ottenuto $B_C = 2W(\beta+1) \simeq 2\frac{\sigma_f}{W}W = 2\sigma_f = 2k_f\sqrt{\mathcal{P}_M} = 2k_f\sqrt{\frac{\Delta_M^2}{12}} = 2k_f\frac{\Delta_M}{2\sqrt{3}} = \frac{\Delta_M k_f}{\sqrt{3}}$, un risultato che è circa pari a 0.58 volte quello precedente. Data le particolarità di $p_M(m)$ uniforme, in questo caso è da preferire il primo risultato.

[31] 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.

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

#### 11.3.3.2 Modulazione a basso indice

Ora l'indice di modulazione $\beta$ si assume piccolo a sufficienza, da far sí 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 $\sigma_\alpha^2$ la varianza della fase modulata e con $\mathcal{P}_\alpha(f)$ il relativo spettro di densità di potenza, pari rispettivamente a

| | $\mathcal{P}_\alpha(f)$ | $\sigma_\alpha^2$ |
|---|---|---|
| **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 $k_\phi$ (o $k_f$) tende a zero, $\mathcal{P}_{\underline{x}}(f)$ si riduce ad un impulso, corrispondente alla portante non modulata. All'aumentare di $k_\phi$ (o $k_f$), aumenta anche $\sigma_\alpha^2$ 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, rappresentati da $\mathcal{P}_\alpha(f)$ e delle sue *autoconvoluzioni*. E' immediato notare come, al crescere di $k_\phi$ (o $k_f$), cresca la banda.

In appendice 11.4.5 è illustrata una tecnica di modulazione per segnali FM modulati a basso indice.

## 11.4 Appendici

### 11.4.1 Calcolo della potenza di un segnale AM BLU

Mostriamo che se $X_{BLU}(t) = \frac{k_a}{\sqrt{2}}(m(t)\cos\omega_0 t - \widehat{m}(t)\sin\omega_0 t)$, allora $\mathcal{P}_x = \frac{k_a^2}{2}\mathcal{P}_m$. Possiamo innanzitutto scrivere che

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

in quanto le componenti a frequenza positiva e negativa di $x(t)$ sono ortogonali (infatti $\int_{-\infty}^{\infty} X^+(f) X^-(f)\, df = 0$), e lo spettro di densità di potenza è una funzione pari della frequenza: $\mathcal{P}_x(f) = \mathcal{P}_x(-f)$. Inoltre, invertendo la relazione $\mathcal{P}_{\underline{x}}(f) = 4\mathcal{P}_{x^+}(f - f_0)$ valida per la densità di potenza dell'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 prima relazione mostrata, fornisce $\mathcal{P}_x = 2\mathcal{P}_{x^+} = \frac{1}{2}\mathcal{P}_{\underline{x}}$.

Nel caso AM-BLU si ha inoltre $\underline{x}(t) = \frac{k_a}{\sqrt{2}}[m(t) + j\widehat{m}(t)]$, tenendo ora conto che $\mathcal{P}_{\underline{x}} = \mathcal{R}_{\underline{x}}(0)$, si ottiene

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

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 ma intuibile). Pertanto si ottiene

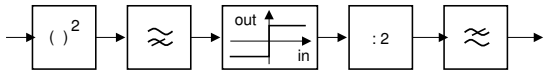
$$\mathcal{P}_x = \frac{1}{2}\frac{k_a^2}{2}\left[\mathcal{P}_m + \mathcal{P}_m\right] = \frac{1}{4}k_a^2 \cdot 2\mathcal{P}_m = \frac{k_a^2}{2}\mathcal{P}_m$$

#### 11.4.1.1 Calcolo della potenza di segnali BLD-PI, PS, PPS

Vale lo stesso procedimento adottato sopra, in cui ora

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

### 11.4.2 Ricostruzione della portante mediante quadratura




Nel caso di una trasmissione BLD-PS, la portante di demodulazione può essere ottenuta mediante la seguente elaborazione. Il segnale $x(t) = m(t)\cos(\omega_0 t + \varphi)$ viene elevato al quadrato producendo

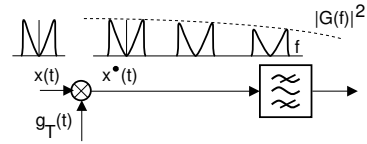
$$\frac{1}{2}m^2(t)\left[1 + \cos(2\omega_0 t + 2\varphi)\right]$$

il cui termine di banda base $\frac{1}{2}m^2(t)$ viene rimosso dal filtro passa alto. Successivamente, lo squadratore produce un'onda quadra a frequenza $2f_0$ che viene divisa per 2 da un contatore binario. Infine, un passa basso provvede ad eliminare le armoniche dell'onda quadra a frequenza $nf_0$, fornendo cosí la portante desiderata, a meno di una ambiguità di segno.

### 11.4.3 Il mixer

Il dispositivo moltiplicatore, presente negli schemi di mo-demodulazione, viene anche chiamato *mixer*, in quanto miscela tra loro due segnali.

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




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

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

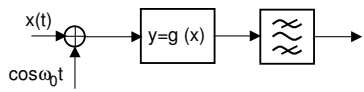
$$\mathcal{P}_{G_T}(f) = |G(f)|^2 \cdot \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{k}f_0\right)$$

Il prodotto di tale segnale per $x(t)$, produce un $x^{\bullet}(t)$ con densità di potenza

$$\mathcal{P}_{x^{\bullet}}(f) = \mathcal{P}_x(f) * \mathcal{P}_{G_T}(f) = \frac{|G(f)|^2}{T}\sum_{n=-\infty}^{\infty} \mathcal{P}_{G_T}\left(f - \frac{n}{k}f_0\right)$$

Pertanto, il desiderato spettro di potenza si ottiene inserendo dopo il moltiplicatore un filtro passa banda centrato su $f_0$, ossia sulla $k-esima$ replica spettrale di $\mathcal{P}_{x\bullet}(f)$. Lo stesso dispositivo può essere usato anche per i moltiplicatori di ricezione: in tal caso, il filtro da usare sarà un passa basso.

Un secondo metodo di realizzare il mixer è con un sommatore, un oscillatore, un dispositivo non lineare, e di nuovo un filtro passa-banda. Il dispositivo non lineare è del tipo




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

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

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

in cui, osservando che i termini $\cos^n \omega_0 t$ sono relativi a termini a frequenza $nf_0$, è possibile ancora una volta estrarre il termine che ci interessa.

### 11.4.4 Trasmissione televisiva

Illustriamo molto brevemente le principali caratteristiche delle trasmissioni televisive broadcast *analogiche*, con riferimento agli standard *dismessi* in Italia.

**Codifica dell'immagine** Una trasmissione televisiva avviene riproducendo 25 diverse immagini (dette *quadri*) al secondo. Ogni immagine è scomposta in 625 linee orizzontali, che vengono trasmesse in due fasi: prima le linee dispari, poi quelle pari. In questo modo un singolo quadro è riprodotto due volte[33] ogni $\frac{1}{25} = 0.04$ secondi (seppure in modo alternato) portando cosí a 50 semiquadri/secondo[34] la frequenza di rinfresco, in modo da impedire i fenomeni di *sfarfallamento* (FLICKER) ottico[35].

La riproduzione di un quadro avviene mediante un *tubo catodico*, il quale dispone posteriormente di un catodo che emette elettroni, che sono accelerati da un segnale di luminanza positivo applicato all'anodo, e che terminano 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 $\frac{semiquadri}{secondo}$.

**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.

[33] 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.

[34] 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.

[35] 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.

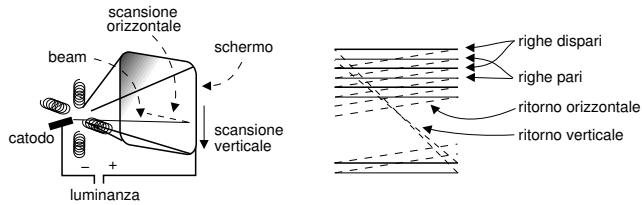



Figura 11.2: Modalità di scansione interallacciata dell'immagine televisiva

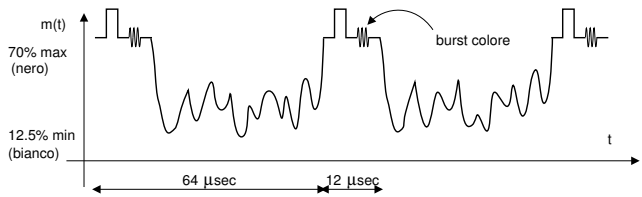



Figura 11.3: Forma d'onda del segnale televisivo

Durante la trasmissione di ogni semiquadro, ogni riga dispone di $\frac{1}{15625} = 64$ $\mu$secondi. Il segnale modulante è sempre positivo, ed associa ai valori piú piccoli la maggiore luminanza[36], trasmettendo in logica negata, in modo che gli impulsi di sincronismo orizzontale siano di ampiezza superiore al livello del nero (il 70 % dell'ampiezza). Il tempo dedicato alla trasmissione della luminanza di una riga è di 52 $\mu$sec, mentre nei restanti 12 il segnale oltrepassa il livello del nero (in modo da rendere invisibile il beam) e quindi un impulso determina il ritorno orizzontale. In figura è anche mostrato un *burst colore* che è presente nelle trasmissioni a colori per sincronizzare la *portante di colore* (vedi di seguito).

**Formato dell'immagine** Ogni singolo quadro è realizzato con un rapporto di aspetto 4:3 (che rappresenta il rapporto tra le dimensioni orizzontale e verticale), e solo 575 delle 625 linee vengono mostrate (infatti 25 linee per ogni semiquadro cadono al difuori dello schermo[37]).

**Occupazione spettrale** Diverse considerazioni[38] hanno portato a stabilire che la banda del segnale televisivo sia di circa $\pm 5$ 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 pressochè 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

[36] In questo modo si riduce mediamente la potenza trasmesa, dato che sono più frequenti scene chiare che scure.

[37] 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.

[38] 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 avere una banda minore od uguale di $\frac{f_c}{2} = 6.5$ MHz.

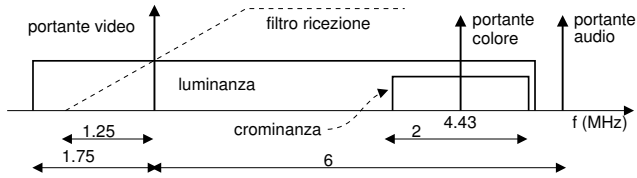



Figura 11.4: Occupazione spettrale di un segnale televisivo

modulazione di ampiezza a banda laterale ridotta, conseguendo un risparmio di banda e contemporaneamente preservando le componenti del messaggio a frequenze piú basse.

La figura 11.4 mostra la situazione, in cui solo parte (1.75 MHz) della banda inferiore del segnale di luminanza viene trasmessa, mentre il filtro di ricezione provvede a realizzare un filtraggio complessivo tale che $\underline{H}(f) + \underline{H}^*(-f) = cost$ (vedi nota 19 a pag. 267).

**Segnale audio** Nella figura 11.4 è rappresentata anche una portante audio, che viene trasmessa oltre la banda del segnale video, mediante una modulazione FM con $\Delta f_{Max} = 25$ KHz.

**Segnale di crominanza** Il requisito che piú di altri ha determinato quale soluzione adottare per effettuare trasmissioni a colori, è che queste dovessero essere correttamente visibili anche da parte di ricevitori in bianco e nero.

Un risultato di colorimetria è che ogni colore è scomponibile nella somma di tre colori fondamentali (verde, rosso e blu), effettivamente operata dagli apparati di acquisizione. La somma[39] della tre componenti fornisce il segnale di luminanza $L$, che viene utilizzato esattamente come per il bianco e nero. Il segnale di crominanza è invece costruito da una coppia di segnali differenza $\begin{cases} \Delta_R = R - L \\ \Delta_B = B - L \end{cases}$, che sono usati per modulare in ampiezza, portante soppressa, una portante di colore, usando $\Delta_R$ come componente in fase e $\Delta_B$ come componente in quadratura[40]. Una analisi più precisa, è fornita al § 18.2.2.

L'occupazione spettrale del segnale di crominanza è ridotta ($\pm 1$ MHz) rispetto a quello di luminanza, in quanto la *risoluzione spaziale* dell'occhio umano è ridotta per stimoli colorati, e quindi $\Delta_R$ e $\Delta_B$ possono variare piú lentamente di $L$.

**Sincronizzazione** Per impedire fenomeni di interferenza tra C.A. di B.F. nella ricezione del segnale di crominanza, occorre effettuare una demodulazione omodina, e l'oscillatore del ricevitore si mantiene coerente con la portante di colore, grazie ai *burst di colore* presenti dopo l'impulso di sincronizzazione orizzontale, costituiti da 8 cicli di portante. Questo segnale ha inoltre lo scopo di segnalare la *presenza* della

[39] 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.

[40] 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.

componente di crominanza: in caso contrario infatti (trasmissione B/N) il ricevitore deve disattivare il circuito del colore, per non produrre deterioramenti dell'immagine.

**Interferenza** La presenza di entrambi i segnali di luminanza e crominanza nella stessa banda sembrerebbe dare luogo a difficili problemi di interferenza. Innanzitutto osserviamo che, come anticipato, il segnale di luminanza è concentrato attorno alla portante video, e dunque arreca un disturbo ridotto[41] 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 $\mu$sec, che determina uno spettro con energia concentrata alle armoniche di $f_r = 15625$ Hz. Pertanto, si colloca la portante di colore a metà tra due armoniche del segnale di luminanza, in modo che le densità spettrali risultino, pur se sovrapposte, intercalate. L'uso di *filtri a pettine*[42] nel ricevitore può quindi ridurre notevolmente l'interferenza.

### 11.4.5 Modulazione FM a basso indice

Riprendiamo qui il caso in cui $\beta \ll 1 \Rightarrow \Delta\alpha \ll 1$, e quindi l'espansione in serie di potenze di $\underline{x}(t)$ può arrestarsi al primo ordine; se il segnale modulante è cosinusoidale, il segnale FM risulta

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

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

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

che confrontiamo con l'espressione

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

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

Il confronto rivela che mentre nell'AM il segnale modulante opera *in fase* alla portante, nell'FM a basso indice opera *in quadratura*. Il risultato esposto costituisce ad ogni modo uno *schema di modulazione* per segnali FM a basso indice, realizzabile *sommando* alla portante, un segnale modulato AM su di una portante in quadratura.

Resta il fatto che uno schema di modulazione del genere produce anche una modulazione AM parassita: quest'ultima è eliminata in ricezione dall'azione congiunta di uno squadratore e di un filtro passa basso.

### 11.4.6 FM broadcast

Illustriamo brevissimamente i parametri delle trasmissioni FM ricevibili mediante "la radio di casa". Nella banda 88-108 MHz operano le radio FM, con spaziatura di 200

[41] 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 ?

[42] Introdotto al § 9.7, l'argomento può essere approfondito presso `http://it.wikipedia.org/wiki/Filtro_comb`

KHz l'una dall'altra. Ad ogni emittente è concessa una deviazione massima della frequenza istantanea pari a $\Delta f = 75$ KHz.

Il trasmettitore viene tarato mediante un $m(t)$ sinusoidale a frequenza di 15 KHz, e $k_f$ regolato in modo da ottenere

$\Delta f = 75$ KHz

In queste condizioni, risulta $\beta = \frac{k_f}{w} = \frac{75}{15} = 5$, e la regola di Carson fornisce

$B_C = 2(k_f + w) = 2(75 + 15) = 180$ KHz

Un esame degli andamenti riportati in Fig. 11.1 mostra che per $\beta = 5$, le $\mathcal{J}_n(\beta) \neq 0$ sono le prime 8, e dunque la "*vera*" banda ha una estensione

$B = 2 \cdot 8w = 16 \cdot 15 \cdot 10^3 = 240$ 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 $\Delta f$ non è piú definita con esattezza, e conviene ricorrere alla definizione di

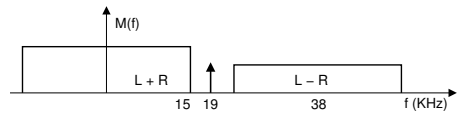
$\beta' = \frac{\sigma_f}{W} = \frac{k_f\sqrt{P_m}}{W} = \beta\sqrt{P_m} = \beta\frac{1}{\sqrt{2}} = 0.707 \cdot \beta$

a cui corrisponde una banda *efficace*

$B = 2W(\beta' + 1) = 2 \cdot 15 \cdot 10^3 \cdot (0.707 \cdot 5 + 1) = 136$ KHz.

Nell'FM stereo, il segnale trasmesso deve essere compatibile con i ricevitori mono, ed allora il segnale modulante è un segnale multiplato FDM, e "composto" da tre "canali":

- La somma di Left + Right (L+R) come segnale di banda base, che consente la compatibilità con gli apparati "mono";
- Il segnale L-R è centrato a frequenza di 38 KHz mediante modulazione AM-BLD;
- A 19 KHz, è presente una portante a cui si concede il 10% di $\mathcal{P}_M$, mentre il restante 90% di $\mathcal{P}_M$ è condiviso tra L+R e L-R. Il tutto è poi modulato FM.




La portante a 19 KHz può essere impiegata per sincronizzare il ricevitore, e generare la portante (a frequenza doppia, di 38 KHz) necessaria a demodulare il canale L-R. Se assente, indica la ricezione di un canale mono.

A prima vista, sembrerebbe che la presenza del canale L-R possa aumentare la massima deviazione di frequenza. In realtà non è cosí, per due motivi:

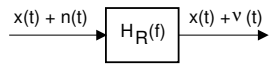
- quando L+R è grande, vuol dire che i due canali sono simili, e dunque L-R è piccolo, e viceversa;
- Il canale L-R, trovandosi a frequenze piú elevate, è caratterizzato da un indice di modulazione inferiore. Infatti, la massima deviazione di frequenza istantanea dipende dalle *ampiezze* di $m(t)$, e non dalla sua banda.

# Capitolo 12

# Prestazioni delle trasmissioni modulate

## 12.1 Il rumore nei segnali modulati

Consideriamo un segnale modulato, affetto da un rumore additivo gaussiano bianco $n(t)$, con densità di potenza




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

Prima di effettuare la demodulazione, il segnale ricevuto transita in un filtro di ricezione $H_R(f)$, che ha lo scopo di limitare la banda del rumore ricevuto, e quindi ridurre l'entità della potenza di rumore in ingresso al ricevitore. Il filtro $H_R(f)$ presenta un modulo costante nella banda del segnale, mentre tende a zero al di fuori di tale banda. In questo modo, il segnale utile $x(t)$ transita inalterato, ed il rumore $n(t)$ viene limitato in banda, producendo $\nu(t)$.

### 12.1.1 Rapporto segnale-rumore e banda di rumore

La qualità del segnale modulato ricevuto può essere descritta dal rapporto

$$SNR_{RF}(f) = \frac{\mathcal{P}_x(f)}{\mathcal{P}_\nu(f)}$$

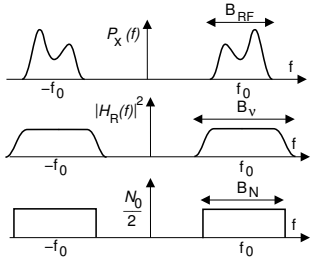
che, nel caso in cui le densità di potenza di segnale varino con la frequenza, mostra come la stessa cosa si può dire anche per il rapporto segnale rumore. Per ciò che riguarda $\mathcal{P}_x(f)$, questa è ottenibile come illustrato al capitolo precedente, una volta noto il segnale modulante e la tecnica di modulazione, mentre $\mathcal{P}_\nu(f)$ risulta pari a

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

D'altra parte, ha senso valutare l'$SNR$ *complessivo*, ovvero il rapporto tra le potenze di segnale e di rumore *totali*. Per ciò che riguarda il segnale, è ben noto che $\mathcal{P}_x = \int_{-\infty}^{\infty} \mathcal{P}_x(f)\,df$; in modo analogo, si ottiene $\mathcal{P}_\nu = \int_{-\infty}^{\infty} \frac{N_0}{2}\,|H_R(f)|^2\,df$, valutandone cioè la potenza che attraversa il filtro di ricezione $H_R(f)$. Data l'impossibilità pratica di realizzare un filtro ideale (rettangolare), $H_R(f)$ è caratterizzato da una banda (a frequenze positive) $B_\nu$, piú estesa di $B_{RF}$ (che è la banda di segnale).

La potenza totale del rumore uscente da $H_R(f)$ risulta pertanto pari a

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




Il termine $B_{RF} \le B_N \le B_v$ rappresenta la cosiddetta *banda di rumore* definita come

$$B_N = \frac{\int_0^{\infty}|H_R(f)|^2\,df}{|H_R(f_0)|^2}$$

ossia come la banda di un filtro ideale (rettangolare) che lascia passare la stessa quantità di rumore.

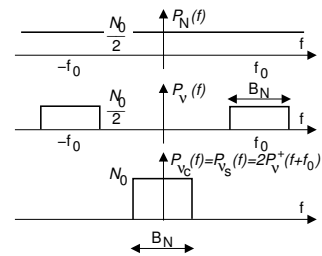
### 12.1.2 Demodulazione di un processo di rumore

Il rumore $\nu(t)$ che esce dal filtro di ricezione $H_R(f)$ è di tipo passa-banda, e può quindi essere descritto nei termini delle sue componenti analogiche di bassa frequenza:

$$\nu(t) = \nu_c(t)\cos\omega_0 t - \nu_s(t)\sin\omega_0 t$$

e, ricordando i risultati ottenuti al § 10.4.2, $\nu_c(t)$ e $\nu_s(t)$ sono entrambe gaussiane se $\nu(t)$ lo è; osserviamo inoltre che nel caso in cui la banda di $\nu(t)$ sia *stretta rispetto a* $f_0$, l'inviluppo complesso $\underline{\nu}(t) = \nu_c(t) + j\nu_s(t)$ evolve *lentamente* rispetto alla velocità di rotazione di $\underline{\nu}(t)\,e^{j\omega_0 t}$.

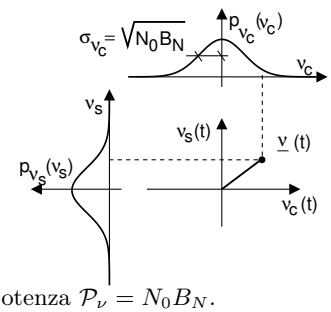
La figura seguente rappresenta la situazione in cui il rumore $n(t)$ in ingresso a $H_R(f)$ sia un processo ergodico bianco con densità di potenza $\mathcal{P}_n(f) = \frac{N_0}{2}$, ed il filtro di ricezione possegga una risposta in frequenza unitaria $|H_R(f_0)|^2 = 1$; in questo caso si ottiene che




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

e quindi

$$\mathcal{P}_{\nu_c}(f) = \mathcal{P}_{\nu_s}(f) = 2\mathcal{P}_\nu^+(f+f_0) = N_0\mathrm{rect}_{B_N}(f)$$




Pertanto, $\nu_c(t)$ e $\nu_s(t)$ risultano essere due processi congiuntamente gaussiani, ergodici, a media nulla ed uguale varianza (e potenza)

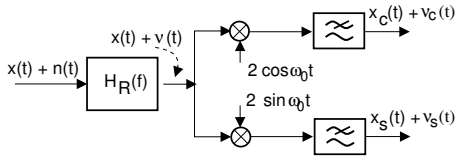
$$\sigma^2_{\nu_c} = \sigma^2_{\nu_s} = \mathcal{P}_\nu = N_0 B_N$$

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

## 12.2 Prestazioni delle trasmissioni AM

Per valutare il rapporto $SNR$ per le diverse tecniche di modulazione AM, esprimiamo il segnale modulato nei termini delle sue componenti analogiche

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




Operando una demodulazione coerente del segnale $x(t)$ affetto da rumore bianco $n(t)$ e filtrato da $H_R(f)$ come in figura, si ottengono le due componenti analogiche per il segnale demodulato: $\begin{cases} d_c(t) &= x_c(t) + \nu_c(t) \\ d_s(t) &= x_s(t) + \nu_s(t) \end{cases}$. Tra la potenza del segnale ricevuto e quella delle sue C.A. di B.F. sussiste[1] la relazione: $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_{x_c} + \frac{1}{2}\mathcal{P}_{x_s}$.

### 12.2.1 Potenza di segnale e di rumore dopo demodulazione. SNR

Nel caso di modulazione AM, siamo interessati alla sola componente in fase, che è sufficiente a fornire $m(t)$; il rapporto tra $\mathcal{P}_{x_c}$ e $\mathcal{P}_{\nu_c}$ fornirà dunque il valore di $SNR$ che stiamo cercando. La tabella che segue mostra i valori delle componenti di segnale e di rumore, assieme alle rispettive potenze espresse in funzione di una medesima potenza di messaggio $\mathcal{P}_m$, per tre casi di modulazione AM, di cui discutiamo individualmente.

| | $x_c(t)$ | $x_s(t)$ | $\mathcal{P}_{x_c}$ | $\mathcal{P}_x$ | $B_N$ | $\mathcal{P}_{\nu_c}$ |
|---|---|---|---|---|---|---|
| **BLD-PS** | $m(t)$ | $0$ | $\mathcal{P}_m$ | $\frac{1}{2}\mathcal{P}_m$ | $2W$ | $2WN_0$ |
| **BLD-PI** | $\sqrt{\eta}(a_p + m(t))$ | $0$ | $\eta\cdot\alpha$ | $\frac{1}{2}\eta\cdot\alpha = \frac{1}{2}\mathcal{P}_m$ | $2W$ | $2WN_0$ |
| **BLU-PS** | $\frac{1}{\sqrt{2}}m(t)$ | $\mp\frac{1}{\sqrt{2}}\widehat{m}(t)$ | $\frac{1}{2}\mathcal{P}_m$ | $\frac{1}{2}\mathcal{P}_m$ | $W$ | $WN_0$ |

Rispetto alla notazione adottata al Capitolo 11, si considera ora il termine $k_a$ inglobato in $m(t)$, e quindi non evidenziato in tabella; inoltre, si pone $\alpha = \left(a_p^2 + \mathcal{P}_m\right)$. Inoltre, la banda di rumore $B_N$ presa in considerazione nella tabella è *la minima* possibile, pari a quella del segnale modulato $B_{RF}$, direttamente legata (nella modulazione AM) a quella del segnale modulante $\pm W$. Pertanto, i risultati che otterremo sono *i migliori* possibili: infatti, se $B_N > B_{RF}$, l'$SNR$ risulterà peggiore. Precisiamo infine che nella valutazione dell'$SNR$ che segue, ci si riferisce sempre ad una medesima potenza ricevuta $\mathcal{P}_x$ e ad una densità di rumore $\mathcal{P}_n(f) = \frac{N_0}{2}$, allo scopo di rendere confrontabili i risultati.

#### 12.2.1.1 BLD-PS

Si ha $\mathcal{P}_{x_c} = \mathcal{P}_m$ e $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$; dopo demodulazione il segnale $d_c(t) = x_c(t) + \nu_c(t)$ presenta dunque una potenza di segnale utile $\mathcal{P}_{x_c} = \mathcal{P}_m = 2\mathcal{P}_x$ (con $\mathcal{P}_x$ pari alla potenza del segnale ricevuto) ed una potenza di rumore $\mathcal{P}_{\nu_c} = 2WN_0$; dunque un

$$SNR = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} = \frac{\mathcal{P}_m}{2WN_0} = \frac{\mathcal{P}_x}{WN_0} = SNR_0$$

in cui nell'ultima eguaglianza si definisce:

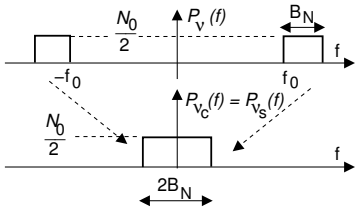
[1] Infatti i segnali $x_c(t)\cos\omega_0 t$ e $x_s(t)\sin\omega_0 t$ risultano ortogonali, e le potenze si sommano.

**SNR di riferimento** La grandezza $SNR_0$ è denominata rapporto segnale-rumore *di riferimento*, ed è il rapporto tra la potenza di segnale *ricevuto* e la potenza di rumore *in una banda pari a quella del messaggio di banda base.*[2]

Osserviamo dunque che la modulazione BLD-PS non altera il rapporto $SNR_0$ (è come se il processo di modulazione fosse assente). Notiamo infine (e questo è valido anche per i casi che seguono) che $SNR$ può riferirsi indifferentemente sia alle potenze di segnale che a quelle disponibili (vedi § 14.2.2.1), in quanto $SNR_0 = \frac{\mathcal{P}_x}{WN_0} = \frac{\mathcal{P}_x}{WN_0}\frac{4R_g}{4R_g} = \frac{W_{d_x}}{W_{d_N}}$.

### 12.2.1.2 BLU-PS

In questo caso, per ottenere una $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$ uguale al caso BLD-PS, le componenti $x_c(t)$ ed $x_s(t)$ devono essere poste pari a $\frac{1}{\sqrt{2}}m(t)$ e $\frac{1}{\sqrt{2}}\widehat{m}(t)$, rispettivamente (vedi § 11.1.4).




A seguito del processo di demodulazione, si ottiene un rumore in banda base che occupa ancora una banda $B_N$, ma possiede una densità uguale a quella del rumore a RF, in quanto i contenuti a frequenze positive e negative non si sovrappongono, come mostrato in figura. Risulta:

$$SNR = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} = \frac{\frac{1}{2}\mathcal{P}_m}{WN_0} = \frac{\mathcal{P}_x}{WN_0} = SNR_0$$

Dunque, si ottengono prestazioni identiche a quelle BLD. Si noti che il risultato è valido solo se $\nu(t)$ è effettivamente limitato alla sola banda $B_{RF}$. Se infatti si fosse adottato un filtro con banda piú larga, come ad esempio un $H_R(f)$ con $B_N = 2W$, si sarebbe ottenuto $\mathcal{P}_{\nu_c}(f) = N_0$, ed $SNR$ risulterebbe dimezzato.

### 12.2.1.3 BLD-PI

Per ottenere una potenza di segnale ricevuto $\mathcal{P}_x = \frac{1}{2}\mathcal{P}_m$ uguale ai due casi precedenti, si considera la ricezione di un segnale

$$x(t) = \sqrt{\eta}\,(a_p + m(t))\cos\omega_0 t \quad \text{in cui} \quad \eta = \frac{\mathcal{P}_m}{a_p^2 + \mathcal{P}_m}$$

ovvero $\eta$ è proprio pari all'efficienza della BLD-PI introdotta al § 11.1.1.4. Nel valutare l'$SNR$, faremo riferimento alla sola componente di messaggio $\sqrt{\eta}m(t)$ del segnale demodulato $\mathcal{P}_x$, che ha potenza $\eta\mathcal{P}_m = 2\eta\mathcal{P}_x$. La quantità $a_p^2$ si riferisce infatti alla portante non modulata, e non fornisce informazione. Si ha pertanto:

$$SNR = \frac{\mathcal{P}_{x_c}}{\mathcal{P}_{\nu_c}} = \frac{2\eta\mathcal{P}_x}{2WN_0} = \eta\frac{2\mathcal{P}_x}{2WN_0} = \eta SNR_0$$

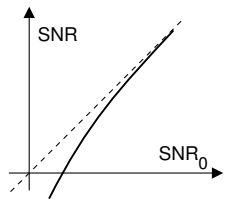
Dunque in questo caso constatiamo che la presenza della portante comporta una riduzione di prestazioni proprio pari all'efficienza $\eta = \frac{\mathcal{P}_m}{a_p^2+\mathcal{P}_m}$.

[2] In virtù di questa definizione, $SNR_0$ è una grandezza che caratterizza le condizioni operative ($\mathcal{P}_x$, $\mathcal{P}_n(f) = \frac{N_0}{2}$ e $W$ sono grandezze *indipendenti*) ma non è legata alla particolare tecnica di modulazione adottata. Pertanto, esprimere $SNR$ in funzione di $SNR_0$ permette il confronto tra i diversi casi *a parità di condizioni operative.*

L'analisi fin qui esposta si riferisce però al caso di demodulazione coerente: invece per BLD-PI si usa il demodulatore di inviluppo! In tal caso, il segnale demodulato è il modulo dell'inviluppo complesso, ovvero

$$d(t) = |\underline{x}(t) + \underline{\nu}(t)| = \sqrt{[\sqrt{\eta}(a_p + m(t)) + \nu_c(t)]^2 + \nu_s^2(t)}$$

Nel caso in cui $\underline{\nu}(t)$ sia piccolo, si può ottenere una approssimazione che ci riconduce al caso precedente. In caso contrario, sorgono termini *prodotto* tra $m(t)$ e $\nu_c(t)$, ed in definitiva l'$SNR$ risulta peggiore (per bassi $SNR_0$) del caso di demodulazione sincrona omodina, come illustrato nella curva riportata a fianco.



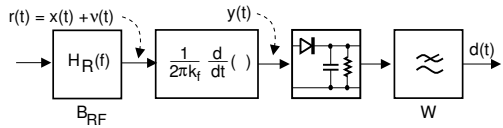

## 12.3 Prestazioni delle trasmissioni FM

Quando si è analizzata la tecnica di modulazione FM, si è fatta piú volte notare la sua caratteristica non lineare. E' lecito aspettarsi che questa caratteristica determini dei risvolti "bizzarri" per quanto riguarda l'$SNR$ del segnale demodulato: e difatti è proprio cosí. Anticipiamone due:

- La potenza del rumore *diminuisce* all'aumentare della potenza ricevuta
- L'$SNR$ *migliora* all'aumentare della banda occupata.

### 12.3.1 Rumore dopo demodulazione FM

Analizziamo innanzitutto il comportamento di un demodulatore a discriminatore[3], quando è presente in ingresso una portante *non modulata* di ampiezza[4] $A = \sqrt{2\mathcal{P}_x}$ sovrapposta ad un rumore gaussiano passa-banda:



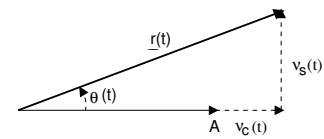

$$r(t) = A\cos\omega_0 t + \nu_c(t)\cos\omega_0 t - \nu_s(t)\sin\omega_0 t$$

A differenza del caso AM, ora il filtro $H_R(f)$ deve lasciar passare una banda di frequenze di estensione almeno pari alla banda che sarebbe occupata dal segnale FM, stimata applicando ad esempio la *regola di Carson* 11.3, ossia $B_C \simeq 2W(\beta + 1)$.

E' immediato verificare come le componenti analogiche di bassa frequenza di

$$\underline{r}(t) = r_c(t) + jr_s(t)$$

valgano $\begin{cases} r_c(t) &= A + \nu_c(t) \\ r_s(t) &= \nu_s(t) \end{cases}$, come mostrato nella figura a fianco. Notiamo come, per piccoli valori (rispetto ad $A$) di $\nu_c(t)$ e $\nu_s(t)$, l'inviluppo complesso ricevuto $\underline{r}(t)$ rimanga "prossimo" a quello ($A$) della portante non modulata. Come noto, $\nu_c(t)$ e $\nu_s(t)$ appartengono a due processi congiuntamente gaussiani, a media nulla e deviazione




[3] Descritto al § 11.2.2 ed applicato alla demodulazione FM al §11.3.1.2

[4] Con questa posizione, la potenza della portante risulta $\frac{(\sqrt{2\mathcal{P}_x})^2}{2} = \frac{2\mathcal{P}_x}{2} = \mathcal{P}_x$.

standard $\sigma_{\nu_c} = \sigma_{\nu_s} = \sqrt{N_0 B_N}$, in cui $B_N \geq B_{RF}$ è la banda di rumore del ricevitore, ed $N_0/2$ è la densità di potenza del rumore in ingresso.

Ricordiamo ora che nel caso FM, il *segnale informativo* è legato alla *derivata* della fase $\theta(t)$. Esprimiamo dunque $r(t)$ mettendo $\theta(t)$ in evidenza:

$$r(t) = \Re\left\{\underline{r}(t)\, e^{j\omega_0 t}\right\} = \Re\left\{|\underline{r}(t)|\, e^{j\theta(t)} e^{j\omega_0 t}\right\} = |\underline{r}(t)| \cos(\omega_0 t + \theta(t))$$

Osserviamo quindi che il termine $|\underline{r}(t)|$ viene rimosso dal limitatore (vedi §11.3.1.2) che usualmente è anteposto al discriminatore. Il segnale $y(t)$ in uscita dal derivatore è quindi dato da

$$y(t) = \left(\frac{f_0}{k_f} + \frac{1}{2\pi k_f}\frac{d}{dt}\theta(t)\right) \sin(\omega_0 t + \theta(t))$$

che viene a sua volta elaborato da parte del demodulatore di inviluppo come fosse un segnale BLD-PI, fornendo in definitiva un segnale demodulato

$$d(t) = \frac{1}{2\pi k_f}\frac{d}{dt}\theta(t)$$

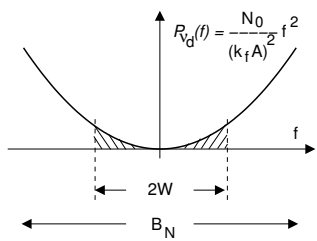
### 12.3.2 Caso di basso rumore

Se $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma^2_{\nu_c} = \sigma^2_{\nu_s} = N_0 B_N$ allora, come osservato, l'inviluppo complesso del rumore ha modulo *abbastanza* piú piccolo di $A$. Pertanto si può scrivere

$$\theta(t) = \arctan\frac{\nu_s(t)}{A + \nu_c(t)} \simeq \arctan\frac{\nu_s(t)}{A} \simeq \frac{\nu_s(t)}{A}$$

e dunque

$$\mathcal{P}_\theta(f) = \frac{1}{A^2}\mathcal{P}_{\nu_s}(f) = \frac{N_0}{A^2}$$




Ricordando ora (vedi § 3.7) che l'operazione di derivata (svolta dal discriminatore) equivale a moltiplicare lo spettro di ampiezza del segnale che si deriva per $j2\pi f$, ovvero moltiplicare la sua densità di potenza per $(2\pi f)^2$, si ottiene che la densità di potenza del segnale demodulato (che in questo caso è tutta dovuta al rumore) risulta pari a

$$\mathcal{P}_{\nu_d}(f) = \frac{1}{(2\pi k_f)^2}(2\pi f)^2 \mathcal{P}_\theta(f) = \left(\frac{f}{k_f}\right)^2 \frac{N_0}{A^2} = \frac{N_0}{(k_f A)^2} f^2$$

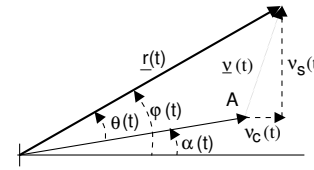
Infine, troviamo che la potenza totale di rumore dopo demodulazione risulta pari a

$$\mathcal{P}_{\nu_d} = \sigma^2_{\nu_d} = 2\int_0^W \frac{N_0}{(k_f A)^2} f^2 df = \frac{2}{3}\frac{N_0}{(k_f A)^2} W^3 \tag{12.1}$$

in cui $W$ è la banda del segnale modulante, ed il rumore è limitato in tale banda in virtú del filtro passa basso posto a valle del discriminatore. Si noti che invece le potenze $\sigma^2_{\nu_c}$ e $\sigma^2_{\nu_s}$ sono relative alla banda $B_N$, pari a quella del segnale modulato.

Notiamo subito la veridicità della prima affermazione fatta ad inizio capitolo: 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 ad ipotizzare $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma^2_{\nu_c} = \sigma^2_{\nu_s} = N_0 B_N$ possiamo osservare che, in presenza di una fase modulante $\alpha(t)$, la fase $\varphi(t)$ dell'inviluppo complesso del segnale ricevuto $r(t)$ è costituita dalla somma di $\alpha(t)$ con l'angolo $\theta(t)$ dovuto al rumore sovrapposto alla portante di ampiezza $A$. Pertanto all'uscita del discriminatore si ottiene




$$d(t) = \frac{1}{2\pi k_f} \frac{d}{dt} (\alpha(t) + \theta(t))$$

Il rapporto $SNR$ è ora definito come $SNR = \frac{\mathcal{P}_d}{\mathcal{P}_{\nu_d}}$, in cui $\mathcal{P}_{\nu_d}$ è la potenza del rumore demodulato, calcolata alla (12.1), e $\mathcal{P}_d$ è la potenza di segnale utile demodulato, pari a $d(t) = \frac{1}{2\pi k_f} \frac{d}{dt} \alpha(t)$. Sappiamo che $\alpha(t) = 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau$, e pertanto la potenza di segnale utile demodulato risulta proprio pari a $\mathcal{P}_d = \mathcal{P}_m = \int_{-W}^{W} \mathcal{P}_m(f)\, df$. Quindi:

$$SNR = \frac{\mathcal{P}_d}{\mathcal{P}_{\nu_d}} = \frac{\mathcal{P}_m}{\frac{2}{3}\frac{N_0}{(k_f A)^2} W^3} = 3\frac{\mathcal{P}_m k_f^2}{W^2 N_0 W} \frac{A^2}{2} = 3\frac{\sigma^2_{f_d}}{W^2} \frac{\mathcal{P}_x}{N_0 W} = 3\beta^2 SNR_0$$

Il risultato ottenuto conferma la seconda affermazione di inizio sezione: si ha un *miglioramento* rispetto all'$SNR$ *di riferimento* $SNR_0$ (e dunque rispetto all'AM) *tanto maggiore* quanto piú è grande l'indice di modulazione $\beta$, ovvero *quanto maggiore è la banda occupata* dal segnale modulato.

**Discussione dei passaggi** Innanzi tutto, è ovvio che $\frac{A^2}{2} = \mathcal{P}_x$ (la potenza ricevuta non cambia ed è sempre uguale a quella della portante non modulata) e che $\frac{\mathcal{P}_x}{N_0 W} = SNR_0$, il rapporto tra potenza ricevuta e potenza di rumore *nella banda del messaggio*. Mostriamo ora che $\mathcal{P}_m k_f^2 = \sigma^2_{f_d}$. Indichiamo con $f_d(t) = f_i(t) - f_0$ la deviazione della frequenza istantanea rispetto ad $f_0$, e cioè pari alla derivata della fase istantanea diviso $2\pi$, meno la frequenza portante:

$$f_d(t) = \frac{1}{2\pi} \frac{d}{dt} \left( 2\pi f_0 t + 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau \right) - f_0 = (f_0 + k_f m(t)) - f_0 = k_f m(t)$$

Pertanto si ha $\sigma^2_{f_d} = k_f^2 \sigma^2_m = k_f^2 \mathcal{P}_m$ se $m(t)$ è a media nulla: praticamente, $\sigma_{f_d}$ rappresenta la deviazione standard della frequenza istantanea, e per questo è una grandezza rappresentativa della larghezza di banda del segnale modulato. Infine, il rapporto $\frac{\sigma_{f_d}}{W} = \beta$ è posto pari all'indice di modulazione perché appunto rappresenta una misura di quanto sia variata l'occupazione di banda del segnale modulato, rispetto alla massima frequenza $W$ presente nel segnale modulante.

**Discussione del risultato** $SNR = 3\beta^2 SNR_0$. Notiamo innanzitutto che se $\beta < \frac{1}{\sqrt{3}} = 0,57$ non si ha miglioramento, anzi si peggiora. Ma con bassi indici di modulazione abbiamo già visto che FM ha un comportamento che può avvicinarsi a quello lineare dell'AM, e dunque ci possiamo non-sorprendere. D'altra parte, $SNR$ può migliorare (e di molto) con $\beta > \frac{1}{\sqrt{3}}$: ad esempio, se $\beta = 5 \Rightarrow 3\beta^2 = 75$ volte meglio, ovvero 17,75 dB di miglioramento! In compenso, la regola di Carson ci dice che la banda occupata aumenta di circa $2(\beta + 1) = 12$ volte quella di banda base... dunque

il miglioramento di SNR avviene *a spese dell'occupazione di banda*. (La cosa non preoccupa per le trasmissioni via satellite, in quanto c'è riuso di frequenze in diversità di spazio).

Potrebbe ora sembrare che si possa aumentare indefinitamente $\beta$ (nei limiti della banda disponibile) per migliorare a piacere l'$SNR$. In realtà non è cosí, dato che ad un certo punto l'analisi effettuata perde validità. Questo accade perché se $\beta$ è troppo elevato, occorre che la banda di rumore del ricevitore sia piú ampia (essendo aumentata la banda del segnale modulato) e perciò non si verifica piú che $\mathcal{P}_x = \frac{A^2}{2} \gg \sigma^2_{\nu_c} = \sigma^2_{\nu_s} = N_0 B_N$. Le conseguenze di questo fatto sono illustrate alla sottosezione successiva.

**Esercizio**

Sia dato un trasmettitore FM con potenza trasmessa 1 Watt e segnale modulante $m(t)$ con banda $\pm B = \pm 10$ MHz. Un collegamento con attenuazione disponibile $A_d = 100$ dB lo interfaccia ad un ricevitore con temperatura di sistema $T_{ei} = 2900$ $^o$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 $\beta$ trovato in 2) non sia troppo piccolo, e quale sia il suo massimo valore;

**4)** Il nuovo valore $\beta'$, volendo dotare il collegamento di un margine pari a 25 dB.

**Soluzione**

**1)** Questa domanda va affrontata dopo lo studio del capitolo sul rumore termico, 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 $F = \frac{T_{e_i}}{T_0} = 10$; pertanto $F_{dB} = 10$ dB. Per proseguire l'esercizio con le nozioni fin qui acquisite, esplicitiamo che $\mathcal{P}_n(f) = \frac{N_0}{2} = \frac{1}{2} k T_{ei} = \frac{1}{2} \cdot 1.38 \cdot 10^{-23} \cdot 2900 \simeq 2 \cdot 10^{-20}$ Watt/Hz.

**2)** $SNR = 3\beta^2 SNR_0 = 3\beta^2 \frac{W_R}{N_0 W} = 3\beta^2 \frac{W_T G_d}{N_0 B}$; il valore numerico di $SNR$ risulta $10^{\frac{SNR_{dB}}{10}} = 10^4$, mentre quello di $A_d$ è $10^{\frac{A_d(dB)}{10}} = 10^{10}$ e quindi $G_d = 1/A_d = 10^{-10}$. Sostituendo i valori, ed invertendo la relazione, si ottiene $\beta_{min} = \sqrt{\frac{SNR \cdot N_0 B}{3 \cdot W_T G_d}} = \sqrt{\frac{10^4 \cdot 4 \cdot 10^{-20} \cdot 10^7}{3 \cdot 10^{-10}}} = 3.65$. Applicando la regola di Carson per la banda: $B_{RF} \simeq 2B \cdot (\beta + 1) = 2 \cdot 10^7 \cdot 4.65 = 9.3 \cdot 10^7 = 93$ MHz.

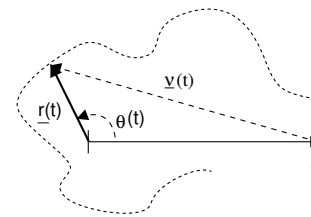
**3)** Perché l'analisi svolta abbia valore, deve risultare $W_R \gg \sigma^2_{n_c} = \sigma^2_{n_s} = N_0 B_N = N_0 B_{RF} = 4 \cdot 10^{-20} \cdot 9.3 \cdot 10^7 = 3.72 \cdot 10^{-12}$ Watt, ma poichè $W_R = \frac{W_T}{A_d} = \frac{1}{10^{10}} = 10^{-10}$, si ha $\frac{W_R}{\sigma^2_{n_c}} = \frac{10^{-10}}{3.72 \cdot 10^{-12}} = 26$. Il valore di 26 soddisfa quindi pienamente l'esigenza di *grande segnale*. Per trovare $\beta_{Max}$, scriviamo allora $\beta_{Max} \Rightarrow W_R = 10 \cdot \sigma^2_{n_c} = 10 \cdot N_0 \cdot B_{RF} = 10 \cdot N_0 \cdot 2B \cdot (\beta_{Max} + 1)$, e dunque $\beta_{Max} = \frac{W_R}{10 \cdot N_0 \cdot 2B} - 1 = \frac{10^{-10}}{8 \cdot 10^{-12}} - 1 = 12.5 - 1 = 11.5$, a cui corrisponde una banda $B_{RF} = 2B \cdot (\beta_{Max} + 1) = 2 \cdot 10^7 \cdot 12.5 = 250$ MHZ, ed un guadagno di $SNR = 10 \lg_{10} 3\beta^2_{Max} \simeq 26$ dB, mentre con $\beta_{min}$ nominale si sarebbe ottenuto $10 \lg_{10} (3 \cdot 3.65^2) = 16$ dB.

**4)** 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 $\beta'$: $SNR = 10^4 = 3\beta'^2 \frac{W_T G_d G'_d}{N_0 B} = 3\beta^2 \frac{W_T G_d}{N_0 B} \frac{\beta'^2}{\beta^2} G'_d$; dunque deve risultare $\frac{\beta'^2}{\beta^2} G'_d = 1$ e quindi $\beta'^2 = \beta^2 \sqrt{\frac{1}{G'_d}} = 3.65 \sqrt{316} = 3.65 \cdot 17.7 = 64.88$ .... non ce la facciamo. Infatti, al piú (con $\beta = \beta_{Max} = 11.5$) si ha un margine di 10 dB.

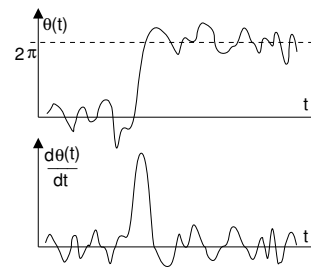
### 12.3.3 Caso di elevato rumore

Qualora il valore efficace del rumore in ingresso al discriminatore sia confrontabile con quello del segnale utile ricevuto, si verifica un effetto soglia all'aumentare del rumore, e l'$SNR$ degrada molto rapidamente.

Riprendendo lo schema che mostra l'inviluppo complesso della portante non modulata $A$, del rumore in ingresso $\underline{\nu}(t)$, e del segnale ricevuto $\underline{r}(t)$, notiamo che se i valori efficaci dei primi due sono comparabili, può verificarsi il caso che $\underline{r}(t)$ ruoti *attorno* all'origine.

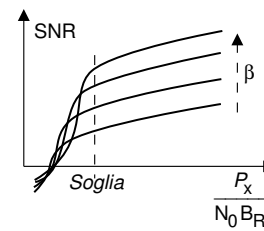



Quando ciò si verifica, a valle del derivatore che è presente nel discriminatore si determina un *click*, ovvero un segnale impulsivo di area pari a $2\pi$, come illustrato nella figura seguente. Questo fatto è facilmente verificabile, ascoltando una radio FM broadcast, che in condizioni di cattiva ricezione 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 $SNR$ di ingresso pari a 10 dB, e per $SNR$ peggiori di tale valore l'effetto aumenta molto rapidamente, cosicchè si parla di *effetto soglia*.

Nel grafico a lato è riportato un tipico andamento dell'$SNR$ dopo demodulazione, con l'indice di modulazione $\beta$ che svolge il ruolo di parametro, e possiamo osservare come con un $SNR$ di ingresso inferiore alla soglia, le prestazioni degradino rapidamente. Si è trovato che demodulando con un PLL, anzichè con un discriminatore, la soglia si riduce a circa 7 dB.




Nella pratica comune il segnale di rumore può essere costituito da una *interferenza* dovuta ad una emittente adiacente (ossia con una portante prossima alla nostra) che *sovramodula*, ovvero adotta un indice di modulazione troppo elevato, ed invade la banda delle emittenti contigue.

### 12.3.4 Enfasi e de-enfasi

Abbiamo osservato che, in presenza di rumore bianco in ingresso, il rumore dopo demodulazione ha un andamento parabolico. Questo comporta che, se il messaggio modulante $m(t)$ avesse un $\mathcal{P}_m(f)$ a sua volta bianco, l'$SNR(f)$ alle frequenze piú elevate sarebbe molto peggiore, rispetto al suo valore per frequenze inferiori. Nella pratica, si possono verificare (ad esempio) i seguenti problemi:

1. Nelle trasmissioni FDM-FM (vedi § 10.1.1.3), in cui piú canali vengono modulati AM, multiplati in frequenza, e ri-modulati congiuntamente in FM a basso indice, i canali agli estremi della banda FDM sono piú rumorosi;
2. nell'FM commerciale, il segnale modulante è molto piú 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 $m(t)$, in modo che anch'esso presenti uno spettro "parabolico", e poi aggiungere una rete di de-enfasi in ricezione (praticamente un integratore, ovvero un passa-basso) tale da ripristinare l'originale sagoma spettrale del segnale e rendere la densità di potenza del rumore nuovamente uniforme.

Con un po di riflessione, ci si accorge che l'uso di una coppia enfasi-deenfasi equivale ad effettuare una trasmissione PM! 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) piú elevate.

# Capitolo 13

# Modulazione numerica

Sono qui discusse le tecniche adottate qualora il segnale da trasmettere in forma modulata non sia analogico, ma costituito da una sequenza numerica, o simbolica. Il contesto applicativo può variare su un ampio ventaglio di casi, come le forme di broadcast numerico (terrestre o satellitare), le reti di accesso WiFi o di telefonia cellulare, i modem dell'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 codici di linea di *banda base* (vedi § 5.2), e dunque è necessario produrre un segnale modulato (vedi cap. 10) per trasporre la banda del segnale in accordo ai vincoli imposti dal canale. Ora non vengono però semplicemente applicate le tecniche esposte al cap. 11, ma queste vengono rese specifiche alla caratteristica del segnale dati dati di essere costituito da sequenze di simboli appartenenti ad un alfabeto finito, da mappare su un insieme finito di punti dello spazio, che per segnali modulati è lo spazio dell'inviluppo complesso.

## 13.1 Modulazione di ampiezza e di frequenza

Iniziamo con l'illustrare soluzioni direttamente riconducibili a queste due tecniche di modulazione analogica.

### 13.1.1 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:

$$x_{BPSK}(t) = a \sin(\omega_0 t + \varphi(t)) \quad \text{dove} \quad \varphi(t) = \sum_{k=-\infty}^{\infty} \varphi_k \text{rect}_{T_b}(t - kT_b) \tag{13.1}$$

con i valori $\varphi_k$ 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

[1] Letteralmente, *slittamento di tasto a due fasi.*

fase (§ 11.3), è facile mostrare come possa essere realizzata mediante una comune modulazione di ampiezza BLD (§ 11.1.1).

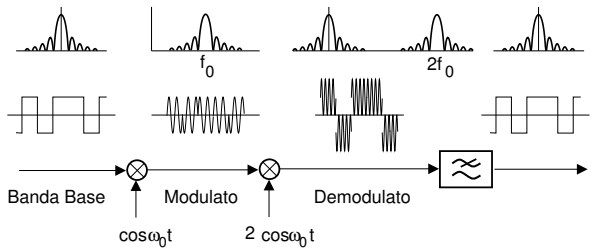
Se definiamo infatti un segnale $m(t)$ come un codice di linea NRZ bipolare (§ 5.2.1), che assume valori $\pm 1$ in corrispondenza delle cifre binarie 0 ed 1, allora il segnale

$$x_{BPSK}(t) = m(t)\cos\omega_0 t$$

è equivalente al precedente, e la sua mo-demodulazione coerente avviene come rappresentato alla figura che segue. 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.




Una buona caratteristica di questa tecnica è l'andamento costante dell'ampiezza della portante modulata, che permette di utilizzare la massima potenza del trasmettitore, appena inferiore al valore che inizia a produrre fenomeni di distorsione (§ 14.6). L'aspetto negativo è l'elevata occupazione di banda (§ 5.1.2.1), dovuta all'uso di forme d'onda rettangolari per $m(t)$ che, nel caso di trasmissione su canali con limitazioni di banda, causano una diminuzione della massima frequenza binaria. Per questi motivi, il metodo è paricolarmente indicato nel caso di collegamenti in cui la potenza di trasmissione è limitata, ma non la banda[3].

Gli aspetti ora illustrati possono essere *sovvertiti* qualora il segnale $m(t)$ sia generato utilizzando forme d'onda $g(t)$ con occupazione di banda limitata, come ad esempio la famiglia a coseno rialzato (§ 5.2.2.3). In quest'ultimo caso la banda occupata a frequenze positive risulta pari a $f_b(1+\gamma)$, doppia rispetto al caso di banda base, a causa della modulazione BLD, e l'ampiezza del segnale modulato *non è più costante*. Mentre infatti in corrispondenza degli istanti $kT_b$ l'ampiezza di $x_{BPSK}(t)$ assume esattamente uno dei valori ($\pm 1$) del segnale dati $m(t)$, nell'intervallo tra due istanti $kT_b < t < (k+1)T_b$ questa dipende della somma di tutte le code delle funzioni $g(t)$ relative ai simboli trasmessi (vedi fig. 5.4 a pag. 74).

### 13.1.2 L-ASK

Ci riferiamo ora al caso in cui si operi una classica AM-BLD (da cui il termine *Amplitude Shift Keying* - ASK) a partire da un segnale dati $m(t)$ multilivello (§ 5.1.2.4), producendo un segnale modulato di espressione

$$x_{L-ASK}(t) = m(t)\cos(2\pi f_0 t) \quad \text{dove} \quad m(t) = \sum_{k=-\infty}^{\infty} a_k \cdot \mathrm{rect}_{T_s}(t - kT_s)$$

[2] Qui e nel seguito assumiamo di disporre di una portante di demodulazione omodina, ossia priva di errori di fase e frequenza, così come di una perfetta temporizzazione di simbolo; le considerazioni al riguardo sono svolte all'appendice 13.6.2.

[3] Come ad esempio i collegamenti satellitari.

in cui $m(t)$ agli istanti $kT_s$ assume valori $a_k$ distribuiti uniformemente, in un intervallo $\Delta$, su $L$ *livelli di ampiezza* centrati sullo zero[4].

L'ampiezza di L-ASK subisce dunque variazioni, come mostrato nella figura a fianco per un caso con $L = 8$, in cui è rappresentato anche un diagramma detto *costellazione*, che rappresenta i valori assunti dall'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 $M = \log_2 L$[5].


Anche qui, nel caso in cui $m(t)$ sia generato mediante forme d'onda con caratteristica a coseno rialzato (§ 5.2.2.3) anziché con un codice di linea NRZ, la densità spettrale assume il noto andamento, e la banda a frequenze positive occupata da $x_{L-ASK}(t)$ risulta pari a

$$B_{L-ASK} = f_s(1+\gamma) = \frac{f_b}{\log_2 L}(1+\gamma) \qquad (13.2)$$


Se consideriamo $\gamma = 0$, possiamo definire:

**Efficienza Spettrale** $\rho$ (o *densità di informazione*) come il rapporto tra la frequenza binaria e la banda occupata

$$\rho_{L-ASK} = \frac{f_b}{B} = \log_2 L$$

che si esprime in *bit/sec/Hz* e rappresenta appunto quanti bit/sec sono trasmessi per ogni Hz utilizzato.

Il valore trovato $\rho = \log_2 L$ rappresenta l'efficienza spettrale dell'L-ASK quando adotta impulsi a a banda minima (ovvero con $\gamma = 1$); per altre forme di modulazione e/o di impulsi si ottengono altri valori, ed il loro confronto esprime la bontà del metodo rispetto all'utilizzo della banda a disposizione. Ad esempio, se confrontiamo il risultato ottenuto ora con quello relativo ad una trasmissione numerica di banda base, notiamo un peggioramento di un fattore 2, dovuto all'uso di una AM-BLD. Come per il caso analogico, la banda può essere dimezzata adottando una AM-BLU, ma vedremo invece tra breve che si preferiscono seguire approcci diversi, come ad esempio PSK e QAM.

### 13.1.3 L-FSK

Qualora si desidera che l'ampiezza del segnale modulato si mantenga strettamente costante, può essere adottata la modulazione FSK (*Frequency Shift Keying),* che associa ad ogni simbolo un *livello di frequenza* $f_k$ che si somma a quello della portante, in

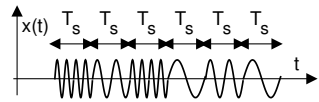
[4] Per chi si sta chiedendo quanto valgono questi livelli, diciamo che il livello i-esimo (con $i = 0, 1, \cdots, L-1$) corrisponde al valore $a^i = i \cdot \frac{\Delta}{L-1} - 1$ . Verificare per esercizio con $\Delta = 2$ ed $L = 4$.

[5] 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.

accordo all'espressione

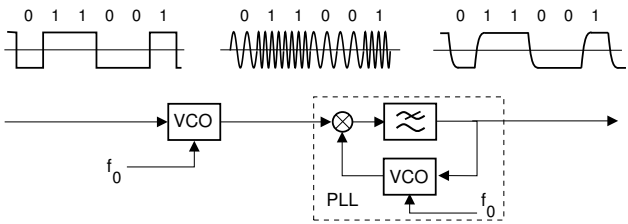
$$x_{FSK}(t) = \cos\left[2\pi\left(f_0 + m(t)\right)t\right] \quad \text{dove} \quad m(t) = \Delta \cdot \sum_{k=-\infty}^{\infty} f_k \cdot \mathrm{rect}_{T_s}(t - kT_s)$$

in cui ogni elemento della sequenza $f_k$ assume uno tra i valori $\{0, 1, 2, \ldots, L-1\}$. Si tratta in pratica di una portante la cui frequenza nominale $f_0$ è alterata di una quantità $\Delta \cdot f_k$ Hz per l'intervallo temporale pari al periodo di simbolo $T_s$, in cui $\Delta$ rappresenta ora la spaziatura (uniforme) tra le frequenze associate agli $L$ livelli. Pertanto l'espressione può essere riscritta come




$$x_{FSK}(t) = \sum_{k=-\infty}^{\infty} \cos\left[2\pi\left(f_0 + \Delta f_k\right)t\right] \cdot \mathrm{rect}_{T_s}(t - kT_s)$$

Il risultato è senza dubbio ad ampiezza costante; se $T_s \gg \frac{1}{f_0}$ si può adottare uno schema di mo-demodulazione basato su di un PLL (vedi § 11.2.1.3 e 11.3.1.1) riportato (per $L = 2$) in figura, in cui all'uscita del passa basso ritroviamo il segnale modulante.




Lo schema è effettivamente utilizzato per modem a bassa velocità e basso costo, ed ha il pregio di funzionare anche in presenza di errori tra l'$f_0$ usata al trasmettitore e quella al ricevitore. Per raggiungere velocità $f_b$ più elevate a parità di $L$, occorre ridurre $T_s$, in modo da aumentare $f_s = \frac{f_b}{\log_2 L}$. In tal caso può essere necessario ricorrere ad un demodulatore più complesso, come per caso seguente.

**FSK ortogonale** Nel caso in cui si realizzi la condizione $\Delta = \frac{n}{2T_s}$ con $n$ intero, le diverse frequenze $f_0 + \Delta f_k$ sono *ortogonali*[6], e può essere adottato un *demodulatore a correlazione*, introdotto al § 9.4.4.2, mostrato a lato, e discusso alla nota[7].

[6] L'ortogonalità tra le forme d'onda associate ai diversi simboli è sinonimo di intercorrelazione nulla (§ 9.1.4), ovvero $\int_0^{T_s} \cos\left[2\pi\left(f_0 + m\Delta\right)t\right] \cos\left[2\pi\left(f_0 + n\Delta\right)t\right] dt = \begin{cases} .5 \cdot T_s & \text{se } n = m \\ 0 & \textit{altrimenti} \end{cases}$.
Infatti, ricordando che $\cos^2\alpha = \frac{1}{2} + \frac{1}{2}\cos 2\alpha$, l'uscita del filtro adattato per $m = n$ vale $\frac{1}{2}\int_0^{T_s}\left(1 + \cos\left(4\pi\left(f_0 + m\Delta\right)t\right)\right)dt$, ed il coseno che viene integrato descrive un numero intero di periodi all'interno dell'intervallo $(0, T_s)$, fornendo quindi un contributo nullo. Se invece $n \neq m$ la funzione integranda non contiene il termine costante, ma solo termini a media nulla.
Si può dimostrare (vedi appendice 13.6.3) che una spaziatura $\Delta = \frac{1}{2T_s}$ garantisce l'ortogonalità solo nel caso in cui tra le forme d'onda *non sussitano ritardi di fase*, così come sopra espresso. Se invece i diversi simboli presentano una fase aleatoria $\phi_k$, ossia hanno espressione $\cos\left[2\pi\left(f_0 + \Delta f_k\right)t + \phi_k\right]$ con $\phi_k$ casuale e diversa per $\forall k$, allora si ottengono segnali incorrelati solo adottando una spaziatura doppia, e cioè $\Delta = \frac{1}{T_s}$. Questo secondo caso è detto di *modulazione incoerente*, per distinguerlo da quello in cui $\phi_k = 0$, detto *coerente*.

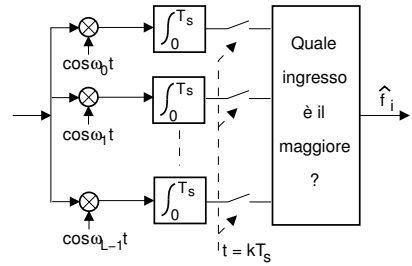
[7] Ognuno dei correlatori del banco esegue l'integrale indicato alla nota precedente, integrando su $T_s$ il prodotto tra il segnale ricevuto e tutte le possibili frequenze $f_0 + m\Delta$ con $m \in$

Nel caso di *modulazione coerente*, la minima banda occupata può essere approssimata[8] come

$$B_{FSK} \simeq \frac{L}{2T_s} \qquad (13.3)$$

(considerando $L$ elevato e dunque $B \gg \Delta$), pertanto l'efficienza spettrale risulta

$$\rho_{FSK} = \frac{f_b}{B} = \frac{f_s \log_2 L}{L/2T_s} = \frac{2\log_2 L}{L}$$



Demodulatore a correlazione

ossia $\frac{L}{2}$ volte peggiore dell' L-ASK. In appendice 13.6.3 è riportato un approfondimento dell'analisi relativa all'FSK ortogonale.

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 $T_s \gg \frac{1}{f_0}$ e demodulatore a PLL) è di facile realizzazione e poco costoso; ad esempio, veniva usato per salvare su *compact cassette* audio i dati degli *home computer* degli anni '70[9]

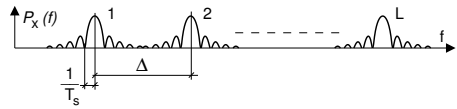
**Se *L=2*** l'efficienza spettrale $\rho = \frac{f_b}{B}$ è uguale all'ASK[10], come verificabile sostituendo $T_s = \frac{1}{f_b}\log_2 L$ nella (13.3). Al contrario, al crescere di $L$ l'efficienza spettrale diviene sempre peggiore.

**La probabilità di errore** può essere resa *piccola a piacere*, nei limiti della teoria dell'informazione[11]. La figura che segue mostra i valori di $E_b/N_0$ necessari per ottenere

$\{0, 1, 2, \ldots, L-1\}$. Se le frequenze sono ortogonali, al termine dell'intervallo di integrazione una sola delle uscite sarà diversa da zero. Il confronto tra i risultati indicato in figura è necessario, perché la presenza di rumore additivo *corrompe* l'ortogonalità tra simboli.

Nel caso di modulazione coerente, sia il trasmettitore che il ricevitore devono rispettare specifiche realizzative più stringenti, dovendo necessariamente realizzarsi un errore di fase nullo tra le frequenze di confronto ed il segnale ricevuto.

[8] In generale, se ogni diversa $f_k$ è equiprobabile, l'FSK ha una densità spettrale del tipo:




Se $L \gg 1$, le $L$ diverse frequenze occupano una banda (circa) uguale a $L \cdot \Delta$; qualora $\Delta = \frac{1}{2T_s} = \frac{f_s}{2}$, la banda risulta $L \cdot \frac{f_s}{2} = \frac{L}{2T_s}$.

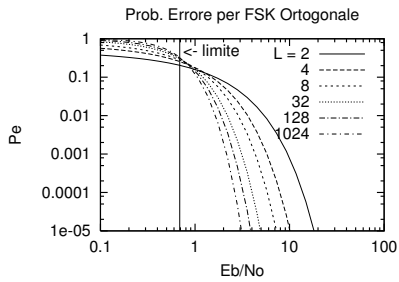
[9] tipo: Sinclair Spectrum , Commodore Vic20 e 64 ... Come noto, le cassette audio soffrono di variazioni di velocità di trascinamento del nastro (*wow & flutter*), ma il PLL non ne risente.

[10] Tranne che, essendo ora presenti solo 2 frequenze, l'approssimazione $B \simeq \frac{L}{2T_s}$ non è più valida.

[11] Ovvero tenendo conto che (vedi §17.2.3) $f_b$ non può superare la capacità di canale (17.29), che a sua volta non può superare il limite $C_\infty$ espresso dalla (17.31). Ma per l'esattezza, l'espressione teorica della probabilità di errore per simbolo risulta essere

$$P_e^{FSK}\,(simbolo) = 1 - \frac{1}{\sqrt{\pi^L}} \int_{-\infty}^{\infty} \mathrm{e}^{-z^2} \left( \int_{-\infty}^{z+\sqrt{\log_2 L \cdot E_b/N_0}} \mathrm{e}^{-y^2} dy \right)^{L-1} dz$$

che deve essere valutata numericamente per ricavare le curve mostrate.




le varie $P_e$ con diversi valori di $L$, e illustra come, all'aumentare di $L$, sia necessaria sempre minor potenza per ottenere la stessa $P_e$, a patto che risulti

$$E_b/N_0 > 1/\log_e 2 = 0,69$$

Questo valore è noto come *limite di Shannon-Hartley*[12] ricavato alla (17.30) a pag. 426.

Il risultato evidenziato merita un ulteriore commento: osserviamo infatti che la banda occupata

$$B_{FSK} \simeq \frac{L}{2T_s} = \frac{L}{2\frac{1}{f_s}} = \frac{L}{2\frac{\log_2 L}{f_b}} = \frac{f_b}{2}\frac{L}{\log_2 L}$$

*aumenta* (a parità di $f_b$) all'aumentare di $L$. Pertanto, per un $E_b/N_0$ assegnato (ovvero con $f_b$, potenza di segnale, e potenza di rumore preassegnate), l'FSK permette di ottenere $P_e$ arbitrariamente piccole, a spese di una occupazione di banda sempre maggiore. L'aumento di $L$ non può però essere qualunque, oltre che per le limitazioni del canale, anche a causa della complessità del ricevitore!

### 13.1.4 Natura di $E_b/N_0$

L'analisi delle prestazioni delle tecniche di modulazione numerica che affronteremo esprime la probabilità di errore per simbolo $P_e$ (*simbolo*) o per bit $P_e$ (*bit*) in funzione della grandezza $\frac{E_b}{N_0}$, introdotta al § 7.5.3, che rappresenta l'equivalente del rapporto segnale rumore *di riferimento* $SNR_0 = \frac{\mathcal{P}_x}{N_0 W}$ definito al §12.2.1.1[13], e quindi consente il confronto tra tecniche diverse. Infatti, considerando che la potenza ricevuta può essere espressa come $\mathcal{P}_x = \frac{E_b}{T_b} = E_b \cdot f_b$, si ottiene

$$SNR_0 = \frac{\mathcal{P}_x}{N_0 f_b} = \frac{E_b}{N_0} \tag{13.4}$$

che dipende unicamente da grandezze che identificano la qualità del collegamento ($\mathcal{P}_x$ ed $N_0$) e la velocità del messaggio ($f_b$) senza alcun riferimento al metodo di modulazione. Al § 7.5.3.1 si deriva la relazione esatta tra $\frac{E_b}{N_0}$ e $SNR$ *effettivo* per un segnale dati a coseno rialzato e multilivello.

### 13.1.5 Prestazioni di L-ASK

Esaminiamo ora la probabilità di errore per un segnale L-ASK in funzione di $E_b/N_0$, al variare del numero di livelli. Abbiamo osservato al § 13.1.2 come questo si ottenibile come modulazione AM BLD di un segnale dati di banda base, ed al § 12.2.1.1 si è mostrato che in questo caso, dopo demodulazione, risulta $SNR = SNR_0$. Pertanto in base alla (13.4) le prestazioni per il caso di un segnale multilivello a coseno rialzato

[12] In appendice 13.6.3 è esposta una motivazione informale del comportamento descritto.

[13] 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.

di banda base sono le stesse[14] di quelle ricavate al § 7.5.5, ovvero una probabilità di errore *per simbolo* pari a

$$P_e^{L-ASK}\,(simbolo) = \left(1 - \frac{1}{L}\right) erfc\left\{\sqrt{3\frac{E_b}{N_0}\frac{log_2 L}{(L^2-1)}}\right\} \tag{13.5}$$

e che si riferisce ad un segnale *a coseno rialzato* con $\gamma = 0$([15]), e pertanto le curve di $P_e\,(bit)$ sono quelle di fig. 7.6 a pag. 147, dove si tiene anche conto dell'uso di un codice di Gray (§ 5.2.2.4) per associare i livelli a configurazioni binarie.

Osserviamo esplicitamente che per $L = 2$ si ottiene

$$P_e^{BPSK}\,(bit) = \frac{1}{2} erfc\left\{\sqrt{\frac{E_b}{N_0}}\right\} \tag{13.6}$$

che rappresenta le stesse prestazioni ottenibili per la modulazione BPSK.

Per completare i confronti, osserviamo che ora all'aumentare di $L$ la banda (13.2) (per $\gamma = 0$)

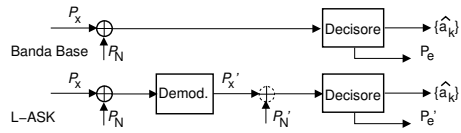
$$B_{L-ASK} = f_s = \frac{f_b}{\log_2 L}$$

si riduce, mentre la $P_e$ aumenta: un comportamento diametralmente opposto all'FSK, e che può tornare utile in presenza di canali con limitazioni di banda ma non di potenza; in tal caso infatti la potenza può essere aumentata per compensare il peggioramento di prestazioni legato all'uso di molti livelli e di una banda ridotta.

## 13.2 Modulazione di fase

Si tratta del caso in cui l'informazione modulante è impressa, anziché sulle ampiezze o sulla frequenza, direttamente sulla fase del segnale modulato.

[14] Forniamo qui dimostrazione forse inutilmente elaborata. Con riferimento alla figura seguente, il calcolo della $P_e$ per l'L-ASK si imposta definendo valori di $E_b$ ed $N_0$ equivalenti a quelli di banda base, ma ottenuti dopo demodulazione, e cioè $E_b' = P_x' T_b$ e $N_0' = P_N'/W$ (infatti, $P_N' = \frac{N_0'}{2} 2W$, con $W = \frac{f_s}{2} = \frac{f_b}{2\log_2 L}$).




L'equivalenza è presto fatta, una volta tarato il demodulatore in modo che produca in uscita la componente in fase $x_c\,(t)$ limitata in banda tra $\pm W$. Infatti in tal caso $P_x' = P_{x_c} = k_a^2 P_M = 2P_x$ e quindi $E_b' = P_x' T_s = 2P_x T_s = 2E_b$; per il rumore si ottiene $N_0' = \frac{P_N'}{W}$ in cui $P_N' = P_{n_c} = \sigma_{n_c}^2 = \sigma_n^2 = \frac{N_0}{2} 4W$ e quindi $N_0' = 2N_0$. Pertanto, il valore $E_b'/N_0'$ su cui si basa ora il decisore è lo stesso $E_b/N_0$ in ingresso al demodulatore.

[15] Se $\gamma \neq 0$, valgono le considerazioni svolte al § 7.5.5.

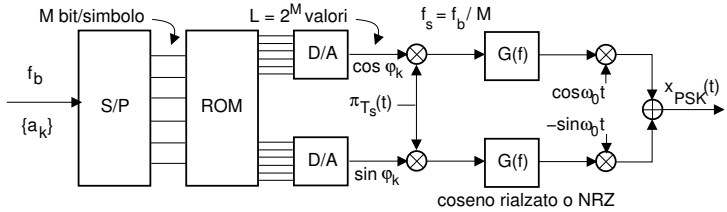



Figura 13.1: Modulatore L-PSK

### 13.2.1 QPSK ed L-PSK

Questi acronimi si riferiscono alla circostanza in cui siano possibili quattro oppure $L > 4$ scelte diverse[16] per la fase, dando luogo ad un segnale modulato con espressione

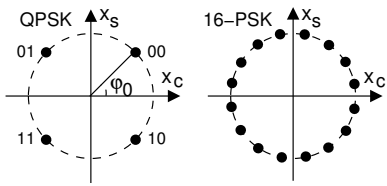
$$x_{L-PSK}(t) = a\cos(\omega_0 t + \varphi(t))$$

e quindi un inviluppo complesso

$$\underline{x}_{L-PSK}(t) = a\mathrm{e}^{j\varphi(t)} = a\cos\varphi(t) + ja\sin\varphi(t)$$

in cui

$$\varphi(t) = \sum_{k=-\infty}^{\infty} \varphi_k \mathrm{rect}_{T_s}(t - kT_s) \quad \text{e} \quad \varphi_k \epsilon \{\varphi_0, \varphi_1, \ldots, \varphi_{L-1}\} \tag{13.7}$$




La generica fase $\varphi_i = \frac{\pi}{L} + i \cdot \frac{2\pi}{L}$ con $i = 0, 1, \cdots L-1$ 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 $t = kT_s$. Se $\varphi(t)$ è realizzata mediante rettangoli come in (13.7), corrisponde ad un codice di linea NRZ ad $L$ livelli. L'espressione 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[17] come illustrato in fig. 13.1: il valori di $\cos\varphi_{\hat{\imath}}$ e $\sin\varphi_{\hat{\imath}}$ per gli $L$ diversi gruppi di $M$ bit sono precalcolati in una memoria, ed usati come ampiezze per realizzare due segnali dati usati quindi come c.a di b.f.

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 distribuzione di potenza assume una sagoma $\frac{\sin x}{x}$ centrata in $f_0$ e con lobo principale di estensione pari ad $f_s = {}^{f_b}/_M$. L'occupazione di banda può essere limitata a $B = f_s(1+\gamma)$ se si realizza $\varphi(t)$ mediante impulsi di Nyquist $g(t)$ a coseno rialzato, potendo così scrivere

$$\underline{x}_{L-PSK}(t) = a\sum_{k=-\infty}^{\infty} \mathrm{e}^{j\varphi_k} \cdot g(t - kT_s)$$

[16] Il caso in $L = 2$ ricade nel BPSK già discusso

[17] che *non* è una modulazione AM-BLU dato che $x_s \neq \widehat{x}_c$

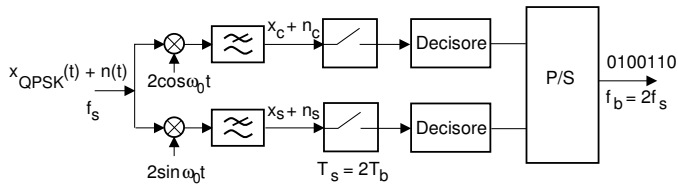



Figura 13.3: Demodulatore QPSK

Contrariamente al caso di $g(t)$ rettangolare[18], adottando una $g(t)$ a coseno rialzato $\underline{x}(t)$ passa dai punti della costellazione solo negli istanti significativi, mentre nell'intervallo tra due istanti segue traiettorie con ampiezza variabile, non rispettando quindi la proprietà di ampiezza costante che una modulazione angolare dovrebbe avere. Pertanto la scelta tra NRZ e coseno rialzato dipende perciò dalla necessità di limitare la dinamica delle ampiezze, oppure l'estensione della banda.

Dal punto di vista dell'efficienza spettrale, quest'ultima è identica a quanto ottenuto per l'ASK con ugual numero di livelli, dato che si ha la medesima frequenza di simbolo $f_s = \frac{f_b}{\log_2 L}$ e dunque $\rho = \frac{f_b}{B} = \frac{f_s \log_2 L}{f_s} = \log_2 L$ (per coseno rialzato con $\gamma = 0$).

Dal punto di vista delle prestazioni, occorre distinguere il caso in cui $L = 4$ (indicato come QPSK = *Quadrature Phase Shift Keying*) da quello con $L$ generico, in quanto sussistono due diverse architetture per il demodulatore.

### 13.2.2 Prestazioni QPSK

In questo caso (PSK con 4 livelli) il demodulatore è costituito da due rami indipendenti in fase e quadratura, operanti a frequenza di simbolo $f_s$ metà di quella binaria. Ognuno dei due rami effettua una decisione per uno dei due bit che compongono il simbolo, e le due decisioni vengono reserializzate, come mostrato in fig. 13.3. Entrambi i rami si comportano pertanto come un demodulatore L-ASK (§ 13.1.2) con L=2, ovvero (a parte una rotazione di fase) un BPSK, e dunque per un segnale a coseno rialzato la probabilità di errore relativa ad ogni singolo ramo è espressa[19] dalla (13.6), fornendo (per $\gamma = 0$):

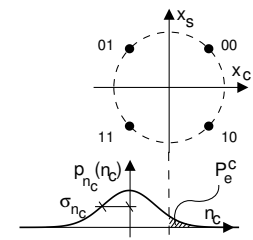



Figura 13.2: Costellazione QPSK e calcolo della $P_e^c$

$$P_e^c = P_e^s = \frac{1}{2}\text{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$$

[18] Se $g(t) = \text{rect}_{T_s}(t)$, $|\underline{x}|$ giace su di un cerchio, spostandosi *istantaneamente* da un punto all'altro della costellazione

[19] In effetti, dovremmo mostrare che l'attuale valore di $E_b/N_0$ è lo stesso del caso BPSK: ma mentre per $N_0$ è evidente che si ha lo stesso valore, sembra che il valore di $E_b$ si dimezzi. Infatti, a partità 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 (fig. 13.2), mentre nel QPSK le fasi formano un angolo di $45^o$ rispetto agli assi, riducendone la distanza dallo zero di $\sqrt{2}$, e riducendo dunque la potenza della c.a di b.f. di un fattore 2, e così pure il valore $E_b$. In realtà, la durata doppia del periodo di simbolo $T_s$ compensa questa riduzione, e dunque $E_b = P_x T_b$ si mantiene invariato.

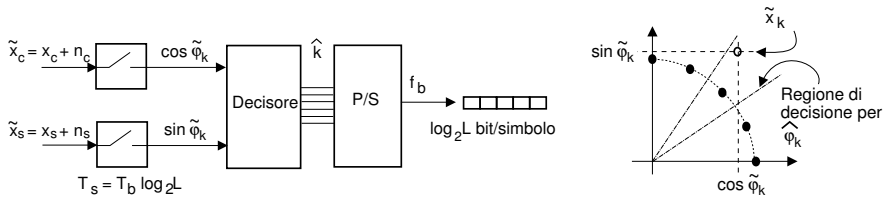



Figura 13.4: Demodulatore L-PSK

che rappresenta la probabilità che l'inviluppo complesso del rumore demodulato sovrapposto al segnale, valutato all'istante di decisione $kT_s$, giaccia nell'area mostrata in figura 13.2. La probabilità di errore in un bit della sequenza re-serializzata risulta quindi

$$P_e^{QPSK}(bit) = P_e^c \cdot Pr\{c\} + P_e^s \cdot Pr\{s\} = \frac{1}{2}(P_e^c + P_e^s) = \frac{1}{2}\mathrm{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$$

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, e si assume che $P_e^{(c)} \cdot P_e^{(s)} \ll P_e^{(c)}$ e quindi trascurabile[20].

Osserviamo quindi come il QPSK consenta di ottenere *le stesse prestazioni* del BPSK, utilizzando solo *metà banda:*

$$B_{QPSK} = f_s = \frac{f_b}{2}$$

Il risultato (relativo al caso $\gamma = 0$) ha una giustificazione intuitiva: osserviamo infatti che dimezzando la banda, si dimezza anche la varianza del rumore gaussiano in ingresso al demodulatore; questo fatto compensa la riduzione di ampiezza delle componenti analogiche di bassa frequenza ricevute nel caso QPSK.

### 13.2.3 Prestazioni L-PSK

In questo caso il demodulatore ha una differente architettura, ed il decisore opera congiuntamente su entrambi i rami, per ottenere la stima del gruppo di $\log_2 L$ bit associati ad una delle possibili fasi $\varphi_k$.

Indicando con $\tilde{x}_{c,s}$ le c.a di b.f. (vedi fig. 13.4) ricevute, la decisione avviene calcolando $\tilde{\varphi}_k = \arctan\frac{\sin\tilde{\varphi}_k}{\cos\tilde{\varphi}_k}$ e stabilendo all'interno di quale regione di decisione $\widehat{\varphi}_k$ cada la fase ricevuta $\tilde{\varphi}_k$. All'aumentare del numero di livelli $L$, la potenza di rumore (che concorre alla probabilità di errore) diminuisce con la stessa legge di riduzione della banda, ovvero con il $\log_2 L$. Al contrario, la spaziatura tra le regioni di decisione diminuisce con legge lineare rispetto ad $L$; pertanto, 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

$$P_e^{L-PSK}(simbolo) = \mathrm{erfc}\left\{\sin\left(\frac{\pi}{L}\right)\sqrt{\frac{E_b}{N_0}\log_2 L}\right\} \tag{13.8}$$

[20] La probabilità di errore per simbolo risulta invece $P_e(simbolo) = P_e^c + P_e^c = \mathrm{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$, considerando di nuovo trascurabile la probabilità di un errore contemporaneo su entrambi i rami.

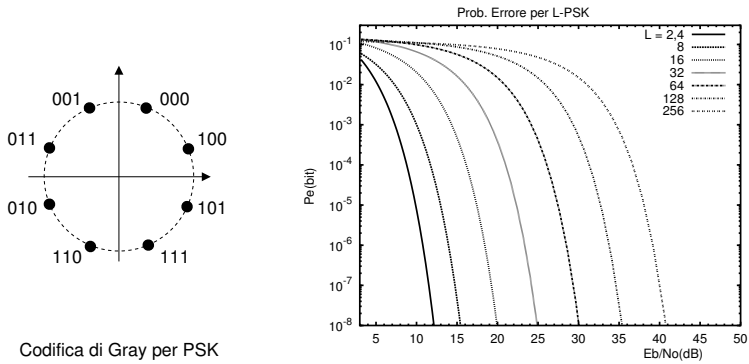



Figura 13.5: Prestazioni L-PSK con codice di Gray

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.

Confrontando il risultato con quello (eq. 13.5) 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 (13.8) è una approssimazione valida se $P_e \ll 1$, e via via più accurata con $L$ crescente.

Nella tabella a fianco è riportato il risultato del confronto, per uno stesso valore di $P_e$, dei valori $\frac{E_b}{N_0}$ necessari per L-PSK (13.8), contro quelli necessari (13.5) per L-ASK: si è eseguito il rapporto tra gli argomenti degli erfc {}, si è elevato al quadrato, indicato come $\Delta$, ed il risultato è espresso in dB. Esaminando il risultato per i diversi valori di $L$, si trova (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. Il risultato (4-5 dB di miglioramento) ha portato a prediligere sempre il PSK rispetto all'ASK.

| $L$ | $\Delta = \frac{1}{3}\left(L^2 - 1\right)\sin^2\frac{\pi}{L}$ | $\Delta_{dB}$ |
|---|---|---|
| 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[21] (mostrato nella figura 13.5), in modo che l'errore tra due livelli provochi l'errore di un solo bit nel gruppo di $\log_2 L$ bit associati a ciascun livello. Le curve di probabilità di errore per bit, riportate anch'esse in fig. 13.5, sono calcolate in questo modo.

[21] vedi il § 5.2.2.4

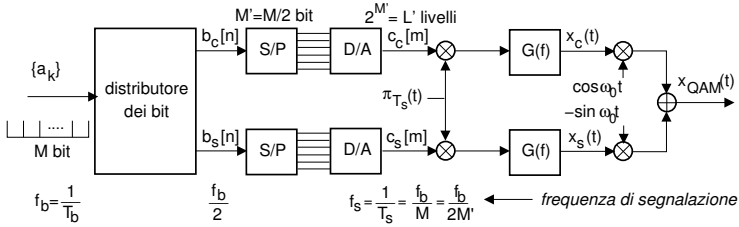



Figura 13.6: Modulatore QAM

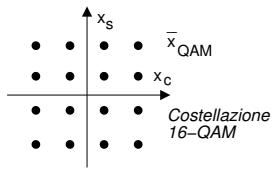
## 13.3 QAM

L'acronimo sta per *Quadrature Amplitude Modulation*, ed individua la tecnica di modulazione che utilizza due portanti in quadratura come il PSK

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

ma a differenza del PSK, ora le componenti di banda base $x_c$ ed $x_s$ non dipendono da una stessa sequenza di fasi, ma sono originate da due flussi di dati distinti.

Con riferimento alla fig. 13.6, osserviamo che sebbene $x_c(t)$ e $x_s(t)$ si ottengono a partire da una medesima sequenza numerica $\{a_k\}$, i suoi bit vengono distribuiti alternativamente sui due rami (sequenze $b_c[n]$ e $b_s[n]$ in fig.) a velocità dimezzata[22], suddividendo un gruppo di $M$ bit in due simboli costituiti da $M' = M/2$ bit. Dalle sequenze $b_c$ e $b_s$ si ottengono quindi i valori $c_c$ e $c_s$ con $L' = 2^{M'} = 2^{M/2} = \sqrt{L}$ livelli, che attraversando il filtro $G(f)$, danno luogo ai segnali di banda base $x_c$ ed $x_s$.




La sequenza di operazioni descritte determina una costellazione *quadrata*, composta da $L = (L')^2$ punti, che rappresentano le coordinate (nel piano dell'inviluppo complesso) in cui $\underline{x}$ è forzato a transitare in corrispondenza degli istanti di Nyquist multipli di $T_s$, che risulta essere pari a

$$T_s = T_b \cdot 2M' = \frac{1}{f_b} 2\log_2 L' = \frac{1}{f_b} 2\log_2 (L)^{1/2} = \frac{1}{f_b}\log_2 L$$

Se $G(f)$ è a coseno rialzato con roll off $\gamma$, 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 al PSK con uguale numero di livelli (di cui condivide quindi anche l'efficienza spettrale). Notiamo che, per come si è impostato lo schema di distribuzione dei bit tra i rami, $L$ deve risultare un quadrato perfetto. Nulla impedisce di elaborare schemi più complessi in cui $L'$ è diverso per i due rami, o più in generale la cui costellazione non sia quadrata. E' appena il caso di notare che l'ampiezza del segnale modulato varia notevolemente tra un simbolo e l'altro.

[22] In pratica, l'indice $n$ si incrementa ogni due incrementi dell'indice $k$.

### 13.3.1 Prestazioni di QAM

Notiamo subito che la "distanza" tra due punti della costellazione QAM è maggiore (a parità di $L$) del caso PSK; pertanto, c'è da aspettarsi un miglioramento delle prestazioni (a parità di $E_b/N_0$) in quanto l'area che individua la probabilità di errore è ridotta. Il valore della probabilità di errore si determina dopo aver osservato che ciascuno dei due rami in fase e quadratura costituisce un segnale ASK multilivello con $L' = \sqrt{L}$.

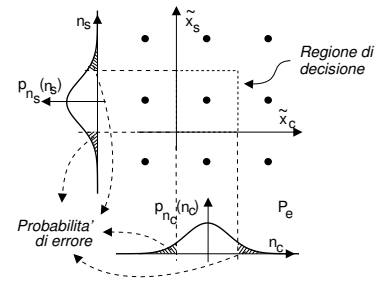
Abbiamo già calcolato che, per tale segnale, il rapporto $E_b/N_0$ dopo demodulazione è esattamente pari a quello del segnale modulato; pertanto otteniamo

$$\begin{aligned} P_\alpha &= P_e^c\,(simbolo) = P_e^s\,(simbolo) = \\ &= \left(1 - \frac{1}{L'}\right) \mathrm{erfc}\left\{\sqrt{3\frac{E_b}{N_0}\frac{\log_2 L'}{(L')^2 - 1}}\right\} \end{aligned}$$

Ricordando ora che $L' = \sqrt{L} = (L)^{1/2}$ e dunque $\log_2 L' = \frac{1}{2}\log_2 L$, si ottiene

$$P_\alpha = \left(1 - \frac{1}{\sqrt{L}}\right) \mathrm{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{N_0}\frac{\log_2 L}{L-1}}\right\}$$

La probabilità di errore (a simbolo) complessiva, cioè la probabilità che il segnale ricevuto $\underline{\tilde{x}} = \underline{x} + \underline{n}$ cada fuori della regione di decisione relativa all'$\underline{x}$ trasmesso, 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 a livelli dei due rami, consente di esprimere la probabilità di errore per bit come




$$P_e\,(bit) = \frac{P_e\,(carattere)}{\log_2 L} = \frac{2}{\log_2 L}\left(1 - \frac{1}{\sqrt{L}}\right) \mathrm{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{N_0}\frac{\log_2 L}{L-1}}\right\}$$

In figura 13.7 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 trasmettere 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?

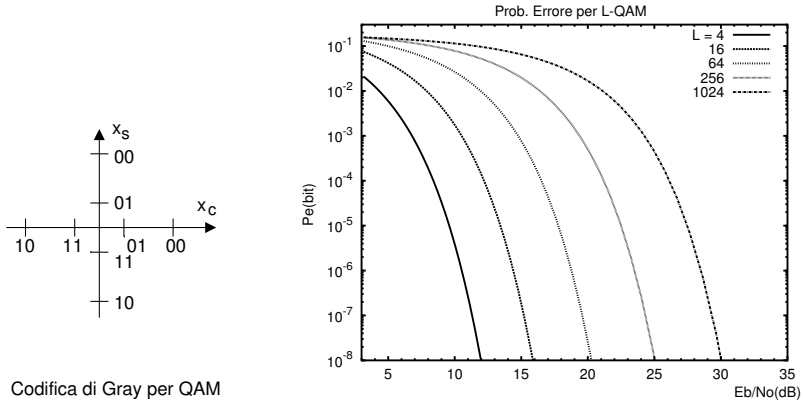



Figura 13.7: Prestazioni L-QAM con codice di Gray

**5)** Nel caso 16-QAM con la $P_e$ del punto 1) e $\mathcal{P}_x = 10^{-3}$ (Volt)$^2$, quale nuova $f_b$ è possibile raggiungere?

**Soluzione**

**1)** Osserviamo che $E_b = \mathcal{P}_x \cdot T_b = \frac{\mathcal{P}_x}{f_b} = \frac{10^{-3}}{10^6} = 10^{-9}$ (Volt)$^2$/Hz, mentre $N_0 = 2\mathcal{P}_N(f) = 4 \cdot 10^{-11}$ (Volt)$^2$/Hz, pertanto $\frac{E_b}{N_0} = 25$ e $\left(\frac{E_b}{N_0}\right)_{dB} = 10 \log_{10} 25 \simeq 14$ dB.

- 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.
- La banda occupata risulta $B = \frac{f_b}{\log_2 L} = \frac{10^6}{4} = 250$ KHz.

**2)** Le curve delle prestazioni per il QAM mostrano che con $E_b/N_0 = 14$ dB e 16 livelli, si ottiene $P_e \simeq 3 \cdot 10^{-6}$.

**3)** le stesse curve mostrano che, con il 16-QAM, la $P_e = 10^{-3}$ si ottiene con $E_b/N_0 = 10.5$ dB, ovvero 14-10.5=3.5 dB in meno, che corrispondono ad una potenza $\mathcal{P}'_x = \frac{\mathcal{P}_x}{10^{0.35}} = \frac{10^{-3}}{2.24} = 4.47 \cdot 10^{-5}$ (Volt)$^2$.

**4)** Dimezzare la banda equivale a raddoppiare $\log_2 L$, ovvero utilizzare un numero di livelli $L = (L')^2 = 256$. 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 eliminato riducendo in ugual 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, mostra che l'argomento sotto radice della $\mathrm{erfc}\{\}$ 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.

## 13.4 Schema riassuntivo delle prestazioni

La tabella seguente mette a confronto le prestazioni ottenibili con le tecniche di modulazione fin qui discusse, per un segnale dati a coseno rialzato.

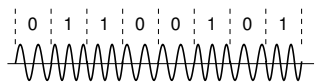
| *Metodo* | $M$ *bit/simbolo* | $P_e(bit)$ *(*) con codifica di Gray* | *Banda* RF | $\rho$ *[bit/sec/Hz]* |
|---|---|---|---|---|
| BPSK | 1 | $\frac{1}{2}\operatorname{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$ | $f_b\,(1+\gamma)$ | $\frac{1}{(1+\gamma)}$ |
| L-ASK | $\log_2 L$ | $\frac{1}{M}\left(1-\frac{1}{L}\right)\operatorname{erfc}\left\{\sqrt{3\frac{E_b}{N_0}\frac{M}{\left(L^2-1\right)}}\right\}$ | $f_b\frac{(1+\gamma)}{M}$ | $\frac{M}{(1+\gamma)}$ |
| L-FSK incorente | $\log_2 L$ | | $f_b\frac{L}{M}$ | $\frac{M}{L}$ |
| L-FSK coerente | $\log_2 L$ | *nota 11 a pag. 299* | $f_b\frac{L}{2M}$ | $\frac{M}{2L}$ |
| QPSK | 2 | $\frac{1}{2}\operatorname{erfc}\left\{\sqrt{\frac{E_b}{N_0}}\right\}$ | $f_b\,(1+\gamma)$ | $\frac{2}{(1+\gamma)}$ |
| L-PSK | $\log_2 L$ | $\frac{1}{M}\operatorname{erfc}\left\{\sin\left(\frac{\pi}{L}\right)\sqrt{\frac{E_b}{N_0}M}\right\}$ | $f_b\frac{(1+\gamma)}{M}$ | $\frac{M}{(1+\gamma)}$ |
| L-QAM | $\log_2 L$ | $\frac{2}{M}\left(1-\frac{1}{\sqrt{L}}\right)\operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{N_0}\frac{M}{L-1}}\right\}$ | $f_b\frac{(1+\gamma)}{M}$ | $\frac{M}{(1+\gamma)}$ |

(*) 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 § 7.5.5.

## 13.5 Altre possibilità

Sebbene il lettore possa pensare di averne già viste più di quante non si fosse mai aspettato, quelle illustrate sono solamente le tecniche *di base*. Senza ora entrare in dettagli troppo elaborati, accenniamo ad alcune ulteriori tecniche:

**MSK** (*Minimum Shift Keying*) simile all'FSK, ma con la variante di mantenere una *continuità di fase* tra simboli contigui, come accade per la famiglia di tecniche dette *Continuous phase modulation o* CPM[23]. Questa caratteristica consente una riduzione della banda occupata, in virtù dell'assenza di brusche variazioni di ampiezza.




**Offset keying** [24] Una variante del QAM e QPSK, in cui i periodi di simbolo per i 2 rami sono sfasati del 50%. La capacità di sincronizzazione del ricevitore risulta migliore, e la banda ridotta.

**Partial response QAM** [25] Il segnale modulato è filtrato, e si introduce deliberatamente una ISI in modo controllato. Migliora l'efficienza spettrale.

[23] http://en.wikipedia.org/wiki/Continuous_phase_modulation
[24] http://en.wikipedia.org/wiki/Phase-shift_keying#Offset_QPSK
[25] http://complextoreal.com/wp-content/uploads/2013/01/qpr.pdf

**Codifica differenziale** Nel caso in cui la portante usata per la demodulazione omodina presenti un errore di fase, si produce un errore sistematico nel processo di decisione. Come soluzione, si può realizzare il decisore in modo che prenda come riferimento di fase, la fase del simbolo precedente, causando un lieve peggioramento di prestazioni. All'appendice § 13.6.1 la questione viene approfondita.

**TCM** (*Trellis Coded Modulation*). Trellis significa *traliccio*, e rappresenta un modo di realizzare una codifica di canale che impone vincoli alle possibili sequenze. Il numero di livelli è artificialmente aumentato, ma i punti della costellazione risultante non sono tutti possibili, anzi solo un ristretto numero lo è, in funzione dei simboli precedenti. Il risultato è un miglioramento della $P_e$ (od una riduzione di $E_b$ necessaria) a spese di una maggiore occupazione di banda.

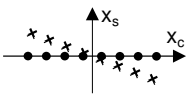
**OFDM** (*Orthogonal Frequency Division Multiplex*). Simile sotto certi aspetti all'FDM, in quanto suddivide la banda in più portanti, che sono però ora attive contemporaneamente. Ogni portante effettua tipicamente una modulazione QAM (con più livelli), e la spaziatura tra portanti è scelta in modo che le loro frequenze risultino ortogonali, annullando così le interferenze tra canali. La realizzazione si basa su *componenti hardware* che effettuano operazioni di FFT (Fast Fourier Transform) per sintetizzare il segnale e demodularlo. Il vantaggio principale è l'assenza di necessità di equalizzazione. Viene impiegato per ottenere velocità di trasmissione molto elevate su mezzi trasmissivi scarsamente condizionati, come nel caso dell'ADSL (vedi § 6.9.4) su linea telefonica. Alla appendice 13.6.4 è esposta una analisi dettagliata della tecnica.

**Spread spectrum** (*Modulazione ad espansione di spettro*). La stessa banda di frequenze è contemporaneamente utilizzata da più trasmissioni differenti, che non interferiscono tra loro perché ognuna utilizza forme d'onda ortogonali a quelle delle altre, e che sono caratterizzate da una occupazione spettrale *molto* superiore a quella minima. La tecnica di trasmissione risultante prende anche il nome di *Multiplazione a Divisione di Codice*. Alla appendice 13.6.5 è presente un approfondimento.

## 13.6 Appendici

### 13.6.1 Codifica differenziale

Qualora la portante con cui si effetta la demodulazione omodina presenti un errore di fase, il piano dell'inviluppo complesso subisce una rotazione[26], causando un errore sistematico nel processo di decisione dovuta allo spostamento dei punti di costellazione ottenuti campionando le c.a. di b.f. (crocette in figura) rispetto a quelli che si otterrebbero nel caso di demodulazione coerente (pallini). Per rimediare al problema, si può estendere il principio usato a pag. 69 per segnali di banda base al caso delle modulazioni numeriche, rendendo la decisione su quale punto di costellazione sia stato ricevuto indipendente dalla fase della portante di demodulazione, ma dipendente invece dalla fase dell'inviluppo complesso osservata per il simbolo precedente. Ciò si realizza modificando il modo in cui vengono determinati i punti




[26] Vedi al riguardo la trattazione svolta ai § 11.2.1.1 e 14.5.4.1.

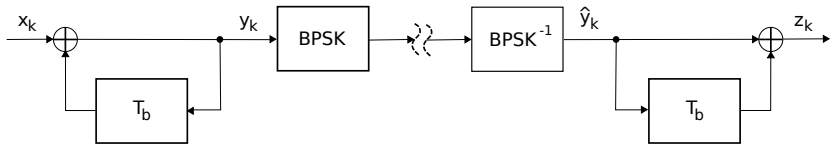



Figura 13.8: Codifica e detezione differenziali

di costellazione da trasmettere, scegliendoli ora in funzione *di una coppia* di simboli consecutivi, anziché di uno solo.

Per illustrare la tecnica, procediamo con un esempio relativo al caso di trasmissione BPSK della sequenza $x_k = 001011011010010$. La fig. 13.8, simile[27] a quella a pag. 413, mostra la sequenza delle operazioni necessarie, e che consistono nel trasformare il messaggio binario $x_k$ in quello $y_k$ in base alla relazione

$$y_k = x_k \oplus y_{k-1}$$

(in cui l'operatore $\oplus$ rappresenta un *or esclusivo*), e quindi effettuare la modulazione BPSK di $y_k$ anziché $x_k$. Dal lato ricevente, il segnale BPSK viene demodulato ottenendo la sequenza $\hat{y}_k$, che viene a sua volta trasformata in $z_k$ in base all'espressione

$$z_k = \hat{y}_k \oplus \hat{y}_{k-1}$$

che, in assenza di errori (ossia $\hat{y}_k = y_k$ per tutti i $k$), permette di ottenere nuovamente i valori del messaggio originario $x_k$ a partire dalla sequenza $z_k$. E' infatti facile verificare che le trasformazioni descritte producono il risultato mostrato a lato, che evidenzia come[28] la trasformazione sia effettivamente invertibile. Notiamo come 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.

$x_k = 001011011010010$
$y_k = 001101101100011$
$z_k = /01011011010010$

Se assumiamo ora di rappresentare lo zero con una fase nulla, e l'uno con una fase di $\pi$, riscriviamo la (13.1) come

$$x_{BPSK}(t) = a\cos(\omega_0 t + \varphi(t)) \quad \text{con} \quad \varphi(t) = \pi \cdot \sum_{k=-\infty}^{\infty} x_k \text{rect}_{T_b}(t - kT_b)$$

e ponendo per semplicità il periodo di bit pari ad un ciclo di portante, otteniamo in fig. 13.9 la forma d'onda BPSK associata alla sequenza originaria $x_k$, posta a confronto con quella relativa invece a $y_k$ ed indicata come DBPSK, in cui la D sta per *differenziale*. Anche se i simboli del segnale DBPSK sono in corrispondenza univoca con i bit della sequenza $y_k$, osserviamo che possono anche essere derivati *direttamente* dall'esame della $x_k$, dato che la fase del DBPSK si inverte per gli $x_k = 1$, e non si inverte per $x_k = 0$. Pertanto, è possibile realizzare un demodulatore che, senza utilizzare la formula che calcola $z_k$, prende le corrette decisioni a riguardo dei bit di $x_k$ basandosi sui *cambiamenti* di fase anziché sui loro valori assoluti.

D'altra parte però lo schema di fig. 13.8 mostra l'uso di un modem BPSK *convenzionale*, delegando le operazioni di differenziazione e sua inversa a circuiti addizionali.

[27] La similitudine non è per nulla casuale, dato che qualora il predittore mostrato a pag. 413 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.

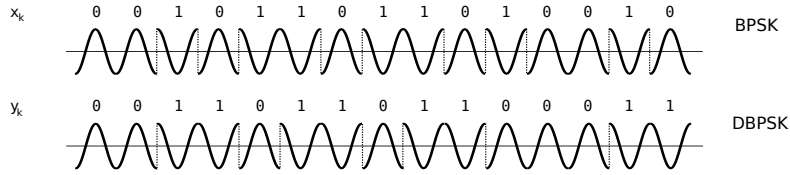



Figura 13.9: Segnale BPSK e BPSK *differenziale*

Verifichiamo quindi che anche in tale forma si riesce a recuperare il segnale originario, nel caso in cui ad esempio la portante di demodulazione presenti un errore di fase di $\pi$[29], tale da causare *l'inversione* di tutti i bit decodificati, producendo così un messaggio $\hat{y}_k =$ 110010010011100. E' facile verificare che anche in questo caso, applicando l'operatore $z_k = \hat{y}_k \oplus \hat{y}_{k-1}$, si ottiene nuovamente la sequenza originaria.

Esaminiamo ora cosa accade in presenza di errori: supponiamo di ricevere un messaggio $y_k =$ 00<u>0</u>101101100011, in cui il terzo bit (sottolineato) è errato. Calcolando $z_k = \hat{y}_k \oplus \hat{y}_{k-1}$ questa volta si ottiene $z_k =$ /00<u>1</u>1011010010 che risulta uguale a $x_k$ tranne che nel terzo e quarto bit. Infatti, dato che $z_k$ dipende dagli indici $k$ e $k-1$ di $y$, l'effetto dell'errore non si propaga oltre il bit successivo a quello errato. Dato quindi che ad ogni errore del decisore si ottengono due bit errati anziché uno, a parità di $E_b/N_0$ il DBPSK è affetto da un tasso di errore circa *doppio* di quello del BPSK.

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 seguente[30] per $L = 4$([31]), ovvero nel caso della modulazione DQPSK. L'inviluppo complesso di tale segnale quindi assumerà, negli istanti di simbolo, valori la cui fase dipende dalla fase del simbolo precedente, incrementata del $\Delta\theta$ mostrato in tabella, consentendo la corretta ricezione anche in presenza di una portante di demodulazione affetta da errori di fase multipli di $\frac{\pi}{2}$. Anche qui se (a causa del rumore) si verifica un errore di ricezione, questo si propaga anche al simbolo successivo.

| $x_{k-1}x_k$ | $\Delta\theta$ |
|---|---|
| 00 | 0 |
| 01 | $\pi/2$ |
| 11 | $\pi$ |
| 10 | $-\pi/2$ |

Nel caso del QAM si può nuovamente applicare una forma di codifica differenziale, ma lo schema di corrispondenza tra gruppi di bit e punti della costellazione è più complesso[32], e non viene qui trattato.

Infine, citiamo l'uso che viene fatto della codifica differenziale nei sistemi di trasmissione multiportante come l'OFDM (§ 13.6.4), in cui il sincronismo di fase viene acquisto in corrispondenza di una sottoportante, ma la risposta in frequenza del canale causa una rotazione del piano dell'inviluppo complesso relativo alle altre sottoportanti. In tale circostanza, la codifica differenziale viene applicata ai gruppi di bit associati a portanti contigue, limitando così l'errore di fase introdotto dal canale, a quello che si verifica in un intervallo di frequenza pari alla spaziatura tra sottoportanti.

[29] E' questo un caso tuu'altro che anomalo, in quanto il segnale BPSK contiene entrambe le fasi in egual misura (se i bit sono equiprobabili) e quindi un circuito come il PLL di pag. 265 può *agganciarsi* indifferentemente all'una o all'altra.

[30] Tratta da *Andrea Goldsmith*, Wireless Communications, pag. 151.

[31] 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 $\Delta\theta_k = k\frac{2\pi}{L-1}$.

[32] Vedi ad es. *Krzysztof Wesolowski*, Introduction to Digital Communication Systems, Wiley, pag. 328.

### 13.6.2 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, ed evitare l'insorgenza di ISI. Le due problematiche posso essere affrontate l'una di seguito all'altra, adottando le soluzioni già esposte[33]. D'altra parte sono ora possibili varianti, che ad esempio usano entrambe le c.a. di b.f.[34], oppure che tentano di acquisire per primo il sincronismo di simbolo, ed usano i valori delle c.a. di b.f. ricevute per effettuare le possibili correzioni alla fase dell'oscillatore di demodulazione[35].

Qualora la portante di demodulazione presenti una ambiguità di fase residua si può applicare la tecnica della codifica differenziale esposta al § 13.6.1, oppure inserire una sequenza di simboli noti (o *flag*) all'inizio della trama trasmissiva, in modo che il confronto tra i valori attesi e quelli ricevuti permetta di correggere tale ambiguità. Da notare che i *flag* o *trailer* ad inizio trama possono essere vantaggiosamente usati anche da schemi di recupero del clock del tipo di quelli a pag. 5.5.2.

### 13.6.3 FSK ortogonale

A pagina 297 è stata introdotta la modulazione FSK, e nelle note si è iniziata la discussione relativa alle condizioni di ortogonalità tra le frequenze di confronto ed il segnale ricevuto; prendiamo qui in considerazione segnali del tipo generale $\cos\left[2\pi\left(f_0+\Delta f_k\right)t+\phi_k\right]$, in cui è incluso un errore di fase aleatorio tra simboli, in modo da esaminare le differenze tra il caso di modulazione coerente ed incoerente.

Iniziamo dunque con lo sviluppare l'espressione dell'integrale di intercorrelazione $\rho=\int_0^{T_s}\cos\left[2\pi\left(f_0+m\Delta\right)t\right]\cos\left[2\pi\left(f_0+n\Delta\right)t+\phi\right]dt$ facendo uso della relazione $\cos\alpha\cos\beta=\frac{1}{2}\left[\cos\left(\alpha+\beta\right)+\cos\left(\alpha-\beta\right)\right]$ e riferendoci per semplicità al caso di due frequenze contigue (ponendo $m=0$ ed $n=1$):

$$\rho = \frac{1}{2}\int_0^{T_s}\left\{\cos\left[2\pi\left(2f_0+\Delta\right)t+\phi\right]+\cos\left[2\pi\Delta t-\phi\right]\right\}dt = \tag{13.9}$$

$$= \frac{1}{2}\int_0^{T_s}\cos\left[2\pi\left(2f_0+\Delta\right)t+\phi\right]dt+\frac{1}{2}\int_0^{T_s}\cos\left[2\pi\Delta t-\phi\right]dt \tag{13.10}$$

Per quanto riguarda il primo integrale, esso assume un valore nullo se $2f_0+\Delta=\frac{k}{T_s}$, perché in tal caso in un intervallo $T_s$ entrano un numero intero di periodi, ed il coseno ha valor medio nullo. Concentriamoci allora sul valore di $\Delta$ che annulla anche il secondo integrale, che riscriviamo facendo uso della relazione $\cos\left(\alpha-\beta\right)=$

[33] Per il recupero della portante si possono usare circuiti del tipo di § 11.4.2, mentre l'uso del PLL (§ 11.2.1.3) non è possibile a causa della assenza di residui di portante. Una volta acquisto il sincronismo di frequenza, quello di simbolo può essere ottenuto mediante schemi operanti in banda base, come quelli a pag. 5.5.2.

[34] Come nel caso del recupero di portante basato sul *Costas loop*, vedi ad es. http://goo.gl/WkFRcP.

[35] Vedi as es. http://en.wikipedia.org/wiki/Carrier_recovery#Decision-directed

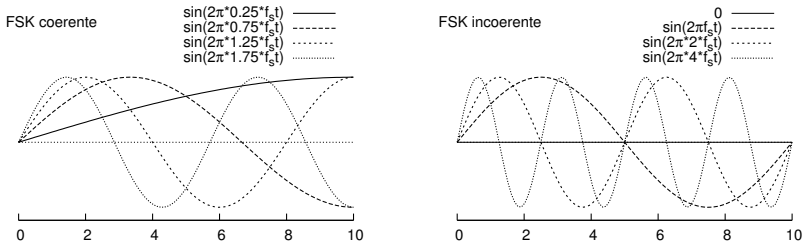



Figura 13.10: Forme d'onda ortogonali nei casi di modulazione *coerente* ed *incoerente*

$\cos\alpha\cos\beta + \sin\alpha\sin\beta$:

$$\begin{aligned}
&\int_0^{T_s} \cos(2\pi\Delta t - \phi)\, dt = \\
= &\int_0^{T_s} [\cos(2\pi\Delta t)\cos\phi + \sin(2\pi\Delta t)\sin\phi]\, dt = \\
= &\left.\frac{\sin(2\pi\Delta t)}{2\pi\Delta}\right|_0^{T_s} \cdot \cos\phi - \left.\frac{\cos(2\pi\Delta t)}{2\pi\Delta}\right|_0^{T_s} \cdot \sin\phi = \\
= &\; T_s \left[\frac{\sin(2\pi\Delta T_s)}{2\pi\Delta T_s} \cdot \cos\phi + \frac{1-\cos(2\pi\Delta T_s)}{2\pi\Delta T_s} \cdot \sin\phi\right] \qquad (13.11)
\end{aligned}$$

Osserviamo ora che, nel caso in cui $\phi = 0$, il secondo termine della (13.11) si annulla per qualunque $\Delta$. Esaminiamo quindi ora solamente il primo termine, individuando così il risultato relativo al caso di

**Modulazione coerente** Il termine $\frac{\sin(2\pi\Delta T_s)}{2\pi\Delta T_s}$ 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 il primo termine della (13.10) si annulli, deve sussistere la relazione $2f_0 + \Delta = 2f_0 + \frac{f_s}{2} = \frac{k}{T_s} = kf_s$, che fornisce la condizione $f_0 = f_s\frac{2k-1}{4}$, ossia $f_0$ deve essere scelta come uno tra i valori $\frac{1}{4}f_s$, $\frac{3}{4}f_s$, $\frac{5}{4}f_s$, $\frac{7}{4}f_s, \ldots$. Notiamo come la spaziatura $\frac{f_s}{2}$ tra i possibili valori per la portante, coincida con quella tra le frequenze di segnalazione. Pertanto la parte sinistra della figura 13.10 rappresenta, disegnate in un intervallo pari a $T_s$, sia le portanti che possono essere usate, sia le prime frequenze che è possibile adottare per un modulazione FSK *coerente* basata sul valore minimo di $f_0$ pari a $\frac{1}{4}f_s$.

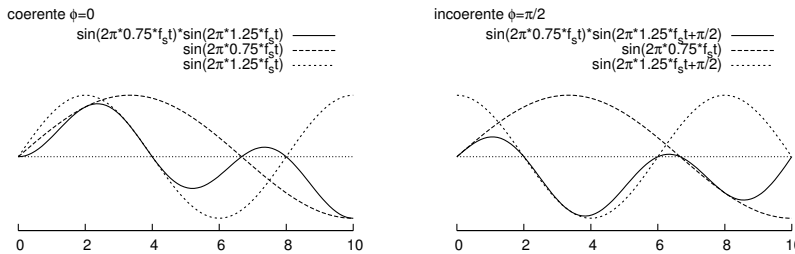
Nel caso in cui $f_0$ non assuma uno dei valori individuati, il primo termine di (13.10) non si annulla, ma se $f_0 \gg \frac{1}{T_s}$, risulta trascurabile rispetto al secondo. Pertanto, nel caso di trasmissioni su canali di tipo passa-banda, la scelta di $f_0$ non è più determinante; d'altra parte, la spaziatura tra le frequenze di segnalazione pari a $\frac{f_s}{2}$ produce comunque il risultato che due frequenze di segnalazione contigue, accumulano in un intervallo $T_s$ una differenza di fase di mezzo periodo.

**Modulazione incoerente** In questo caso si ha $\phi \neq 0$. In generale la (13.11) presenta entrambi i termini; mentre il primo (come ora esaminato) si annulla per $\Delta = \frac{k}{2T_s}$, il

secondo invece è nullo solo se $\Delta = \frac{k}{T_s}$. Questa circostanza determina il risultato che occorre ora adottare una spaziatura tra portati doppia della precedente, e pari cioè a $\Delta = f_s$.

Tornando ad esaminare la (13.10), il suo primo termine ora si annulla ora 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 sia identica al caso precedente, ma la spaziatura necessaria alle frequenze di segnalazione si sia ora dimezzata. La circostanza che sia adesso ammessa anche una portante a frequenza nulla consente ora di tracciare la parte destra della figura 13.10, che mostra le prime frequenze di segnalazione che è possibile adottare per una modulazione FSK *incoerente* basata sul valore minimo di $f_0 = 0$.

**Verifica grafica** La figura che segue mostra il risultato del prodotto di due frequenze ortogonali distanti $\frac{f_s}{2}$ e calcolate in assenza di errore di fase (a sinistra) e con un errore di fase pari a $\phi = \frac{\pi}{2}$. Si può notare come in questo secondo caso si perda l'ortogonalità tra i segnali, essendo il risultato prevalentemente negativo.




**Discussione sull'ottimalità per $L \to \infty$** Osserviamo innanzitutto che il ricevitore a correlazione commette errore nel caso in cui il rumore sovrapposto al segnale di ingresso sia casualmente "simile" ad una delle cosinusoidi utilizzate per la trasmissione. In tal caso, l'uscita dell'integratore relativo alla frequenza "simile" può superare quella relativa alla frequenza trasmessa.

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.

Chiaramente, all'aumentare di $L$ aumenta proporzionalmente la complessità del ricevitore, che deve disporre di un numero di correlatori crescente. Pertanto, le prestazioni ideali per $L$ che tende ad infinito rivestono solamente un interesse teorico.

### 13.6.4 OFDM

La sigla sta per ORTHOGONAL FREQUENCY DIVISION MULTIPLEX, ossia *multiplazione a divisione di frequenza ortogonale*. Si tratta della tecnica di modulazione numerica adottata per le trasmissioni ADSL[36] DVB-T, e WIFI, ed ha la particolarità di utilizzare in modo ottimo la banda del canale, e di ridurre l'operazione di equalizzazione ad un prodotto tra vettori.

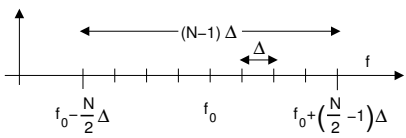
[36] ADSL = *Asymmetric Digital Subscriber Loop*, dove il *Subscriber Loop* rappresenta il circuito di utente che si realizza tra apparecchio e centrale quando si solleva il telefono. Vedi § 6.9.4.

#### 13.6.4.1 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[37] della modulazione FSK, in cui tutte le diverse frequenze

$$f_n = f_0 + \Delta \cdot \left( n - \frac{N}{2} \right) \tag{13.12}$$




con $n = 0, 1, ..., N-1$, sono utilizzate contemporaneamente, ed ognuna realizza una modulazione numerica anche a più livelli (es. QPSK o QAM) con impulso NRZ rettangolare. Indicando ora 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

$$x_{OFDM}(t) = \sum_k \mathrm{rect}_T(t - kT) \sum_{n=0}^{N-1} \left( a_{n_c}^k \cos \omega_n (t - kT) - a_{n_s}^k \sin \omega_n (t - kT) \right) \tag{13.13}$$

$$= \sum_{k=-\infty}^{\infty} \delta(t - kT) * \left( \mathrm{rect}_T(t) \sum_{n=0}^{N-1} \left( a_{n_c}^k \cos \omega_n t - a_{n_s}^k \sin \omega_n t \right) \right) \tag{13.14}$$

in cui la prima sommatoria (su $k$) identifica gli istanti di simbolo, e la seconda (su $n$) le diverse portanti.

E' facile osservare[38] che tale segnale presenta un inviluppo complesso rispetto a $f_0$ pari a

$$\underline{x}_{OFDM}(t) = \sum_{k=-\infty}^{\infty} \delta(t - kT) * \left( \mathrm{rect}_T(t) \sum_{n=0}^{N-1} \underline{a}_n^k e^{j2\pi \left[ \Delta \left( n - \frac{N}{2} \right) \right] t} \right) \tag{13.15}$$

L'espressione (13.14) non vincola la durata $T$ di un simbolo ad un valore particolare; deve però risultare $T \geq T_0 = \frac{1}{\Delta}$, in quanto il ricevitore opera sul segnale una finestratura temporale di estensione $T_0 = \frac{1}{\Delta}$ allo scopo di rendere ortogonali tra loro[39] le frequenze $f_n = f_0 + \Delta \cdot \left( n - \frac{N}{2} \right)$, e mettere in grado il ricevitore di calcolare i valori $\underline{a}_n^k$ per tutti gli $n$ presenti all'istante $t = kT$, mediante un ricevitore concettualmente simile a quello a correlazione presentato a pag. 298.

[37] 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.

[38] Osserviamo innanzitutto che per un segnale

$$x(t) = \cos \omega_1 t = \frac{1}{2} \left( e^{j\omega_1 t} + e^{-j\omega_1 t} \right)$$

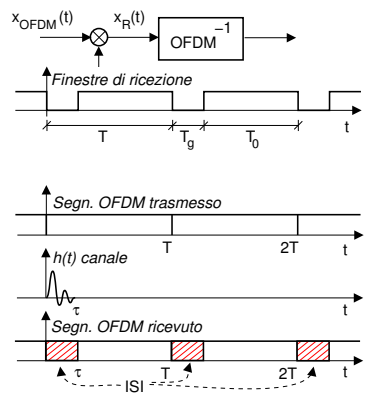
risulta $x^+(t) = \frac{1}{2} e^{j\omega_1 t}$, e quindi il suo inviluppo complesso $\underline{x}(t)$ calcolato rispetto ad $f_0$ vale

$$\underline{x}(t) = 2x^+(t) e^{-j\omega_0 t} = 2 \frac{1}{2} e^{j\omega_1 t} e^{-j\omega_0 t} = e^{j(\omega_1 - \omega_0)t}$$

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 (13.12) si ottiene $f_n - f_0 = f_0 + \Delta \cdot \left( n - \frac{N}{2} \right) - f_0 = \Delta \cdot \left( n - \frac{N}{2} \right)$ e quindi la (13.15).

[39] Come mostrato per il caso *incoerente* discusso al § 13.6.3

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"[40] risiede nel fatto che, in presenza di un canale non perfetto, la parte iniziale di ogni simbolo risulta corrotta (vedi figura) da una interferenza intersimbolica (ISI) dovuta al risultato della convoluzione tra la coda del simbolo precedente e l'$h(t)$ del canale.




Consideriamo ora un solo simbolo (fissiamo $k = 0$ e consideriamo l'origine dei tempi ritardata di $T_g$) ricevuto nell'intervallo $T_0 = \frac{1}{\Delta} \leq T$, con inviluppo complesso

$$\underline{x}_{T_0}(t) = \mathrm{rect}_{T_0}(t) \cdot \sum_{n=0}^{N-1} \underline{a}_n \mathrm{e}^{j2\pi\left[\Delta\left(n-\frac{N}{2}\right)\right]t} \tag{13.16}$$

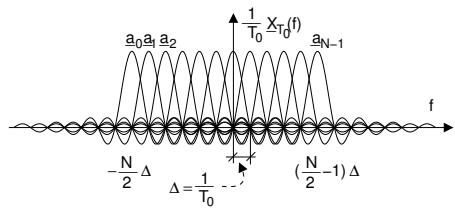
e calcoliamone la trasformata per determinare l'occupazione di banda:

$$\underline{X}_{T_0}(f) = T_0 \,\mathrm{sinc}(fT_0) * \sum_{n=0}^{N-1} \underline{a}_n \delta\left(f - \Delta\left(n - \frac{N}{2}\right)\right) = \tag{13.17}$$

$$= T_0 \sum_{n=0}^{N-1} \underline{a}_n \,\mathrm{sinc}\left(\left(f - \Delta\left(n - \frac{N}{2}\right)\right) T_0\right) \tag{13.18}$$

Otteniamo allora il risultato mostrato in figura a lato, dove si evidenzia come ogni funzione *sinc* risulti moltiplicata per uno dei coefficienti $\underline{a}_n$, che potrebbero quindi essere ri-ottenuti in ricezione campionando (in modo complesso) $\underline{X}(f)$su frequenze spaziate di $\Delta$.

Dalla (13.18) si ottiene la densità di potenza $\mathcal{P}_{\underline{x}_R}(f)$ dell'inviluppo complesso $\underline{x}_R(t)$ ricevuto e finestrato, di cui $\underline{X}_{T_0}(f)$ rappresenta la trasformata di un generico periodo principale, dopo aver specificato il numero di bit $M_n$ e la potenza $\mathcal{P}_n$ assegnate alla portante *n-esima*, vincolate a fornire




$$\sum_{n=0}^{N-1} M_n = M \quad \text{e} \quad \sum_{n=0}^{N-1} \mathcal{P}_n = \mathcal{P}$$

in cui $M$ è il numero di bit/simbolo trasportati dall'insieme delle portanti e $\mathcal{P}$ è la potenza di $x_{OFDM}(t)$. Essendo le portanti ortogonali nel periodo $T_0$, possiamo applicare la relazione $\mathcal{P}(f) = \sigma_a^2 \frac{\mathcal{E}(f)}{T}$ (vedi § 9.2.4) alle singole componenti e sommare i contributi (vedi nota 11 a pag. 24). Nel caso in cui la sequenza $\{\underline{a}_n\}$ sia a valori

[40] 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.

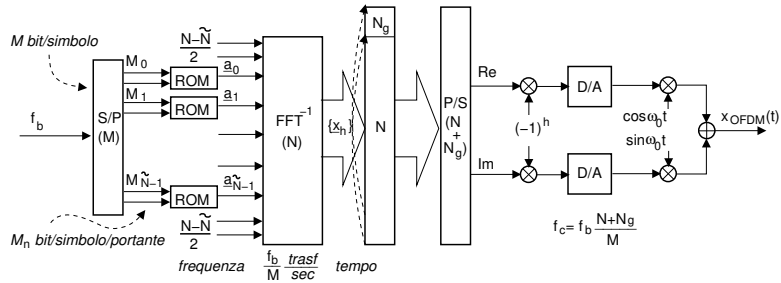



Figura 13.11: Architettura di un modulatore OFDM numerico

indipendenti ed uniformemente distribuiti su di una costellazione QAM quadrata di lato $2\sqrt{3\mathcal{P}_n \frac{\sqrt{L_n}-1}{\sqrt{L_n}+1}}$ (vedi nota[41]) in cui $L_n = 2^{M_n}$, si ottiene che $\sigma^2_{\underline{a}_n} = E\left\{\underline{a}_n^2\right\} = 2\mathcal{P}_n$, permettendo di scrivere la densità di potenza di $\underline{x}_R(t)$ in ingresso al demodulatore come

$$\mathcal{P}_{\underline{x}_R}(f) = \frac{1}{T}\sum_{n=0}^{N-1} 2\mathcal{P}_n T_0^2 \,\mathrm{sinc}^2\left(\left(f - \Delta\left(n - \frac{N}{2}\right)\right)T_0\right)$$

a cui corrisponde una potenza complessiva[42] pari a

$$\mathcal{P}_{\underline{x}_R} = \frac{1}{T}\sum_{n=0}^{N-1} 2\mathcal{P}_n T_0 = 2\frac{T_0}{T}\sum_{n=0}^{N-1}\mathcal{P}_n$$

Infine, la potenza totale di $x_r(t)$ risulta

$$\mathcal{P}_{x_R} = \mathcal{P}^+_{x_R} + \mathcal{P}^+_{x_R} = 2\frac{1}{4}\mathcal{P}_{\underline{x}_R} = \frac{T_0}{T}\sum_{n=0}^{N-1}\mathcal{P}_n$$

in cui è evidenziata la perdita di potenza legata alla presenza del preambolo.

#### 13.6.4.2 Architettura di modulazione

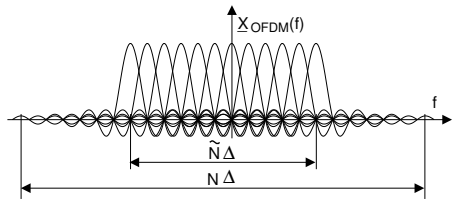
Una caratteristica fondamentale della modulazione OFDM è quella di essere realizzata senza *oscillatori e integratori*, ma completamente tramite circuiti digitali.

Con riferimento alla figura 13.11, il flusso binario a frequenza $f_b$ viene parallelizzato per formare simboli ad $L = 2^M$ livelli a frequenza $f_s = \frac{f_b}{M} = \frac{f_b}{\log_2 L}$. Questi $M$ bit/simbolo sono suddivisi in $\tilde{N}$ gruppi di $M_n$ $(n = 0, 1, ..., \tilde{N}-1)$ bit ciascuno, con $M = \sum_{n=0}^{\tilde{N}-1} M_n$. Ogni gruppo di $M_n$ bit produce un punto di costellazione $\underline{a}_n$ scelto tra $L_n = 2^{M_n}$ punti possibili.

[41] Al § 9.9.4 si è mostrato che se gli $a_n$ sono v.a. indipendenti e distribuite uniformemente su $L'$ livelli tra $\pm A$, si ottiene $\sigma_a^2 = \frac{A^2}{3}\frac{L'+1}{L'-1}$. Nel caso di una costellazione QAM quadrata ad $L$ livelli si ha $L' = \sqrt{L}$, e se le realizzazioni sui rami in fase e quadratura sono indipendenti risulta $\sigma^2_{\underline{a}_n} = E\left\{(a_{n_c} + j a_{n_s})^2\right\} = 2\sigma_a^2 = \frac{2A^2}{3}\frac{\sqrt{L}+1}{\sqrt{L}-1}$; volendo eguagliare tale valore a $2\mathcal{P}_n$, occorre quindi scegliere $A = \sqrt{3\mathcal{P}_n \frac{\sqrt{L}-1}{\sqrt{L}+1}}$.

[42] Si è fatto uso del risultato $\int_{-\infty}^{\infty} T_0^2 \mathrm{sinc}^2(fT_0)\,df = T_0$.

La sequenza $\{\underline{a}_n\}$ viene *arricchita* con $N - \tilde{N}$ valori nulli (metà all'inizio e metà alla fine) ottenendo una nuova sequenza $\{\underline{a}_n\}$ di $N$ valori, in modo che la sommatoria in (13.18) 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 presi a frequenza $f_c = N\cdot\Delta\ \frac{campioni}{secondo}$. Il blocco indicato come $\mathrm{FFT}^{-1}$ svolge proprio questa operazione: esso infatti esegue efficentemente[43] il calcolo




$$\sum_{n=0}^{N-1}\underline{a}_n \mathrm{e}^{j2\pi\frac{n}{N}h} = \frac{1}{(-1)^h}\underline{x}_{T_0}(hT_c) \tag{13.19}$$

Il risultato della $\mathrm{FFT}^{-1}$ è quindi una sequenza di coefficienti complessi $\{\underline{x}_h\}$, che a meno di un segno alterno sono uguali ai campioni dell'inviluppo complesso $\underline{x}_{T_0}(t)$ fornito dalla (13.16) relativamente ad un simbolo. Il preambolo da trasmettere durante il tempo di guardia $T_g$ si ottiene *aggiungendo* in testa a $\{\underline{x}_h\}$ un gruppo di campioni prelevati dalla coda[44].

Infine, le parti reale ed immaginaria di $\{\underline{x}_h\}$ sono inviate ad una coppia di convertitori D/A operanti a $f_c = \frac{N+N_g}{T} = \frac{N}{T_0} = N\Delta$ in modo da ottenere le c.a. di b.f., utilizzate per produrre il segnale $x_{OFDM}(t)$ mediante una coppia di modulatori in fase e quadratura.

---

[43] La (13.19) è in qualche modo simile alla formula di ricostruzione (2.3) (vedi pag. 17) per il segnale periodico limitato in banda $\pm\frac{N}{2}F$

$$x(t) = \sum_{m=-N/2}^{N/2} X_m \mathrm{e}^{j2\pi mFt}$$

che calcolata per $t = hT_c = \frac{h}{NF}$ fornisce $x(hT_c) = \sum_{m=-N/2}^{N/2} X_m \mathrm{e}^{j2\pi\frac{m}{N}h}$. Ponendo ora $n = m + \frac{N}{2}$ e $Y_n = X_{n-\frac{N}{2}}$ otteniamo

$$x(hT_c) = \sum_{n=0}^{N-1} Y_n \mathrm{e}^{j2\pi\frac{n-\frac{N}{2}}{N}h} = \mathrm{e}^{-j\pi h}\sum_{n=0}^{N-1} Y_n \mathrm{e}^{j2\pi\frac{n}{N}h}$$

Osservando ora che $\mathrm{e}^{-j\pi h} = (-1)^h$ e confrontando con la (13.16) si ottiene la (13.19). La coppia di relazioni

$$X_n = \frac{1}{N}\sum_{h=0}^{N-1} x_h \mathrm{e}^{-j2\pi\frac{h}{N}n} \quad \text{e} \quad x_h = \sum_{n=0}^{N-1} X_n \mathrm{e}^{j2\pi\frac{n}{N}h}$$

sono chiamate *Discrete Fourier Transform* (DFT) diretta e inversa, in quanto costituiscono la versione discreta della trasformata di Fourier (vedi § 4.2), e consentono il calcolo di una serie di campioni in frequenza a partire da campioni nel tempo e viceversa.

La FFT (*Fast Fourier Transform*) esegue le stesse operazioni, ma organizza i calcoli sfruttando le proprietà di periodicità degli esponenziali complessi, in modo da realizzare una mole di calcoli non superiori a $N\cdot\log_2 N$ per trasformate ad $N$ punti. Questo risultato è possibile solamente se $N$ è una potenza di 2, e quindi la modulazione OFDM opera necessariamente su $N = 2^H$ portanti, con $H$ intero.

[44] In effetti la (13.19) 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*.

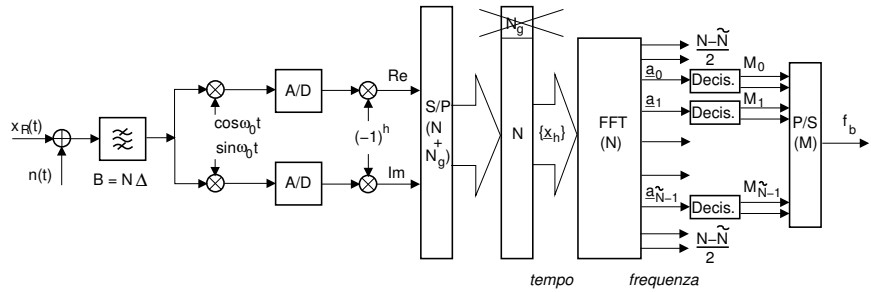



Figura 13.12: Architettura di un demodulatore OFDM numerico

#### 13.6.4.3 Efficienza dell'OFDM

Come vedremo tra breve, 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) è della stessa natura di quella introdotta dal roll-off $\gamma$ di un impulso a coseno rialzato, in quanto ha lo scopo di evitare che si verifichino fenomeni di interferenza tra simboli. Osserviamo che l'efficienza migliora all'aumentare di $T$ e di $N$, dato che $T_g$ ed $N - \tilde{N}$ sono fissi.

#### 13.6.4.4 Architettura di demodulazione

Per ottenere gli elementi della sequenza $\{\underline{a}_n\}$ e quindi il gruppo di $M$ bit che hanno originato il simbolo, si adotta l'architettura mostrata in figura 13.12 che svolge una azione del tutto inversa a quella del modulatore.

Innanzitutto il ricevitore deve acquisire il sincronismo di frequenza e di simbolo (vedi § 13.6.4.10) per determinare l'inizio della ricezione di un singolo blocco di campioni. Il segnale ricevuto viene quindi 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, gli $f_c \cdot T_g = N_g$ campioni del preambolo sono rimossi, ed una FFT permette di ottenere i valori

$$\frac{1}{N}\sum_{h=0}^{N-1} \underline{x}_h \mathrm{e}^{-j2\pi\frac{h}{N}n} = \underline{X}_{T_0}\left(\left(n - \frac{N}{2}\right)\Delta\right) = \underline{a}_n \tag{13.20}$$

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 nuovamente serializzato per produrre gli $M$ bit che hanno dato origine al simbolo.

#### 13.6.4.5 Prestazioni

Il calcolo della $P_e$ per bit si basa su quello relativo alle probabilità di errore $P_{e_n}$ condizionato alle singole portanti. Dato che la portante *n-esima* trasporta $M_n$ bit/simbolo, la probabilità che un bit generico provenga dalla portante *n-esima* risulta pari a $Pr(n) = \frac{M_n}{M}$ e quindi la probabilità che sia errato è pari a

$$P_e = \sum_{n=0}^{\tilde{N}-1} Pr(n) P_{e/n} = \frac{1}{M} \sum_{n=0}^{\tilde{N}-1} M_n P_{e_n} \tag{13.21}$$

**Calcolo della $P_e$ per portante** La $P_{e_n}$ dipende dal numero di livelli $L_n = 2^{M_n}$ scelto per la portante n-esima, e dal rapporto $\left(\frac{E_b}{N_0}\right)_n$ locale.

Per determinare il valore di $P_{e_n}$ conviene applicare i risultati trovati al § 13.3.1 per la modulazione QAM, particolarizzati al caso attuale, in cui si adottano impulsi rettangolari di durata $T_0 = \frac{1}{\Delta}$. Attribuendo ai punti delle costellazioni gruppi di bit secondo la codifica di Gray, risulta

$$P_{e/n} = \frac{2}{\log_2 L_n} P_{\alpha_n} \quad \text{in cui} \quad P_{\alpha_n} = \left(1 - \frac{1}{\sqrt{L_n}}\right) \operatorname{erfc}\left\{\sqrt{\frac{3}{2} SNR_n \frac{1}{L_n - 1}}\right\} \tag{13.22}$$

è la probabilità di errore su di uno dei rami (in fase od in quadratura) della *n-esima* costellazione QAM, come ottenuta in § 7.5.3 per il caso di banda base[45].

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 $\mathcal{P}_{\underline{N}_n}$.

**Potenza di rumore per portante** Per quanto riguarda $\mathcal{P}_{\underline{N}_n}$, si tratta di applicare la (13.20) alla sequenza $\left\{(-1)^h \underline{n}(hT_c)\right\}$ dei campioni dell'inviluppo complesso del rumore, e determinare il valore

$$\mathcal{P}_{\underline{N}_n} = E\left\{(\underline{N}_n)^2\right\} = \sigma^2_{\underline{N}_n} \quad \text{in cui} \quad \underline{N}_n = \frac{1}{N}\sum_{h=0}^{N-1} (-1)^h \underline{n}(hT_c)\, \mathrm{e}^{-j2\pi \frac{h}{N} n}$$

in virtù del fatto che i valori $\underline{n}(hT_c)$ sono a media nulla, che (con $n$ fissato) la FFT ne effettua una combinazione lineare con coefficienti $\mathrm{e}^{-j2\pi \frac{h}{N} n}$, e che essendo $\underline{n}(t)$ ergodico è possibile scambiare medie temporali e di insieme. Sviluppando

$$(\underline{N}_n)^2 = \underline{N}_n \underline{N}^*_n = \frac{1}{N^2}\sum_{h=0}^{N-1}\sum_{k=0}^{N-1} (-1)^{h-k} \underline{n}(hT_c)\, \underline{n}^*(kT_c)\, \mathrm{e}^{-j2\pi \frac{h-k}{N} n}$$

[45] Si consideri che il valore $L$ presente al § 7.5.3 è pari alla radice di $L_n = 2^{M_n}$ della *n-esima* costellazione OFDM.

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[46]

$$\begin{aligned}
\mathcal{P}_{\underline{N}_n} &= \frac{1}{N^2}\sum_{h=0}^{N-1}\sum_{k=0}^{N-1}\mathcal{R}_{\underline{N}}\,((h-k)\,T_c)\,\mathrm{e}^{j\pi(h-k)}\mathrm{e}^{-j2\pi\frac{h-k}{N}n} = \\
&= \frac{1}{N}\sum_{m=-(N-1)}^{N-1}\frac{N-|m|}{N}\mathcal{R}_{\underline{N}}\,(mT_c)\,\mathrm{e}^{j2\pi\frac{mT_c}{2T_c}}\,\mathrm{e}^{-j2\pi\frac{m}{N}n} = \\
&= \frac{1}{N}\sum_{m=-(N-1)}^{N-1} z\,(m)\,\mathrm{e}^{-j2\pi\frac{m}{N}n} \qquad (13.23)
\end{aligned}$$

in cui l'ultima riga semplifica l'espressione introducendo la sequenza $\{z\,(m)\}$ di lunghezza $N$, che si ottiene campionando

$$z\,(t) = \left(1-\frac{|t|}{NT_c}\right)\mathcal{R}_{\underline{N}}\,(t)\,\mathrm{e}^{j2\pi\frac{t}{2T_c}} \qquad (13.24)$$

agli istanti $t = mT_c$ con $T_c = \frac{1}{N\Delta}$.

Mostriamo ora come, per $N$ sufficientemente elevato, la (13.23) possa essere calcolata in funzione dei campioni di $Z\,(f) = \mathcal{F}\,\{z\,(t)\}$, ed in particolare di come risulti

$$\mathcal{P}_{\underline{N}_n} \simeq \Delta\cdot\,Z\,(f)|_{f=n\Delta} \simeq 4\Delta\cdot\mathcal{P}_N\,(f_n)$$

Analizzando i termini che compaiono in (13.24), osserviamo che il prodotto $\mathcal{R}_{\underline{N}}\,(t)\,\mathrm{e}^{j2\pi\frac{t}{2T_c}}$ ha trasformata pari a $\mathcal{P}_{\underline{N}}\,(f)$, translata 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 $\left(1-\frac{|t|}{NT_c}\right) = tri_{2NT_c}\,(t) = tri_{\frac{2}{\Delta}}\,(t)$ possiede come noto trasformata $\mathcal{F}\left\{tri_{\frac{2}{\Delta}}\,(t)\right\} = \frac{1}{\Delta}sinc^2\left(\frac{f}{\Delta}\right)$; 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

$$Z\,(f) = \mathcal{P}_{\underline{N}}\left(f-\frac{N\Delta}{2}\right) * \frac{1}{\Delta}sinc^2\left(\frac{f}{\Delta}\right) \simeq \mathcal{P}_{\underline{N}}\left(f-\frac{N\Delta}{2}\right)$$

avendo approssimato $\frac{1}{\Delta}sinc^2\left(\frac{f}{\Delta}\right)$ come un impulso di area unitaria, per $N\Delta$ grande rispetto a $\Delta$.

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

[46] 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 $\mathcal{R}_{\underline{N}}\,(T_c)\,\mathrm{e}^{j\pi}\mathrm{e}^{-j2\pi\frac{1}{N}n}$ e $\mathcal{R}_{\underline{N}}\,(-T_c)\,\mathrm{e}^{-j\pi}\mathrm{e}^{j2\pi\frac{1}{N}n}$, $N-2$ 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.

$z(m) = z(mT_c)$ che compaiono nella (13.23); in particolare, per $N$ sufficientemente elevato, si ottiene[47] che

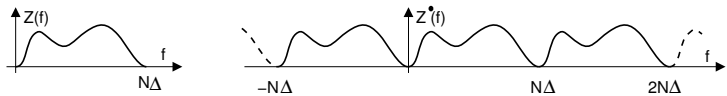
$$\begin{aligned}\mathcal{P}_{\underline{N}_n} &= \frac{1}{N}\sum_{m=-(N-1)}^{N-1} z(m)\,\mathrm{e}^{-j2\pi\frac{m}{N}n} \simeq \Delta\cdot Z(f)\big|_{f=n\Delta} = \\ &= \Delta\cdot\mathcal{P}_{\underline{N}}\left(n\Delta-\frac{N\Delta}{2}\right) = \Delta\cdot\mathcal{P}_{\underline{N}}\left(\Delta\left(n-\frac{N}{2}\right)\right) = \\ &= 4\Delta\cdot\mathcal{P}_N^{+}\left(f_0+\Delta\left(n-\frac{N}{2}\right)\right) = 4\Delta\cdot\mathcal{P}_N(f_n) = 2\Delta\cdot\mathcal{N}_0(f_n)\end{aligned}$$

in cui si è tenuto conto che $\mathcal{P}_{\underline{N}}(f) = 4\mathcal{P}_N^{+}(f+f_0)$ e si è indicata la densità di potenza in ingresso come $\mathcal{P}_N(f) = \frac{\mathcal{N}_0(f)}{2}$.

**Prestazioni per portante** Siamo finalmente in grado di scrivere

$$SNR_n = \frac{\mathcal{P}_{\underline{R}_n}}{\mathcal{P}_{\underline{N}_n}} = \frac{2\frac{T_0}{T}\alpha_n\mathcal{P}}{2\Delta\mathcal{N}_0(f_n)} = \frac{T_0}{T}\alpha_n\frac{T_0\mathcal{P}}{\mathcal{N}_0(f_n)} = \frac{T_0}{T}\alpha_n\frac{E_s}{\mathcal{N}_0(f_n)} = \frac{T_0}{T}\alpha_n M_n\frac{E_{b_n}}{\mathcal{N}_0(f_n)}$$

[47] 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 $H(f) = \frac{1}{N\Delta}rect_{N\Delta}\left(f-\frac{N\Delta}{2}\right)$ restituisce il segnale originario. Scriviamo pertanto

$$z(t) = z^{\bullet}(t) * h(t) = \sum_{m=-\infty}^{\infty} z(mT_c)\,\delta(t-mT_c) * sinc(N\Delta t)\,\mathrm{e}^{j\pi N\Delta t}$$

ed effettuiamone la trasformata:

$$\begin{aligned}Z(f) &= \mathcal{F}\left\{\sum_{m=-\infty}^{\infty} z(mT_c)\,\delta(t-mT_c)\right\}\cdot\frac{1}{N\Delta}rect_{N\Delta}\left(f-\frac{N\Delta}{2}\right) \\ &= \left[\sum_{m=-\infty}^{\infty} z(mT_c)\,\mathrm{e}^{-j2\pi\frac{m}{N\Delta}f}\right]\cdot\frac{1}{N\Delta}rect_{N\Delta}\left(f-\frac{N\Delta}{2}\right)\end{aligned}$$

che, calcolata alle frequenze $f = n\Delta$ con $n = 0, 1, ..., N-1$ fornisce

$$Z(f)\big|_{f=n\Delta} = \frac{1}{N\Delta}\sum_{m=-\infty}^{\infty} z(mT_c)\,\mathrm{e}^{-j2\pi\frac{m}{N}n}$$

Se ora non disponiamo di tutti i campioni $z(mT_c)$, ma solo degli $2N-1$ valori con $m = -(N-1), ..., 0, 1, ..., N-1$, la relazione precedente si applica ad un nuovo segnale $z'(t) = z(t)\cdot rect_{2NT_c}(t)$, 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

$$Z'(f) = Z(f) * \mathcal{F}\{rect_{2NT_c}(t)\} \simeq Z(f) * \delta(f) = Z(f)$$

in cui l'approssimazione è lecita per $N$ elevato.

avendo posto $E_s = T_0\mathcal{P}$ pari all'energia di un simbolo di durata $T_0 = \frac{1}{\Delta}$. L'energia per bit risulta dunque $E_b = \frac{E_s}{M}$, mentre per la portante *n-esima* si ha $E_{b_n} = \frac{E_s}{M_n}$. La $P_e$ per portante risulta quindi

$$P_{e/n} = \frac{2}{M_n}\left(1 - \frac{1}{\sqrt{L_n}}\right)\operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{T_0}{T}\frac{E_{b_n}}{\mathcal{N}_0(f_n)}\frac{\alpha_n M_n}{L_n - 1}}\right\} \tag{13.25}$$

**Caso di rumore bianco** Se $\mathcal{P}_N(f)$ non dipende da $f$, possiamo scrivere

$$\mathcal{P}_N^+(f) = \frac{\mathcal{N}_0}{2} rect_{N\Delta}(f - f_0)$$

e semplificare la (13.25), 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 (13.23): infatti, risulta

$$\mathcal{R}_{\underline{N}}(t) = \mathcal{F}^{-1}\{\mathcal{P}_{\underline{N}}(f)\} = \mathcal{F}^{-1}\left\{4\mathcal{P}_N^+(f + f_0)\right\} = 2\mathcal{N}_0 N\Delta sinc(N\Delta t)$$

e dunque $\mathcal{R}_{\underline{N}}(t) = 0$ con $t = mT_c = \frac{m}{N\Delta}$ per $m \neq 0$. Ciò permette di scrivere in definitiva

$$\mathcal{P}_{\underline{N}_n} = \frac{1}{N}\mathcal{R}_{\underline{N}}(0) = \frac{1}{N}2\mathcal{N}_0 N\Delta = 2\Delta \cdot \mathcal{N}_0$$

**Confronto con la portante singola** Proviamo a verificare se la modulazione OFDM è vantaggiosa in termini di prestazioni, per una medesima occupazione di banda ed a parità di potenza. Nel caso in cui il tempo di guardia $T_g = T - T_0$ sia nullo, in presenza di rumore bianco, e scegliendo un intervallo di simbolo $T_0 = \frac{1}{\Delta}$ da cui derivare $M^{OFDM} = T_0 \cdot f_b$, $M_n = \frac{M^{OFDM}}{\tilde{N}}$ e $\alpha_n = \frac{1}{\tilde{N}}$, si ottengono valori $\frac{E_{b_n}}{\mathcal{N}_0}$ uguali per le diverse portanti, a cui corrisponde il miglior valore di

$$P_e^{OFDM} = P_{e/n} = \frac{2\tilde{N}}{M^{OFDM}}\left(1 - \frac{1}{\sqrt{L_n}}\right)\operatorname{erfc}\left\{\sqrt{\frac{3}{2}\frac{E_b}{\mathcal{N}_0}\frac{1}{\tilde{N}}\frac{M^{OFDM}}{L_n - 1}}\right\}$$

ottenuta tenendo conto che $E_{b_n}M_n = E_s = E_b M^{OFDM} = E_b \log_2 L^{OFDM}$.

Nel caso in cui si adotti una modulazione a portante singola con impulso a coseno rialzato e roff-off $\gamma = \frac{N}{\tilde{N}} - 1$, si determina una occupazione di banda pari a $B = f_L(1 + \gamma)$ 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 in questo caso 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 $L = 2^{M^{QAM}} = 2^{\frac{M^{OFDM}}{\tilde{N}}} = L_n$.

*E allora dov'è la convenienza ? E' il tema delle prossime sottosezioni.*

#### 13.6.4.6 Equalizzazione

Consideriamo il caso in cui la trasmissione attraversi un canale descritto 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 (13.18) si altera, ed i valori $\underline{a}_n$ restituiti dalla (13.20) si modificano in $\underline{b}_n = \underline{a}_n \cdot \underline{H}\left(f - \Delta\left(n - \frac{N}{2}\right)\right)$. Come anticipato, l'equalizzazione è pertanto ridotta ad eseguire un semplice prodotto scalare tra il vettore dei valori $\{\underline{b}_n\}$ e quello dei valori $\left\{\frac{1}{\underline{H}\left(f-\Delta\left(n-\frac{N}{2}\right)\right)}\right\}$, ovviamente purché si si conosca $\underline{H}(f)$, od una sua stima.

**Codifica differenziale** Nel caso in cui la distorsione non sia eccessiva, si può evitare del tutto lo stadio di equalizzazione, e ricorrere ad una codifica differenziale (§ 13.6.1). In presenza di distorsione di fase infatti, il piano dell'inviluppo complesso subisce, per ogni portante consecutiva, una rotazione pari alla differenza della fase di $\underline{H}(f)$ in corrispondenza delle due frequenze contigue. Se questa rotazione non è eccessiva, si può prendere come riferimento di fase il risultato della demodulazione della portante precedente.

#### 13.6.4.7 Sensibilità alla temporizzazione

Nel caso in cui il ricevitore non acquisisca una perfetta sincronizzazione di simbolo, il calcolo della FFT su di un gruppo di campioni presi a partire dalla coda del preambolo non altera per nulla il risultato[48], in virtù della caratteristica di periodicità dello stesso.

#### 13.6.4.8 Ottimalità

Come stiamo per mostrare, questa proprietà è intimamente legata alla possibilità dell'OFDM di assegnare valori di potenza differenti alle diverse portanti.

La trasmissione numerica con una $f_b$ elevata, eseguita utilizzando una sola portante, deve necessariamente occupare una banda molto ampia; nel caso in cui $H(f)$ presenti una elevata distorsione di ampiezza, l'equalizzazione della stessa causa una colorazione del rumore in ingresso al demodulatore, ed un peggioramento delle prestazioni. Un problema analogo nasce nel caso in cui il rumore non sia bianco, ad esempio perchè derivante da un segnale interferente.

In entrambi i casi, la teoria di Shannon che definisce[49] una *capacità di canale* pari a $C = W \log_2\left(1 + \frac{\mathcal{P}_r}{W N_0}\right)$ in presenza di un rumore bianco, con $\mathcal{P}_N(f) = \frac{N_0}{2}$ e con una potenza ricevuta $\mathcal{P}_r$ in una banda (positiva) $W$, si modifica nel seguente modo, per tenere conto dell'andamento incostante di $\mathcal{P}_r(f)$ e $\mathcal{P}_N(f)$. Considerando il canale scomposto in infinite sottobande entro le quali le densità di potenza possono ritenersi costanti, l'espressione della capacità diviene ora:

$$C = \sup_{\mathcal{P}_R(f)} \int_{f \in I_f} \log_2\left(1 + \frac{\mathcal{P}_r(f)}{\mathcal{P}_N(f)}\right) df \tag{13.26}$$

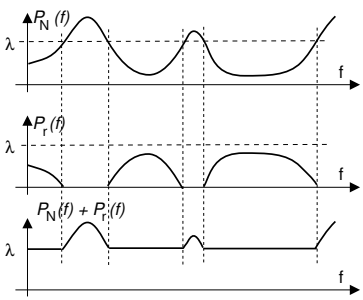
[48] Infatti *non siamo* nelle condizioni di demodulazione coerente dell'FSK, e le portanti del simbolo OFDM ricevuto mantengono ortogonalità purchè finestrate su di un periodo $T_0 = \frac{1}{\Delta}$.

[49] Come discusso ai § 17.2.3 e 17.2.4, il risultato della teoria di Shannon asserisce che è possibile conseguire una velocità di trasmissione $f_b = C$ con probabilità di errore nulla, ma non indica come fare. Una soluzione che vi si avvicina è quella di adottare una codifica di canale a ridondanza elevata, capace di correggere un elevato numero di errori.

in cui $\mathcal{P}_r(f)$ viene fatto variare in tutti i modi possibili e tali che $\int_{f\in I_f} \mathcal{P}_r(f)\,df = \mathcal{P}_r$, con $\mathcal{P}_r(f) \geq 0$, ed $I_f$ rappresenta l'insieme delle frequenze in cui è presente il segnale: $I_f = \{f : \mathcal{P}_r(f) > 0\}$. La (13.26) asserisce quindi che, nel caso in cui $\mathcal{P}_N(f)$ in ingresso al canale non sia bianco, le migliori prestazioni ($f_b$ che tende a $C$) si ottengono solamente sagomando la densità di potenza del segnale ricevuto in modo opportuno.

Per determinare l'andamento ottimo di $\mathcal{P}_r(f)$ si ricorre allora al *calcolo delle variazioni* basato sui *moltiplicatori di Lagrange*[50], che in questa sede non affrontiamo, e che fornisce la condizione

$$\mathcal{P}_r(f) + \mathcal{P}_N(f) = \begin{cases} \lambda & \text{se} \quad \mathcal{P}_N(f) < \lambda \\ \mathcal{P}_N(f) & \text{se} \quad \mathcal{P}_N(f) \geq \lambda \end{cases} \tag{13.27}$$




detta anche del *riempimento d'acqua* perché asserisce che (vedi figura) il segnale debba essere presente in misura maggiore nelle regioni di frequenza dove il rumore è sufficientemente ridotto. La costante $\lambda$ è scelta in modo tale da ottenere $\int \mathcal{P}_r(f)\,df = \mathcal{P}_r$.

In un sistema di modulazione numerica a singola portante, $\mathcal{P}_r(f)$ non può essere modificato a piacere, in quanto il suo andamento deve essere quello legato alla particolare caratteristica di Nyquist $G(f)$ scelta per ottenere una ricezione priva di ISI. Nel caso dell'OFDM invece, la potenza assegnata a ciascuna portante può essere variata liberamente, e se la $\mathcal{P}_r(f)$ che realizza le condizioni (13.27) può essere comunicata al modulatore, è possibile avvicinarsi alla velocità massima permessa dalla (13.26).

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 (13.27), e quindi trasmettere più bit $M_n$ sulle portanti $n$ per le quali $\mathcal{P}_n$ è maggiore.

#### 13.6.4.9 Codifica

Abbiamo appena mostrato come, conoscendo la $H(f)$ e la $\mathcal{P}_N(f)$ del canale, sia possibile equalizzare $\mathcal{P}_x(f) = \frac{\mathcal{P}_r(f)}{|H(f)|^2}$ e al contempo soddisfare (13.27) e rendere massima la $f_b$. Nel caso di collegamenti tempo-varianti però, la $H(f)$ non è nota, ed anche se lo fosse non esiste garanzia che rimanga costante. In tal caso allora non ha senso determinare una distribuzione ottima della potenza e dei bit sulle portanti, mentre invece occorre aggiungere della ridondanza al segnale trasmesso mediante un codice di canale, allo scopo di correggere i bit errati.

Osserviamo ora che, nel caso di una modulazione a portante singola, in presenza di una $H(f)$ tempo-variante, il processo di equalizzazione è particolarmente complesso in quanto deve *inseguire* le variazioni di $H(f)$. Se l'equalizzazione non è perfetta, insorge ISI e la trasmissione diviene rapidamente cosí piena di errori da renderne impossibile la correzione anche adottando codici di canale.

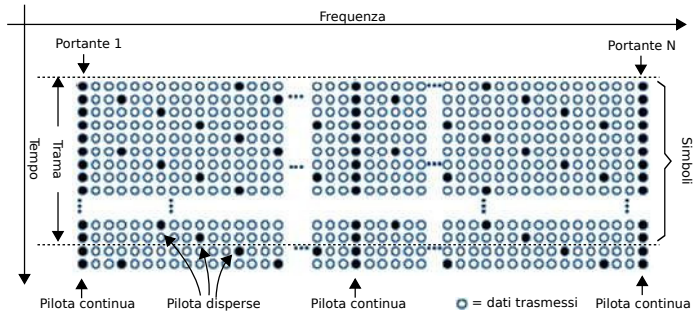
[50] Vedi ad es. http://it.wikipedia.org/wiki/Metodo_dei_moltiplicatori_di_Lagrange

Nel caso dell'OFDM, al contrario, l'andamento di $H(f)$ determina un peggioramento di prestazioni solamente per quelle portanti per le quali $H(f)$ si è ridotto[51]. Pertanto, l'applicazione di un codice di canale al blocco di $M$ bit che costituisce un simbolo, seguito da una operazione di scrambling, consente al lato ricevente di recuperare l'informazione trasmessa anche nel caso in cui per alcune portanti si determini un elevato tasso di errore.

La trasmissione OFDM in cui è presente una codifica di canale prende il nome di trasmissione COFDM (*Coded* OFDM).

#### 13.6.4.10 Portanti pilota

Fin qui abbiamo assunto che il ricevitore OFDM mostrato in fig. 13.12 operi in condizioni di sincronismo sia per quanto riguarda la portate di demodulazione, che per quanto gli intervalli di simbolo. A questo scopo alcune delle sottoportanti - dette *pilota* - non sono usate per trasmettere dati, ma sono mantenute costantemente attive, con potenza di poco superiore, allo scopo di facilitare la sincronizzazione in frequenza. In figura è rappresentato il caso per il DVB-T, in cui ogni riga rappresenta le portanti di un simbolo, e quelle pilota si trovano in posizione fissa; sono inoltre mostrate delle *portanti disperse* (SCATTERED) le cui posizioni evolvono ciclicamente di simbolo in simbolo, e consentono di acquisre un sincronismo sia di simbolo che di trama.




### 13.6.5 Sistemi a spettro espanso

Si basano su di una *manipolazione* del messaggio da trasmettere in modo che questo occupi una banda molto maggiore di quella originaria, e sulla manipolazione inversa in ricezione: tale caratteristica è quindi indicata con il termine di *Spread Spectrum*. Sebbene questa doppia operazione non produca nessun vantaggio effettivo nei riguardi delle prestazioni ottenibili nel caso in cui la ricezione sia disturbata dalla sola presenza di rumore additivo gaussiano, si ottengono invece i seguenti altri benefici

- la presenza di eventuali altre trasmissioni e/o disturbi nella stessa banda del segnale espanso causa una potenza interferente ridotta
- la densità spettrale del segnale trasmesso si confonde con quella del rumore, rendendo la trasmissione stessa difficilmente rilevabile da parte di soggetti ostili
- l'ignoranza della effettiva manipolazione adottata rende la trasmissione indecifrabile da parte di soggetti ostili

[51] Si consideri ad esempio il caso in cui $H(f)$ ha origine da un fenomeno di cammini multipli, che determina un andamento di $H(f)$ oscillante in frequenza.

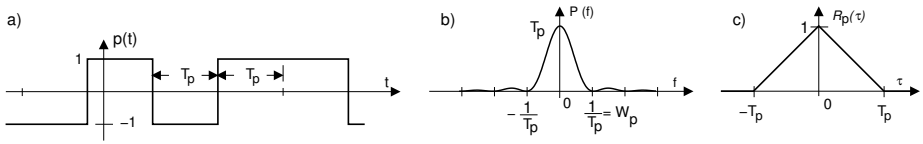



Figura 13.13: a) - sequenza pseudonoise; b) densità di potenza; c) autocorrelazione

#### 13.6.5.1 Sequenze pseudo-casuali

La manipolazione che produce l'espansione spettrale si basa a sua volta sull'utilizzo di una sequenza cosiddetta *pseudo-noise* (PN), le cui caratteristiche statistiche si avvicinano a quelle di un rumore stazionario bianco (e dunque a valori incorrelati), tranne per il fatto che tali valori non sono casuali ma deterministici, in modo che la loro ripetizione ciclica rende la sequenza PN riproducibile dal lato ricevente. La fig. 13.13a mostra una parte di un possibile segnale dati $p(t)$ pseudo casuale, bipolare, di durata $L \cdot T_p$

$$p(t) = \sum_{k=0}^{L-1} a_k g(t - kT_p - \theta) \tag{13.28}$$

realizzato mediante impulsi rettangolari $g(t) = rect_{T_p}(t)$ di durata $T_p$ chiamati *chip*, con polarità stabilita dagli $L$ valori $a_k = \pm 1$ scelti in modo da produrre media nulla e varianza unitaria ($\overline{a_k} = 0$, $\overline{a_k^2} = 1$), mentre la correlazione $\mathcal{R}_a(n)$ tende a zero con $n \neq 0$ ([52]), mimando così la proprietà di indipendenza statistica. Al § 9.9.3 abbiamo mostrato che un segnale del genere presenta uno spettro di densità di potenza

$$\mathcal{P}_p(f) = \sigma \frac{|\mathcal{E}_G(f)|^2}{T_p} = T_p \mathrm{sinc}^2(fT_p) \tag{13.29}$$

mostrato in fig. 13.13b, e per il quale la frequenza $W_p = \frac{1}{T_p}$ ne approssima l'occupazione di banda. Dalla (13.29) consegue che l'autocorrelazione di $p(t)$ si esprime come[53]

$$\mathcal{R}_p(\tau) = \mathcal{F}^{-1}\{\mathcal{P}_p(f)\} = \mathrm{tri}_{2T_p}(\tau)$$

mostrata in fig. 13.13c, e che appunto si azzera per $\tau > T_p$. Le sequenze pseudo-noise utilizzate realmente non soddisfano in pieno tutti questi requisiti, ma vi si avvicinano molto; un approfondimento sulle loro modalità di generazione può essere iniziato visitando ad es. Wikipedia[54].

[52] Data la sequenza deterministica $a_k = \{a_0, a_1, \cdots a_{L-1}\}$ di lunghezza $L-1$, la correlazione tra coppie di elementi a distanza $n$ è definita come

$$\mathcal{R}_a(n) = \frac{1}{L-n} \sum_{k=0}^{L-n-1} a_k a_{k+n}$$


Considerando invece la sequenza periodica ottenuta ripetendo gli $a_k$, possiamo definire la stessa grandezza come $\mathcal{R}_a(n) = \frac{1}{L}\sum_{k=0}^{L-1} a_k a_{(k+n) \bmod L}$

[53] Si veda l'esempio a pag. 43

[54] http://en.wikipedia.org/wiki/Pseudorandom_binary_sequence

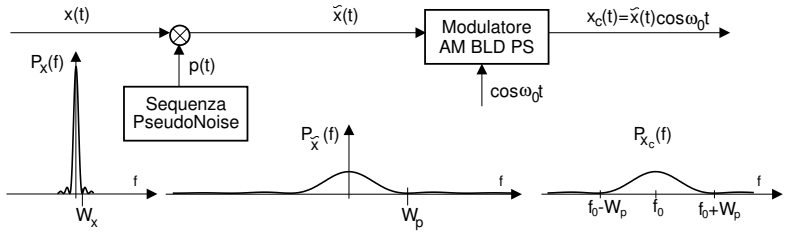



Figura 13.14: Generazione di un segnale modulato DSSS

### 13.6.5.2 Sequenza diretta

La moltiplicazione del segnale PN (13.28) per un segnale di banda base $x(t)$ realizza la tecnica di espansione spettrale nota come *Direct Sequence Spread Spectrum* (o DSSS) il cui effetto, sebbene valido per $x(t)$ qualsiasi, viene ora discusso con riferimento ad un segnale numerico[55] binario NRZ, con frequenza binaria che determina una densità di potenza $\mathcal{P}_x(f)$ del tipo di (13.29) ma con banda $W_x \ll W_p$; il risultato del prodotto $x(t)\,p(t)$ quindi modula AM-BLD-PS una portante a frequenza $f_0$. La fig. 13.14 indica con $\tilde{x}(t) = x(t)\,p(t)$ il segnale *allargato*, la cui potenza è la stessa [56] $\mathcal{P}_x$ di $x(t)$, che ora risulta però *spalmata* sulla banda $W_p$ di $p(t)$. Infatti, la densità di potenza $\mathcal{P}_{\tilde{x}}(f)$ è il risultato della convoluzione in frequenza[57]

$$\mathcal{P}_{\tilde{x}}(f) = \mathcal{P}_x(f) * \mathcal{P}_p(f) \simeq \int_{-W_x}^{W_x} \mathcal{P}_x(\lambda)\,\mathcal{P}_p(f-\lambda)\,d\lambda$$

in cui la definizione degli estremi di integrazione tiene conto del fatto che $\mathcal{P}_x(f) \approx 0$ per $|f| > W_x$. Considerando ora che $W_p \gg W_x$, notiamo che per $|\lambda| \leq W_x$ si ha $\mathcal{P}_p(f-\lambda) \simeq \mathcal{P}_p(f)$, e quindi

$$\mathcal{P}_{\tilde{x}}(f) \approx \left[\int_{-W_x}^{W_x} \mathcal{P}_x(\lambda)\,d\lambda\right] \mathcal{P}_p(f) = \mathcal{P}_x \mathcal{P}_p(f)$$

**Guadagno di processo** Il rapporto di espansione spettrale $\frac{W_p}{W_x}$ tra la banda del segnale allargato e quella del segnale di partenza varia tipicamente tra 10 e 10000 volte, ossia tra 10 e 40 dB, e viene indicato anche come *guadagno di processo* (o *processing gain*), in quanto come vedremo rappresenta una misura del miglioramento dell'SNR nel caso di presenza di segnali interferenti.

**Despreading** Per proseguire nell'analisi, consideriamo lo schema di ricevitore mostrato in fig. 13.15, nella cui parte sinistra è mostrato il segnale modulato ricevuto

[55] Quando il messaggio $x(t)$ è di natura numerica, esso è rappresentato da un flusso binario a velocità $f_b = 1/T_b$, e l'operazione di *spreading* si ottiene scegliendo $T_p = T_b/L \ll T_b$. Moltiplicando modulo due ogni bit del messaggio per la sequenza di chip della PN, si ottiene di fatto una *sequenza di sequenze* PN, ognuna con segno invertito o meno a seconda del valore dei singoli bit del messaggio. Il segnale dati viene poi realizzato a partire dalla sequenza risultante a velocità $f_b \cdot W_p$.

[56] Considerando $x(t)$ realizzazione di un processo ergodico indipendente da $p(t)$, la potenza di $\tilde{x}(t)$ risulta $\overline{\tilde{x}^2} = E\left\{x^2(t)\,p^2(t)\right\} = \overline{x^2} = \mathcal{P}_x$

[57] l'autocorrelazione del prodotto di processi indipendenti è pari al prodotto delle autocorrelazioni, e quindi si applica teorema di Parseval

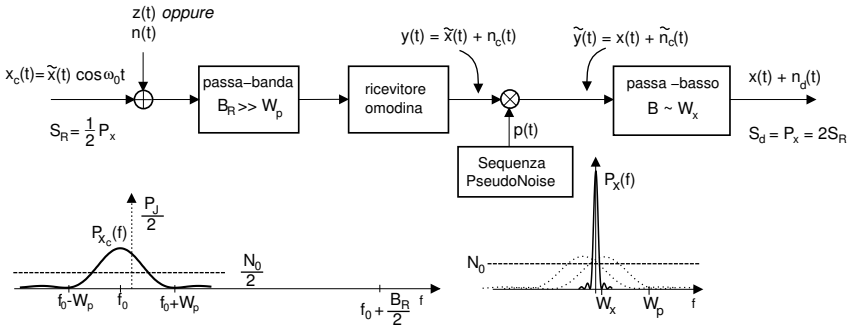



Figura 13.15: Ricevitore DSSS con rumore additivo $n(t)$ o interferenza $z(t)$

$x_c(t) = \tilde{x}(t)\cos\omega_0 t$ con potenza[58] $S_R = \frac{1}{2}\mathcal{P}_x$, a cui si sovrappone un disturbo $n(t)$, ed insieme attraversano il filtro passabanda di ricezione caratterizzato da una banda di rumore $B_R \gg W_p \gg W_x$, dato che deve lasciar passare l'intero spettro *allargato*, compresi i suoi lobi laterali. Il fatto che la potenza $S_R$ sia spalmata su tutta la banda $B_R$ rende il segnale utile poco distinguibile dal livello di rumore, e dunque difficilmente intercettabile.

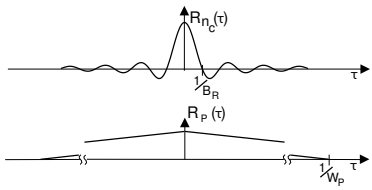
Dopo demodulazione omodina si ottiene il nuovo segnale di banda base $y(t) = \tilde{x}(t) + n_c(t)$ in cui $n_c(t)$ è la componente in fase del rumore. A questo punto avviene l'operazione di *despreading* che si avvale della possibilità per il ricevitore di generare la stessa sequenza PN usata in trasmissione, in forma temporalmente sincrona, in modo da poter scrivere

$$\tilde{y}(t) = [\tilde{x}(t) + n_c(t)]\,p(t) = x(t)\,p^2(t) + n_c(t)\,p(t) = x(t) + \tilde{n}_c(t)$$

Pertanto, mentre il messaggio $x(t)$ è di nuovo quello precedente all'allargamento, il rumore ha subito a sua volta un effetto di allargamento. Un successivo filtraggio passa-basso con banda $W_x$ pari a quella di segnale produce infine il risultato $y_d(t) = x(t) + n_d(t)$, in cui il segnale utile ha potenza $S_d = \mathcal{P}_x = 2S_R$, mentre al termine di disturbo additivo $n_d(t)$ è stata rimossa la potenza che cade al difuori della banda di segnale.

**Prestazioni in presenza di rumore** Nel caso in cui $n(t)$ sia rumore bianco con densità di potenza $\mathcal{P}_n(f) = {}^{N_0}/_2$, la sua componente in fase dopo demodulazione omodina ha densità[59] $\mathcal{P}_{n_c}(f) = N_0 rect_{B_R}(f)$ 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 dopo despreading, con l'aiuto della figura a lato osserviamo che l'autocorrelazione di $\tilde{n}_c(t)$ è pari a $\mathcal{R}_{\tilde{n}_c}(\tau) =$

[58] $\tilde{x}(t)\cos(\omega_0 t + \varphi)$ con $\varphi$ 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.

[59] vedi § 12.1.2

$\mathcal{R}_{n_c}(\tau)\,\mathcal{R}_p(\tau)$, e che $\mathcal{R}_{n_c}(\tau) \simeq 0$ con $|\tau| \gg \frac{1}{B_R} \ll \frac{1}{W_p}$, mentre $\mathcal{R}_p(\tau) \simeq 1$ con $|\tau| \ll T_p = \frac{1}{W_p}$: pertanto possiamo scrivere $\mathcal{R}_{\tilde{n}_c}(\tau) \simeq \mathcal{R}_{n_c}(\tau)$ e quindi

$$\mathcal{P}_{\tilde{n}_c}(f) \simeq \mathcal{P}_{n_c}(f) = N_0 rect_{B_R}(f)$$

La componente di rumore $n_d(t)$ in uscita dall'ultimo passa basso ha quindi una potenza $N_d = 2N_0W_x$, permettendo di valutare il rapporto segnale-rumore dopo demodulazione come

$$\left(\frac{S}{N}\right)_d = \frac{2S_R}{2N_0W_x} = \frac{S_R}{N_0W_x}$$

ossia proprio pari all'$SNR$ *di riferimento* (pag. 288), mostrando come la concatenazione delle operazioni di spreading e despreading *non alteri* le prestazioni del processo di modulazione nei confronti del rumore bianco.

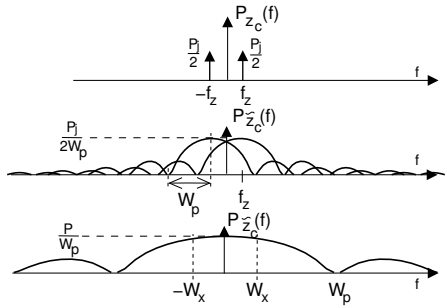
**Prestazioni in presenza di un tono interferente** Mostriamo che se il termine di disturbo additivo $z(t)$ occupa una banda relativamente stretta in rapporto a $B_R$, allora la sua potenza dopo demodulazione risulterà *ridotta* di un fattore pari al guadagno di processo $W_p/W_x$. Ad esempio, consideriamo il caso di un tono interferente (*jammer*), in cui

$$z(t) = \sqrt{2\mathcal{P}_j}\cos(\omega_0 + \omega_z)\,t$$

con potenza $\overline{z^2} = \mathcal{P}_j$ alla frequenza $f_0 + f_z$. In tal caso $z_c(t) = \sqrt{2\mathcal{P}_j}\cos\omega_z t$ e

$$\mathcal{P}_{z_c}(f) = \frac{\mathcal{P}_j}{2}\left[\delta(f - f_z) + \delta(f + f_z)\right] \tag{13.30}$$

che moltiplicato per $p(t)$ produce $\mathcal{P}_{\tilde{z}_c}(f) = \mathcal{P}_{z_c}(f) * \mathcal{P}_p(f)$, mostrato alla riga centrale della figura a seguente[60]. 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 è




$$\overline{z_d^2} = \int_{-W_x}^{W_x} \mathcal{P}_{\tilde{z}_c}(f)\,df \leq 2W_x\frac{\mathcal{P}_j}{W_p}$$

e dunque il rapporto segnale-interferente diviene

$$\left(\frac{S}{z_d^2}\right)_d \geq 2S_R\frac{W_p}{2W_x\mathcal{P}_j} = \frac{S_R}{\mathcal{P}_j}\frac{W_p}{W_x}$$

mostrando quindi un miglioramento pari proprio al guadagno di processo.

[60] Il risultato si ottiene tenendo conto delle eq. (13.29) e (13.30)

**Accesso multiplo** Una frequente applicazione delle tecniche spread spectrum consiste nel permettere la comunicazione *contemporanea* di una pluralità di soggetti, possibile qualora ognuno di essi adotti una diversa sequenza PN: la tecnica prende allora il nome di CDMA (*Code Division Multiple Access*). Mostriamo che in tal caso per ogni comuncazione l'effetto delle altre si riduce ad un modesto innalzamento del rumore di fondo, tanto più piccolo quanto minore è il valore della intercorrelazione tra i codici PN utilizzati.

Dopo la demodulazione, il termine interferente $z(t)$ causato da $N$ diversi utenti, ognuno con un diverso codice $p_n(t)$ e segnale dati $x_n(t)$, può essere scritto come

$$z(t) = \sum_{n=1}^{N} A_n x_n(t - \tau_n) p_n(t - \tau_n) \cos\theta_n$$

in cui $A_n$, $\tau_n$ e $\cos\theta_n$ sono ripettivamente ampiezza, ritardo di simbolo e fase della portante relativi all'$n$-esimo utente. Assumendo ora eguali le ampiezze del segnale utile $x(t)$ e degli interferenti, dopo il despreading otteniamo

$$\tilde{y}(t) = x(t) + \left[\sum_{n=1}^{N} x_n(t - \tau_n) p_n(t - \tau_n) \cos\theta_n\right] p(t)$$

Se rappresentiamo ora il filtro passa basso di fig. 13.15 come un integratore esteso ad un periodo di bit, ovvero un filtro adattato al segnale NRZ, il valore della sua uscita campionata al termine della durata del $k$-esimo periodo di bit risulta

$$\begin{aligned}\tilde{y}(t_k) &= x(t_k) + \sum_{n=1}^{N}\left[\cos\theta_n \int_{kT_b}^{(k+1)T_b} x_n(t - \tau_n) p_n(t - \tau_n) p(t)\, dt\right] \\ &= x(t_k) + z(t_k)\end{aligned}$$

in cui $z(t_k)$ rappresenta il termine di interferenza complessiva da parte di tutti gli altri $N$ utenti; dato che i valori di $x_n$ possono essere $\pm 1$, l'integrale calcola in effetti l'intercorrelazione $\mathcal{R}_{pp_n}(\tau_n)$ (§ 9.1.4) tra $p(t)$ e le altre $p_n(t)$, calcolata per un ritardo $\tau_n$. Pertanto, scegliendo la famiglia di sequenze pseudo-noise in modo che esibiscano una intercorrelazione molto ridotta (in teoria nulla, se le PN fossero *ortogonali*), è possibile rendere trascurabile l'effetto degli interferenti.

# Capitolo 14

# Transito dei segnali nei circuiti

Trattiamo qui della descrizione dei circuiti elettrici in termini di sistemi fisici, e delle modifiche che questi generano nei segnali in transito.

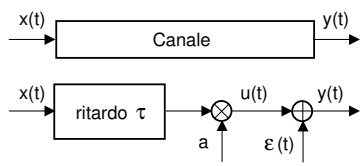
**Elaborazione e Distorsione** Quando un segnale viene "manipolato" di proposito, si dice che questo è *elaborato*. Se viceversa il segnale si altera per una causa indipendente dalla volontà, allora il segnale subisce una *distorsione*[1].

**Canale perfetto** Per valutare l'entità della distorsione, stabiliamo un criterio con cui distinguere la componente di segnale *utile* dal *disturbo*.

Come è noto, la ricezione di un segnale identico a quello trasmesso, tranne che per un fattore di scala ed un ritardo temporale, non altera la sostanza del messaggio: pertanto, un canale che presenti una risposta impulsiva

$$h(t) = a\delta(t-\tau)$$

viene indicato come *canale perfetto*, ed il segnale ricevuto $y(t) = u(t) = ax(t-\tau)$ *è tutto utile*. Se invece non vale quest'uguaglianza, viene definito *disturbo additivo* la differenza $\varepsilon(t) = y(t) - ax(t-\tau) = y(t) - u(t)$ ([2]). Infine, il rap-




[1] L'elaborazione di un segnale è indicata anche come suo "processamento" (dall'inglese PROCESSED=trattato). In altri contesti non "comunicazionistici" la terminologia può essere ancora più varia, come ad esempio.... le alterazioni prodotte sul suono di uno strumento musicale sono indicate come *effetti* ed il segnale risultante è "effettato" (!).

[2] Nella pratica, i valori $a$ e $\tau$ non si conoscono, mentre invece possiamo disporre di coppie di segnali $(x(t), y(t))$. I valori vengono dunque definiti come quelli che *rendono SNR massimo* ovvero $\mathcal{P}_\varepsilon$ minimo. Considerando segnali di potenza, ossia processi stazionari ergodici, si ha

$$\begin{aligned}\mathcal{P}_\varepsilon(a,\tau) &= E\left\{(y(t) - ax(t-\tau))^2\right\} = E\left\{y^2(t)\right\} + a^2 E\left\{x^2(t)\right\} - 2aE\{y(t)\,x(t-\tau)\} = \\ &= \mathcal{P}_y + a^2\mathcal{P}_x - 2a\mathcal{R}_{xy}(\tau)\end{aligned}$$

in cui si è operata la sostituzione $E\{y(t)\,x(t-\tau)\} = \mathcal{R}_{yx}(-\tau) = \mathcal{R}^*_{xy}(\tau) = \mathcal{R}_{xy}(\tau)$.

Il valore di $a$ che rende minimo $\mathcal{P}_\varepsilon(a,\tau)$ si ottiene eguagliando a zero la derivata: $\frac{\partial}{\partial a}\mathcal{P}_\varepsilon(a,\tau) = 2a\mathcal{P}_x - 2\mathcal{R}_{xy}(\tau) = 0 \Rightarrow a_{opt} = \frac{\mathcal{R}_{xy}(\tau)}{\mathcal{P}_x}$, che sostituita nell'espressione di $\mathcal{P}_\varepsilon$ fornisce

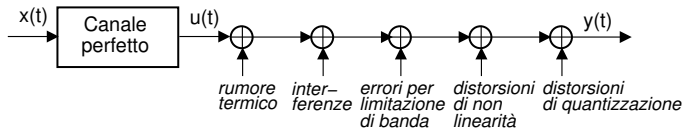
$$\mathcal{P}_\varepsilon(\tau) = \mathcal{P}_y + \left(\frac{\mathcal{R}_{xy}(\tau)}{\mathcal{P}_x}\right)^2\mathcal{P}_x - 2\frac{\mathcal{R}_{xy}(\tau)}{\mathcal{P}_x}\mathcal{R}_{xy}(\tau) = \mathcal{P}_y - \frac{(\mathcal{R}_{xy}(\tau))^2}{\mathcal{P}_x} = \mathcal{P}_y\left(1 - \frac{(\mathcal{R}_{xy}(\tau))^2}{\mathcal{P}_x\mathcal{P}_y}\right)$$

porto $\frac{\mathcal{P}_u}{\mathcal{P}_\varepsilon}$ tra la potenza del segnale utile e quella del disturbo, prende il nome di *rapporto segnale rumore* ($SNR$).

**Potenza di segnale e grandezze elettriche** La caratterizzazione energetica dei segnali è stata finora svolta *a prescindere* dalla natura fisica degli stessi: ovvero, non si è mai specificato se si trattasse di tensioni o correnti, né si sono indicate le impedenze in gioco. Trattando ora di grandezze elettriche, le potenze di segnale, di tensione o di corrente, saranno misurate in $(Volt)^2$ o in $(Ampere)^2$ rispettivamente.

**Esempio** Sia $x(t)$ un segnale di tensione. La sua potenza $\mathcal{P}_x$ ha unità di misura $\left[V^2\right]$, mentre la sua densità di potenza $\mathcal{P}_x(f)$ si esprime in $\left[\frac{V^2}{Hz}\right]$.

**Fonti di disturbo** Per quanto riguarda $\mathcal{P}_u$, questa viene determinata a partire dalla conoscenza di $\mathcal{P}_x$ applicando i risultati relativi ai trasferimenti energetici mostrati in questo capitolo alle condizioni reali determinate dai mezzi trasmissivi descritti al capitolo 15. D'altra parte, il termine di errore $\varepsilon(t)$ può essere dovuto al rumore additivo di natura termica (Cap. 16), a quello di natura interferente (ad es., vedi pag. 360), o di quantizzazione (§ 7.4), così come può insorgere a causa delle distorsioni lineari (§ 14.5) e/o di non linearità (§ 14.6) introdotte da un canale non perfetto.

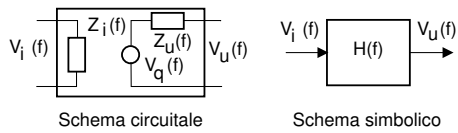



Spesso però queste diverse fonti di disturbo sono analizzate in forma separata e indipendente, ottenendo in definitiva un modello di ricezione in cui per ognuna di esse si è in grado di ottenere un valore di $SNR$ dovuto a quella sola fonte. In tal caso, l'$SNR$ complessivo è ottenuto applicando il risultato mostrato al § 14.7.1.

## 14.1 Caratterizzazione dei circuiti

**Numero di porte** Le coppie di morsetti a cui applicare o da cui prelevare un segnale sono anche denominate *porte*. In questo senso i *generatori* ed le *impedenze* costituiscono reti ad *una* porta, mentre un sistema fisico dotato di relazione ingresso-uscita è una rete *due porte*.

**Modello di rappresentazione** Un circuito può essere rappresentato mediante il suo *modello circuitale*, in cui sono evidenziati generatori, resistenze, impedenze, generatori controllati..., oppure il suo *schema simbolico*, in cui sono solo mostrate le relazioni funzionali tra i segnali in transito.




Il valore di $\mathcal{P}_\varepsilon$ evidentemente è minimo per quel valore di $\tau = \tau_{opt}$ che rende massima $(\mathcal{R}_{xy}(\tau))^2$, ovvero per quella traslazione temporale che rende "più simili" i segnali di ingresso ed uscita.

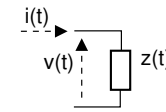
**Proprietà delle reti due porte** Le proprietà di *linearità, permanenza, realizzabilità ideale e fisica, stabilità*, già definite al § 9.5 per i sistemi fisici, possono essere verificate o meno nelle reti due porte.

## 14.2 Bipoli

**Passivi** Non contengono generatori, e sono caratterizzati dalle relazioni esistenti tra la tensione ai loro capi e la corrente che vi scorre (entrante). Il legame tra le due grandezze è una *convoluzione*

$$v(t) = i(t) * z(t)$$

in cui si suppone $i(t)$ la causa e $v(t)$ l'effetto. La trasformata di Fourier fornisce $V(f) = I(f) \cdot Z(f)$ in cui $Z(f)$ prende il nome di *impedenza*, e può scriversi nei termini di parte reale ed immaginaria:



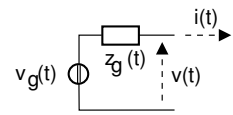

$$Z(f) = R(f) + jX(f)$$

in cui $R(f)$ (*resistenza*) è una funzione *pari* di $f$ (e sempre positiva), mentre $X(f)$ (*reattanza*) è dispari: pertanto, $Z(f) = Z^*(-f)$ e quindi $z(t)$ è reale. Allo stesso tempo, è definita *l'ammettenza*

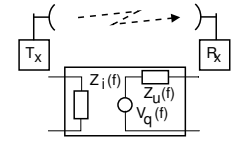
$$Y(f) = \frac{1}{Z(f)} = \frac{R(f) - jX(f)}{|Z(f)|^2}$$

e la corrispondente $y(t) = \mathcal{F}^{-1}\{Y(f)\}$, che permette di scrivere $i(t) = v(t) * y(t)$.

**Attivi** Sono bipoli al cui interno è presente un generatore. Per il teorema di THEVENIN,[3] qualunque circuito può essere ridotto ad un generatore di tensione con in serie una impedenza (vedi figura), in cui $V_g(f)$ rappresenta la tensione a vuoto, ossia quando $I(f) = 0$ (considerata uscente nei bipoli attivi).

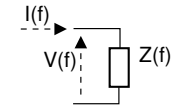



**Esempio** Una antenna trasmittente è schematizzabile come un bipolo passivo, di impedenza pari all'impedenza di ingresso dell'antenna, che assorbe la potenza erogata dal trasmettitore. Una antenna ricevente è schematizzabile come un generatore di tensione con in serie la propria impedenza di uscita, e trasferisce allo stadio di ingresso del ricevitore la potenza ricevuta per via elettromagnetica.




### 14.2.1 Potenza assorbita da un bipolo

Se ad un bipolo passivo di impedenza $Z(f)$ è applicato un segnale di tensione con spettro di densità di potenza $\mathcal{P}_v(f)$, la potenza *dissipata* sul bipolo, indicata come $\mathcal{W}_z(f)$ per distinguerla da quella di segnale, ha densità




$$\mathcal{W}_z(f) = \mathcal{P}_v(f) \cdot \Re\{Y(f)\} = \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} \qquad \left[\frac{V^2}{\Omega \cdot Hz}\right] = \left[\frac{Watt}{Hz}\right] \tag{14.1}$$

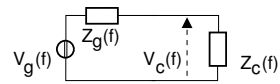
[3] Vedi ad es. `http://it.wikipedia.org/wiki/Teorema_di_Thévenin`

La dimostrazione della relazione illustrata è fornita in appendice 14.7.2. La dipendenza di $Y(f)$ dalla frequenza svolge pertanto *una azione filtrante*, e la potenza totale assorbita (o dissipata) su $Z(f)$ vale

$$\mathcal{W}_z = \int_{-\infty}^{\infty} \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} df \qquad [Watt]$$

### 14.2.2 Connessione tra generatore e carico

La tensione ai capi del carico è valutabile applicando la *regola del partitore*:




$$V_c(f) = V_g(f) \frac{Z_c(f)}{Z_c(f) + Z_g(f)}$$

ossia $V_c(f) = V_g(f) H(f)$ con $H(f) = \frac{Z_c(f)}{Z_c(f)+Z_g(f)}$. La potenza di segnale ai capi del carico vale $\mathcal{P}_{v_c}(f) = \mathcal{P}_{v_g}(f) |H(f)|^2$, e la potenza dissipata su $Z_c(f)$ risulta

$$\begin{aligned} \mathcal{W}_{z_c}(f) &= \mathcal{P}_{v_c}(f) \frac{R_c(f)}{|Z_c(f)|^2} = \mathcal{P}_{v_g}(f) \left| \frac{Z_c(f)}{Z_c(f) + Z_g(f)} \right|^2 \frac{R_c(f)}{|Z_c(f)|^2} = \\ &= \mathcal{P}_{v_g}(f) \frac{R_c(f)}{|Z_c(f) + Z_g(f)|^2} \end{aligned} \tag{14.2}$$

Osserviamo dunque che la potenza dissipata dal carico dipende da $Z_c(f)$, che compare sia a denominatore, che a numeratore con $R_c(f)$. Ci chiediamo allora quale sia il valore di $Z_c$ che realizza il *massimo trasferimento di potenza* tra generatore e carico, sfruttando così appieno la *potenzialità* del generatore detta *potenza disponibile*.

#### 14.2.2.1 Potenza disponibile e massimo trasferimento di potenza

La $\mathcal{W}_{z_c}(f)$ espressa da (14.2) risulta massimizzata qualora il suo denominatore viene reso minimo, e in appendice 14.7.3 si mostra che ciò avviene, per qualunque valore di $R_c(f)$, se si pone $X_c = -X_g$. Pertanto, per $Z_g(f)$ fissato, occorre porre anche $R_c = R_g$, ottenendo così il risultato cercato:

$$\begin{aligned} &se \qquad Z_c(f) = Z_g^*(f) \\ &allora \qquad \mathcal{W}_{z_c}(f) = \max_{Z_c(f)} \{\mathcal{W}_{z_c}(f)\} = \frac{\mathcal{P}_{v_g}(f)}{4R_g(f)} = \mathcal{W}_{d_g}(f) \end{aligned} \tag{14.3}$$

Il valore $\mathcal{W}_{d_g}(f) = \frac{\mathcal{P}_{v_g}(f)}{4R_g(f)}$ prende il nome di spettro di potenza *disponibile* del generatore, dipende solo dai suoi parametri $\mathcal{P}_{v_g}(f)$ e $R_g(f)$, e rappresenta la *massima* potenza ceduta ad un carico che è *adattato* per il *massimo trasferimento di potenza*[4].

La potenza disponibile $\mathcal{W}_{d_g}(f)$ è pertanto *una grandezza caratteristica* del generatore; la potenza effettivamente ceduta ad un carico generico $Z_c(f) \neq Z_g^*(f)$, risulta inferiore a $\mathcal{W}_{d_g}(f)$ di una quantità

$$\alpha(f) = \frac{4R_g(f) R_c(f)}{|Z_g(f) + Z_c(f)|^2}$$

(vedi appendice 14.7.4) e quindi in generale si ha $\mathcal{W}_{z_c}(f) = \alpha(f) \mathcal{W}_{d_g}(f)$.

[4] 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 $Z_g(f) \to 0$.

#### 14.2.2.2 Assenza di distorsioni lineari

Abbiamo già osservato come la tensione ai capi del carico abbia valore $V_c(f) = V_g(f) \cdot \frac{Z_c(f)}{Z_c(f)+Z_g(f)} = V_g(f) H(f)$. Ci chiediamo ora quali condizioni debbano sussistere affinché $H(f)$ si comporti come un *canale perfetto*, ovvero risulti $|H(f)| = cost$ e $\arg\{H(f)\} = 2\pi f\tau$: tali condizioni sono anche indicate come *assenza di distorsioni lineari*. Il risultato cercato si ottiene qualora si ponga

$$Z_c(f) = \alpha Z_g(f) \quad \text{con } \alpha \text{ reale}$$

infatti in tal caso risulta $H(f) = \frac{\alpha Z_g(f)}{(1+\alpha)Z_g(f)} = \frac{\alpha}{1+\alpha}$, ossia $H(f)$ costante. La condizione $Z_c(f) = \alpha Z_g(f)$ prende il nome di *adattamento di impedenza*, a volte ristretta al caso in cui $\alpha = 1$.

#### 14.2.2.3 $Z_g(f)$ reale

Notiamo che massimo trasferimento di potenza ed assenza di distorsioni lineari possono sussistere *congiuntamente,* a patto che $Z_g(f) = R_g$, ovvero che sia il generatore che il carico siano caratterizzati da una impedenza reale.

## 14.3 Reti due porte

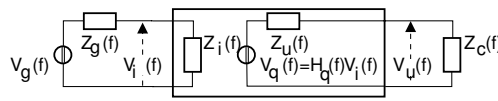
Come anticipato, un circuito accessibile mediante due coppie di morsetti è detto rete due porte, e può essere rappresentata secondo almeno due diversi formalismi: il *modello circuitale* e lo *schema simbolico*.

### 14.3.1 Modello circuitale

In figura è mostrato un possibile modello circuitale[5] 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

$$V_u(f) = V_q(f) H_u(f)$$

[5] 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 temini di impedenze, ammettenze, e parametri ibridi.

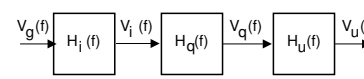
Il caso qui trattato è quello di un modello ibrido, con la particolarità di non presentare influenze esplicite dell'uscita sull'ingresso. Qualora il circuito che si descrive presenti una dipendenza, ad esempio di $Z_i$ da $Z_c$, o $Z_u$ da $Z_g$, questo deve risultare nell'espressione della grandezza dipendente. Viceversa, qualora il circuito presenti in ingresso un generatore controllato da una grandezza di uscita, il modello non è più applicabile.

dipende da quella del generatore controllato $V_q(f)$ mediante il rapporto di partizione $H_u(f) = \frac{Z_c(f)}{Z_u(f)+Z_c(f)}$. Combinando queste relazioni, si ottiene che la funzione di trasferimento complessiva $H(f)$ risulta:

$$V_u(f) = V_g(f) H_i(f) H_q(f) H_u(f) = V_g(f) H(f) \tag{14.4}$$

La relazione mostra come $H(f)$ dipenda, oltre che dalla risposta in frequenza intrinseca della rete $H_q(f)$, anche dalle condizioni di adattamento che si realizzano in ingresso ed in uscita.

### 14.3.2 Schema simbolico




Lo stesso modello circuitale descritto può essere rappresentato equivalentemente mediante lo schema simbolico rappresentato a lato, in cui sono evidenziate le tre funzioni di trasferimento sopra ricavate, e che operano sui segnali indicati. Lo schema simbolico ha il vantaggio di trascendere dal modello circuitale adottato, e di rendere del tutto evidente come la funzione di trasferimento complessiva abbia origine dal prodotto di tre termini di cui solo uno ($H_q(f)$) rappresenta strettamente la rete.

### 14.3.3 Trasferimento energetico

Applicando ora la (14.1) alla potenza ceduta al carico $Z_c(f)$ dal generatore controllato $V_q(f)$, e tenendo conto della (14.4), si ottiene:

$$\mathcal{W}_c(f) = \mathcal{P}_{v_u}(f) \frac{R_c(f)}{|Z_c(f)|^2} = \mathcal{P}_{v_g}(f) |H(f)|^2 \frac{R_c(f)}{|Z_c(f)|^2}$$

Proseguiamo ora l'analisi cercando di individuare una relazione di trasferimento energetico che possa rappresentare caratteristiche esclusive della rete.

**Guadagno di tensione** E' definito come il rapporto tra tensione di uscita e di ingresso:

$$G_v(f) = \frac{V_u(f)}{V_i(f)} = H_q(f) H_u(f)$$

Evidentemente, dipende dalle condizioni di chiusura della rete.

**Guadagno di potenza** E' il rapporto tra la potenza ceduta al carico e quella assorbita all'ingresso della rete:

$$\begin{aligned} G_{\mathcal{W}}(f) &= \frac{\mathcal{W}_c(f)}{\mathcal{W}_i(f)} = \mathcal{P}_{v_g}(f) |H(f)|^2 \frac{R_c(f)}{|Z_c(f)|^2} \cdot \frac{1}{\mathcal{P}_{v_g}(f)} \frac{|Z_g(f) + Z_i(f)|^2}{R_i(f)} = \\ &= |H(f)|^2 \frac{R_c(f)}{R_i(f)} \cdot \left| \frac{Z_g(f) + Z_i(f)}{Z_c(f)} \right|^2 = \\ &= |H_q(f)|^2 \cdot \frac{R_c(f)}{R_i(f)} \cdot \left| \frac{Z_i(f)}{Z_u(f) + Z_c(f)} \right|^2 \end{aligned}$$

ed evidentemente è ancora funzione di $Z_c(f)$ ([6]). Notiamo ora che, qualora il carico sia adattato per il massimo trasferimento di potenza ($Z_c(f) = Z_u^*(f)$), la potenza ceduta a $Z_c(f)$ (e quindi $G_{\mathcal{W}}(f)$) è massima, e la dipendenza di $G_{\mathcal{W}}(f)$ da $Z_c(f)$ decade, risultando

$$G_{\mathcal{W}_{Max}}(f) = |H_q(f)|^2 \cdot \frac{|Z_i(f)|^2}{4R_i(f)\,R_u(f)} \tag{14.5}$$

**Guadagno disponibile** Il rapporto tra la potenza disponibile di uscita, e quella disponibile del generatore posto in ingresso della rete (indipendentemente dal fatto che l'ingresso della rete presenti o meno le condizioni per il massimo trasferimento di potenza) è detto *guadagno disponibile*, e risulta:

$$\begin{aligned} G_d(f) &= \frac{\mathcal{W}_{d_u}(f)}{\mathcal{W}_{d_g}(f)} = \frac{\mathcal{P}_{v_q}(f)}{4R_u(f)} \cdot \frac{4R_g(f)}{\mathcal{P}_{v_g}(f)} = \\ &= \mathcal{P}_{v_g}(f)\,|H_i(f)|^2\,|H_q(f)|^2 \cdot \frac{R_g(f)}{R_u(f)} \cdot \frac{1}{\mathcal{P}_{v_g}(f)} = \\ &= |H_i(f)|^2\,|H_q(f)|^2 \cdot \frac{R_g(f)}{R_u(f)} \end{aligned} \tag{14.6}$$

La relazione trovata mostra la dipendenza di $G_d(f)$ dalle condizioni di chiusura in ingresso; se l'impedenza $Z_g(f)$ del generatore è scelta in modo da conseguire il massimo trasferimento di potenza $Z_g(f) = Z_i^*(f)$, la dipendenza decade ed $|H_i(f)|^2 = \left|\frac{Z_i(f)}{Z_i(f)+Z_i^*(f)}\right|^2 = \frac{|Z_i(f)|^2}{4R_i^2(f)}$; considerando inoltre che $R_g(f) = R_i(f)$, la (14.6) diviene:

$$G_{d_{Max}}(f) = |H_q(f)|^2 \cdot \frac{|Z_i(f)|^2}{4R_u(f)\,R_i(f)} \tag{14.7}$$

Quest'ultima quantità è chiamata *guadagno disponibile* DELLA RETE DUE PORTE ed è quella che appunto dipende solo dai parametri della rete stessa. Confrontando (14.7) con (14.5) notiamo che $G_{d_{Max}}(f)$ coincide con $G_{\mathcal{W}_{Max}}(f)$. Confrontando (14.7) con (14.6), troviamo che $G_d(f) = |H_i(f)|^2\,G_{d_{Max}}(f)\,\frac{4R_g(f)R_i(f)}{|Z_i(f)|^2}$. Considerando ora che $|H_i(f)|^2\,\frac{1}{|Z_i(f)|^2} = \left|\frac{Z_i(f)}{Z_i(f)+Z_g(f)}\right|^2 \frac{1}{|Z_i(f)|^2} = \frac{1}{\left|Z_i(f)+Z_g(f)\right|^2}$, otteniamo

$$G_d(f) = \frac{4R_g(f)\,R_i(f)}{|Z_g(f)+Z_i(f)|^2} \cdot G_{d_{Max}}(f)$$

che ci consente di valutare $G_d(f)$ nelle reali condizioni di chiusura in ingresso, a partire da $G_{d_{Max}}(f) = G_{\mathcal{W}_{Max}}(f)$ che dipende solo dalla rete.

**Collegamento generatore-carico mediante rete due porte**

- Considerando generatore e porta di ingresso della rete adattati per il massimo trasferimento di potenza, la densità di potenza disponibile in uscita risulta

$$\mathcal{W}_{d_u}(f) = G_{d_{Max}}(f)\,\mathcal{W}_{d_g}(f)$$

[6] L'ultimo passaggio tiene conto che (omettendo la dipendenza da $f$):

$$|H|^2 \cdot \left|\frac{Z_g+Z_i}{Z_c}\right|^2 = \left|\frac{Z_i}{Z_i+Z_g} H_q \frac{Z_c}{Z_c+Z_u}\right|^2 \left|\frac{Z_g+Z_i}{Z_c}\right|^2 = |H_q|^2 \cdot \left|\frac{Z_i}{Z_u+Z_c}\right|^2$$

e dunque l'uscita della rete due porte si comporta come un generatore equivalente, caratterizzato da una nuova $\mathcal{W}_{d_u}(f)$ ed una diversa impedenza interna $Z_u(f)$.

- Nel caso in cui in ingresso non si verifichi il massimo trasferimento di potenza, $G_d(f)$ si riduce di un fattore $\beta(f) = \frac{4R_g(f)R_i(f)}{|Z_g(f)+Z_i(f)|^2}$, e dunque la nuova potenza disponibile di uscita risulta

$$\mathcal{W}_{d_u}(f) = \beta(f) \cdot G_{d_{Max}}(f)\, \mathcal{W}_{d_g}(f) = \frac{4R_g(f)\, R_i(f)}{|Z_g(f) + Z_i(f)|^2} \cdot G_{d_{Max}}(f)\, \mathcal{W}_{d_g}(f)$$

- Nel caso infine in cui il carico $Z_c(f)$ in uscita alla rete non sia adattato, quest'ultimo assorbe una potenza inferiore a $\mathcal{W}_{d_u}(f)$ e pari a (vedi Appendice 14.7.4)

$$\mathcal{W}_c(f) = \alpha(f) \cdot \mathcal{W}_{d_u}(f) = \frac{4R_u(f)\, R_c(f)}{|Z_u(f) + Z_c(f)|^2} \cdot \mathcal{W}_{d_u}(f)$$

**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 $A_d(f)$ [dB] $= -G_d(f)$ [dB].

**Reti in cascata** Se più reti sono connesse tra loro l'una di seguito all'altra, e si verificano per ciascuna coppia le condizioni di massimo trasferimento di potenza tra lo stadio di uscita di una e quello di ingresso della successiva, il guadagno disponibile complessivo è il prodotto dei singoli guadagni disponibili: $G_{d_{Tot}} = G_{d1} \cdot G_{d2} \cdot \ldots \cdot G_{dN}$.

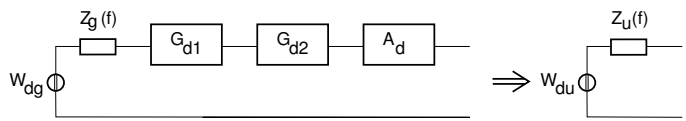
**Esempio** Alla figura seguente è mostrato un generatore con potenza disponibile $W_{dg}$ collegato ad una serie di tre reti due porte; l'effetto complessivo è quello di un nuovo generatore *di uscita* con potenza disponibile $W_{du}$ pari al prodotto di quella del generatore originario, moltiplicata per i guadagni disponibili delle reti attraversate, tenendo anche eventualmente conto delle attenuazioni supplementari:

$$W_{du} = W_{dg} \cdot G_{d1} \cdot G_{d2} \cdot \frac{1}{A_d} \cdot \frac{1}{A_s}$$

che può essere egualmente valutato operando in decibel, come

$$W_{du}\,[dBW] = W_{dg}\,[dBW] + G_{d1}\,[dB] + G_{d1}\,[dB] - A_d\,[dB] - A_s\,[dB]$$

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. 335, puntualizziamo 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.

## 14.4 Misure di potenza in deciBel

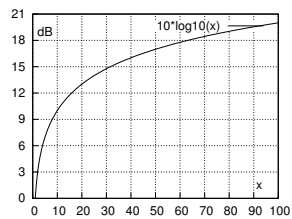
Ci sono almeno due buoni motivi matematici per misurare le grandezze in unità logaritmiche: la prima è che in tal modo si rappresentano in modo uniforme anche grandezze dalla dinamica molto elevata, e la seconda è che prodotti e rapporti, si trasformano in somme e sottrazioni. Inoltre, c'è almeno un buon motivo fisiologico, e cioè che la sensibilità dei nostri sensi segue naturalmente una legge logaritmica, ossia è necessario uno stimolo che aumenta in progressione geometrica, per produrre una sensazione che aumenta linearmente. Ciò posto, va anche detto che l'esperienza di insegnamento mostra come, anche se le misure in dB sono qui per aiutarci nei calcoli, esse sono anche uno degli argomenti in cui lo studente medio tende più facilmente a perdersi. Proviamo quindi a fare un pò di ordine!

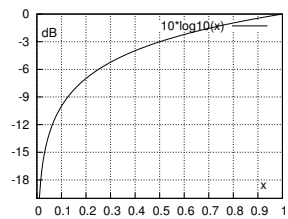
#### 14.4.0.1 La misura logaritmica

Data una qualisiasi grandezza $\alpha$, la sua misura in decibel[7] è definita come

$$\alpha_{dB} = 10 * \log_{10} \alpha \tag{14.8}$$

e descrive le relazioni mostrate nella figura sottostante, a sinistra per valori $\alpha > 1$, ed a destra per $\alpha < 1$, a cui corrispondono rispettivamente valori in decibel positivi e negativi. Inoltre, è mostrata una tabella con alcune corrispondenze che possono comunemente ricorrere: ad esempio, dato che $\log_{10} 2 = 0.30102...$, un valore $\alpha$ pari a 2 equivale a circa 3 dB.




| $\alpha$ | $\alpha_{dB}$ |
|---|---|
| 0 | -$\infty$ |
| $10^{-3}$ | -30 |
| 1 | 0 |
| 2 | $\sim$3 |
| 5 | $\sim$7 |
| 10 | 10 |
| $10^n$ | n·10 |

Nota una grandezza espressa in dB, si può risalire al suo valore naturale, mediante l'ovvia relazione inversa

$$\alpha = 10^{\frac{\alpha_{dB}}{10}}$$

#### 14.4.0.2 Misura relativa dei rapporti

Per esprimere un rapporto $R = \frac{\alpha}{\beta}$ molto grande o molto piccolo, si ricorre spesso alla *scala logaritmica* definita dai dB, calcolando direttamente il rapporto in tali termini, ovvero eseguendo la differenza tra le grandezze $\alpha$ e $\beta$ espresse in dB, in quanto

$$R_{dB} = 10 \cdot \log_{10} \frac{\alpha}{\beta} = 10 \cdot \log_{10} \alpha - 10 \cdot \log_{10} \beta = \alpha_{dB} - \beta_{dB} \tag{14.9}$$

Se le due grandezze $\alpha$ e $\beta$ sono omogenee, come ad esempio due potenze di segnale $\mathcal{P}_x$ e $\mathcal{P}_y$ espresse in $V^2$, o due potenze $\mathcal{W}_x$ e $\mathcal{W}_y$ espresse in $Watt$, allora il loro rapporto è

[7] Un decibel, per come è definito, è la decima parte del Bel. Chissà, forse dopo che definirono il Bel, si accorsero che era troppo grande ? :-)

un *numero puro*, e le sua misura in dB esprime *di quanti dB* il numeratore è maggiore (o minore) del denominatore. Conoscendo una delle due grandezze, ed il valore del loro rapporto, si può ovviamente risalire al valore dell'altra, ovvero ad esempio

$$\alpha_{dB} = R_{dB} + \beta_{dB} \tag{14.10}$$

ma, perché questa ovvia relazione possa avere una utilità pratica, occore sapere cosa rappresenta $\beta$, dopodiché potremo concludere che $\alpha$ rappresenta la stessa cosa, ma $R_{dB}$ decibel più grande. Per questo, si definisce la

#### 14.4.0.3 Musura assoluta delle grandezze

La (14.8) può essere usata per esprimere il *valore assoluto* di una grandezza, assieme alla sua unità di misura, se viene pensata come una applicazione della (14.9), ponendo il denominatore pari *all'unità di misura stessa*. Così ad esempio, la potenza $\mathcal{W}_x$ di $\alpha\, Watt$ viene espressa come

$$\mathcal{W}_{x\,dBW} = 10 \cdot \log_{10} \frac{\alpha_{Watt}}{1\, Watt}\, dBW \tag{14.11}$$

ovvero, misurandola in dB *sopra il Watt*. Quindi, una potenza misurata in $dBW$ può ricondursi alla corrispondente potenza in $Watt$, calcolando

$$\beta_{Watt} = 10^{\frac{\beta_{dBW}}{10}}$$

ed allo stesso modo, si può finalmente applicare la (14.10) per ottenere una grandezza effettiva:

$$\alpha_{dBW} = R_{dB} + \beta_{dBW}$$

Ovviamente, se qui $\beta$ fosse stato *riferito al milliWatt* (e quindi misurato in dBm), anche per $\alpha$ si sarebbe ottenuta la medesima unità di misura.

#### 14.4.0.4 Misura delle densità

Spesso non si ha a che fare con una potenza complessiva, bensì con una *densità* espressa in $\frac{V^2}{Hz}$ o $\frac{W}{Hz}$ (a seconda se si tratti di potenza di segnale o fisica), ovvero con i suoi multipli e sottomultipli ($MHz$, $mWatt$...). Anche in questo caso, è possibile applicare la (14.8) per esprimere in unità logaritmiche la grandezza, purché intesa nel senso della (14.11), ossia indicando l'unità di misura di partenza, individuando così delle grandezze *assolute* misurate in $\frac{dBV^2}{Hz}$, $\frac{dBW}{Hz}$, $\frac{dBW}{MHz}$, $\frac{dBm}{MHz}$....

Quando poi si tratta di applicare una formula come quelle di progetto per i collegamenti, come ad es. la (15.10), occorre prestare attenzione a mantenere congruità dimensionale tra le grandezze usate, eventualmente convertendo dall'una e all'altra.

#### 14.4.0.5 Corrispondenze tra grandezze

Passare da una unità di misura in dB all'altra è molto semplice, basta infatti ricordare le equivalenti relazioni nelle unità lineari rispettive, ed aver presente la misura in dB dei rapporti più comuni. Per questo, ad esempio

- zero $dBW$ equivalgono a 30 $dBm$, dato che 1 Watt = 1000 mWatt;
- -60 $\frac{dBW}{Hz}$ equivalgono a -30 $\frac{dBm}{Hz}$, per lo stesso motivo;
- 20 $\frac{dBW}{MHz}$ equivalgono a -40 $\frac{dBW}{Hz}$, dato che 1 MHz = $10^6$ Hz

## 14.5 Distorsioni lineari

Sono quelle derivanti dal passaggio del segnale $x(t)$ attraverso un canale che non è perfetto, ossia un sistema fisico la cui risposta in frequenza $H(f) = H(f)\,e^{j\varphi_h(f)}$ *non ha* modulo costante e/o fase lineare. Questo determina una *modifica* della forma d'onda a causa della dipendenza delle componenti frequenziali del segnale di uscita $Y(f) = |Y(f)|\,e^{j\varphi_y(f)}$ oltre che da quelle in ingresso, anche dai valori di $H(f)$ come descritto a pag. 223. Queste distorsioni sono dette *lineari* in quanto risultato di una operazione *lineare* come è la convoluzione, verificando così il principio di sovrapposizione degli effetti, e rendendo la distorsione invertibile mediante un dispositivo di equalizzazione. D'altra parte, risultando $|Y(f)| = |X(f)| \cdot |H(f)|$ e $\varphi_y(f) = \varphi_x(f) + \varphi_h(f)$, è possibile descrivere l'effetto della distorsione lineare considerando separatamente la risposta in frequenza $|H(f)|$ e quella in fase $\varphi_h(f)$, anche adottando per queste ultime rappresentazioni particolarmente utili a descrivere in modo complessivo l'entità delle distorsioni stesse.

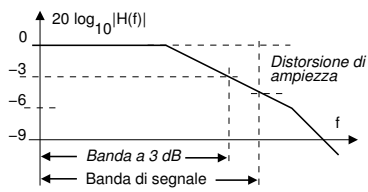
### 14.5.1 Guadagno di potenza in dB

E' definito dall'espressione

$$G_{dB}(f) = 10\log_{10}|H(f)|^2 = 20\log_{10}|H(f)|$$

Qualora l'occupazione di banda del segnale in transito si estenda su di una regione entro la quale $G_{dB}(f)$ non è costante, l'escursione di $G_{dB}(f)$ in tale banda è indicata come *distorsione lineare di ampiezza*, e quantificata appunto in *dB*.

Al contrario, qualora si specifichi quale debba essere la massima distorsione lineare in *dB*, mediante il grafico si individua la banda entro la quale $G_{dB}(f)$ si mantiene all'interno della fascia consentita, ottenendo così il valore della banda per distorsione assegnata (ad esempio, la *banda a 3 dB*, corrispondente alla frequenza di taglio, è definita per questa via).




### 14.5.2 Tempo di ritardo di gruppo

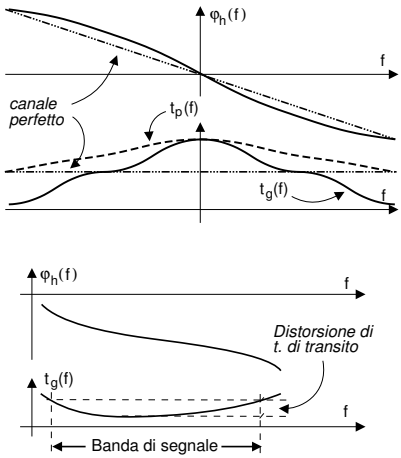
Le conseguenze dello scostamento della risposta in fase $\varphi_h(f)$ rispetto alle condizioni di canale perfetto possono essere sinteticamente espresse mediante la quantità

$$t_g(f) = -\frac{1}{2\pi}\frac{d}{df}\varphi_h(f) \tag{14.12}$$

che rappresenta il ritardo subito da *un gruppo di frequenze vicine a* $f$. Per chiarire il senso di questa definizione, consideriamo un segnale AM-BLD-PS $x(t) = a(t)\cos(2\pi f_0 t)$ la cui densità di potenza è come noto concentrata attorno a $f_0$: in appendice 14.7.5 viene mostrato che, in prima approssimazione, questo si presenta in uscita da $H(f)$ con espressione

$$y(t) = a(t - t_g(f_0))\cos(2\pi f_0 (t - t_p(f_0))) \tag{14.13}$$

dove $t_p(f) = -\frac{\varphi_h(f)}{2\pi f}$ rappresenta il ritardo della portante, mentre $t_g(f)$ è fornito dalla (14.12). Nel caso di canale perfetto si avrebbe $\varphi_h(f) = -2\pi f\tau$, e quindi $t_g = t_p = \tau$ per qualunque frequenza; in caso contrario i due valori possono differire, come mostrato




nella figura a lato, in cui notiamo che $t_g(f) \simeq \tau$ alle frequenze per cui $\varphi_h(f)$ viaggia parallela alla risposta in fase del canale perfetto, mentre risulta $t_p(f) = \tau$ quando $\varphi_h(f)$ la interseca.

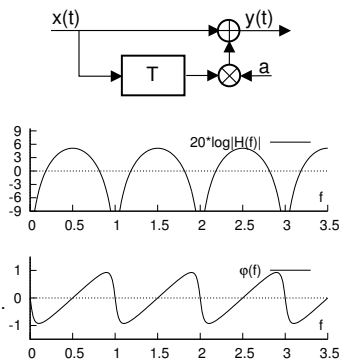
La rappresentazione di $\varphi_h(f)$ nei termini della sua derivata normalizzata $t_g(f) = -\frac{1}{2\pi}\frac{d}{df}\varphi(f)$, misurabile strumentalmente, è utilizzata per valutare in modo grossolano l'entità della *disorsione di fase*, per questo detta anche *distorsione di tempo di transito*, espressa nei termini del massimo scarto tra i valori di $t_g(f)$ nell'ambito della banda di segnale, come mostrato in figura. Ovviamente, minore è questa differenza, e minore risulta l'entità della distorsione subita.

**Esempio** Il filtro trasversale in figura rappresenta un collegamento radio in cui si verifica una eco dovuta a riflessione. Risulta[8]:

$$|H(f)|^2 = 1 + a^2 + 2a\cos 2\pi fT$$

e dunque è presente sia distorsione lineare di ampiezza che di fase (mostrate in figura per $\alpha = .8$ e $T = 1$).

In particolare, $|H(f)|^2$ è periodica di periodo $f = \frac{1}{T}$ e dunque può produrre una forte attenuazione (per $a \simeq 1$) anche a frequenze elevate. Se poi $T$ cambia (perché si sposta il corpo riflettente, oppure si spostano trasmettitore o ricevitore) allora si ha a che fare con un canale tempo-variante, il cui studio non è per ora affrontato.




### 14.5.3 Segnali di banda base

E' noto che l'orecchio umano non è sensibile alle spettro di fase[9] e dunque distorsioni di fase non modificano la qualità del segnale audio - mentre alterazioni di modulo sì.

Se il segnale è numerico, la risposta di fase è importante, perché altrimenti i diversi ritardi di fase alterano l'arrivo delle componenti spettrali dell'impulso elementare, che si deforma e perde la caratteristica di Nyquist. In conseguenza, insorge il fenomeno di interferenza intersimbolica (ISI) e aumenta la probabilità di errore.

### 14.5.4 Segnali modulati

Abbiamo mostrato al § 10.3.1.1 che l'inviluppo complesso $\underline{y}(t) = \frac{1}{2}\underline{x}(t) * \underline{h}(t)$ presenta C.A di B.F. pari a

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

[8] L'espressione di $|H(f)|^2$ è stata ricavata al § 9.7.1. Per la fase (mostrata in figura), osservando che $H(f) = 1 + ae^{-j2\pi fT}$ e che $\varphi(f) = \arctan\frac{\Im}{\Re}$, si ottiene $\varphi(f) = \arctan\frac{a\sin 2\pi fT}{1+\cos 2\pi fT}$.

[9] Al contrario, è sensibile alle sue variazioni: queste ultime sono infatti elaborate dal cervello per estrarne informazioni relative al "movimento" dei suoni percepiti. Confrontando i ritardi differenti e variabili dei segnali pervenuti alle orecchie, si può comprendere se la sorgente degli stessi è in movimento.

Si è anche osservato come, nel caso in cui $H(f)$ risulti a simmetria coniugata rispetto ad $f_0$, si ottiene $h_s(t) = 0$. In tal caso, l'effetto si riduce a quello che si avrebbe in banda base, filtrando il messaggio modulante mediante $h_c(t)$. Pertanto, l'effetto può essere rimosso *equalizzando* i segnali di banda base $y_c(t)$ e $y_s(t)$ con un filtro $H_e(f) = \frac{ae^{j2\pi f\tau}}{H_c(f)}$.

#### 14.5.4.1 Segnali a banda stretta

Nel caso in cui un segnale modulato $x(t)$ presenti una occupazione di frequenza $B$ molto piccola rispetto alla frequenza portante $f_0$, si assume spesso che $H(f)$ nella banda di segnale non vari di molto: ad esempio dal punto di vista del modello circuitale, ciò corrisponde a realizzare le condizioni di adattamento di impedenza in forma approssimata ponendo $Z_g(f) = Z_i(f_0)$ e $Z_c(f) = Z_u(f_0)$, dato che per fequenze $|f - f_0| < \frac{B}{2}$ con $B \ll f_0$, le impedenze $Z_i(f)$ e $Z_u(f)$ non variano di molto. Se invece si preferisce optare per le condizioni di massimo trasferimento di potenza, o più in generale per ciò che riguarda $H_q(f)$, la condizione $B \ll f_0$ spesso permette comunque di approssimare sia modulo che fase come pressoché costanti e pari al valore assunto per $f = f_0$, ossia $H(f) \simeq H(f_0) = G_0 e^{j\phi_0}$. Questa approssimazione permette di trascurare l'effetto delle distorsioni lineari, che in questo caso equivalgono ad una semplice rotazione degli assi dell'inviluppo complesso, reversibile scegliendo opportunamente la fase della portante di demodulazione.

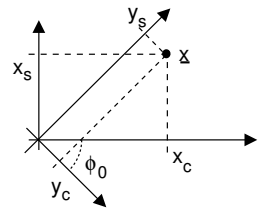
Nelle ipotesi poste, risulta infatti $\underline{H}(f) = 2H^+(f + f_0) = 2G_0 e^{j\phi_0}$ e quindi $\underline{h}(t) = \mathcal{F}^{-1}\{\underline{H}(f)\} = 2G_0 e^{j\phi_0}\delta(t)$; pertanto l'inviluppo complesso di $h(t)$ è un impulso di area complessa $2G_0 e^{j\phi_0} = 2G_0(\cos\phi_0 + j\sin\phi_0)$. All'uscita del canale $H(f)$ troviamo quindi

$$\begin{aligned} \underline{y}(t) &= \frac{1}{2}\cdot\underline{x}(t) * \underline{h}(t) = (x_c(t) + jx_s(t)) * G_0(\cos\phi_0 + j\sin\phi_0)\,\delta(t) = \\ &= G_0[(x_c(t)\cos\phi_0 - x_s(t)\sin\phi_0) + j(x_c(t)\sin\phi_0 + x_s(t)\cos\phi_0)] \end{aligned}$$

che identifica la trasformazione subita come una rotazione

$$\begin{bmatrix} y_c(t) \\ y_s(t) \end{bmatrix} = \begin{bmatrix} \cos\phi_0 & -\sin\phi_0 \\ \sin\phi_0 & \cos\phi_0 \end{bmatrix} \begin{bmatrix} x_c(t) \\ x_s(t) \end{bmatrix}$$

con una matrice dei coefficienti costante. Il risultato della rotazione è esemplificato in figura.




#### 14.5.4.2 Modulazione di ampiezza

**BLD-PS** In questo caso $x_s(t) = 0$ ed allora $\begin{cases} y_c(t) = \frac{1}{2}[x_c(t) * h_c(t)] \\ y_s(t) = \frac{1}{2}[x_c(t) * h_s(t)] \end{cases}$. Adottando, ad esempio, una demodulazione omodina, si ottiene un segnale demodulato pari a $d(t) = x_c(t) * h_c(t)$, equivalente al caso di distorsione lineare di banda base.

**BLD-PI** Il problema maggiore con la portante intera può verificarsi se $H(f_0)$ è molto ridotto (ad esempio a causa di una attenuazione selettiva, esemplificata all'esempio di pag. 344), perché in tal caso il demodulatore inviluppo non funziona più.

**BLU** In questo caso il segnale modulato contiene ambedue le C.A. di B.F., e dunque la presenza di distorsioni lineari provoca il fenomeno noto come *intermodulazione tra componenti analogiche di bassa frequenza,* in quanto in entrambe ($y_c(t)$, $y_s(t)$), si trovano entrambe ($x_c(t)$, $x_s(t)$), mescolate tra loro tramite ($h_c(t)$, $h_s(t)$).

#### 14.5.4.3 Modulazione angolare

Qualora un segnale modulato angolarmente attraversi un canale che presenta distorsioni lineari (di modulo, di fase, od entrambe), si verificano due fenomeni indicati come conversione PM-AM e PM-PM. Si manifesta infatti una modulazione di ampiezza sovrapposta, ed anche la modulazione di fase presenta delle alterazioni. Mentre la modulazione AM "parassita" può essere rimossa da un limitatore in ricezione, quella di fase no; inoltre quest'ultima presenta anche termini non-lineari e dunque non eliminabili mediante equalizzazione.

### 14.5.5 Calcolo dell'SNR

Se il segnale ricevuto ha subito distorsioni di ampiezza, lo spettro di densità di potenza si è deformato, ed il calcolo dell'$SNR$ al ricevitore deve tenere conto dell'effetto filtrante introdotto dal canale attraversato, ossia:

$$SNR = \frac{\mathcal{P}_y}{\mathcal{P}_N} = \frac{\int_B \mathcal{P}_y(f)\,df}{\int_B \mathcal{P}_N(f)\,df} = \frac{\int_B \mathcal{P}_x(f)\,|H(f)|^2\,df}{N_0 B}$$

in cui si è indicata con $B$ la banda a frequenze positive occupata dal segnale.

### 14.5.6 Equalizzazione

Qualora la funzione di trasferimento complessiva del mezzo trasmissivo e delle sue chiusure $H(f) = H_i(f)\,H_q(f)\,H_u(f)$ non soddisfi la condizione di canale perfetto, l'equalizzazione si ottiene inserendo elementi filtranti presso il trasmettitore ($H_T(f)$) ed il ricevitore ($H_R(f)$), in modo da realizzare $H_T(f)\,H(f)\,H_R(f) = a\mathrm{e}^{-j2\pi f\tau}$.

Se non sono presenti entrambi i filtri, ma uno solo di essi, si può scegliere di utilizzare solamente $H_T(f)$, riservando ad $H_R(f)$ l'unico scopo di filtrare il rumore esterno alla banda di segnale, come fatto ai cap. 12 e 15. Se invece l'equalizzazione è tutta demandata ad $H_R(f)$, questa può essere realizzata in modo *adattativo*, modificandone cioè l'andamento senza conoscere l'$H(f)$ da correggere[10], sfruttando conoscenze a riguardo della densità di potenza del segnale in arrivo, od in altri modi[11], come ad es. la tecnica mostrata al § 5.2.3.1. D'altra parte, attribuire il compito dell'equalizzazione ad $H_R(f)$, può causare una "colorazione" del rumore in ingresso al ricevitore, il cui effetto dovrebbe essere analizzato per verificarne le conseguenze sul segnale ricevuto[12].

[10] L'operazione può rendersi necessaria qualora il canale cambi da un collegamento all'altro, come ad esempio nel transito in una rete commutata, od in una comunicazione radiomobile.

[11] Conoscendo la densità di potenza del segnale in arrivo, è possibile generare un segnale "differenza" tra quello che ci si aspetta e quel che invece si osserva, e controllare in modo automatico un filtro in modo da ridurre la differenza tra le densità di potenza al minimo.

Nel caso in cui occorra correggere una distorsione di fase (ad es. nelle trasmissioni numeriche), non ci si può basare sul solo spettro di potenza, e possono essere previste fasi di "apprendimento" dell'equalizzatore, durante le quali il segnale trasmesso è noto al ricevitore, e si può costruire un segnale differenza basato direttamente sulla forma d'onda.

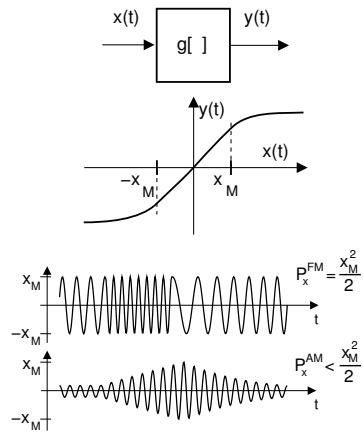
[12] Vedi ad es. le osservazioni alla fine del § 7.6.2.

## 14.6 Distorsioni di non linearità

Descrivono il deterioramento subito dal segnale nel transito attraverso un dispositivo dal comportamento non-lineare, per il quale cioè la relazione ingresso-uscita è del generico tipo $y(t) = g[x(t)]$; espandendo la relazione ingresso-uscita $y = g[x]$, ed arrestando lo sviluppo al terzo ordine, otteniamo:

$$y(t) = G\left[x(t) + \alpha x^2(t) + \beta x^3(t)\right] \tag{14.14}$$

Un caso *tipico* di questo fenomeno si osserva allo stadio finale di un amplificatore di potenza[13] che, per ampiezze del segnale di ingresso maggiori di $\pm x_M$, presenta fenomeni di *saturazione* dei valori di uscita (vedi figura a lato). Notiamo subito come il coefficiente $\beta$, possa rappresentare questo comportamento.




Nel caso in cui il dispositivo che introduce saturazione sia attraversato da segnali con modulazione AM, il fenomeno è particolarmente grave, dato che questi presentano valori di ampiezza direttamente dipendenti da quelli del segnale modulante. Per evitare di operare in regione non lineare, la potenza del segnale in ingresso all'elemento non lineare deve quindi essere ridotta (questa operazione è chiamata *back-off*), e conseguentemente la trasmissione avviene ad un livello di potenza $\mathcal{P}_x^{AM}$ inferiore a quello consentito ($\frac{x_M^2}{2}$) dall'amplificatore.

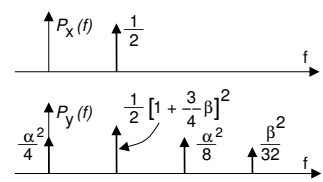
Nel caso di trasmissione FM invece, abbiamo visto che il segnale modulato mantiene sempre la stessa ampiezza, e dunque si può effettuare la trasmissione a piena potenza; in altre parole, una volta fissato il livello di trasmissione, non occorre ricorrere ad un amplificatore sovradimensionato. Esaminiamo ora come quantificare l'effetto delle distorsioni non lineari.

### 14.6.1 Ingresso sinusoidale

Ponendo $x(t) = A\cos\omega_0 t$ la (14.14) si riscrive come[14]

$$\begin{aligned} y(t) &= G\left[A\cos\omega_0 t + \alpha A^2\cos^2\omega_0 t + \beta A^3\cos^3\omega_0 t\right] = \\ &= GA\left[\frac{\alpha A}{2} + \left(1 + \frac{3}{4}\beta A^2\right)\cos\omega_0 t + \frac{\alpha A}{2}\cos 2\omega_0 t + \frac{\beta A^2}{4}\cos 3\omega_0 t\right] \end{aligned}$$

che corrisponde allo spettro di densità di potenza disegnato a lato (che è uno spettro *unilatero*, e calcolato per $A = G = 1$), dove si osserva la comparsa di termini a frequenza multipla di quella di ingresso, oltre che di una componente continua.




Nella pratica, i valori di $\alpha$ e $\beta$ non sono noti, e le relazioni ottenute sono usate per derivarli, ponendo in

[13] come ad esempio è il caso dei TWTA introdotti a pag. 387

[14] Si fa uso delle relazioni $\cos^2\alpha = \frac{1}{2} + \frac{1}{2}\cos 2\alpha$ e $\cos^3\alpha = \frac{3}{4}\cos\alpha + \frac{1}{4}\cos 3\alpha$.

ingresso una sinusoide di potenza nota, ed osservando la potenza delle sue armoniche presenti in uscita.

**Fattori di intermodulazione** Le caratteristiche tecniche che accompagnano gli amplificatori riportano, invece di $\alpha$ e $\beta$, i valori dei *fattori di intermodulazione* $\mu_2$ e $\mu_3$ (detti di seconda e di terza armonica), ottenuti utilizzando appunto un ingresso sinusoidale e misurando le potenze $\mathcal{P}_I$, $\mathcal{P}_{II}$ e $\mathcal{P}_{III}$ alla frequenza in ingresso ed alla sua seconda e terza armonica, e derivando[15] da queste le quantità

$$\mu_2^2 \doteq \frac{\mathcal{P}_{II}}{\mathcal{P}_I^2} \quad \text{e} \quad \mu_3^2 \doteq \frac{\mathcal{P}_{III}}{\mathcal{P}_I^3}$$

da cui si ottengono i coefficienti $\alpha$ e $\beta$ mediante le relazioni $\alpha \simeq 0.7 \cdot \mu_2 \cdot G$, $\beta = 2 \cdot \mu_3 \cdot G^2$ come mostrato alla nota[16].

Scrivendo $\mathcal{P}_{II} = \mu_2^2 \mathcal{P}_I^2$ e $\mathcal{P}_{III} = \mu_3^2 \mathcal{P}_I^3$, osserviamo che per piccoli valori di $\mathcal{P}_I$, la distorsione prodotta sia da $\mathcal{P}_{II}$ che da $\mathcal{P}_{III}$ è trascurabile; all'aumentare di $\mathcal{P}_I$, $\mathcal{P}_{II}$ cresce con il quadrato, mentre $\mathcal{P}_{III}$ con il cubo, e pertanto è quest'ultima componente che poi predomina.

### 14.6.2 Ingresso aleatorio

Nel caso in cui l'ingresso dell'elemento non lineare sia un processo gaussiano, la densità spettrale in uscita può ottenersi come $\mathcal{F}$-trasformata della funzione di autocorrelazione dell'uscita: in virtù di alcune proprietà[17] dei momenti di variabili aleatorie gaussiane, si ottiene in questo caso:

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

ovvero compaiono termini di distorsione di "$2^a$ e $3^a$ armonica" che hanno origine dalla convoluzione della densità di potenza del segnale utile con se stesso.

[15] Le relazioni mostrate si ottengono scrivendo

$$\begin{aligned}
\mathcal{P}_I &= \frac{G^2A^2}{2}\left(1+\frac{3}{4}\beta A^2\right)^2 \simeq \frac{G^2A^2}{2} \quad \left(\text{se } \beta \ll \frac{4}{3A^2}\right)\\
\mathcal{P}_{II} &= \frac{G^2A^4\alpha^2}{8} = \frac{G^4A^4}{4}\frac{1}{G^2}\frac{\alpha^2}{2} = \mathcal{P}_I^2\mu_2^2\\
\mathcal{P}_{III} &= \frac{G^2A^6\beta^2}{32} = \frac{G^6A^6}{8}\frac{1}{G^4}\frac{\beta^2}{4} = \mathcal{P}_I^3\mu_3^2
\end{aligned}$$

[16] Scrivendo $\mu_2^2 \doteq \frac{\mathcal{P}_{II}}{\mathcal{P}_I^2} = \frac{G^2A^4\alpha^2}{8} \cdot \frac{4}{2G^2}$ si ottiene $\mu_2 = \frac{\alpha}{\sqrt{2}G}$ e quindi $\alpha \simeq 0.7 \cdot \mu_2 \cdot G$; allo stesso modo da $\mu_3^2 = \frac{\mathcal{P}_{III}}{\mathcal{P}_I^3} = \frac{G^2A^6\beta^2}{32} \cdot \frac{8}{G^6A^6} = \frac{\beta^2}{4G^4}$ si ha $\mu_3 = \frac{\beta}{2G^2}$ e dunque $\beta = 2 \cdot \mu_3 \cdot G^2$.

[17] Dato che l'uscita ha espressione $y(t) = G\left[x(t) + \alpha x^2(t) + \beta x^3(t)\right]$, il calcolo di $\mathcal{R}_y(\tau) = E\{y(t)\, y(t+\tau)\}$ si sviluppa calcolando i momenti misti $m_x^{(i,j)}(\tau) = E\left\{x^i(t)\, x^j(t+\tau)\right\}$. Se $x(t)$ è un processo gaussiano a media nulla, accade che $m_x^{(i,j)}(\tau) = 0$ se $i+j$ è dispari, mentre in caso contrario si applica il risultato per il valore atteso del prodotto di più v.a. estratte in tempi diversi:

$$E\{x_1 \cdot x_2 \cdot \ldots \cdot x_n\} = \sum \left(E\{x_{p_1} \cdot x_{p_2}\} \cdot E\{x_{p_3} \cdot x_{p_4}\} \cdot \ldots \cdot E\left\{x_{p_{n-1}} \cdot x_{p_n}\right\}\right)$$

in cui la somma è estesa a tutte le possibili permutazioni non equivalenti di $(1, 2 \ldots, n)$ (sono equivalenti se accoppiano con ordine diverso o in posizione diversa le stesse v.a.). Ad esempio, per quattro v.a. si ha:

$$E\{x_1 \cdot x_2 \cdot \ldots \cdot x_n\} = E\{x_1 \cdot x_2\} \cdot E\{x_3 \cdot x_4\} + E\{x_1 \cdot x_3\} \cdot E\{x_2 \cdot x_4\} + E\{x_1 \cdot x_4\} \cdot E\{x_2 \cdot x_3\}$$

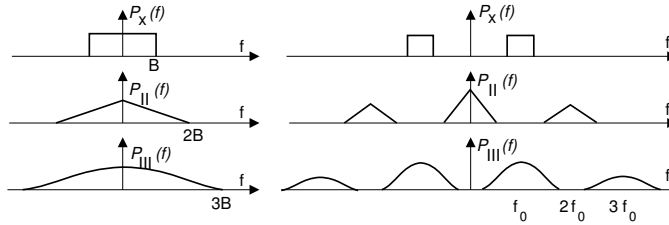



Figura 14.1: Densità spettrale per non linearità; a sin. per banda base, a ds per segnale modulato

Nel caso in cui il processo $x(t)$ sia limitato in banda contigua all'origine, i termini $\mathcal{P}_{II}(f)$ e $\mathcal{P}_{III}(f)$ hanno una banda rispettivamente doppia e tripla di quella di $\mathcal{P}_x(f)$ (vedi fig. 14.1). Nel caso di segnali modulati, oltre ad un allargamento di banda, avviene un fatto diverso e degno di commento: $\mathcal{P}_{II}(f)$ giace in bande diverse da quelle di $\mathcal{P}_x(f)$, pertanto può *essere non considerato* fonte di disturbo. $\mathcal{P}_{III}(f)$ invece ha una componente anch'essa centrata su $f_0$, e dunque è solo questa la fonte disturbo. La potenza totale delle due componenti di disturbo nella banda di $\mathcal{P}_x(f)$ risulta inoltre pari a

$$\mathcal{P}_{II} = 4\mu_2^2\mathcal{P}_I^2 \quad \text{e} \quad \mathcal{P}_{III} = 32\mu_3^2\mathcal{P}_I^3$$

In definitiva, vi sono almeno tre buone ragioni per *tenere d'occhio* il valore di $\beta$, che è causa delle distorsioni di terza armonica:

1. è il coefficiente che tiene conto dei fenomeni di saturazione;
2. produce interferenza "in banda" per i segnali modulati;
3. produce interferenza "fuori banda" che danneggia le trasmissioni a frequenza tripla.

### 14.6.3 Effetto sulla modulazione FM

Se consideriamo un segnale $x(t) = \cos[\omega_0 t + \varphi(t)]$, l'effetto della non linearità produce il segnale

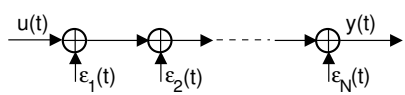
$$\begin{aligned} y(t) &= G\left[\frac{\alpha}{2} + \left(1 + \frac{3}{4}\beta\right)\cos[\omega_0 t + \varphi(t)] + \frac{\alpha}{2}\cos[2\omega_0 t + 2\varphi(t)] + \right. \\ &\quad \left. + \frac{\beta}{4}\cos[3\omega_0 t + 3\varphi(t)]\right] \end{aligned}$$

Osserviamo che i termini a frequenza $2\omega_0$ e $3\omega_0$, nonché il livello in continua, possono essere eliminati mediante un filtro passa-banda centrato in $f = f_0 = \frac{\omega}{2\pi}$; dopo tale operazione, la modulazione di fase $\varphi(t)$ è proprio quella impressa dal modulatore, e pertanto i fenomeni non lineari *non hanno conseguenze sulla FM* (tranne che per le interferenze causate ai canali vicini !)

Il risultato appena illustrato è stato sfruttato nei ponti radio progettati per trasmettere un segnale FDM in FM. Si usa un basso indice di modulazione (risparmiando banda) e si trasmette a piena potenza (senza backoff). La potenza del segnale modulato *non dipende* dal numero di canali contemporaneamente attivi.

## 14.7 Appendici

### 14.7.1 Valutazione dell'$SNR$ dovuto a diverse fonti di disturbo




Consideriamo la situazione in figura, in cui il segnale utile $u(t)$ è affetto da diverse cause di disturbo $\varepsilon_i(t)$ indipendenti tra loro, per ognuna delle quali è noto il relativo $SNR_i = {}^{\mathcal{P}_u}/_{\mathcal{P}_{\varepsilon_i}}$. In tal caso l'effetto di tutte le cause contemporaneamente attive determina un valore di $SNR$ complessivo definito come $SNR_T = {}^{\mathcal{P}_u}/_{\sum_{i=1}^{N} \mathcal{P}_{\varepsilon_i}}$; dato che $\mathcal{P}_{\varepsilon_i} = {}^{\mathcal{P}_u}/_{SNR_i}$ si ottiene

$$SNR_T = \frac{\mathcal{P}_u}{\mathcal{P}_u \sum_{i=1}^{N} \frac{1}{SNR_i}} = \frac{1}{\sum_{i=1}^{N} \frac{1}{SNR_i}}$$

Questo risultato ricorda quello della impedenza equivalente a più impedenze poste in parallelo, il che porta a descrivere l'$SNR$ complessivo come *il parallelo* degli $SNR$. Una applicazione di questo risultato verrà esposta al § 16.3.4.

### 14.7.2 Potenza assorbita da un bipolo

La dimostrazione inizia definendo una potenza *istantanea* assorbita dal bipolo come $w(t) = v(t)\, i(t) = v(t) \cdot (v(t) * y(t))$. La potenza *media* (nel tempo) allora risulta

$$\begin{aligned}
\mathcal{W}_z &= \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} w(t)\, dt = \\
&= \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} v(t) \left[ \int_{-\infty}^{\infty} v(t-\tau)\, y(\tau)\, d\tau \right] dt = \\
&= \int_{-\infty}^{\infty} y(\tau) \left[ \lim_{\Delta t \to \infty} \frac{1}{\Delta t} \int_{-\Delta t/2}^{\Delta t/2} v(t)\, v(t-\tau)\, dt \right] d\tau \\
&= \int_{-\infty}^{\infty} y(\tau)\, \mathcal{R}_v(-\tau)\, d\tau = \int_{-\infty}^{\infty} Y(f)\, \mathcal{P}_v(f)\, df = \\
&= \int_{-\infty}^{\infty} \left[ \Re\{Y(f)\} + j \Im\{Y(f)\} \right] \mathcal{P}_v(f)\, df = \\
&= \int_{-\infty}^{\infty} \mathcal{P}_v(f) \frac{R(f)}{|Z(f)|^2} df
\end{aligned}$$

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 proprità 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* $\mathcal{W}_z(f)$.

### 14.7.3 Condizioni per il massimo trasferimento di potenza

Svolgiamo per intero la dimostrazione delle (14.3). Verifichiamo subito che, mantenendo $Z_g(f)$ fisso e per qualunque valore di $R_c(f)$, la potenza ceduta ad un carico espressa dalla (14.2): $\mathcal{W}_{z_c}(f) = \mathcal{P}_{v_g}(f)\frac{R_c(f)}{|Z_c(f)+Z_g(f)|^2}$ risulta massima se il suo denominatore è il più piccolo possibile, e ciò si verifica quando $X_c(f) = -X_g(f)$, ed in tal caso risulta

$$\mathcal{W}_{z_c}(f) = \mathcal{P}_{v_g}(f)\,\frac{R_c(f)}{(R_c(f)+R_g(f))^2} = \frac{\mathcal{P}_{v_g}(f)}{R_c(f)+2R_g(f)+{}^{(R_g(f))^2}/_{R_c(f)}} \tag{14.15}$$

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:

$$\frac{d}{dR_c}\left(R_c + 2R_g + \frac{R_g^2}{R_c}\right) = 1 - \left(\frac{R_g}{R_c}\right)^2 = 0 \tag{14.16}$$

che fornisce la condizione $R_c(f) = \pm R_g(f)$ in cui il valore negativo viene scartato mentre quello positivo, assieme alla condizione $X_c(f) = -X_g(f)$ determina la condizione $Z_c(f) = Z_g^*(f)$ espressa alla (14.3). Volendo verificare che la (14.16) individui effettivamente un minimo e non un massimo del denominatore di (14.15), se ne può eseguire la derivata seconda, ottenendo

$$\frac{d^2}{dR_c^2}\left(R_c + 2R_g + \frac{R_g^2}{R_c}\right) = \frac{d}{dR_c}\left[1 - \left(\frac{R_g}{R_c}\right)^2\right] = 2\frac{R_g^2}{R_c^3}$$

che verifichiamo immediatamente essere sempre positiva.

### 14.7.4 Potenza ceduta ad un carico $Z_c(f) \neq Z_g^*(f)$

Avendo a disposizione un generatore di segnale di potenza disponibile $\mathcal{W}_d(f)$ ed impedenza interna $Z_g(f)$ assegnate, la tensione a vuoto ai suoi capi ha densità di potenza (di segnale) pari a $\mathcal{P}_v(f) = \mathcal{W}_d(f) 4R_g(f)$. Collegando al generatore un carico generico $Z_c(f)$, la potenza dissipata da quest'ultimo risulta pari a $\mathcal{W}_{z_c}(f) = \mathcal{P}_v(f) \frac{R_c(f)}{|Z_g(f)+Z_c(f)|^2}$. Il rapporto tra la densità di potenza effettivamente ceduta a $Z_c(f)$, e quella che le sarebbe ceduta se questa fosse adattata per il massimo trasferimento di potenza, fornisce la perdita di potenza subìta:

$$\frac{\mathcal{W}_{z_c}(f)}{\mathcal{W}_d(f)} = \mathcal{W}_d(f) 4R_g \frac{R_c(f)}{|Z_g(f)+Z_c(f)|^2} \cdot \frac{1}{\mathcal{W}_d(f)} = \frac{4R_g(f) R_c(f)}{|Z_g(f)+Z_c(f)|^2} = \alpha(f)$$

Pertanto, se $Z_c(f) \neq Z_g^*(f)$, su $Z_c(f)$ si dissipa una potenza pari a $\mathcal{W}_{z_c}(f) = \alpha(f)\mathcal{W}_d(f)$. Il medesimo risultato è valido anche per l'analisi dell'accoppiamento tra il generatore equivalente di uscita di una rete due porte ed un carico.

**Esempio** Consideriamo un generatore con $Z_g(f)$ resistiva e pari a 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 $R_c(f) = 50\ \Omega$ ed $X_c(f) = 2\pi f L = 50\ \Omega$ per $f = f_0$ (da cui $L = \frac{50}{2\pi 10^6} = 7.96\,\mu H$).

Essendo la banda di segnale $2W \ll f_0$, approssimiamo la dipendenza da $f$ di $X_c(f)$ come una costante. In queste ipotesi, la potenza effettivamente ceduta al carico risulta $\mathcal{W}_{z_c} = \alpha \mathcal{W}_d$, con

$$\alpha = \frac{4R_g R_c}{|Z_g + Z_c|^2} = \frac{4 \cdot 50 \cdot 50}{|50 + 50 + j50|^2} = \frac{10000}{12500} = 0.8$$

e quindi $\mathcal{W}_{z_c}$ = 0,8 Watt ovvero, in dBm: $10 \log_{10} 0.8$ = -0.97 dbW = 29.03 dBm.

Il valore $\alpha_{dB} = 10 \log_{10} \alpha$ = 0,97 dB rappresenta il valore della perdita di potenza causata dal mancato verificarsi delle condizioni di massimo trasferimento di potenza, e può essere tenuto in conto come una attenuazione supplementare al collegamento, in fase di valutazione del *link budget* (vedi capitolo seguente).

### 14.7.5 Derivazione del tempo di ritardo di gruppo

Dimostriamo qui il risultato (14.13). Consideriamo di operare nello spazio dell'inviluppo complesso, e di disporre di un canale affetto dalla sola distorsione di fase, ovvero con risposta in frequenza $H(f) = 1 \cdot e^{j\varphi(f)}$; la $\mathcal{F}$-trasformata di $\underline{y}(t)$ si scrive allora come $\underline{Y}(f) = \frac{1}{2}\underline{X}(f)\underline{H}(f)$, dove per $x(t) = a(t)\cos(2\pi f_0 t)$ si ha $\underline{X}(f) = X_c(f) = A(f)$, mentre $\underline{H}(f) = 2H^+(f + f_0) = 2e^{j\varphi(f+f_0)} = 2e^{j\varphi_{pb}(f)}$ dove il pedice $_{pb}$ simboleggia

la caratteristica *passa-basso* dell'inviluppo complesso. Sviluppando $\varphi_{pb}(f)$ in serie di Maclaurin e troncandola al primo termine si ottiene[18] per $f$ prossimo a zero

$$\varphi_{pb}(f) \simeq -2\pi\left(f_0 t_p(f_0) + f t_g(f_0)\right)$$

e quindi

$$\begin{aligned}\underline{Y}(f) &= A(f)\,\mathrm{e}^{j\varphi_{pb}(f)} = A(f)\,\mathrm{e}^{-j2\pi\left(f_0 t_p(f_0)+f t_g(f_0)\right)} \\ &= \mathrm{e}^{-j2\pi f_0 t_p(f_0)}\cdot A(f)\,\mathrm{e}^{-j2\pi f t_g(f_0)}\end{aligned}$$

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

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

a cui corrisponde[19] il segnale modulato espresso dalla (14.13).

[18] Infatti sussistono i seguenti passaggi

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

[19] Infatti al termine $\underline{z}(t) = \mathrm{e}^{-j2\pi f_0 t_p}$ corrispondono le componenti a frequenza positiva e negativa

$$z^{\pm}(t) = \frac{1}{2}\underline{z}(t)\,\mathrm{e}^{\pm j2\pi f_0 t} = \frac{1}{2}\mathrm{e}^{-j2\pi f_0 t_p}\mathrm{e}^{\pm j2\pi f_0 t} = \frac{1}{2}\mathrm{e}^{\pm j2\pi f_0\left(t-t_p\right)}$$

da cui otteniamo in modo semplice il corrispondente segnale

$$z(t) = z^{+}(t) + z^{-}(t) = \frac{1}{2}\left(\mathrm{e}^{j2\pi f_0\left(t-t_p\right)} + \mathrm{e}^{-j2\pi f_0\left(t-t_p\right)}\right) = \cos\left(2\pi f_0\left(t-t_p\right)\right)$$

# Capitolo 15

# Collegamenti e mezzi trasmissivi

In questo capitolo sono caratterizzati i mezzi trasmissivi cavo, radio e fibra ottica, applicando agli stessi le considerazioni fin qui svolte. Le caratteristiche dei diversi mezzi intervengono nella analisi del dimensionamento dei sistemi, e quando possibile sono discusse tecniche particolari di realizzazione.

## 15.1 Dimensionamento di un collegamento

Introduciamo subito alcune definizioni e concetti, che aiutano a meglio analizzare il problema del dimensionamento di un collegamento. I parametri fondamentali del collegamento sono espressi dalla potenza disponibile del trasmettitore $W_{dT}$, dalla minima potenza che occorre ricevere $W_{R_{Min}}$ (spesso indicata come *sensibilità* del ricevitore) e dall'attenuazione disponibile $A_d$ del mezzo di trasmissione che si intende utilizzare.

**Determinazione di $W_{R_{Min}}$** Si ottiene in base alla conoscenza del livello di rumore $\frac{N_0}{2}$ in ingresso al ricevitore (vedi cap. 16) e dell'$SNR$ (cap. 12) o della $P_e$ (vedi eq. (7.11) a pag. 146 e cap. 13) che si desidera ottenere. Nel caso analogico, se si richiede un valore $SNR = \alpha SNR_0 = \alpha \frac{W_R}{N_0 W}$, si ottiene[1]

$$W_{R_{Min}} = N_0 W \cdot \frac{SNR}{\alpha}$$

mentre in quello numerico, la necessità di conseguire un valore di $P_e^{bit}$ consente di determinare il valore di $\frac{E_b}{N_0} = \frac{W_R}{f_b N_0}$, e quindi

$$W_{R_{Min}} = N_0 f_b \cdot \frac{E_b}{N_0}$$

Benché la valutazione delle prestazioni svolta ai predecenti capitoli consideri potenze *di segnale*, lo stesso valore $SNR$ esprime anche un rapporto tra potenze *disponibili*, dato che sia segnale che rumore hanno origine da generatori che condividono la stessa impedenza interna (vedi § 16.2). Infatti

$$\frac{W_{dR}}{W_{dN}} = \frac{P_R}{4R_g} \frac{4R_g}{P_N} = \frac{P_R}{P_N}$$

[1] Come definito al § 12.2.1.1, $SNR_0$ dipende solo dalle caratteristiche del collegamento, mentre il coefficiente $\alpha$ rappresenta la dipendenza dal tipo di modulazione adottata, e differisce da uno nei casi di modulazione FM, AM-PI e AM-PPS.

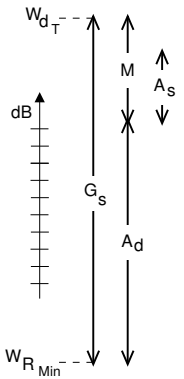
così come l'$SNR$ non varia se, anziché le potenze *disponibili*, si considerano quelle assorbite da un carico (lo stadio di ingresso del ricevitore), dato che segnale e rumore subiscono il medesimo rapporto di partizione (vedi § 14.2.2).

**Guadagno di Sistema** Il rapporto

$$G_s = \frac{W_{dT}}{W_{R_{Min}}}$$

è detto *Guadagno di sistema*[2] e rappresenta il massimo valore di attenuazione *disponibile* $A_d$ che è possibile superare. La differenza in decibel $G_{s_{dB}} = W_{dT}\,[dBW] - W_{R_{Min}}\,[dBW]$ 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 $\geq 0$, 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, avere origine dal fallimento delle condizioni per il massimo trasferimento di potenza, oppure essere causate da un fenomeno piovoso in un collegamento radio, o dipendere dalla perdita di segnale dovuta alla giunzione tra tratte in fibra ottica....

**Grado di servizio** Nel capitolo 8 il concetto di grado di servizio è stato associato al valore di probabilità con cui può verificarsi un fenomeno di blocco in un elemento di commutazione. Un concetto del tutto analogo sussiste, qualora le attenuazioni supplementari siano grandezze aleatorie, e la loro somma possa superare il valore del margine a disposizione: in tal caso, la potenza ricevuta si riduce sotto la minima $W_{R_{Min}}$, ed il collegamento "va *fuori specifiche*". Pertanto, in sede di dimensionamento di un collegamento, indichiamo con grado di servizio la percentuale di tempo per la quale si mantiene $W_R > W_{R_{Min}}$, ovvero la probabilità che le attenuazioni supplementari *non* superino il margine, ossia

$$Grado\ di\ Servizio = Pr\left\{\sum A_{s_{dB}} < M_{dB}\right\}$$

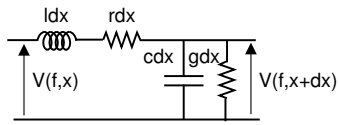
**Esempio** Un grado di servizio del 99.99 % equivale a poco meno di 1 ora l'anno di fuori servizio, e corrisponde a richiedere che $Pr\left\{\sum A_{s_{dB}} > M_{dB}\right\} = 10^{-4}$.

[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.

## 15.2 Collegamenti in cavo

La descrizione completa delle caratteristiche e delle prestazioni dei cavi in rame è una materia molto vasta, di cui forniamo di seguito solo alcuni risultati, strettamente legati agli aspetti di telecomunicazione, il più rilevante dei quali è senz'altro il manifestarsi dell'effetto pelle, che determina (per $f > 100$ KHz) una attenuazione in dB proporzionale a $\sqrt{f}$. La sezione è completata da una breve catalogazione dei cavi usati per telecomunicazioni.

### 15.2.1 Costanti distribuite, grandezze derivate, e condizioni generali




Un conduttore elettrico uniforme e di lunghezza infinita, è descritto in base ad un modello a costanti distribuite, espresso in termini delle *costanti primarie* costituite dalla resistenza $r$, la conduttanza $g$, la capacità $c$ e l'induttanza $l$ per unità di lunghezza. La teoria delle linee uniformi definisce quindi due grandezze derivate dalle costanti primarie: l'*impedenza caratteristica* $Z_0(f)$ e la *costante di propagazione* $\gamma(f)$.

**Impedenza caratteristica** E' definita come

$$Z_0(f) = R_0(f) + jX_0(f) = \sqrt{\frac{r + j2\pi fl}{g + j2\pi fc}} \tag{15.1}$$

e rappresenta il rapporto tra $V(f)$ ed $I(f)$ in un generico punto del cavo, permettendo di scrivere

$$I(f) = \frac{V(f)}{Z_0(f)}$$

**Costante di propagazione** E' definita come

$$\gamma(f) = \alpha(f) + j\beta(f) = \sqrt{(r + j2\pi fl)(g + j2\pi fc)} \tag{15.2}$$

mentre la grandezza $\mathrm{e}^{-\gamma(f)d}$ rappresenta il rapporto dei valori di tensione presenti tra due punti di un cavo di lunghezza infinita, distanti $d$, permettendo di scrivere

$$V(f, x + d) = \mathrm{e}^{-\gamma(f)d} V(f, x)$$

**Condizioni di chiusura** Qualora il cavo di lunghezza $d$ sia chiuso ai suoi estremi su di un generatore con impedenza $Z_g(f)$ e su di un carico $Z_c(f)$, risultano definiti i *coefficienti di riflessione* del generatore e del carico:

$$r_g(f) = \frac{Z_g(f) - Z_0(f)}{Z_g(f) + Z_0(f)} \quad \text{e} \quad r_c(f) = \frac{Z_c(f) - Z_0(f)}{Z_c(f) + Z_0(f)} \tag{15.3}$$

Osserviamo subito che nel caso in cui $Z_g(f) = Z_c(f) = Z_0(f)$, risulta $r_g(f) = r_c(f) = 0$.

**Quadripolo equivalente** L'impedenza vista dai morsetti di *ingresso* e di *uscita* di un cavo, interposto tra generatore e carico, vale rispettivamente

$$Z_i(f) = Z_0(f)\frac{1 + r_c(f)\cdot e^{-2d\gamma(f)}}{1 - r_c(f)\cdot e^{-2d\gamma(f)}} \quad \text{e} \quad Z_u(f) = Z_0(f)\frac{1 + r_g(f)\cdot e^{-2d\gamma(f)}}{1 - r_g(f)\cdot e^{-2d\gamma(f)}} \quad (15.4)$$

Allo stesso tempo, la funzione di trasferimento *intrinseca* risulta

$$H_q(f) = 2\frac{e^{-d\gamma(f)}}{1 - r_g(f)\cdot r_c(f)\cdot e^{-2d\gamma(f)}} \quad (15.5)$$

**Condizioni di adattamento** Nel caso in cui $Z_g(f) = Z_c(f) = Z_0(f)$, come sappiamo, il quadripolo si comporta in modo perfetto. In tal caso, risultando $r_g(f) = r_c(f) = 0$, si ottiene che $Z_i(f) = Z_u(f) = Z_0(f)$ e $H_q(f) = \frac{V_q(f)}{V_i(f)} = 2e^{-d\gamma(f)}$: il cavo si comporta allora come se avesse lunghezza infinita. In tal caso, inoltre, risulta che $H_i(f) = \frac{1}{2}$ ed $R_g(f) = R_u(f)$; pertanto il guadagno disponibile risulta

$$G_d(f) = |H_i(f)|^2 |H_q(f)|^2 \frac{R_g(f)}{R_u(f)} = \frac{1}{4}\left|2e^{-d[\alpha(f)+j\beta(f)]}\right|^2 = e^{-2d\alpha(f)}$$

**Condizione di Heaviside** Nel caso in cui i valori delle costanti primarie siano tali da risultare $r \cdot c = l \cdot g$, relazione nota come *condizione di Heaviside*, le (15.1) e (15.2) si modificano, e si ottiene

$$\gamma(f) = \sqrt{rg} + j2\pi f\sqrt{lc} \qquad e \qquad Z_0(f) = \sqrt{\frac{r}{g}} = \sqrt{\frac{l}{c}} = R_0$$

e pertanto, risultando $\alpha(f)$ costante e $\beta(f)$ linearmente crescente con la frequenza, si realizzano le condizioni di un canale perfetto; dato inoltre che l'impedenza caratteristica $Z_0(f) = R_0$ è solo resistiva ed indipendente dalla frequenza, diviene semplice realizzare la condizione di adattamento $Z_g(f) = Z_c(f) = R_0$, il che determina al contempo anche il massimo trasferimento di potenza, e implica che $r_g(f) = r_c(f) = 0$, e quindi

$$H_q(f) = 2e^{-d\alpha(f)}e^{-jd\beta(f)} = 2e^{-d\sqrt{rg}}e^{-jd2\pi f\sqrt{lc}}$$

In definitiva, la funzione di trasferimento complessiva in questo caso vale

$$H(f) = H_i(f)\,H_q(f)\,H_u(f) = \frac{1}{2}2e^{-d\sqrt{rg}}e^{-jd2\pi f\sqrt{lc}}\frac{1}{2} = \frac{1}{2}e^{-d\sqrt{rg}}e^{-jd2\pi f\sqrt{lc}}$$

equivalente quindi ad un canale perfetto con guadagno $G = \frac{1}{2}e^{-d\sqrt{rg}}$ e ritardo $t_R = d\sqrt{lc}$; al contempo, l'attenuazione disponibile risulta indipendente da $f$, e pari a

$$A_d(f) = 1/G_d(f) = e^{2d\sqrt{rg}}$$

### 15.2.2 Trasmissione in cavo

In generale, le costanti primarie del cavo non soddisfano le condizioni di Heaviside, e le impedenze di chiusura non sono adattate. In tal caso si ha $r_g(f) \neq 0$ e/o $r_c(f) \neq 0$, e devono essere applicate le (15.4) e (15.5).

**Cavo molto lungo** Se il cavo è sufficientemente lungo da poter porre $e^{-2d\gamma(f)} \ll 1$, ossia $\left|e^{-2d\gamma(f)}\right| = e^{-2d\alpha(f)} \ll 1$, le (15.4) divengono $Z_i(f) = Z_u(f) \simeq Z_0(f)$, mentre la (15.5) si semplifica in $H_q(f) = 2e^{-d\gamma(f)}$; nel caso generale risulta pertanto

$$G_d(f) = |H_q(f)|^2 \cdot |H_i(f)|^2 \cdot \frac{R_g(f)}{R_u(f)} = 4 \cdot e^{-2d\alpha(f)} \cdot |H_i(f)|^2 \cdot \frac{R_g(f)}{R_u(f)}$$

che evidenzia due cause di distorsione lineare, di cui la prima dipende dal disadattamento di impedenze in ingresso ed uscita: qualora invece si realizzi la condizione $Z_g(f) = Z_c(f) = Z_0(f)$, si ottiene

$$A_d(f) = \frac{1}{G_d(f)} = e^{2d\alpha(f)}$$

La seconda causa di distorsione dipende dal comportamento non perfetto di $H_q(f) = 2e^{-d\gamma(f)}$, che secondo la teoria può essere completamente corretto, solo nel caso in cui le costanti primarie soddisfino le condizioni di Heaviside. In pratica, però, il risultato è diverso, perchè.... le "costanti primarie" *non sono costanti !!!*

**Effetto pelle** Si tratta di un fenomeno legato all'addensamento del moto degli elettroni verso la superficie del cavo, al crescere della frequenza. Per questo motivo, si riduce la superficie del conduttore realmente attraversata da corrente elettrica, a cui corrisponde un aumento della resistenza per unità di lunghezza $r$. Si può mostrare che, per frequenze maggiori di 50-100 KHz, la resistenza per unità di lunghezza $r$ aumenta proporzionalmente a $\sqrt{f}$, e quindi si può scrivere $\alpha(f) = \alpha_0\sqrt{f}$, in cui la costante $\alpha_0$ 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 sè tutte le costanti coinvolte, prende il nome di *attenuazione chilometrica*, ed è espresso in db/Km, ad una determinata frequenza (ad es. 1 MHz). Pertanto, poiché nell'applicare la formula occorre mantenere congruenza dimensionale, si ottiene in definitiva:

$$A_d(f)|_{dB} = A_0(f_R) \cdot d_{Km} \cdot \sqrt{\frac{f}{f_R}}$$

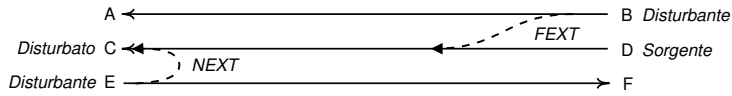
in cui $f_R$ rappresenta la frequenza di riferimento per la quale è disponibile il valore di $A_0$. Questo risultato può essere usato come formula di progetto, e mette in evidenza come l'attenuazione in dB dei cavi sia linearmente proporzionale alla lunghezza[3].

**Equalizzazione** In presenza di effetto pelle, la funzione di trasferimento intrinseca $H_q(f) = 2e^{-d\gamma(f)}$ presenta una dipendenza da $f$ tutt'altro che perfetta, causando potenzialmente distorsioni lineari sui segnali in transito. Un problema analogo insorge anche in assenza di effetto pelle, qualora si manifesti un disadattamento di impedenze ed il cavo non sia sufficentemente lungo (vedi appresso).

[3]Questa circostanza è comune con le trasmissioni in fibra ottica, ed è legato alla presenza nel mezzo di una componente dissipativa, in questo caso la resistenza.

Se la banda di segnale è sufficientemente estesa da causare una distorsione lineare non trascurabile, o se la particolare natura del segnale (ad es. numerico) richiede la presenza di un ritardo strettamente costante con $f$, è necessario prevedere uno stadio di equalizzazione.

**Diafonia** La diafonia, indicata in inglese con il termine di *crosstalk*, consiste nei fenomeni di interferenza tra i messaggi trasportati su cavi disposti in prossimità reciproca, e dovuti a fenomeni di induzione elettromagnetica ed accoppiamenti elettrostatici. Il fenomeno è particolarmente rilevante in tutti i casi in cui molti cavi giacciono affasciati in una medesima canalizzazione, condividendo un lunghezza significativa di percorso. Nel caso di telefonia analogica, la diafonia può causare l'ascolto indesiderato di altre comunicazioni[4]; nel caso di trasmissioni numeriche o di segnali modulati, la diafonia produce un disturbo additivo supplementare, che peggiora le prestazioni espresse in termini di probabilità di errore o di SNR.




Con riferimento allo schema della figura soprastante, consideriamo un collegamento D-C su cui gravano due cause di interferenza di diafonia: il collegamento da E ad F produce il fenomeno di *paradiafonia* (in inglese NEXT, *near end crosstalk*), mentre il collegamento da B ad A produce il fenomeno di *telediafonia* (FEXT, *far end crosstalk*). Nel primo caso, il segnale disturbante ha origine in prossimità del punto di prelievo del segnale disturbato, mentre nel secondo ha origine in prossimità del punto di immissione.

L'entità del disturbo è quantificata mediante un valore di attenuazione di diafonia tra le sorgenti disturbanti e l'estremo disturbato. La circostanza che, nei rispettivi punti di immissione, i segnali disturbanti hanno la stessa potenza della sorgente che emette il segnale disturbato, permette di definire lo *scarto di 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[5] $W_E^{next} = W_E - A_{EC} = W_D - A_{EC} = W_C + A_{DC} - A_{EC} = W_C - \Delta A_{EC}$ , ossia di $\Delta A_{EC}$ dB inferiore al segnale utile. Una definizione del tutto analoga risulta per la telediafonia (FEXT), per la quale il livello di potenza del segnale disturbante proveniente da B ed osservato al punto C risulta $W_B^{fext} = W_C - \Delta A_{BC}$ in cui lo *scarto di telediafonia* ha il valore

$$\Delta A_{BC}|_{dB} = A_{BC}|_{dB} - A_{DC}|_{dB}$$

[4] ... le famose *interferenze* telefoniche, praticamente scomparse con l'avvento della telefonia numerica (PCM), da non confondere con ... *le intercettazioni*.

[5] Omettiamo di indicare di operare in dB per chiarezza di notazione.

#### 15.2.2.1 Casi limite

**Cavo a basse perdite** E' un modello applicabile per tutte quelle frequenze per cui risulti $r \ll 2\pi f l$ e $g \ll 2\pi f c$. In tal caso le (15.1) e (15.2) 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 $Z_g = Z_c = R_0$, che determina

$$H_q(f) = 2\mathrm{e}^{-jd2\pi f\sqrt{lc}}$$

quindi il cavo non presenta distorsioni di ampiezza, ha una attenuazione trascurabile, e manifesta una distorsione di fase lineare in $f$, realizzando quindi le condizioni di canale perfetto.

**Cavo corto** E' il caso di collegamenti interni agli apparati, o tra un trasmettitore-ricevitore e la relativa antenna. La ridotta lunghezza del cavo permette di scrivere

$$\mathrm{e}^{-d\gamma(f)} = \mathrm{e}^{-d\alpha(f)}\mathrm{e}^{-jd\beta(f)} \simeq \mathrm{e}^{-jd\beta(f)}$$

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. (15.3) mostra come risulti $r_c(f) = \pm 1$ rispettivamente, e la prima delle (15.4) diviene

$$Z_i(f) = Z_0(f)\frac{1 \pm \mathrm{e}^{-j2d\beta(f)}}{1 \mp \mathrm{e}^{-j2d\beta(f)}}$$

e si vede che per quei valori (ricorrenti) di frequenza $f$ che rendono $\mathrm{e}^{-jd\beta(f)} = \pm 1$ ([6]), l'impedenza di ingresso del cavo può risultare infinita o nulla.

Evidentemente, le distorsioni lineari prodotte in questo caso hanno un andamento del tutto dipendente dalle particolari condizioni operative, e dunque la loro equalizzazione deve prevedere componenti in grado di adattarsi alla $H_q(f)$ del caso[7]. 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[8].

[6] Ovvero, tali che $|\cos 2d\beta(f) - j\sin 2d\beta(f)| = 1$, e quindi $d\beta(f) = k\frac{\pi}{2}$ con $k = 0, 1, 2, \ldots$

[7] Può ad esempio rendersi necessario "tarare" un trasmetttore radio, la prima volta che lo si collega all'antenna.

[8] E' questa la fase in cui il modem che usiamo per collegarci al provider internet emette una serie di orribili suoni....

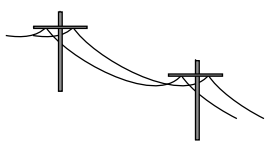
### 15.2.3 Tipologie di cavi per le telecomunicazioni

Descriviamo i principali tipi di cavo utilizzati, per i quali forniamo in tabella i valori tipici delle grandezze essenziali, nelle condizioni illustrate nel testo che segue.

| **Tipo di cavo** | $A_0$ $[dB/Km]$ | $Z_0$ $[\Omega]$ | $r$, $g$, $l$, $c$ per 1 Km |
|---|---|---|---|
| Linee aeree | 0.036 ad 1 KHz<br>0.14 a 100 KHz | 600 | 5, $10^{-6}$, $2{\cdot}10^{-3}$, $5{\cdot}10^{-9}$ |
| Coppie ritorte | 1.2 ad 1 KHz<br>6 a 100 KHz<br>20 a 1 MHz | $600e^{-j\frac{\pi}{4}}$ | 100, $5{\cdot}10^{-5}$, $10^{-3}$, $5{\cdot}10^{-8}$ |
| Coax 1.2/4.4 mm | 5.3 ad 1 MHz | 75 con politene | 89, $1.88{\cdot}10^{-7}$, $.26{\cdot}10^{-6}$, $10^{-10}$ |
| " 2.6/9.5 mm | 2.3 ad 1 MHz | 50 con aria | 41, ", ", " |
| " 8.4/38 mm | .88 ad 1 MHz | $\frac{138}{\sqrt{\varepsilon_r}}\log_{10}\frac{D}{d}$ | 1.45, ", ", " |

#### 15.2.3.1 Coppie simmetriche

**Linee aeree** Sono costituite da una coppia di conduttori nudi, di bronzo od acciaio rivestito in rame, con diametro $\phi$ da 2 a 4 mm, sostenuti da una palificazione che li mantiene a distanza di 15 - 30 cm. L'uso delle linee aeree è andato estinguendosi con il tempo, ma rimane largamente diffuso nei paesi più poveri.

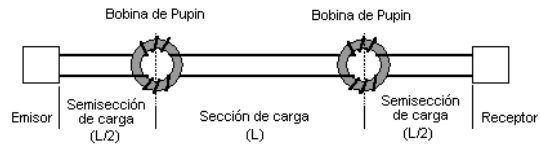


I valori riportati in tabella sono riferiti a conduttori con $\phi = 3$ mm, a frequenza di 1 KHz; la $r$ già a 100 KHz cresce al valore di 20 Ω/Km, mentre la conduttanza $g$ a 100 KHz e con tempo molto umido, può crescere fino a decine di volte il suo valore nominale ad 1 KHz. I valori riportati mostrano come le condizioni di Heaviside non siano rispettate, in quanto $rc \gg lg$, anche se lo scarto è inferiore rispetto al caso delle coppie ritorte.

L'impedenza caratteristica riportata in tabella, di circa 600 Ω, è ottenuta applicando il modello a basse perdite, con le costanti primarie indicate.

**Coppie ritorte** Sono costituite da una coppia di conduttori in rame, con $\phi$ da 0.4 ad 1.3 mm, rivestiti di materiale isolante, ed avvolti tra loro secondo eliche con passo grande rispetto al diametro. Un numero variabile di tali coppie (tra qualche decina e qualche centinaio) sono poi raggruppate assieme, e rivestite con guaine protettive isolanti o metalliche; il risultato dell'operazione è interrato o sospeso mediante una fune in acciaio. L'uso delle coppie ritorte, nato allo scopo di realizzare il collegamento tra utente e centrale telefonica, si è esteso al cablaggio di reti locali (LAN) con topologia a stella (IEEE 802.3); in tale contesto, i cavi sono indicati come UTP (*unshielded twisted pair*).

I valori riportati in tabella sono riferiti a conduttori con $\phi = .7$ mm, a frequenza di 1 KHz; la $r$ a 100 KHz è circa doppia. La $g$ dipende sostanzialmente dall'isolante utilizzato, mentre l'aumento di $c$ è evidentemente legato alla vicinanza dei conduttori. Anche in questo caso, risulta $rc \gg lg$, e dunque le condizioni di Heaviside non sono verificate. Nel passato, si è fatto largo uso dell'espediente di innalzare artificialmente $l$, collocando ad intervalli regolari una induttanza "concentrata" (le cosiddette bobine *Pupin*), realizzando così nella banda del canale telefonico un comportamento approssimativamente perfetto.



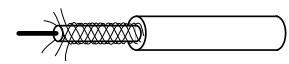

Al crescere della frequenza, però, le bobine Pupin producono un effetto passa basso, aumentando di molto il valore di attenuazione; attualmente, se le stesse coppie ritorte sono utilizzate per la trasmissione di segnali numerici PCM, le bobine Pupin sono state rimosse, ed al loro posto inseriti ripetitori rigenerativi.

L'impedenza caratteristica riportata in tabella, di circa 600 Ω, è valida a frequenze audio, con cavi $\phi = .7$ mm. Prevalendo l'aspetto capacitivo, al crescere della frequenza $Z_0$ si riduce a 100-200 Ω, con fase di -10 gradi. L'attenuazione chilometrica riportata, è sempre relativa al caso $\phi = .7$ mm; per diametri di 1.3 mm si ottengono valori circa dimezzati, mentre con $\phi = .4$ mm il valore di $A_0$ risulta maggiore.

Come ultima osservazione, illustriamo come l'avvolgimento della coppia su se stessa ha lo scopo di ridurre i disturbi di diafonia. Infatti, se il passo dell'elica è diverso tra le coppie affasciate in unico cavo, le tensioni e correnti indotte da una coppia su di un'altra non interessano sempre lo stesso conduttore, ma entrambi in modo alternato. L'avvolgimento della coppia disturbante, inoltre, produce una alternanza dei conduttori in vicinanza della coppia disturbata, aggiungendo una ulteriore alternanza del verso del fenomeno di disturbo. Con questi accorgimenti, si trovano attenuazioni di diafonia a frequenze vocali, dell'ordine di 80-90 dB su 6 Km. All'aumentare della frequenza, e della lunghezza del percorso comune, l'attenuazione di diafonia diminuisce (e quindi l'interferenza aumenta), fino a mostrare valori di 60-70 dB a 750 KHz su 1.6 Km.

#### 15.2.3.2 Cavo coassiale

Un conduttore centrale è ricoperto di dielettrico, su cui è avvolto il secondo conduttore, intrecciato a formare una sorta di calza, e racchiuso a sua volta in una guaina isolante. La particolare conformazione del cavo lo rende molto più resistente ai fenomeni di interferenza; indicando con $\phi$ 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. 362. 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.



Finchè $D/\phi = 3.6$, l'impedenza caratteristica dipende solo dal dielettrico, con l'espressione generale fornita in tabella, ottenendo i valori di 50 e 75 Ω con dielettrico aria e polietilene rispettivamente.

I valori delle costanti primarie riportati in tabella sono ottenuti facendo uso delle seguenti relazioni: $r = 8.4 \cdot 10^{-8}\sqrt{f}\left(\frac{1}{D} + \frac{1}{\phi}\right)\ \Omega/m$; $l = 0.46 \log_{10}\frac{D}{\phi}\ \mu H/m$; $g = 152 \cdot 10^{-12}\frac{f\varepsilon_r \tan\delta}{\log_{10}\frac{D}{\phi}}\ S/m$; $c = \frac{24.2\cdot\varepsilon_r}{\log_{10}\frac{D}{\phi}}\ pF/m$; in cui si è posto $f$ (in Hz nelle formule) pari a 1 MHz, $D$ e $d$ sono espressi in metri, $\varepsilon_r$ è la costante dielettrica, e $\tan\delta$ è l'angolo di perdita del dielettrico; nel caso del polietilene, risulta $\varepsilon_r = 2.3$, $\tan\delta = 3 \cdot 10^{-4}$.

**Esercizio** Si desidera effettuare una trasmissione FDM di 120 canali telefonici, ognuno modulato AM-BLU, su di un cavo coassiale, nella banda di frequenze $1 \div 1.48$ MHz. Desiderando una potenza ricevuta per ogni canale di almeno 1 mW, e disponendo di un trasmettitore in

grado di erogare 10 W, determinare la massima lunghezza del collegamento, supponendo verificate le condizioni di adattamento agli estremi del cavo, con impedenza caratteristica resistiva, ed attenuazione chilometrica $A_0 = 5.3$ dB/Km ad 1MHz. Di quanto dovrebbe aumentare la potenza trasmessa $W_{dT}$ per raddoppiare la lunghezza?

**Soluzione** Supponendo tutti i canali contemporaneamente attivi, la potenza trasmessa per ciascuno di essi risulta pari a

$$W_{dT}^{(n)} = \frac{10}{120} = 83.3 \text{ mW, con } n = 1, 2, \ldots, 120.$$

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

$$G_s^{(120)}\Big|_{dB} = 10\log_{10}\frac{W_{dT}^{(120)}}{W_{R_{Min}}} = 10\log_{10}\frac{83.3}{1} = 19.2 \text{ dB},$$

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

$$A_{d_{Tot}}^{(120)}\Big|_{dB} = A_d^{(120)}\,(dB/Km) \cdot L_{Km} = 19.2 \text{ dB},$$

da cui si ricava per la massima lunghezza

$$L_{Km} = \frac{A_{d_{Tot}}^{(120)}\big|_{dB}}{A_d^{(120)}(dB/Km)} = \frac{19.2}{6.46} = 2.97 \text{ Km},$$

che come vediamo è imposta dal canale più attenuato.

Per il primo canale, invece, si ha $A_d^{(1)}\,(dB/Km) = A_0$, e dunque

$$A_{d_{Tot}}^{(1)}\Big|_{dB} = A_0\,(dB/Km) \cdot L_{Km} = 5.3 \cdot 2.97 = 15.74 \text{ dB}.$$

La differenza tra $G_s|_{dB}$ (uguale per tutti i canali) e $A_{d_{Tot}}^{(1)}\Big|_{dB}$ rappresenta il margine di sistema per il primo canale, pari a

$$M = G_s - A_d = 19.2 - 15.74 = 3.46 \text{ dB}.$$

La stessa quantità, è anche uguale alla *distorsione lineare di ampiezza* del cavo nella banda del segnale.

Nel caso in cui si voglia superare una lunghezza doppia, anche $A_{d_{Tot}}^{(120)}\Big|_{dB}$ raddoppia, e per mantenere $W_{R_{Min}} = 1$ mW, deve raddoppiare anche $G_s^{(120)}\Big|_{dB}$. Pertanto la nuova potenza/canale risulta

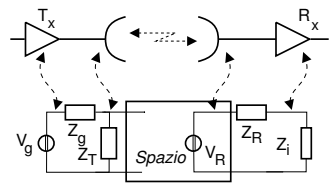
$$W_{dT}^{'}\,(dBm) = W_{R_{min}}\,(dBm) + G_s^{'}\,(dB) = 0 + 2G_s\,(dB); \text{ quindi}$$

$$W_{dT}^{'}\,(mW) = 10^{\frac{W_{dT}^{'}(dBm)}{10}} = 10^{\frac{2G_s(dB)}{10}} = 10^{\frac{2\cdot 19.2}{10}} = 10^{3.84} = 6918.3$$

milliWatt, cioè 6.91 Watt/canale, e quindi $6.91 \cdot 120 = 830$ Watt complessivi !

## 15.3 Collegamenti radio

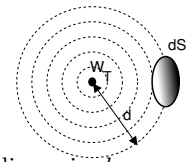
I segnali modulati occupano in genere una banda molto stretta attorno alla portante, tanto da poterli assimilare ad una singola sinusoide (vedi § 14.5.4.1). Pertanto, le condizioni di massimo trasferimento di potenza tra amplificatore finale e antenna trasmittente ($Z_g = Z_T^*$) e tra antenna ricevente e stadio di ingresso al ricevitore ($Z_R = Z_i^*$) danno luogo, nella banda di segnale, ad una componente di distorsione lineare che non dipende dalla




frequenza (modulo e fase costante, vedi § 14.5.4.1), e questo corrisponde (a parte una rotazione di fase) all'assenza di distorsioni lineari. Tutta la potenza disponibile fornita dall'amplificatore finale $W_{dT} = \frac{V_{Teff}^2}{4R_g}$ viene ceduta all'antenna, e da questa allo spazio. In realtà $Z_T$ dipende dalla geometria dello spazio circostante; perciò l'amplificatore del trasmettitore Tx va *accordato* dopo aver posizionato l'antenna.

### 15.3.1 Trasduzione elettromagnetica

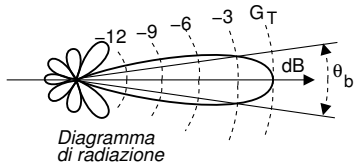
**Antenna isotropa** Se l'antenna trasmittente irradia allo stesso modo in tutte le direzioni, $W_{dT}$ si distribuisce su di una sfera; dunque una superficie $dS$, posta a distanza $d$, è attraversata da una potenza pari a

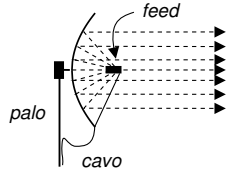



$$dW = W_{dT}\frac{dS}{4\pi d^2} \quad \text{[Watt]} \qquad (15.6)$$

Si noti che il denominatore rappresenta la superficie di una sfera di raggio $d$.

**Antenna direttiva** Sono antenne che hanno direzioni privilegiate di emissione. Ad esempio, le antenne paraboliche dispongono di un *illuminatore* o FEED[9] posto in corrispondenza del *fuoco* della parabola stessa, la cui superficie riflette le onde elettromagnetiche in modo che si propaghino in forma pressoché parallela[10].




$W_{dT}$ si distribuisce quindi *non* sfericamente, e la direzione di propagazione massima esibisce un guadagno $G_T$ rispetto all'antenna isotropa, mentre l'intensità di campo irradiato spazialmente è descritta da un *diagramma di radiazione*. Il valore di $G_T$ dipende dal rapporto tra le dimensioni dell'antenna e quelle della lunghezza d'onda $\lambda$ secondo la relazione

$$G_T = 4\pi\frac{A}{\lambda^2} \qquad (15.7)$$

avendo indicato con $A$ l'area dell'antenna.

Può essere definita una *larghezza del fascio* (BEAM WIDTH), che misura l'angolo $\theta_b$ entro cui la potenza irradiata è superiore alla metà della massima potenza presente nella direzione privilegiata[11]. Ovviamente minore è $\theta_b$, e maggiore è $G_T$.

[9] Dall'inglese *to feed* = alimentare.

[10] Il processo di focalizzazione parabolica, comunemente usato ad esempio nei *fanali* degli autoveicoli, era ben noto ad un certo siracusano...

[11] Si tratta di un concetto del tutto analogo alla "frequenza di taglio a 3 dB", ma applicata ad un dominio spaziale con geometria radiale.

**Antenna ricevente** Se una antenna identica a quella trasmittente viene usata (dall'altro lato del collegamento) per ricevere, questa mantiene lo stesso guadagno $G_R = G_T$ e lo stesso $\theta_b$. Si definisce allora la sua *area efficace* come il valore

$$A_e = G_R \frac{\lambda^2}{4\pi} \tag{15.8}$$

legato alla forma e dimensione dell'antenna, a meno di un fattore di efficienza $\rho$ ([12]). Perciò una stessa antenna ($A_e$ fisso) aumenta il suo guadagno (e stringe il *beam*) all'aumentare della frequenza, ovvero al diminuire di $\lambda = \frac{c}{f}$ ([13]).

### 15.3.2 Bilancio energetico

**Potenza ricevuta** Usando l'area efficace dell'antenna ricevente (15.8) per intercettare parte della potenza irradiata (15.6), 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

$$\left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4\pi df}{c}\right)^2$$

è chiamato *attenuazione di spazio libero*, che dipende da $f^2$. In realtà ai fini del bilancio energetico, la dipendenza dalla frequenza si elide con quella relativa al guadagno delle antenne: $G_T = A_e \frac{4\pi}{\lambda^2} = A_e \frac{4\pi f^2}{c^2}$ ([14]).

**Attenuazione disponibile** Il rapporto

$$A_d = \frac{W_{dT}}{W_{dR}} = \left(\frac{4\pi df}{c}\right)^2 \frac{1}{G_T G_R} \tag{15.9}$$

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

$$A_d\,(\mathsf{dB}) = 32.4 + 20\log_{10} f\,(\mathsf{MHz}) + 20\log_{10} d\,(\mathsf{Km}) - G_T\,(\mathsf{dB}) - G_R\,(\mathsf{dB}) \tag{15.10}$$

[12] 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 (15.7), a meno di non usare appunto il valore di area efficace.

[13] 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.

[14] Mantenendo fissa la dimensione delle antenne, si ottiene il risultato che trasmissioni operanti a frequenze più elevate permettono di risparmiare potenza. Purtroppo però, guadagni superiori a 30-40 dB (corrispondenti a piccoli valori di $\theta_b$) sono controproducenti, per i motivi esposti al §15.3.3.1.

nota come *equazione di Friis*. Osserviamo che, a differenza della trasmissione in cavo, l'attenuazione cresce con il quadrato della distanza, e quindi con il suo logaritmo quando espressa in decibel. Infatti ora l'attenuazione è dovuta esclusivamente all'aumentare della superficie su cui si distribuisce la potenza irradiata, e non a fenomeni dissipativi, come accade invece per cavo e fibra ottica.

Come vedremo tra breve, i collegamenti radio terrestri, casalinghi, e mobili, sono affetti da una serie di fenomeni tali che la (15.10) si limita a rappresentare solo un particolare aspetto del problema. Per contro, il sistema di telecomunicazione che meglio rappresenta le condizioni di spazio libero è quello tra terra ed satellite, per il semplice fatto che non vi sono frapposti ostacoli, e che approfondiamo brevemente alla appendice 15.5.2.

### 15.3.3 Condizioni di propagazione e attenuazioni supplementari

In virtù di molteplici fenomeni, il calcolo della potenza ricevuta non è così banale come sembra a prima vista. In particolare, al valore $A_d$ (dB) devono essere sommate (in decibel) tutte le attenuazioni *in più*:

**Perdite di accoppiamento** Dovute al mancato verificarsi delle condizioni di massimo trasferimento di potenza: ammontano a qualche dB.

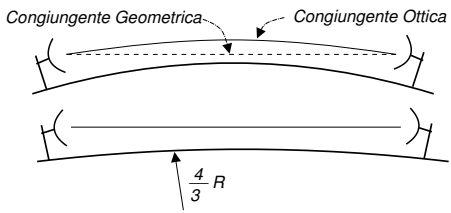
**Assorbimento terrestre** Quando l'antenna è distante dal suolo meno di qualche lunghezza d'onda, l'energia si propaga per onda superficiale, ovvero la crosta terrestre fa da conduttore. Questa forma di propagazione provoca una attenuazione supplementare che aumenta con la frequenza, tanto che già a 3 MHz raggiunge i 25 dB ogni 10 Km (equivalente ad una riduzione di potenza di $10^{2.5} = 316$ volte). Le *onde medie* (0,3-3 MHz) sono meno attenuate, ed ancora meno le *onde lunghe* (10-300 KHz) che viaggiano appunto via terra.

#### 15.3.3.1 Condizioni di visibilità

Come ricavabile anche dall'espressione dell'area efficace, all'aumentare della frequenza si possono ottenere antenne di dimensioni ridotte e contemporaneamente di elevato guadagno. Allo stesso tempo, per evitare l'assorbimento terrestre, occorre posizionare l'antenna in alto (in cima ad una torre), e trasmettere per *onda diretta*, condizione nota anche come LOS o *line of sight*.

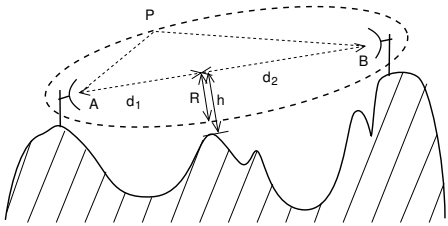
A causa della curvatura terrestre, esiste una altezza minima da rispettare: ad esempio con torri da 60 metri si raggiungono distanze (in visibilità) di 50 Km. Ovviamente, il problema si presenta in pianura. Tratte più lunghe richiedono torri più alte, ma anche guadagni di antenna maggiori (e quindi antenne più grandi e più direttive). Questa non è però una soluzione molto praticabile, in quanto in presenza di vento forte le antenne "grandi" possono spostarsi e perdere il puntamento; inoltre, il costo delle torri aumenta esponenzialmente con l'altezza.

**Orizzonte radio** Nel calcolare l'altezza delle torri (ed il puntamento delle antenne) si deve considerare anche il fenomeno legato al fatto che l'onda elettromagnetica, propagandosi, *si piega* verso gli strati dell'atmosfera con indici di rifrazione maggiori (ossia verso terra). Pertanto, i calcoli vengono effettuati supponendo che il raggio terrestre sia




4/3 volte quello reale. Inoltre, l'indice di rifrazione (che aumenta verso il basso) può variare con l'ora e con le condizioni climatiche: pertanto, anche in questo caso, le antenne con guadagni elevati (e molto direttive) possono andare fuori puntamento.

**Ellissoidi di Fresnel** Nella propagazione elettromagnetica occorre tenere conto dei fenomeni di diffrazione, che *deviano* nella zona in *ombra*[15] le onde radio che transitano in prossimità di ostacoli. Pertanto, la determinazione dell'orizzonte radio deve 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 definito come il luogo dei punti $P$ per i quali la somma delle distanze $d(A,P) + d(P,B)$ è pari a $d(A,B) + \frac{\lambda}{2}$, in cui $\lambda = \frac{c}{f}$ è la lunghezza d'onda della trasmissione a frequenza $f$.



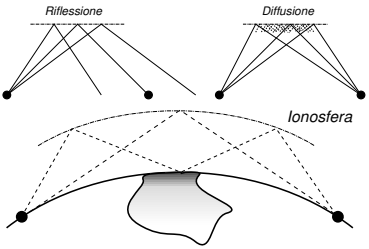

Suddividendo la distanza $d(A,B)$ tra i due fuochi $A$ e $B$ in due segmenti $d_1$ e $d_2$, individuati dalla posizione dell'ostacolo, si trova che il raggio dell'ellissoide è pari a

$$R = \sqrt{\frac{\lambda}{\frac{1}{d_1} + \frac{1}{d_2}}}$$

che, nel caso $d_1 = d_2 = \frac{d(A,B)}{2}$, assume il valore massimo $R_M = \frac{1}{2}\sqrt{\lambda d}$. Qualora si determini la condizione $h < R$, il collegamento subisce una attenuazione supplementare che aumenta al diminuire di $\frac{h}{R}$, ed è maggiore per gli *spigoli vivi*, fino ad arrivare ad una decina di dB.

#### 15.3.3.2 Diffusione e riflessione atmosferica

Tra 0,1 e 10 GHz si può verificare il fenomeno della *diffusione troposferica* (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* (la fascia oltre gli 80 Km), dove strati ionizzati causano *riflessioni* del segnale, e consentono la trasmissione anche tra luoghi non in visibilità[16], ma con il rischio di cammini multipli. E' questo il caso tipico delle *onde corte*.

Per frequenze sotto il MHz la propagazione è per *onda di terra,* e l'assorbimento terrestre impedisce di coprire grandi distanze (tranne che per le *onde lunghe,* meno

[15] 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.

[16] Anche, ma non solo, in concorso con la riflessione operata da masse d'acqua, come mostrato in figura.

attenuate). Anche in questo caso può verificarsi la diffusione troposferica, specie *di notte*.

#### 15.3.3.3 Assorbimento atmosferico

Per lunghezze d'onda di dimensione comparabile a quella delle molecole di ossigeno, si produce un fenomeno dissipativo *di assorbimento*; le frequenze interessate sono quelle superiori a 30 GHz, con un massimo di 20 dB/Km a 60 GHz([17]). Inoltre, il vapor d'acqua (con molecole di dimensioni maggiori) produce una attenuazione supplementare di 1-2 dB/Km (al massimo) a 22 GHz[18]. Sotto i 10 GHz non si verifica né assorbimento da ossigeno, né da vapore.

#### 15.3.3.4 Dimensionamento di un collegamento soggetto a pioggia

In caso di pioggia, si manifesta una ulteriore causa di assorbimento atmosferico, detto appunto *da pioggia*, che costituisce la principale fonte di attenuazione supplementare per frequenze superiori a 10 GHz. L'attenuazione supplementare da pioggia aumenta con la frequenza portante, con l'intensità di precipitazione e con l'estensione della zona piovosa lungo il tragitto radio; questi ultimi due fattori sono evidentemente elementi aleatori, e per questo il dimensionamento mira a stabilire quale sia il margine necessario a garantire un grado di servizio prefissato. Il margine necessario, è pertanto pari al valore di attenuazione supplementare che viene superata con una probabilità minore o uguale al grado di servizio.

Una formula sperimentale che consente di determinare il valore in dB dell'attenuazione supplementare che viene superata con probabilità $p$ è:

$$A_s\left(r_0,\, d,\, p\right) = K \cdot r_0^{\alpha} \cdot d \cdot \beta\left(d\right) \cdot \gamma\left(p\right) \quad \text{[dB]}$$

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 $\alpha$ sono costanti che caratterizzano l'entità dell'interazione dell'onda radio con la pioggia, in funzione della frequenza portante e di altre condizioni climatiche ed ambientali, i cui valori medi sono riportati nella tabella che segue.

| $f_0$(GHz) | 10 | 15 | 20 | 25 | 30 | 35 |
|---|---|---|---|---|---|---|
| $\alpha$ | 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 $\gamma\left(p\right) = 6.534 \cdot 10^{-3} \cdot p^{-(.718+.043 \cdot \log_{10} p)}$, che vale 1 per $p = 10^{-4}$, permette di tener conto del grado di servizio che si vuole ottenere. Infine, $\beta\left(d\right) = 1/\left(1 + .0286 \cdot d\right)$ è un fattore correttivo che tiene conto del fatto che *non piove lungo tutto* il collegamento. I grafici in fig. 15.1, mostrano l'andamento del termine $K \cdot r_0^{\alpha} \cdot d \cdot \beta\left(d\right)$ per diversi valori di $f_0$ ed $r_0$, in funzione dell'estensione del collegamento; infine, è riportato il grafico della funzione $\gamma\left(p\right)$ 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

[17] L'elevata attenuazione chilometrica presente a 60 GHz può essere sfruttata nei sistemi di comunicazione cellulare, allo scopo di riusare una stessa banda di frequenze anche a breve distanza.

[18] L'assorbimento di potenza da parte delle molecole d'acqua per onde eletromagnetiche a 22 GHz è il principio su cui si basa il forno a microonde.

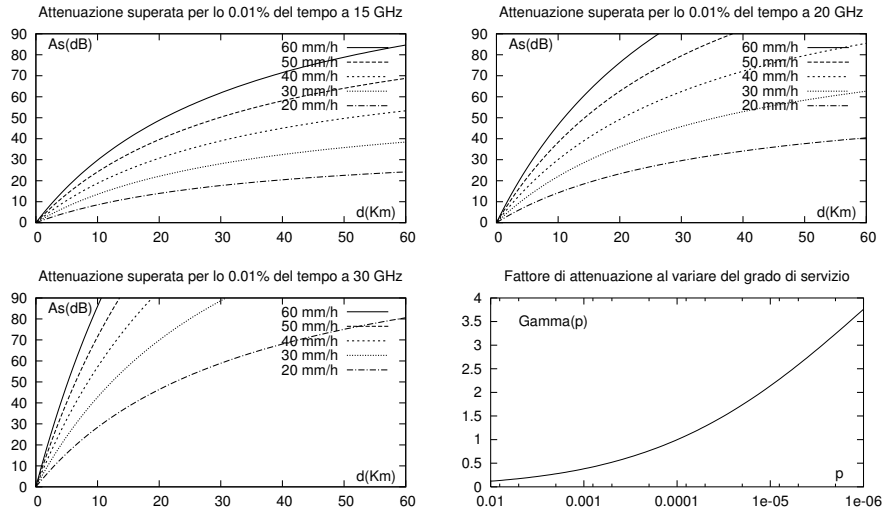



Figura 15.1: Curve di attenuazione supplementare per pioggia

a causa dell'eccesso di potenza ricevuta, qualora non siano presenti le attenuazioni supplementari: può essere allora utilizzato un canale di ritorno nell'altra direzione, in modo da regolare la potenza del trasmettitore.

#### 15.3.3.5 Cammini multipli

Oltre i 30 MHz, nonostante la direttività delle antenne, alcuni raggi obliqui possono incontrare superfici riflettenti (laghi o masse d'acqua), oppure brusche variazioni dell'indice di rifrazione, che causano la riflessione totale del raggio, e la ricezione di una 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

$$h(t) = \sum_{n=1}^{N} a_n \delta(t - T_n) \tag{15.11}$$

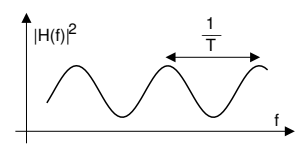
in cui i valori $T_k$ sono i ritardi con cui si presentano le diverse eco, ognuna caratterizzata da una ampiezza $a_k$, in accordo allo schema di filtro trasversale presentato al § 9.7.

La corrispondente risposta in frequenza

$$H(f) = \sum_{n=1}^{N} a_n e^{-j2\pi f T_n}$$

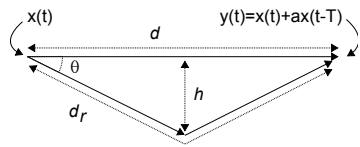
può produrre delle distorsioni lineari; ricordiamo infatti (vedi § 9.7.1) che il modello di cammino multiplo per una sola eco con ritardo $T$, presenta una

$$|H(f)|^2 = 1 + a^2 + 2a\cos 2\pi f T$$



periodica in frequenza con periodo $f = \frac{1}{T}$, ed osserviamo che all'aumentare di $T$, le oscillazioni di $|H(f)|^2$ si infittiscono[19], e dunque aumenta la possibilità che $|H(f)|^2$ vari di molto nella banda del segnale, causando distorsioni lineari che devono essere equalizzate.

**Esempio** Consideriamo la geometria descritta in figura, in cui un collegamento di portata $d$ subisce un fenomeno di riflessione a metà della sua lunghezza, da parte di una superficie riflettente che dista $h$ dalla congiungente, e ricaviamo l'espressione del ritardo $T$. Ricordando che $tempo = \frac{spazio}{velocita'}$ e

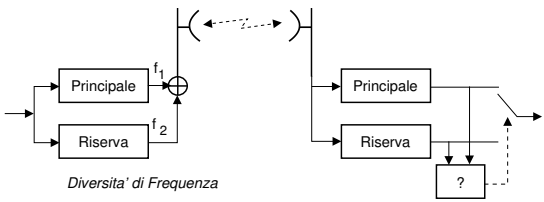



indicando con $d_r$ la distanza percorsa dall'onda riflessa, otteniamo che la differenza tra i tempi di arrivo dell'onda diretta e riflessa vale $T = \frac{1}{c}(d_r - d)$; inoltre, dalla trigonometria risulta che $\frac{d}{2} = \frac{d_r}{2}\cos\theta$. Combinando le due relazioni, otteniamo che $T = \frac{d}{c}\left(\frac{1}{\cos\theta} - 1\right)$, in cui $\theta = \arctan\frac{h}{d/2} = \arctan 2\frac{h}{d}$. Attualizzando il risultato ad uno scenario in cui $d = 1$ Km ed $h = 100$ metri, si ottiene $\theta = 11^o 31'$, $\cos\theta = 0.98$, e $T = 3,3$ $\mu$secondi.

Qualora la banda del segnale sia invece sufficientemente piccola rispetto a $\frac{1}{T}$, e si possa considerare $|H(f)|^2$ costante, la presenza di cammini multipli può comunque dar luogo ad attenuazione, che prende il nome di *flat fading*[20] .

### 15.3.3.6 Collegamenti in diversità

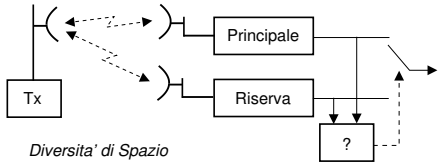
Il *fading* prodotto da cammini multipli può essere rilevante, specialmente se due repliche del segnale giungono al ricevitore con ampiezze molto simili. Il problema può essere risolto prevedendo una ridondanza degli apparati.




**Diversità di frequenza** Lo stesso collegamento è operato su due diverse portanti; nel caso in cui la trasmissione operata mediante una delle portanti subisca attenuazione, quella che utilizza l'altra portante ne è probabilmente esente (o viceversa).

Se la banda è particolarmente affollata, la stessa configurazione può essere adottata per fornire una ridondanza $N : 1$. Ad esempio, in una trasmissione multiplata FDM, la portante di riserva viene assegnata al canale del banco FDM che presenta la maggiore attenuazione.

**Diversità di spazio** Adottando due diverse antenne riceventi in posizioni differenti, la differenza di percorso $T$ tra cammini multipli è differente per le due




[19] Ad esempio, desiderando $\frac{1}{T} > 1$ MHz, si ottiene $T_{Max} = 1$ $\mu$sec; se l'onda radio si propaga alla velocità $c = 3 \cdot 10^8$ m/sec, la massima differenza di percorso vale $\Delta_{max} = c \cdot T_{Max} = 3 \cdot 10^8 \cdot 10^{-6} = 300$ metri.

[20] Il termine *fading* si traduce come *affievolimento*, ma è spesso usato in inglese, cosicchè le distorsioni lineari per segnali a banda stretta sono dette *condizioni di fading piatto.*

antenne, e dunque la risposta in frequenza $|H(f)|^2 = 1+a^2+2a\cos 2\pi fT$ ha una diversa periodicità nei due casi. Pertanto, anche se un ricevitore subisce una attenuazione selettiva, l'altro ricevitore ne è esente.

### 15.3.4 Collegamenti radiomobili

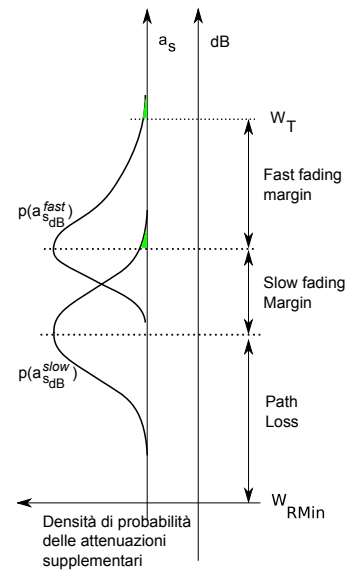
Le condizioni di propagazione per terminali radiomobili, come nel caso della telefonia cellulare, presentano diversi aspetti particolari. Innanzitutto, l'antenna del terminale mobile è molto vicina al suolo, e ciò comporta la presenza di una eco fissa da terra, quasi sempre il mancato rispetto delle condizioni di Fresnel[21], ed una attenuazione supplementare da assorbimento terrestre. Inoltre, specialmente in ambito urbano, si verifica un elevato numero di cammini multipli e diffrazioni, che perdipiù variano nel tempo in conseguenza dello spostamento del terminale. Infine, l'uso condiviso di una stessa banda di frequenze radio da parte di una moltitudine di terminali, determina la necessità di riusare le stesse frequenze in regioni differenti, e l'attuazione di meccanismi di codifica di canale per ridurre gli effetti delle interferenze e del fading variabile[22]. Analizziamo brevemente i primi due fenomeni, fornendo modelli matematici per tenere conto delle conseguenze delle attenuazioni supplementari e dei fenomeni di multipath variabile, rimandando la discussione sulle tecniche di accesso multiplo ad una prossima edizione.

#### 15.3.4.1 Determinazione del margine

La figura a lato mostra come viene determinato il margine di sistema per un collegamento radiomobile, mettendo in evidenza i contributi discussi nel seguito. Oltre ad una componente di attenuazione *nominale* indicata come *path loss*, occorre considerare due componenti aleatorie di attenuazione supplementare, indicate come *slow* e *fast fading*.

Il valore del *path loss* risulta maggiore di quello di $A_d$ (eq. 15.9) a causa delle condizioni di propagazione non ideali, e viene analizzato al § 15.3.4.2.

Lo *slow fading* tiene conto dei fenomeni lentamente variabili nel tempo, come l'ombreggiatura (*shadowing*) legata alla frapposizione di rilievi, edifici, alberi ed oggetti. L'attenuazione supplementare conseguente $a_{s_{dB}}^{slow}$ non varia di molto con il movimento del ricevitore, ed al § 15.3.4.3 si mostra come il suo valore sia una v.a. gaussiana in dB con varianza $\sigma_{SF}^2$, consentendo di determinare lo *slow*




[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] Mentre il fading produce una attenuazione variabile al segnale, la stessa variabilità delle condizioni di propagazione può portare a livelli interferenza variabili, causati da altre trasmissioni nella stessa banda. La variabilità temporale della qualità del segnale ricevuto, in particolare quella *veloce* (vedi appresso), produce errori a *burst*, che possono essere corretti mediante codifica di canale ed interleaving (vedi § 5.3.3.1).

*fading margin* $M_{dB}^{slow}$ come quel valore di $a_{s_{dB}}^{slow}$ che viene superato con probabilità sufficentemente bassa.

Il *fast fading* tiene invece conto degli innumerevoli cammini multipli presenti in ambito urbano ed *indoor*, che possono produrre una attenuazione supplementare maggiore del caso precedente, e variare rapidamente a seguito di spostamenti anche modesti. Al § 15.3.4.4 si fornisce una descrizione statistica anche per questo caso, permettendo anche ora di determinare un margine $M_{dB}^{fast}$ tale da rendere trascurabile la probabilità che $a_{s_{dB}}^{fast}$ ecceda il suo valor medio per più di $M_{dB}^{fast}$. Il margine complessivo viene quindi posto pari alla somma dei due margini, come mostrato in figura.

#### 15.3.4.2 Path loss

La dipendenza della attenuazione dal quadrato della distanza presente in (15.9), si riferisce al caso ideale di spazio libero; campagne di misura evidenziano invece che l'esponente di $d$ aumenta fino al valore 4, a seconda del tipo di ambiente (urbano, rurale) e dell'altezza dell'antenna ricevente. Pertanto, il termine $20\log_{10} d\,(Km)$ che compare in (15.10) deve essere sostituito con $n \cdot 10\log_{10} d\,(Km) + \alpha$, e quindi ora l'espressione da usare è

$$A_d\,(\mathsf{dB}) = 32.4 + 20\log_{10} f\,(\mathsf{MHz}) + n \cdot 10\log_{10} d\,(Km) + \alpha - G_T\,(\mathsf{dB}) - G_R\,(\mathsf{dB})$$

in cui $n$ ed $\alpha$ 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 $\alpha$ può variare da 7 a 15 dB con antenna fissa alta 30 e 10 metri rispettivamente, e subire un incremento di quasi 30 dB passando da un ambiente aperto ad un ambito urbano.

#### 15.3.4.3 Slow fading

La stima delle grandezze $n$ ed $\alpha$ ora introdotte è svolta *mediando* i risultati di diverse misure condotte nel territorio che si intende caratterizzare: infatti per territori diversi, si riscontrano valori di attenuazione complessiva diversi, anche per uguali valori di $d$. Questo fenomeno è indicato come *slow fading*, poiché non si presenta muovendosi di poco in una stessa zona, in quanto dipende dalla orografia del territorio e dalla natura degli oggetti limitrofi. Non conoscendo a priori in che zona ci si trovi, l'effetto dello *slow fading* (SF) si manifesta come una attenuazione supplementare $a_s$ aleatoria, che risulta avere un andamento gaussiano in dB[23] (per questo detto *lognormale*) e cioè del tipo

$$p_{A_s}\,(a_s(dB)) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(a_s(dB))^2}{2\sigma_{SF}^2}}$$

[23]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. 157) $\underline{x} = a + \underline{r}$, 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 $a \gg |\underline{r}|$

dove $\sigma^2_{SF}$ varia tra 6 ed 8 dB per elevazioni dell'antenna tra 5 e 15 metri[24]. Per velocità del mobile non superiori ai 15 Km/h, si può assumere $a_s$ costante in frequenza per qualche MHz, e nel tempo per poche centinaia di millisecondi.

**Esempio** Un collegamento radio tra antenne omnidirezionali poste a $d = 20$ KM e con portante $f_0 = 27$ MHz è usato da una trasmissione per cui occorre ricevere una potenza di almeno $W_R = -50$ dBm. Determinare la potenza $W_T^{slib}$ che occorre trasmettere in condizioni di *spazio libero*, e la nuova potenza $W_T^{sfad}$ necessaria a garantire una probabilità di fuori servizio pari al 5%, in presenza di una attenuazione supplementare di *slow fading* caratterizzata da $\sigma^2_{SF} = 8$ dB. Utilizziamo la (15.10) per calcolare

$$\begin{aligned} A_d\,(\mathsf{dB}) &= 32.4 + 20\log_{10} f\,(\mathsf{MHz}) + 20\log_{10} d\,(\mathsf{Km}) - G_T\,(\mathsf{dB}) - G_R\,(\mathsf{dB}) = \\ &= 32.4 + 20\log_{10} 27 + 20\log_{10} 20 = 32.4 + 28.6 + 26 = 87\,\mathsf{dB} \end{aligned}$$

da cui si ottiene

$$W_T^{slib}(dBm) = W_R(dBm) + A_d(dB) = -50 + 87 = 37\,\mathsf{dBm}$$

pari a 7 dBW ovvero $10^{0.7} = 5$ Watt. Lo slow fading 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. 140 per cui $0.05 = \frac{1}{2}\mathrm{erfc}\left(\frac{a_s(dB)}{\sqrt{2}\sigma_{SF}}\right)$, e quindi graficamente si ottiene $\frac{a_s(dB)}{\sqrt{2}\sigma_{SF}} = 1.5$, da cui $a_s(dB) = 1.5 \cdot \sqrt{2} \cdot \sqrt{8} = 1.5 \cdot 1.41 \cdot 2.82 = 6$ dB, che rappresenta il margine cercato, e che ci consente di calcolare la nuova $W_T^{sfad}$ come $W_T^{sfad}(dBW) = W_T^{slib}(dBW) + 6dB = 13$ dBW, ovvero $10^{1.3} = 20$ Watt.

### 15.3.4.4 Fast fading

Rappresenta le rapide e profonde fluttuazioni nel livello del segnale radio, osservate durante *il movimento*. Queste fluttuazioni sono causate dalle variazioni dei ritardi di fase con cui i cammini multipli giungono al ricevitore: spostandosi infatti di $\frac{\lambda}{2}$([25]) si può passare da una situazione di somma coerente ad una completa opposizione di fase.

Qualora la ricezione avvenga principalmente in *assenza di visibilità*, ed in presenza di un numero elevato di cammini multipli, i valori del modulo dell'inviluppo complesso del segnale $\rho\,(t) = |\underline{x}\,(t)|$ che giunge ad un ricevitore in movimento sono adeguatamente rappresentati da una v.a. di Rayleigh (vedi appendice 15.5.1), la cui d.d.p. ha

---

possiamo scrivere

$$\begin{aligned} a_s(dB) &= 10\log_{10}\frac{1}{|a+\underline{r}|^2} = -10\log_{10}\left((a+r_c)^2 + r_s^2\right) = \\ &= -10\log_{10}\frac{a^2}{a^2}\left(a^2 + 2ar_c + r_c^2 + r_s^2\right) = \\ &= 10\left(\log_{10} a^2 + \log_{10}\left(1 + \frac{2r_c}{a} + \frac{|\underline{r}|^2}{a^2}\right)\right) = \\ &\simeq 10\left(\log_{10} a^2 + \frac{2r_c}{a}\right) = 10\log_{10} a^2 + 20\frac{r_c}{a} \end{aligned}$$

in quanto $\log(1+\alpha) \simeq \alpha$ con $\alpha \ll 1$, e quindi $a_s\,(dB)$ ha media $10\log_{10} a^2$ (compresa nel *path loss*) ed esibisce una d.d.p. gaussiana, la stessa di $r_c$.

[24] All'aumentare dell'altezza dell'antenna, si estende l'area di copertura della stessa, ma in ambito urbano questo corrisponde ad una maggiore variabilità delle effettive condizioni operative.

[25] A frequenza di 1 Ghz, si ha $\lambda \simeq 30$ cm. Questo fenomeno può essere facilmente sperimentato tentando di sintonizzare un televisore dotato di *antenna interna*, ed osservando come la qualità del segnale vari sensibilmente anche per piccoli spostamenti del ricevitore; ma un esempio più attuale, è la *ricerca del campo* per poter telefonare.

espressione

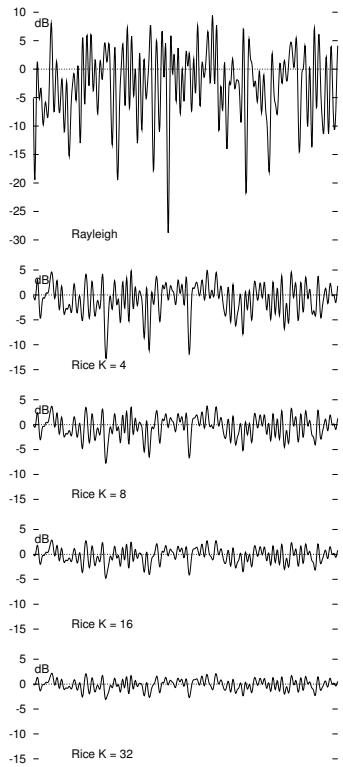
$$p_P(\rho) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$$

con $\rho \geq 0$, avendo indicato con $\sigma^2$ la potenza delle componenti in fase e quadratura di $\underline{x}(t)$[26]. Sotto tali ipotesi, la potenza istantanea $s$ ricevuta, legata a $\rho^2$, ha d.d.p. esponenziale negativa[27]:

$$p_S\left(s = \rho^2\right) = \frac{1}{2\sigma^2} \exp\left(-\frac{s}{2\sigma^2}\right)$$

e quindi (vedi § 8.2.1) presenta frequentemente valori molto bassi e prossimi a zero, che in dB danno luogo a *profonde* attenuazioni (*deep fades*).

Alla figura a lato è mostrato l'andamento in dB del livello di segnale ricevuto, relativo alle condizioni di ricezione *medie*, nei casi di *fading di Rayleigh* oppure *di Rice*[28] per diversi valori del *fattore di Rice* $K$ definito come il rapporto $K = \frac{a^2}{2\sigma^2}$ tra la potenza $\frac{a^2}{2}$ dell'onda diretta e quella $\sigma^2$ della componente dovuta al multipath. Come si può notare, il fading di Rayleigh produce attenuazioni più profonde, mentre in presenza di una forte componente diretta, l'ampiezza del fading si riduce sensibilmente.




#### 15.3.4.5 Dimensione di cella e velocità di trasmissione

L'analisi svolta finora è relativa al caso di *fading piatto*, ovvero il caso in cui l'effetto dei cammini multipli si riduce ad una $H(f) = ae^{j\varphi}$ costante complessa per tutti i valori di $f$, in conseguenza della ipotesi introdotta al § 15.5.1 che il segnale non vari di molto nell'intervallo temporale $\Delta\tau = \tau_{max} - \tau_{min}$ tra l'arrivo della prima e dell'ultima replica, detto anche *dispersione temporale*. Quest'ultima posizione consente di definire *banda di coerenza*[29] il valore

$$\Delta f_c = \frac{1}{\Delta\tau}$$

[26] Le componenti in fase e quadratura $x_c(t)$ e $x_s(t)$ sono costituite dalla somma di quelle di tutti i cammini multipli, e per questo motivo possono essere assunte di tipo gaussiano, a media nulla e varianza $\sigma^2$.

[27] Impostando il cambiamento di variabile $s = \rho^2$, si possono applicare le regole viste al § 7.6.4, 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:

$$p_S(s) = p_P\left(\sqrt{s}\right) \cdot \frac{d}{ds}\rho(s) = \frac{\sqrt{s}}{\sigma^2} \exp\left(-\frac{\left(\sqrt{s}\right)^2}{2\sigma^2}\right) \cdot \frac{1}{2\sqrt{s}} = \frac{1}{2\sigma^2} \exp\left(-\frac{s}{2\sigma^2}\right)$$

[28] Come anticipato alla nota 23, il fading è detto *di Rice* quando uno dei cammini multipli prevale sugli altri.

[29] Vedi ad es. `http://www.tlc.polito.it/~perotti/it/tcr/book.pdf`

che costituisce un riferimento contro cui confrontare la banda $W$ del segnale $x(t)$ da trasmettere. Se $W < \Delta f_c$ siamo appunto nelle condizioni di fading *piatto*, mentre se $W > \Delta f_c$ le componenti spettrali di $x(t)$ subiscono alterazioni statisticamente indipendenti, i cammini multipli causano un effetto filtrante, ed il canale corrispondente viene detto *selettivo in frequenza*, poiché alcune frequenze sono più attenuate di altre. Inoltre, 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 $W < \Delta f_c$ implica anche che $T_s \simeq \frac{1}{W} > \frac{1}{\Delta f_c} = \Delta\tau$, ossia che la dispersione temporale risulta inferiore al periodo di simbolo, limitando gli effetti dell'ISI. Dato che la correzione degli effetti di distrosione lineare e ISI causate dai canali selettivi in frequenza richiede al ricevitore complesse operazioni di equalizzazione, si preferisce per quanto possibile operare in condizioni di fading piatto, quindi occupare una banda $W < \Delta f_c$, e limitare conseguentemente la velocità di segnalazione $f_s$.

Per celle molto grandi, la differenza di percorso tra cammini multipli può essere notevole, determinando una banda di coerenza ridotta, e quindi una bassa velocità di trasmissione. Riducendo la dimensione di cella, è possibile invece aumentare la velocità, dato che le differenze di ritardo si riducono. Per questo motivo, se celle con raggio di chilometri e dispersioni di ritardo di oltre 10 $\mu$sec necessitano di equalizzazione anche per trasmissioni a 64 kbps, comunicazioni *indoor* con dispersioni di ritardo inferiori ad 1 $\mu$sec possono presentare *flat fading* per velocità superiori al Mbps. Celle di dimensione minima, dette anche *picocelle*, presentano dispersioni temporali di solo qualche decina di picosecondi, permettendo di operare a molti Mbps anche senza equalizzazione.

## 15.4 Collegamenti in fibra ottica

Una fibra ottica è realizzata in vetro o silicio fuso, ovvero qualunque materiale dielettrico trasparente alla luce, tanto che può essere realizzata anche in plastica. Il suo utilizzo è quello di trasportare energia luminosa in modo guidato. Una caratteristica che deriva direttamente dalla sua natura, è l'immunità della fibra ottica ai disturbi di natura elettromagnetica; tale proprietà impedisce fenomeni di interferenza (diafonia), così come non permette di prelevare segnale dall'esterno (intercettazione).

**Il segnale luminoso** Le lunghezze d'onda delle radiazioni elettromagnetiche nel campo del visibile sono comprese tra 50 nm (1 nm=$10^{-9}$ metri) dell'ultravioletto fino a circa 100 $\mu$m dell'infrarosso, che corrispondono a frequenze (ricordando ancora che $f = \frac{c}{\lambda}$) che vanno da $6 \cdot 10^{15}$ Hz a $3 \cdot 10^{12}$ 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 portante. Se $f_0 = 1$ GHz, si ha 1 MHz di banda; ma se $f_0 = 10^{14}$, si ha una banda di 100 GHz!

| Ultravioletto | → | Infrarosso | |
|---|---|---|---|
| $50 \cdot 10^{-9}$ | → | $10^{-4}$ | $\lambda$ [metri] |
| $6 \cdot 10^{15}$ | ← | $3 \cdot 10^{12}$ | $f$ [Hz] |

### 15.4.1 Trasmissione ottica

Anche se sono teoricamente possibili schemi di modulazione analogici, le fibre ottiche sono usate prevalentemente per trasportare informazione di natura *numerica* secondo

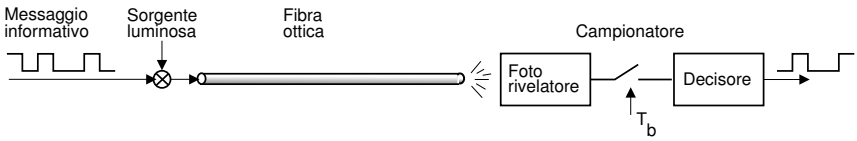
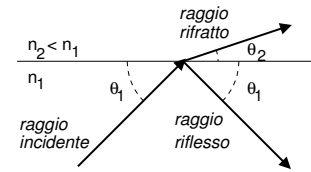



Figura 15.2: Schema di trasmissione in fibra ottica

lo schema di fig. 15.2, in cui la luce emessa da una sorgente è accesa o spenta (ovvero modulata in ampiezza con uno schema ON/OFF). All'altro estremo della fibra, un fotorivelatore effettua una rivelazione incoerente dell'energia luminosa, che viene nuovamente convertita in un segnale elettrico. Le prime fibre ottiche risalgono al 1970, e fornivano attenuazioni dell'ordine di 20 dB/Km. Attualmente si sono raggiunti valori di attenuazione di 0.2 dB/Km, pari ad un quarto di quella dei migliori cavi coassiali. D'altra parte, a differenza del rame, il materiale utilizzato per le fibre (vetro o silicio) è largamente disponibile in natura. Inoltre, a parità di diametro, una fibra ottica trasporta un numero anche 1000 volte maggiore di comunicazioni rispetto ad un cavo coassiale, fornendo quindi anche un risparmio di spazio.

**Propagazione luminosa e indice di rifrazione** Lo spazio libero è il mezzo di propagazione in cui la luce viaggia più velocemente. Il rapporto tra $c = 3 \cdot 10^8$ m/sec, e la velocità di propagazione $v$ in un mezzo trasparente, è l'*indice di rifrazione* $n$ del mezzo stesso: $n = {}^{c}/_{\nu}$, risultando $n \geq 1$. Ad esempio, se $n = 2$ allora la velocità è la metà.

Quando un raggio luminoso incontra un mezzo con diverso indice $n$ (ad esempio, da $n_1$ ad $n_2 < n_1$) una parte di energia si riflette con angolo $\theta_1$ uguale a quello incidente, e la restante parte continua nell'altro mezzo, ma con diverso angolo $\theta_2 < \theta_1$. Risulta




$$\frac{\cos\theta_1}{\cos\theta_2} = \frac{n_2}{n_1}$$

e dunque il raggio rifratto è più inclinato nel mezzo con $n$ inferiore (dove viaggia più veloce). Esiste un valore $\theta_c = \arccos \frac{n_2}{n_1}$ sotto il quale non si ha rifrazione, ma tutto il raggio viene riflesso.

E' proprio su questo fenomeno che si basa l'attitudine delle fibre ottiche di trasportare energia luminosa. La fibra ottica è infatti costituita da un nucleo (*core*) centrale con indice di rifrazione $n_1$, circondato da un rivestimento (*cladding*) con indice $n_2 < n_1$; entrambi racchiusi in una guaina (*jacket*) di materiale opaco (vedi fig. 15.3).

Quando una sorgente luminosa è posta davanti alla fibra, l'energia si propaga mediante diversi *modi di propagazione,* definiti nel contesto della meccanica quantistica, e identificabili in chiave di ottica geometrica come i diversi angoli di incidenza. Il *modo principale* è quello che si propaga lungo l'asse rettilineo, mentre i *modi secondari* sono quelli con angolo $< \theta_c$, che si riflettono completamente al confine tra core e cladding. I modi associati ad angoli più elevati di $\theta_c$ vengono progressivamente assorbiti dalla guaina, e dunque non si propagano.

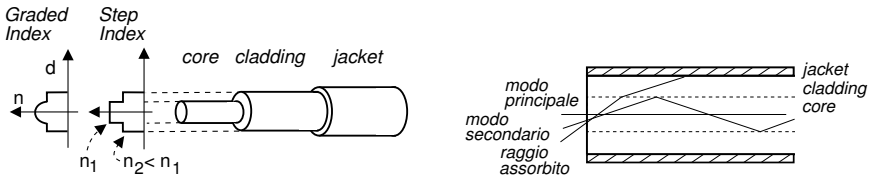
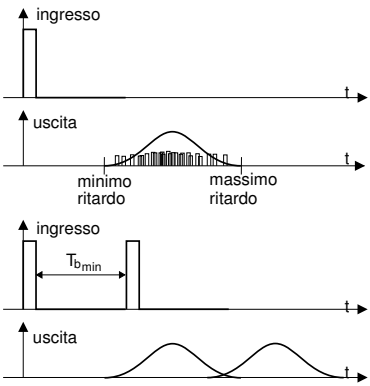



Figura 15.3: Tecnologia della fibra ottica e modi di propagazione

Il valore $\Delta = \sqrt{n_1^2 - n_2^2} = n_1 \sin\theta_c$ prende il nome di *apertura numerica*, e permette di risalire al massimo angolo di incidenza mediante la relazione $\theta_c = \arcsin \frac{\Delta}{n_1}$. Come si vede, $\Delta$ è tanto più piccolo quanto più $n_1$ ed $n_2$ sono simili; al diminuire di $\Delta$, si riduce anche la potenza luminosa che viene immessa nella fibra ottica, ma si ottiene il beneficio illustrato di seguito.

**Dispersione modale** Questo fenomeno è dovuto al fatto che i modi propagazione relativi agli angoli di incidenza più elevati percorrono di fatto *più strada,* e dunque impiegano più tempo per giungere a destinazione.

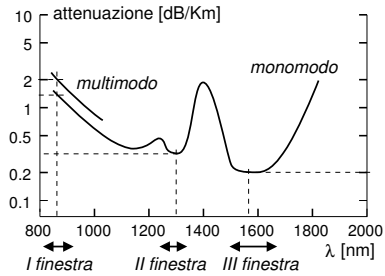
Pertanto, ogni singolo impulso luminoso presente in ingresso produce in uscita più impulsi distanziati nel tempo, uno per ogni modo di propagazione. Dato che inoltre avviene un continuo scambio di energia tra i diversi modi, si ottiene che l'uscita sarà un segnale con una maggiore estensione temporale (vedi figura). L'entità della *dispersione temporale* (differenza tra ritardo max e min) sarà tanto maggiore quanto più il collegamento è lungo, e quanti più modi partecipano alla propagazione. L'effetto più appariscente del fenomeno descritto consiste nella limitazione della massima frequenza con cui gli impulsi luminosi possono essere posti in ingresso alla fibra; impulsi troppo vicini risulterebbero infatti affetti da interferenza intersimbolica (ISI) ed in pratica indistinguibili in uscita. Pertanto la massima frequenza di segnalazione in una fibra ottica, dipende dalla *lunghezza* della fibra stessa.




Si chiamano *fibre multimodo* le fibre ottiche in cui sono presenti più modi di propagazione. Queste sono del tipo STEP INDEX se $n$ cambia in modo brusco, o GRADED INDEX se il *core* ha un indice graduato. Nel secondo caso la dispersione temporale è ridotta; infatti quando i modi secondari attraversano la sezione periferica del core, incontrano un indice di rifrazione $n$ ridotto, e quindi viaggiano più veloci. Una diversa (e drastica) soluzione al problema della dispersione temporale, è fornita dalle fibre ottiche *monomodo*: queste sono realizzate con un core di diametro così piccolo[30], da consentire alla sorgente luminosa di immettere luce nella fibra solo con angolo di incidenza nullo, e quindi di permettere la propagazione del solo modo principale.

Ovviamente le ultime due soluzioni (graded index e fibra monomodo) si sono rese possibili grazie ai progressi nei processi di fabbricazione.

[30]Si passa dai 50 $\mu$m per le fibre multimodo, a circa 8 $\mu$m nel caso monomodo.




| *Finestra* | $\lambda$ [$\mu$m] | $A_d$ [dB/Km] |
|---|---|---|
| *I* | 0.8÷0.9 | 1.2 (monomodo) |
| | | 2 (multimodo) |
| *II* | 1.2÷1.3 | 0.35 |
| *III* | 1.5÷1.7 | 0.2 |

Figura 15.4: Dipendenza della attenuazione chilometrica dalla lunghezza d'onda

**Attenuazione** In modo simile ai cavi elettrici, anche le fibre ottiche sono mezzi dissipativi, in quanto parte dell'energia in transito viene assorbita dalla fibra stessa e trasformata in calore. I fenomeni di assorbimento sono legati alla presenza di impurità chimiche, che possono ridurre la trasparenza oppure avere dimensioni (a livello molecolare) comparabili con le lunghezze d'onda in gioco.

Per questi motivi, la caratteristica di attenuazione chilometrica ha un andamento (vedi fig. 15.4 ) fortemente dipendente da $\lambda$, e sono stati individuati 3 intervalli di lunghezze d'onda (detti *finestre*) per i quali l'assorbimento è ridotto, ed in cui sono effettuate le trasmissioni ottiche.

La prima finestra (con attenuazione maggiore) è stata l'unica disponibile agli inizi, a causa dell'assenza di trasduttori affidabili a frequenze inferiori, ed è tuttora usata per collegamenti economici e scarsamente critici. La seconda finestra ha iniziato ad essere usata assieme alle fibre monomodo, grazie all'evoluzione tecnologica dei trasduttori, mentre l'uso della III finestra si è reso possibile dopo essere riusciti a limitare la *dispersione cromatica* delle fibre (vedi appresso).

Un'altra fonte di attenuazione può avere origine dalle *giunzioni* tra tratte di fibre ottiche: l'uso di connettori produce una perdita di $0.4 \div 1$ dB, ed i giunti meccanici $\simeq 0.2$ dB oppure anche 0,05 dB se ottimizzati per via strumentale. Si possono infine *fondere* tra loro le fibre, con perdite tra 0,01 e 0,1 dB.

### 15.4.2 Dimensionamento del collegamento

**Dispersione cromatica** Dopo aver ridotto od eliminato il fenomeno di dispersione modale, si è individuata una ulteriore causa di dispersione temporale dell'energia immessa nella fibra ottica: il problema si verifica se il segnale di ingresso non è perfettamente monocromatico, ovvero sono presenti diverse lunghezze d'onda. Dato che il valore dell'indice di rifrazione dipende dalla lunghezza d'onda, $\lambda$ diverse si propagano con velocità differenti e raggiungono l'altro estremo della fibra in tempi successivi. La dispersione cromatica nominale $D_0$ della fibra si misura in $\left[\frac{psec}{Km \cdot nm}\right]$, e dà luogo ad una dispersione temporale $D = D_0 \cdot L \cdot \Delta\lambda$ direttamente proporzionale alla lunghezza $L$ della fibra, ed alla estensione della *gamma cromatica* $\Delta\lambda$ della sorgente[31]. Per ridurre il fenomeno è possibile:

[31] Il fenomeno della dispersione cromatica è l'equivalente ottico della distorsione di fase (o distorsione di ritardo) introdotta al § 14.5 per i segnali elettrici.

- utilizzare una lunghezza d'onda $\lambda$ per la quale la dispersione cromatica è ridotta. Ad esempio, una fibra di silicio *normale* produce una dispersione cromatica 15 volte inferiore a 1.3 $\mu$m che non a 1.5 $\mu$m;
- scegliere una sorgente con la minima estensione cromatica $\Delta\lambda$ possibile.

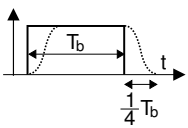
**Trasduttori elettro-ottici** Quelli usati per primi sono stati gli economici LED (*Light Emitting Diode*), che richiedono una circuiteria di interfaccia semplice, sono poco sensibili alle condizioni ambientali, e quindi risultano affidabili. D'altra parte, i LED emettono luce su più lunghezze d'onda, mentre per limitare la dispersione cromatica (e quindi raggiungere frequenze di segnalazione più elevate) occorre ricorrere ai *Diodi Laser* (LD).

I LASER forniscono anche una maggiore potenza, e quindi divengono indispensabili per coprire distanze maggiori[32]. D'altra parte sono più costosi, hanno vita media ridotta rispetto ai LED, e richiedono condizioni di lavoro più controllate. Notiamo inoltre che una fibra ottica posta inizialmente in opera mediante sorgenti LED, può essere potenziata (in termini di banda) sostituendo il LED con il LASER.

| *Sorgente* | $\lambda$ *(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 |

L'uso di sorgenti che operano in III finestra, che (presentando una attenuazione ridotta) permette di operare con tratte più lunghe, obbliga in generale a ridurre la frequenza di segnalazione, a causa della maggiore dispersione cromatica. Quest'ultima limitazione è stata rimossa da un particolare tipo di fibra, detta *dispersion shifted*, che presenta un minimo di dispersione cromatica $D_0$ in III finestra anziché in II, e che raggiunge valori migliori di 3.5 psec/Km·nm.

**Prodotto banda-lunghezza e codici di linea** Come anticipato, la dispersione cromatica $D$ risulta proporzionale alla lunghezza del collegamento $L$ ed all'estensione cromatica $\Delta\lambda$ della sorgente. Se pensiamo di effettuare una trasmissione con codici NRZ e periodo $T_b = \frac{1}{f_b}$, ed imponiamo che la dispersione temporale sia non maggiore di $\frac{1}{4}T_b$, si ottiene




$$D_0 \cdot L \cdot \Delta\lambda \leq 0.25 \cdot T_b \tag{15.12}$$

in cui $D_0$ è la dispersione cromatica *nominale* [psec/Km·nm], $L$ è la lunghezza [Km], $\Delta\lambda$ è l'estensione cromatica della sorgente [nm], e $T_b$ è la durata di un bit [psec]. Associando ora il concetto di *banda* $B$ alla frequenza di segnalazione $f_b = \frac{1}{T_b}$, la relazione (15.12) può essere riscritta in modo da evidenziare il *prodotto della banda per la lunghezza* $PBL$, che è pari al valore

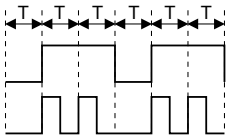
$$PBL_{NRZ} = f_b \cdot L = \frac{.25}{D_0 \cdot \Delta\lambda} \quad [Tbps \cdot Km]$$

[32] 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.

che è una grandezza dipendente dalla coppia fibra-sorgente[33]. Inserendo i valori di $\Delta\lambda$ (della sorgente) e $D_0$ (della fibra), 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[34] rispetto al caso precedente:

$$PBL_{RZ} = \frac{.5}{D \cdot \Delta\lambda} = 2 \cdot PBL_{NRZ}$$

Qui sotto è mostrato il confronto tra i codici RZ e NRZ, e la tabella dei valori di $PBL$ associati alle coppie fibra-sorgente indicate.




| *Fibra* | *Sorgente* | $\lambda$ *[nm]* | $PBL$ *[Gbps·Km]* |
|---|---|---|---|
| Multimodo | LED | 820 | 0.013 |
| Graded Index | LED | 820 | 2 |
| Monomodo | LASER | 1300 | 250 |
| Monomodo | LASER | 1550 | 36.8 |
| Monomodo *DS* | LASER | 1550 | 179 |

**Esercizio** Determinare la lunghezza massima di un collegamento in fibra ottica monomodo, operante con $\lambda$=1.3 $\mu$m, e che garantisca una velocità $f_b$=417 Mbps, assumendo un guadagno di sistema di 42 dB (ovvero disponendo di una potenza di trasmissione 42 dB maggiore della minima potenza necessaria in ricezione).

**Soluzione** Dal grafico di fig. 15.4 si trova che per $\lambda$=1300 nm, l'attenuazione chilometrica è di 0,35 dB/Km, che determina una $A_d = 0,35 \cdot L_{Km}$ dB. Imponendo ora $G_s \geq A_d$ si ottiene una lunghezza non superiore a $\frac{42}{0,35}$=120 Km, che identifica il *Limite di Attenuazione* del collegamento. Verifichiamo quindi che non intervenga un limite più stringente a causa della dispersione cromatica. Supponendo di utilizzare la sorgente laser in grado di conseguire un $PBL$ di 250 Gbps·Km, si ottiene una lunghezza massima pari a $\frac{250.000}{417}$=600 Km, che costituisce il *Limite di Dispersione*.

**Massima lunghezza di tratta** L'esercizio svolto ha lo scopo di mostrare la metodologia di progetto per un collegamento in fibra ottica, in cui vengono calcolati entrambi i limiti di *Attenuazione* e di *Dispersione*, e la massima lunghezza del collegamento è determinata dal vincolo più stringente. Nel caso dell'esercizio, la lunghezza è determinata dal limite di attenuazione, ed il progetto può essere rivisto utilizzando una sorgente *meno pura* per risparmiare, oppure una sorgente *più potente* per aumentare il guadagno di sistema e conseguentemente migliorare il limite di attenuazione. In questo secondo caso, può essere opportuno prestare attenzione al fatto che, aumentando la potenza di emissione, la purezza cromatica della sorgente può degradare (in quanto si verifica un aumento di $\Delta\lambda$ dovuto a fenomeni non lineari) con un conseguente peggioramento del limite di dispersione; è pertanto possibile ricercare la soluzione di migliore compromesso tra potenza di emissione e purezza spettrale.

Qualora non si riesca a rientrare nelle specifiche di progetto con una unica tratta, occorrerà suddividere il collegamento in più tratte, collegate da *ripetitori rigenerativi*, oppure ripartire la banda su più fibre poste in parallelo, ovvero bilanciarsi tra queste due soluzioni.

[33] In questo senso, il prodotto *banda-lunghezza* costituisce un parametro di sistema che tiene conto di un concorso di cause. Un pò come il concetto di *tenuta di strada* di una autovettura, che dipende da svariati fattori, come il peso, i pneumatici, la trazione, il fondo stradale....

[34] Tuttavia, il dimezzamento della durata di un bit causa una perdita di potenza di 3 dB, in base alle considerazioni riportate a pag. 382.

**Trasduttori ottico-elettrici** La conversione del segnale uscente dalla fibra ottica meriterebbe una ampia trattazione approfondita, ma qui ci limitiamo a riferire esclusivamente poche cose fondamentali.

Il trasduttore utilizzato fin dall'inizio, economico ed affidabile, è il diodo P-I-N[35], che però non è adatto all'impiego con $\lambda$ più elevate. Un secondo tipo di trasduttore molto usato è il diodo APD[36] (*Avalanche Photo Detector*), caratterizzato da un "effetto valanga" che lo rende più sensibile di 10-15 dB rispetto ai P-I-N; d'altra parte però gli APD sono più delicati, e più sensibili alla temperatura.

La tabella che segue riporta i valori di sensibilità $W_R$ (ossia la minima potenza che è necessario ricevere) di diversi fotorivelatori, necessaria a conseguire[37] una probabilità di errore per bit $P_e = 10^{-11}$.

| *Fotorivelatore* | $\lambda$ *[nm]* | *Sensibilità [dBm]* | $f_b$ *[Mbps]* |
|---|---|---|---|
| Si P-I-N | 850 | -48 | 50 |
| Si APD | 850 | -58 | 50 |
| InGaAs P-I-N | 1310 | -35 | 420 |
| InGaAs APD | 1310 | -43 | 420 |
| InGaAs APD | 1550 | -37.5 | 678 |

**Dipendenza della sensibilità dalla durata del simbolo** Nella tabella è riportato anche il valore della frequenza di segnalazione $f_b$ a cui si riferisce la sensibilità, in quanto le prestazioni conseguite dal decisore che si trova a valle del trasduttore dipendono, come noto, 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[38]

$$W_R\left(f_b^{'}\right)\ [\text{dBm}] = W_R\left(f_b\right)\ [\text{dBm}] + 10\log\frac{f_b^{'}}{f_b}$$

### 15.4.3 Multiplazione a divisione di lunghezza d'onda - WDM

Nel campo delle trasmissioni ottiche, per le quali è spesso sottinteso che si sta trasportando un segnale già multiplato TDM, si aggiunge una nuova possibiità di multiplazione, la WDM (*Wavelength Division Multiplex*), in cui si usano diverse lunghezze d'onda (nella stessa finestra) per diverse comunicazioni contemporanee.

Il modo più semplice ed intuitivo di realizzare la multiplazione di lunghezza d'onda è di adottare dei *rifrattori prismatici*, realizzando un circuito ottico del tipo illustrato in figura[39]. I dispositivi di multiplazione di forma d'onda di questo tipo vengono detti *passivi* e *reversibili*, in quanto non necessitano di alimentazione, ed

[35] `http://it.wikipedia.org/wiki/Diodo_PIN`

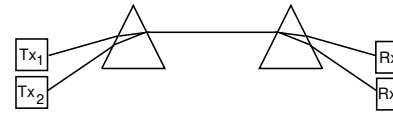
[36] `http://it.wikipedia.org/wiki/Fotodiodo_a_valanga`

[37] La consuetudine del dimensionamento dei collegamenti in fibra otttica 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!

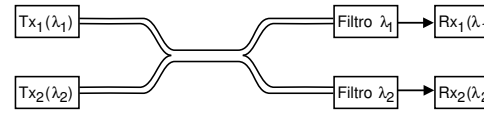
[38] Questo metodo di calcolo della sensibilità ad una $f_b$ diversa da quella delle tabelle è approssimato, in quanto nei trasduttori avvengono fenomeni non-lineari che legano il livello di potenza del rumore alla potenza di segnale ricevuta. Trascurando questo effetto, si può applicare l'espressione sopra riportata.

[39] Si sfrutta il principio "dell'arcobaleno" (o dei *Pink Floyd* ?), in quanto uno stesso materiale (il

uno stesso apparato può indifferentemente svolgere una funzione e la sua inversa. La passività del WDM rende questa tecnica attraente, qualora si pensi di distribuire fibre ottiche di casa in casa: ognuno avrebbe una sua lunghezza d'onda $\lambda_i$, e la fibra sarebbe una per tutto il condominio.




Se le $\lambda_i$ sono troppo vicine tra loro (con una spaziatura dell'ordine di 0.015 nm), allora i prismi non riescono più nel compito di separazione geometrica, e conviene ricorrere ad una separazione della potenza (si fa uscire parte di segnale luminoso da *jacket*) ed un filtraggio (realizzato otticamente mediante gelatine) di ognuna delle $\lambda_i$. Così facendo però si perde molta potenza.




### 15.4.4 Ridondanza e pericoli naturali

Le fibre vengono normalmente interrate, e per questo sono esposte ai pericoli di essere mangiate da talpe e topi, o di essere interrotte a causa di lavori stradali od agricoli. Quelle sottomarine sono a rischio per via di squali e reti a strascico. E' più che opportuno prevedere una adeguata ridondanza (vedi § 6.6.3), in modo che in caso di interruzione di un collegamento sia possibile deviare tutto il traffico su di un altro collegamento.

### 15.4.5 Sonet e SDH

SONET è l'acronimo di *Synchronous Optical Network*, lo standard americano definito allo scopo di permettere l'interconnessione diretta tra reti in fibra ottica, perfettamente interoperabile con l'equivalente europeo *Synchronous Digital Hierarchy* (SDH); l'unità di multiplazione fondamentale per i due casi è pari a 51.84 Mbps per Sonet, ovvero 3 volte tanto (155,52 Mbps) per l'Europa. Per approfondimenti, vedi § 6.4.

| *Data Rate [Mbps]* | *Sigla CCITT* |
|---|---|
| 51.84 | |
| 155.52 | STM-1 |
| 466.56 | STM-3 |
| 622.08 | STM-4 |
| 933.12 | STM-6 |
| 1244.16 | STM-8 |
| 1866.24 | STM-12 |
| 2488.32 | STM-16 |

### 15.4.6 Dalle fibre ottiche alle reti ottiche

Lo sviluppo tecnologico permette attualmente[40] di realizzare dispositivi in grado di operare direttamente sul segnale ottico WDM, e cioè

- amplificatori ottici - questa funzione realizzata direttamente nel dominio ottico abbatte la complessità ed i ritardi legati alla doppia conversione altrimenti necessaria;
- multiplatori add and drop - per inserire o estrarre una specifica $\lambda$;

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.

[40] A titolo integrativo, si propongono i riferimenti a
`http://it.wikipedia.org/wiki/Amplificatore_ottico`,
`http://en.wikipedia.org/wiki/Wavelength-division_multiplexing#Dense_WDM`,
`http://www.fiber-optics.info/articles/dense_wavelength-division_multiplexing`

- convertitori di lunghezza d'onda - per convertire una $\lambda$ ad un'altra;
- commutatori di lunghezza d'onda - per estrarre una $\lambda$ da un segnale WDM ed inserirla in un altro.

Questi dispositivi permettono la realizzazione di trasmissioni DWDM (*Dense Wavelength Division Multiplex*), che ospitano quasi un centinaio di diverse $\lambda$ nella stessa fibra, e che hanno trovato impiego dapprima nei collegamenti a lunga distanza, favoriti dagli amplificatori ottici, e quindi nelle sezioni di rete via via più periferiche, sfruttando i dispositivi di commutazione ed accesso, ed infine hanno trovato un ruolo anche nei meccanismi di protezione dai guasti.

Un ulteriore importante risultato della trasmissione DWDM è che, ospitando differenti tributari ad alta velocità su diverse $\lambda$, 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.

## 15.5 Appendici

### 15.5.1 Fading piatto e veloce

Procediamo con l'analizzare gli effetti che il fenomeno dei cammini multipli produce nei confronti dell'inviluppo complesso $\underline{x}(t)$ del segnale che lo subisce. La (15.11) ci permette di scrivere[41] l'inviluppo complesso $\underline{y}(t)$ del segnale ricevuto come

$$\underline{y}(t) = \sum_{n=1}^{N} a_n \underline{x}(t-\tau_n)\, e^{-j2\pi f_0 \tau_n} \tag{15.13}$$

in cui $\tau_n$ è il ritardo dell'$n$-esimo cammino, ed $a_n$ la rispettiva ampiezza. A questo punto la teoria prevede due possibili approssimazioni, a seconda se tra l'arrivo della prima replica (ritardata di $\tau_{min}$) e l'arrivo dell'ultima (ritardata di $\tau_{max}$) il segnale $x(t)$ non vari di molto (e cioé $x(t-\tau_{min}) \simeq x(t-\tau_n) \simeq x(t-\tau_{max})$)[42], oppure invece vari in modo significativo. Il primo caso è tipico dei segnali *a banda stretta* e dei canali che presentano *fading piatto*, mentre nel secondo caso il segnale si dice *a banda larga* ed il canale *selettivo in frequenza*.

Nel caso di segnale a banda stretta la (15.13) si può riscrivere come

$$\begin{aligned} \underline{y}(t) &\simeq \underline{x}(t) \sum_{n=1}^{N} a_n e^{-j2\pi f_0 \tau_n} = \underline{x}(t) \sum_{n=1}^{N} a_n (\cos\varphi_n - j \sin\varphi_n) \\ &= (X + jY)\, \underline{x}(t) \end{aligned} \tag{15.14}$$

in cui la costante complessa $X + jY$ riassume l'effetto delle diverse repliche, ed i valori $\varphi_n = 2\pi f_0 \tau_n$ rappresentano i ritardi di fase della portante per i diversi cammini (nota

[41] La (15.13) si ottiene considerando un generico segnale modulato $x(t) = a(t)\cos 2\pi f_0 t$: per ognuna delle repliche

$$x_n(t) = x(t-\tau_n) = a(t-\tau_n)\cos 2\pi f_0 (t-\tau_n) = a(t-\tau_n)\cos(2\pi f_0 t - 2\pi f_0 \tau_n)$$

l'inviluppo complesso rispetto ad $f_0$ può essere scritto come

$$\underline{x}_n(t) = a(t-\tau_n)\, e^{-j2\pi f_0 \tau_n} = \underline{x}(t-\tau_n)\, e^{-j2\pi f_0 \tau_n}$$

[42] Si consideri che il risultato dell'esempio di pag. 371 valuta i ritardi in gioco dell'odine di grandezza dei microsecondi.

41). A partire da valori della portante $f_0$ dell'ordine dell'inverso dei ritardi $\frac{1}{\tau_n}$, e tanto più per $f_0$ più elevate[43], bastano piccole variazioni di ritardo per produrre $\varphi_n$ del tutto indipendenti ed uniformemente distribuite tra 0 e $2\pi$. D'altra parte, anche i valori $a_n$ possono considerarsi variabili aleatorie indipendenti, e se il loro numero $N$ è elevato, sussitono le condizioni per considerare $X$ ed $Y$ della (15.14) realizzazioni di due v.a. indipendenti e gaussiane a media nulla ed uguale varianza. A pag. 156 si mostra come in questo caso il modulo $\rho = \sqrt{X^2 + Y^2}$ costituisca una v.a. di RAYLEIGH, che nella eq. (15.14) rappresenta il fattore moltiplicativo di $\underline{x}(t)$ rispetto al caso senza cammini multipli. A pag. (7.6.4.2) infine si definisce la v.a. di RICE, che nel nostro caso rappresenta l'alterazione di $\underline{x}(t)$ nel caso in cui uno dei cammini multipli abbia una ampiezza preponderante su quella degli altri.

### 15.5.2 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 metereologici, di radiolocalizzazione (il GPS, ma non solo), per ponti radio televisivi, di telefonia, broadcast. Senza molto togliere alla generalità dell'esposizione, questa procede illustrando l'ultimo caso citato, detto DVB (*Digital Video Broadcast*), in cui il satellite semplicemente ritrasmette verso una estesa area geografica i segnali ricevuti da terra, come mostrato in figura 15.5, assieme all'*ipsogramma*[44] relativo.

**Studio di produzione** Non volendo assolutamente entrare qui negli innumerevoli dettagli che andrebbero illustrati, limitiamoci a descrivere i passi necessari a generare il segnale inviato al satellite:

- si effettua la codifica digitale MPEG2 del segnale televisivo, ottenendo un flusso numerico chiamato PS (*Program Stream*);
- più PS sono pacchettizzati e multiplati 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. 78) a tre stadi, in cui è prima applicato un codice di *Reed-Solomon*, poi un *interleaver*, e quindi un codificatore *convoluzionale*, rendendo il segnale particolarmente robusto agli errori di trasmissione sia singoli che a burst;

[43] Se ad esempio i ritardi $\tau_n$ sono dell'ordine di $10^{-6}$, l'ipotesi è valida per $f_0 > 1$ MHz.

[44] 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 l'andamento del livello di potenza in un collegamento.

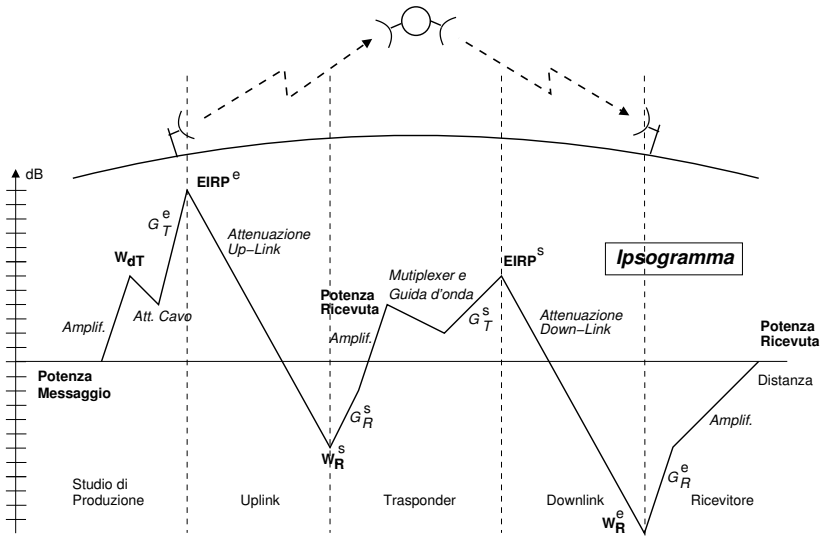



Figura 15.5: Andamento del livello di potenza in dB per un collegamento satellitare

- il nuovo flusso numerico è modulato QPSK (a due bit per simbolo) con codifica *di Gray*, sagomando i simboli con un filtro a coseno rialzato con $\gamma = 0.35$ ripartito tra trasmettitore e ricevitore finale, ossia adottando un formatore di impulsi a *radice di coseno rialzato* (vedi §5.2.2.3).

**Uplink** Il *collegamento in salita* (UPLINK) è quello mediante il quale lo studio di produzione invia al satellite l'MPEG-TS che deve essere re-distributo. 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$^e$ *(effective irradiated power)* rappresenta la potenza effettivamente irradiata[45], che si riduce notevolmente nella trasmissione da terra a satellite (*Up-Link*). Nel caso, ad esempio, in cui la portante sia di 2 GHz e l'orbita sia geostazionaria[46] (36.000 Km da terra), l'attenuazione di spazio libero dell'Up-Link (eguale a quella del *Down-Link* da satellite a terra) è di circa 190 dB.

**Transponder** Il segnale ricevuto, di potenza $W_R^s$, attraversa l'antenna ricevente del satellite di guadagno $G_R^s$, e l'amplificatore seguente eleva ulteriormente il livello del segnale tramesso, che subisce alcune perdite nel collegamento con l'antenna trasmit-

[45] 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.

[46] 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, 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.

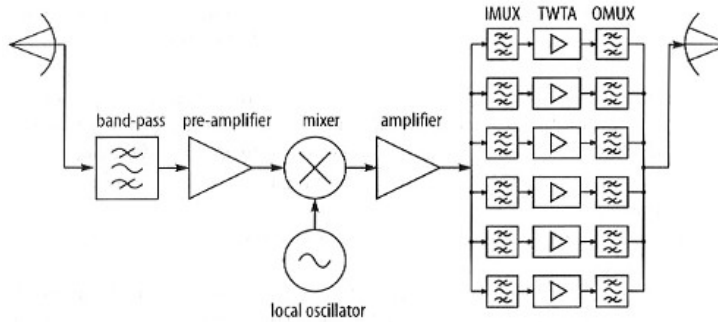
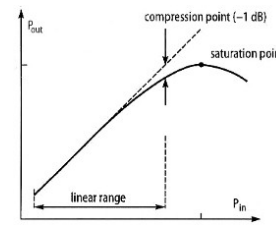



Figura 15.6: Elaborazione di bordo per un trasponder DVB satellitare

tente 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 tramesso costituirebbe un rilevante termine di *interferenza* per il lato ricevente del satellite, nonostante l'elevata direttività delle antenne, dando così luogo ad un fenomeno di *diafonia*[47]. La Fig. 15.6 mostra come il segnale a banda larga (che trasporta molteplici canali) ricevuto da terra viene prima filtrato alla banda del segnale utile, quindi amplificato una prima volta, poi miscelato con un oscillatore locale[48], ed infine amplificato una seconda volta[49].

Quindi, i singoli canali FDM che compongono il segnale sono separati tra loro mediante il banco di filtri passa-banda indicati come IMUX (*input multiplexer*), e amplificati individualmente mediante dei TWTA[50] che, se spinti alla massima potenza, presentano una caratteristica ingresso-uscita non lineare (vedi § 14.6), mostrata alla figura a lato.




Nel caso di trasmissioni modulate angolarmente, la distorsione in ampiezza è ben tollerata, e quindi si può mantenere limitato il *back-off* necessario; d'altra parte, le componenti frequenziali spurie prodotte dalla non linearità devono essere rimosse per non provocare disturbo alle altre comunicazioni, e questo è il compito del banco di filtri passa banda OMUX (*output multiplexer*) posti di seguito ai TWTA.

**Footprint e Downlink** L'antenna trasmittente del satellite sagoma il proprio diagramma di radiazione in modo da concentrare la potenza trasmessa in una ben determi-

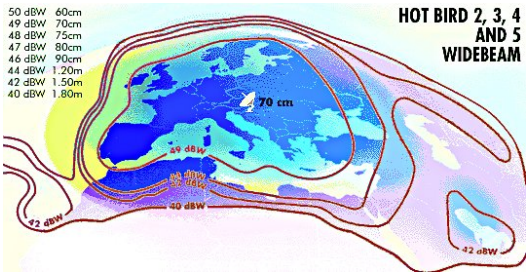
[47] Le considerazioni sulla diafonia si applicano altrettanto bene anche al caso di ripetitori terrestri.

[48] Come descritto al § 11.2.3, l'oscillatore locale deve avere una frequenza $f_e$ tale che $f_d = f_u - f_e$, in modo cheil segnale di downlink sia centrato ad una frequenza pari alla differenza tra quella di uplink e quella di eterodina.

[49] La suddivisione della amplificazione in due stadi a frequenza diversa previene fenomeni di reazione positiva.

[50] *Travelling Wave Tube Amplifier*, ovvero tubi amplificatori ad onda progressiva: `http://it.wikipedia.org/wiki/Travelling_wave_tube`.

nata area della terra, dando luogo alla cosiddetta *footprint* (impronta) raffigurata a lato, in cui le curve isomere individuano sia il livello di potenza ricevuto, che il diametro (e quindi il guadagno) necessario per l'antenna ricevente.




La tecnica che permette di distribuire la potenza emessa secondo una geometria diversa da una simmetria radiale prende il nome di *beamforming*, e si basa sull'utilizzo di più antenne trasmittenti, in modo da realizzare un *phased array* (vedi `http://en.wikipedia.org/wiki/Phased_array`). Ad ogni antenna dell'array perviene lo stesso segnale modulato, ma con una fase tale da creare uno schema di interferenza con le altre antenne dell'array, in modo che alla distanza di ricezione, si determini la distribuzione spaziale desiderata.

Dal lato del ricevitore terrestre arriva dunque un segnale di potenza $W_R^e$, che ha subito l'attenuazione del down-link; questo è quindi riportato ad un livello di potenza appropriato, sia grazie al guadagno di antenna, che per mezzo di uno stadio di amplificazione.

**Temperatura di antenna** Come illustrato al § 15.3, una antenna ricevente è schematizzabile come un generatore controllato, ed al Cap. 16 si mostra come la sua impedenza interna sia la fonte del rumore additivo gaussiano in ingresso al ricevitore, caratterizzato da una densità di potenza disponibile $\mathcal{W}_{dn}(f) = \frac{1}{2}kT_g$, 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[51].

**Ricevitore a terra** La figura 15.7 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 GH, 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[52] che collega l'antenna al ricevitore casalingo. Quindi, si ritrova un schema simile a quello del trasponder, ovvero amplificatore-mixer-amplificatore, in cui questo secondo stadio eterodina centra il canale desiderato alla frequenza intermedia di 479.5 MHz.

**Polarizzazione** Chi ha provato a sintonizzare un ricevitore TV satellitare, si sarà accorto che tra le varie opzione possibili, si può indicare anche il *tipo di polarizzazione*,

[51] 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.

[52] Come descritto nel paragrafo che discute dell'*effetto pelle* (pag. 359), l'attenuazione in dB del cavo aumenta con l'aumentare della radice della frequenza.

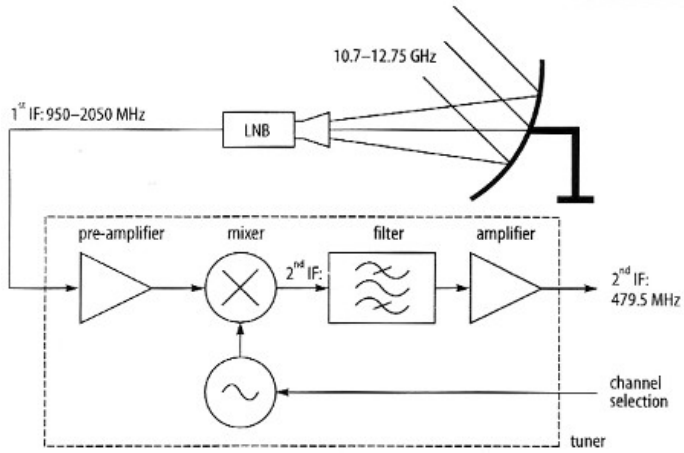



Figura 15.7: Ricevitore satellitare DVB

orizzontale o verticale. Questo termine si riferisce all'orientamento rispetto all'orizzonte del piano su cui si muove 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 frequenza possono essere effettuate due trasmissioni contemporanee.

### 15.5.3 Allocazione delle frequenze radio

L'assegnazione generale dello spettro radio ai diversi utilizzi è riportata in tabella 15.1, che non pretende di essere completa né tantomeno 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 e' presente il broadcast FM.

**UHF:** Numerati da 14 a 69 a partire dalla portante video di 471.25 MHz, fino a 801.25 MHZ, spaziati di 6 MHz.

Per le stesse frequenze, sono state attivate la trasmissioni televisive in *digitale terrestre.*, ad eccezione dei canali da 61 a 69, che saranno assegnati agli operatori di telefonia mobile di $4^a$ generazione.

**Bande di frequenza Radar** Oltre alle bande HF, VHF ed UHF, le trasmissioni radar che operano in SHF ed EHF distinguono tra i seguenti intervalli di frequenze:

| *Intervallo* | $\lambda$ | *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 15.1: Allocazione delle frequenze radio

| 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 |

# Capitolo 16

# Rumore termico

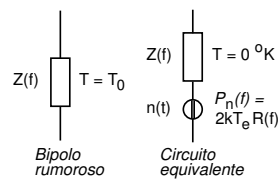
Descriviamo ora la natura e le fonti del processo di rumore sempre presente negli apparati di telecomunicazione, e di come questo sia tenuto in considerazione nel progetto degli stessi.

## 16.1 Rumore nei bipoli passivi

Ai capi di un resistore $R$ a temperatura $T$ è presente una *tensione a vuoto* $n(t)$, realizzazione di un processo gaussiano a media nulla, che è l'effetto del moto caotico degli elettroni all'interno della resistenza[1]. Lo spettro di densità di potenza della tensione a vuoto ha espressione[2]

$$\mathcal{P}_n(f) = 2R\frac{\hbar f}{\mathrm{e}^{\frac{\hbar f}{kT}} - 1} \simeq 2kTR \quad [\mathrm{Volt}^2]$$

in cui $k = 1.38 \cdot 10^{-23}$ Joule/$^o$K è la *costante di Boltzman* ed $\hbar = 6.62 \cdot 10^{-34}$ Joule·sec è la costante *di Plank*: questi valori[3] fanno sì che l'approssimazione $\mathcal{P}_n(f) \simeq 2kTR$ sia valida ad ogni frequenza di interesse.




In un bipolo passivo di impedenza $Z(f) = R(f) + jX(f)$, solamente la parte reale (componente resistiva) concorre a generare il processo di rumore termico, che pertanto possiede una densità di potenza *di segnale* $\mathcal{P}_n(f) \simeq 2KTR(f)$. Nel caso in cui il bipolo contenga *più* resistori a temperature diverse, si può definire una temperatura equivalente $T_e(f)$; un bipolo passivo equivale pertanto allo stesso bipolo non rumoroso (a temperatura zero), con in serie un generatore di rumore con densità di potenza $\mathcal{P}_n(f) \simeq 2kT_e(f)R(f)$. Questo generatore equivalente, è

[1] Possiamo pensare che gli elettroni, qualora si trovino in maggior misura in una metà della resistenza, producano una differenza di potenziale negativa in quella direzione. Allo zero assoluto (- 273 $^o$C) il moto caotico degli elettroni cessa, e si annulla così la tensione di rumore. Di qui l'aggettivo *termico* per descrivere il fenomeno.

[2] Si tratta di una forma della legge di Plank, vedi
`http://en.wikipedia.org/wiki/Thermal_noise#Noise_at_very_high_frequencies`

[3] Espandendo $\mathrm{e}^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots$ si ottiene che per $x \ll 1$ risulta $\mathrm{e}^x \simeq 1 + x$, e quindi $\mathrm{e}^{\frac{\hbar f}{kT}} \simeq 1 + \frac{\hbar f}{kT}$.

quindi descritto da una *potenza disponibile di rumore*

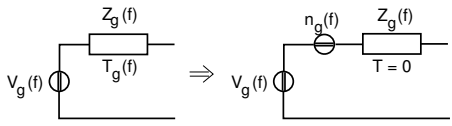
$$\mathcal{W}_{dn}(f) = \frac{\mathcal{P}_n(f)}{4R(f)} = \frac{1}{2}kT_e(f) \left[\frac{\text{Watt}}{\text{Hz}}\right]$$

Nel caso in cui $T_e(f) = T_0 = 290\ ^oK$ (temperatura ambiente), il termine $kT_0 = N_0 = 2\mathcal{W}_{dn}(f)$ assume i valori riportati di seguito per diverse unità di misura, da adottare in alternativa, allo scopo di rendere la grandezza omogenea con le altre che compaiono nelle formule di progetto:

$$kT_0 \quad = \quad \text{-204 [dBW/Hz]} \quad = \quad \text{-174 [dBm/Hz]} \quad = \quad \text{-114 [dBm/MHz]}$$

Ad esempio, all'uscita di un filtro passa-banda ideale non rumoroso[4] di estensione 1 MHz, si ha una potenza disponibile di rumore pari a $10^{-11.4}$ mW .

## 16.2 Rapporto segnale rumore dei generatori




Un generatore di tensione $V_g(f)$, che possiede una propria impedenza interna $Z_g(f)$ a temperatura equivalente $T_g(f)$, produce anch'esso un processo di rumore in virtù della componente reale $R_g(f) = \Re\{Z_g(f)\}$ di $Z_g(f)$, e $Z_g(f)$ può quindi schematizzarsi con il circuito equivalente mostrato in figura. Pertanto, oltre alla potenza disponibile di segnale $\mathcal{W}_{dg}(f) = \frac{\mathcal{P}_g(f)}{4R_g(f)}$, troviamo anche una potenza disponibile di rumore $\mathcal{W}_{d_n}(f) = \frac{1}{2}kT_g(f)$, e dunque un rapporto segnale rumore disponibile

$$SNR_g(f) = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)}$$

che come osserviamo dipende da $f$, sia a causa di $\mathcal{W}_{dg}(f)$ che di $T_g(f)$. Infine, notiamo che lo stesso valore di $SNR_g$ è esprimibile anche come rapporto tra le potenze *di segnale* anziché *disponibili*: infatti

$$SNR_g(f) = \frac{\mathcal{P}_g(f)}{4R_g(f)} \cdot \frac{1}{\frac{1}{2}kT_g(f)} = \frac{\mathcal{P}_g(f)}{2kT_g(f)R_g(f)} = \frac{\mathcal{P}_g(f)}{\mathcal{P}_n(f)}$$

## 16.3 Rumore nelle reti due porte

Se colleghiamo un generatore rumoroso a temperatura $T_g$ all'ingresso di una rete due porte a temperatura $T_Q$, è lecito aspettarsi all'uscita della rete un processo di rumore dipendente sia dal generatore che dalla rete, e la cui potenza disponibile $\mathcal{W}_{dn_u}(f)$ può essere espressa in funzione di una temperatura equivalente di uscita $T_{e_u}(f)$, tale che

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}kT_{e_u}(f)$$

[4] Si intende dire che il filtro non introduce altro rumore oltre quello di natura termica.

D'altra parte a $T_{e_u}(f)$ concorrono sia la temperatura del generatore $T_g(f)$, che la rete con una propria $T_{Q_u}(f)$ "equivalente di uscita"; scriviamo dunque

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}k\cdot[T_g(f)\,G_d(f) + T_{Q_u}(f)]$$

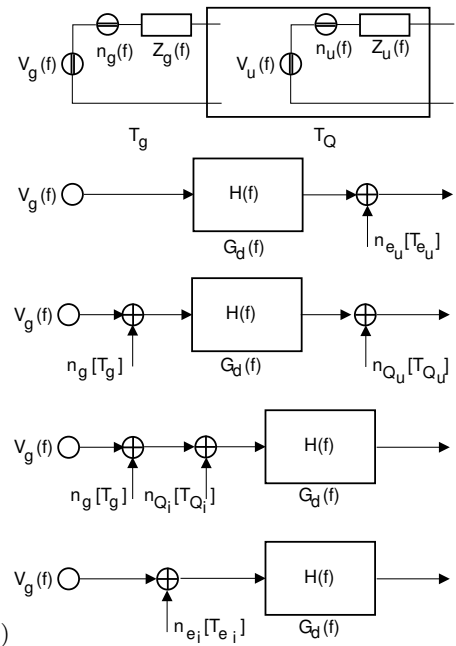
in cui la potenza disponibile in ingresso alla rete (che ha guadagno disponibile $G_d(f)$) è riportata in uscita, moltiplicata per $G_d(f)$. Se effettuiamo l'operazione inversa per il contributo di rumore dovuto a $T_{Q_u}$, otteniamo $\mathcal{W}_{dn_u}(f) = \frac{1}{2}kG_d(f)\cdot[T_g(f) + T_{Q_i}(f)]$ (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

$$T_{e_i}(f) = T_g(f) + T_{Q_i}(f) = T_g(f) + \frac{T_{Q_u}(f)}{G_d(f)} \tag{16.1}$$

è detta anche *temperatura di sistema* $T_s = T_{e_i}$, poiché riporta in ingresso alla rete tutti i contributi al rumore di uscita, dovuti sia al generatore che alla rete. Siamo però rimasti con un problema irrisolto: che dire a riguardo di $T_{Q_i}$ *e* $T_{Q_u}$?




### 16.3.1 Reti passive

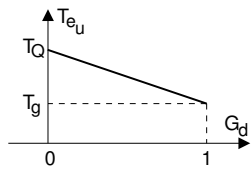
Supponiamo ora tutti i componenti della rete due porte alla stessa temperatura $T_Q$. In questo caso si può mostrare che risulta

$$\begin{cases} T_{Q_u}(f) &= [1 - G_d(f)]\,T_Q \\ T_{Q_i}(f) &= \frac{T_{Q_u}(f)}{G_d(f)} = [A_d(f) - 1]\,T_Q \end{cases}$$

in modo da poter scrivere:

$$\begin{cases} T_{e_u}(f) &= G_d(f)\,T_g(f) + T_{Q_u}(f) = G_d(f)\,T_g(f) + [1 - G_d(f)]\,T_Q \\ T_{e_i}(f) &= \frac{T_{e_u}(f)}{G_d(f)} = T_g(f) + [A_d(f) - 1]\,T_Q \end{cases}$$

Questo risultato evidenzia come per una rete passiva (con $0 \le G_d \le 1$), 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.



#### 16.3.1.1 Rapporto SNR in uscita

Se si valuta il rapporto segnale rumore in uscita alla rete, otteniamo

$$SNR_u(f) = \frac{\mathcal{W}_{dg}(f) G_d(f)}{\frac{1}{2}kT_{e_i}(f) G_d(f)} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}k \cdot [T_g(f) + [A_d(f) - 1] T_Q]}$$

Ricordando che il generatore in ingresso presenta un $SNR_i(f) = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)}$, possiamo valutare il peggioramento prodotto dalla presenza della rete:

$$\begin{aligned} \frac{SNR_i(f)}{SNR_u(f)} &= \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}kT_g(f)} \cdot \frac{\frac{1}{2}k \cdot [T_g(f) + [A_d(f) - 1] T_Q]}{\mathcal{W}_{dg}(f)} = \\ &= 1 + \frac{T_Q}{T_g(f)} \cdot [A_d(f) - 1] \end{aligned} \tag{16.2}$$

#### 16.3.1.2 Fattore di rumore per reti passive

Il coefficiente (16.2) $F(f) = 1 + \frac{T_Q}{T_g(f)} \cdot [A_d(f) - 1] \geq 1$ è chiamato *fattore di rumore*[5] della rete passiva, e rappresenta il peggioramento dell'$SNR$ dovuto alla sua presenza, potendo infatti scrivere

$$SNR_u(f) = \frac{SNR_i(f)}{F(f)} \leq SNR_i(f)$$

Notiamo subito che se $T_g(f) = T_Q$, allora $F = A_d$: pertanto una rete passiva che si trova alla stessa temperatura del generatore, esibisce un fattore di rumore pari all'attenuazione. Infatti, mentre la potenza disponibile di rumore è la stessa (essendo generatore e rete alla stessa temperatura), il segnale si attenua di un fattore $A_d$.

### 16.3.2 Reti attive

In questo caso il rumore introdotto dalla rete ha origine *non solo* dai resistori, e dunque *non è più vero* che $T_{Q_u}(f) = [1 - G_d(f)] T_Q$. Inoltre, il guadagno disponibile può assumere valori $G_d > 1$. In questo caso, si può esprimere l'$SNR$ in uscita dalla rete come

$$SNR_u(f) = \frac{\mathcal{W}_{dg}(f) G_d(f)}{\frac{1}{2}k [G_d(f) T_g(f) + T_{Q_u}(f)]} = \frac{\mathcal{W}_{dg}(f)}{\frac{1}{2}k \cdot [T_g(f) + T_{Q_i}(f)]}$$

ed il peggioramento individuato in (16.2) come

$$\frac{SNR_i(f)}{SNR_u(f)} = 1 + \frac{T_{Q_i}(f)}{T_g(f)} = F(f, T_g) \tag{16.3}$$

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

[5] 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.

#### 16.3.2.1 Fattore di rumore per reti attive

Viene posto pari a

$$F(f) = 1 + \frac{T_{Q_i}(f)}{T_0}$$

e rappresenta il peggioramento di $SNR$ causato dalla rete quando il generatore è a temperatura ambiente $T_0 = 290\ ^oK = 17\ ^oC$. 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. Questo ci permette dunque il calcolo di $T_{Q_i}(f) = T_0[F(f) - 1]$ che, sostituito nella (16.1), permette finalmente di valutare la temperatura di sistema come

$$T_{e_i}(f) = T_g(f) + T_0[F(f) - 1]$$

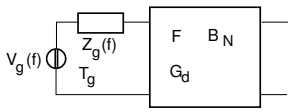
ed il peggioramento dell'$SNR$ come

$$\frac{SNR_i(f)}{SNR_u(f)} = 1 + \frac{T_0}{T_g(f)}[F(f) - 1]$$

**Riassunto**

- il fattore di rumore è definito come il peggioramento di $SNR$ dovuto alla presenza della rete tra generatore e carico, quando il generatore è a temperatura $T_0 = 290\ ^oK = 17\ ^oC$;
- dal fattore di rumore si deriva la temperatura di sistema $T_{e_i}(f) = T_g(f) + T_0[F(f) - 1]$;
- Se $T_g = T_0$ allora $T_{e_i}(f) = F(f)T_0$, e dunque la temperatura di sistema $T_{e_i}$ è $F(f)$ volte quella del generatore;
- Se la rete non è rumorosa si ottiene $F = 1$ (pari a 0 dB);
- Se la rete è passiva allora $F(f) = [A_d(f) - 1]\frac{T_Q}{T_0} + 1$, e se $T_Q = T_0$ allora $F = A_d$.

**Esempio** Sia data una rete due porte con assegnati guadagno disponibile $G_d$, banda di rumore $B_N$ e fattore di rumore $F$. Valutare il rapporto segnale rumore disponibile in uscita nei due casi in cui il generatore si trovi ad una generica temperatura $T_g$ oppure a $T_0$.




**Soluzione** Sappiamo che la densità di potenza disponibile di rumore in uscita vale

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}kT_{e_i}G_d = \frac{1}{2}k \cdot [T_g + T_{Q_i}] \cdot G_d$$

in generale $F = 1 + \frac{T_{Q_i}}{T_0}$ e quindi $T_{Q_i} = T_0(F - 1)$, dunque

$$\mathcal{W}_{dn_u}(f) = \frac{1}{2}k \cdot [T_g + T_0(F - 1)] \cdot G_d$$

Pertanto, la potenza disponibile di rumore si ottiene integrando la densità sulla banda di rumore

$$\mathcal{W}_{dn_u} = k \cdot [T_g + T_0(F - 1)] \cdot G_d B_N$$

che, nel caso in cui $T_g = T_0$, si riduce a $\mathcal{W}_{dn_u} = kT_0FG_dB_N$. 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:

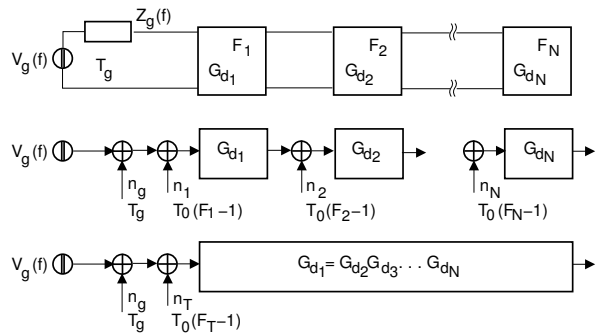
$$\begin{aligned} SNR_u &= \frac{SNR_i}{F\left(T_g\right)} = \frac{\mathcal{W}_{dg}}{kT_gB_N} \cdot \frac{1}{1+\frac{T_{Q_i}}{T_g}} = \frac{\mathcal{W}_{dg}}{kT_gB_N} \cdot \frac{1}{1+\frac{T_0(F-1)}{T_g}} = \\ &= \frac{\mathcal{W}_{dg}}{k\left[T_g + T_0\left(F-1\right)\right]B_N} = \frac{\mathcal{W}_{dg}}{kT_{e_i}B_N} \end{aligned}$$

### 16.3.3 Fattore di rumore per reti in cascata

Sappiamo che il guadagno disponibile dell'unica rete due porte equivalente alle $N$ reti poste in cascata, è pari al prodotto dei singoli guadagni, ovvero $G_d = \Pi_{n=1}^{N} G_{d_n}$. 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

$$T_{Q_i}^{(T)} = T_{Q_i}^{(1)} + T_{Q_i}^{(2)}\frac{1}{G_{d_1}} + T_{Q_i}^{(3)}\frac{1}{G_{d_1}G_{d_2}} + \cdots + T_{Q_i}^{(N)}\frac{1}{\Pi_{n=1}^{N-1}G_{d_n}}$$

in cui, sostituendo le espressioni per i $T_{Q_i}^{(n)}$ si ottiene

$$T_{Q_i}^{(T)} = T_0 \cdot \left[F_1 - 1 + \frac{F_2-1}{G_{d_1}} + \frac{F_3-1}{G_{d_1}G_{d_2}} + \cdots + \frac{F_N-1}{\Pi_{n=1}^{N-1}G_{d_n}}\right]$$

Applicando la definizione $F^{(T)} = 1 + \frac{T_{Q_i}^{(T)}}{T_0}$, si ottiene

$$F^{(T)} = F_1 + \frac{F_2-1}{G_{d_1}} + \frac{F_3-1}{G_{d_1}G_{d_2}} + \cdots + \frac{F_N-1}{\Pi_{n=1}^{N-1}G_{d_n}}$$

che costituisce proprio l'espressione cercata:

$$F^{(T)} = F_1 + \sum_{i=2}^{N} \frac{F_i - 1}{\Pi_{j=1}^{i-1} G_{d_j}}$$

Il risultato si presta alle seguenti considerazioni:

- la *prima rete* due porte deve avere $F$ *più piccolo possibile*, in quanto quest'ultimo non può essere ridotto in alcun modo e contribuisce per intero ad $F^{(T)}$;
- la *prima rete* due porte deve avere $G_d$ *più elevato possibile*, in quanto quest'ultimo divide tutti i contributi di rumore delle reti seguenti.

Pertanto l'elemento che determina in modo preponderante il rumore prodotto da una cascata di reti due porte è la *prima* rete della serie, ed il suo progetto deve essere eseguito con cura particolare, anche tenendo conto del fatto che le due esigenze sopra riportate sono spesso in contrasto tra loro. E' inoltre appena il caso di ricordare che l'espressione ottenuta *non è* in dB, mentre spesso $F$ è fornito appunto in dB; pertanto per il calcolo di $F^{(T)}$ occorre prima esprimere tutti gli $F_i$ in unità lineari.

**Esercizio** Un trasmettitore con potenza di 50 mW e portante 30 MHz, modula AM BLD PS un segnale con banda $\pm W = \pm$ 10 KHz. Qualora si desideri mantenere un $SNR$ in ricezione di almeno 25 dB, determinare la distanza che è possibile coprire adottando antenne isotrope, ed un ricevitore caratterizzato da un fattore di rumore $F = 10$ dB.

**Svolgimento** Assumendo che si verifichino le condizioni di massimo trasferimento di potenza, il valore desiderato $SNR = \frac{\mathcal{W}_R}{\mathcal{W}_N}$ può essere ottenuto se $\mathcal{W}_R = \mathcal{W}_N \cdot SNR = 2W \cdot \mathcal{W}_{dN_i} \cdot SNR = W \cdot kT_0F \cdot SNR$, e quindi

$$\mathcal{W}_R(\text{dBm}) = 10\log_{10}10^4(\text{Hz}) - 174(\text{dBm/Hz}) + F_{dB} + SNR_{dB} =$$
$$= 40 - 174 + 10 + 25 = -99 \text{ dBm}.$$

Il guadagno di sistema risulta allora pari a

$$G_s(\text{dB}) = \mathcal{W}_T(\text{dBm}) - \mathcal{W}_R(\text{dBm}) = 10\log_{10}50 + 99 = 17 + 99 = 116 \text{ dB}$$

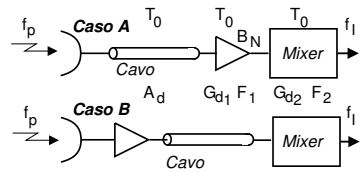
Non prevedendo nessun margine, l'attenuazione dovuta alla distanza è numericamente pari al guadagno di sistema, e pertanto scriviamo

$$A_d = 116 = 32.4 + 20\log_{10} f(\text{MHz}) + 20\log_{10} d(\text{Km}) =$$
$$= 32.4 + 29.5 + 20\log_{10} d(\text{Km})$$

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.

**Esercizio** Una trasmissione video modulata AM-BLU con portante $f_p = 2$ GHz viene ricevuta secondo uno dei due schemi in figura, indicati come caso **A** e **B**.

E' presente una discesa in cavo coassiale con $\phi = 1.2/4.4$ mm lunga 50 metri, un filtro-amplificatore con guadagno disponibile $G_{d_1} = 20$ 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 $T_0 = 290\ ^oK$. Calcolare:



**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 $T_a = 10\ ^oK$ 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 $Z_u = Z_a = 50\ \Omega$, oppure con $Z_u = 50\ \Omega$ e $Z_a = 50 - j\ 50\ \Omega$.

**Svolgimento** Determiniamo innanzitutto l'attenuazione del cavo coassiale, che risulta $A_d(f) = A_0\sqrt{f\,(MHz)}$ dB/Km. Per il diametro indicato risulta $A_0 = 5.3$ dB/Km, ed alla frequenza di 2 GHz si ottiene $A_d(f)_{dB} = 5.3\sqrt{2 \cdot 10^3} = 237$ dB/Km; e quindi in 50 metri si hanno 11.85 $\simeq$ 12 dB. Dato che il cavo è a temperatura $T_0$, risulta anche $F_{cavo} = A_d = 12$ dB. Riassumendo:

| | $A_d = F_{cavo}$ | $F_1$ | $G_{d_1}$ | $F_2$ | $G_{d_2}$ |
|---|---|---|---|---|---|
| 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

$$W_{d_R} = SNR_i \cdot W_{d_N} = SNR_0 \cdot F^A \cdot B_N \cdot kT_0 \text{ e quindi}$$

$$W_{d_R}(\text{dBm}) = SNR_0(\text{dB}) + F^A(\text{dB}) + B_N(\text{dBMHz}) + KT_0(\text{dBm/MHz}) =$$
$$= 50 + 12.6 + 8.45 - 114 = -43 \text{ dBm}$$

**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 $T_a = 10\ ^oK \neq T_0$, 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

$$W_{d_T} = W_{d_R} + A_d = -54.3 + 109.5 = 55.2 \text{ dBm} = 25.2 \text{ dBW} \rightarrow 331 \text{ Watt}$$

**3)** Nel caso di adattamento, la potenza ceduta all'antenna $T_x$ è proprio quella disponibile del generatore, e quindi si ha $W_{d_T} = \frac{\sigma_g^2}{4R}$, 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 (14.1))

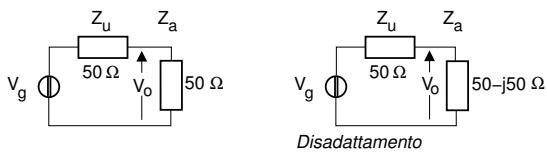
$$\mathcal{W}_T = \mathcal{P}_{v_o} \frac{R_a}{|Z_a|^2} = P_{v_o} \frac{50}{50^2+50^2} = P_{v_o} \cdot 10^{-2}$$

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.

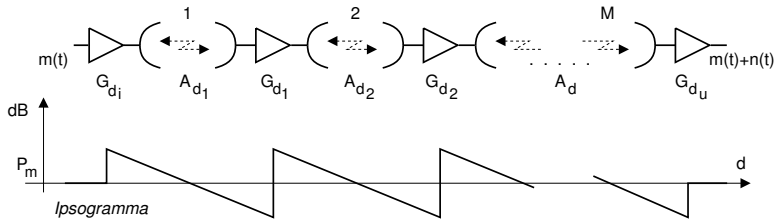
Evidentemente, il disadattamento produce un innalzamento del valore efficace, se si vuol mantenere la stessa potenza di uscita.




### 16.3.4 Rumore nei ripetitori

Analizziamo il problema con riferimento ad un collegamento radio, anche se la trattazione può essere estesa ad altre tecniche trasmissive. Esaminiamo il caso in cui il collegamento da effettuare sia molto lungo, tanto da impedirne la realizzazione mediante un'unica *tratta*, od a causa dell'eccessiva attenuazione disponibile, oppure per la mancanza di condizioni di visibilità. In tal caso, occorre suddividere il collegamento in più *tratte* (in numero di $M$) dimensionate in modo tale che ognuna ripristini il livello di segnale ad un valore pari a quello in ingresso alla tratta stessa (tranne ovviamente l'ultima).

Tra ogni coppia di tratte si trova un *ripetitore,* che amplifica il segnale in misura pari al proprio guadagno disponibile $G_{d_i} = \frac{1}{A_{d_i}}$, e pari cioè all'inverso dell'attenuazione disponibile della tratta precedente. Il rumore termico accumulato alla fine del




collegamento può calcolarsi con i metodi tradizionali, ma considerando che il livello di segnale è lo stesso per tutti i ripetitori, si ritrova il risultato ottenuto al § 14.7.1, come ora illustreremo. Faremo quindi notare l'influenza della distorsioni *di non linearità* nel progetto.

#### 16.3.4.1 Rumore termico accumulato

Osservando solamente il segnale in uscita dall'ultimo ripetitore, si può definire un $SNR$ complessivo come $SNR_T = \frac{\mathcal{P}_m}{\mathcal{P}_n}$. D'altra parte, il rumore complessivo è dovuto ai contributi di rumore introdotti dai singoli ripetitori: essendo questi ultimi indipendenti tra loro, la potenza di rumore accumulata è la somma delle singole potenze di rumore:

$$\mathcal{P}_n = \sigma_n^2 = E\left\{n^2(t)\right\} = E\left\{\left(\sum_i n_i(t)\right)^2\right\} = E\left\{\sum_i n_i^2(t)\right\} = \sum_i \sigma_{n_i}^2 = \sum_{i=1}^{M} \mathcal{P}_{n_i}$$

Osserviamo ora che per ogni singolo ripetitore può essere definito un proprio $SNR_i = \frac{\mathcal{P}_{m_i}}{\mathcal{P}_{n_i}}$, e quindi $\mathcal{P}_{n_i} = \frac{\mathcal{P}_{m_i}}{SNR_i}$. Pertanto l'$SNR$ complessivo risulta: $SNR_T = \frac{\mathcal{P}_m}{\sum_i \frac{\mathcal{P}_{m_i}}{SNR_i}}$. Notiamo ora che, essendo il rumore complessivo riferito ad un livello di segnale di riferimento, lo stesso deve avvenire per i singoli contributi $\mathcal{P}_{n_i}$, cosicché nell'ultima espressione occorre considerare $\mathcal{P}_{m_i} = \mathcal{P}_m$ con $\forall i$, fornendo in definitiva

$$SNR_T = \frac{\mathcal{P}_m}{\mathcal{P}_m \sum_i \frac{1}{SNR_i}} = \frac{1}{\sum_i \frac{1}{SNR_i}}$$

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 la resistenza equivalente di un parallelo di resistenze; l'analogia evidenzia, tra l'altro, che se una tratta è considerevolmente peggiore delle altre, $SNR_T$ dipenderà essenzialmente da questa.

Il risultato a cui siamo giunti ha validità più generale del caso illustrato, e può essere invocato ogni volta che un sistema di comunicazione è affetto da più cause di disturbo additivo indipendenti tra loro, per ognuna delle quali si sia separatamente in grado di giungere ad una espressione di $SNR$.

Proseguiamo l'analisi ipotizzando ora che tutte le tratte siano uguali tra loro, ovvero con eguali $A_d$ e $G_d$, uguali temperature di rumore, ed uguali $SNR_i$. In tal caso si ottiene

$$SNR_T = \frac{1}{\frac{M}{SNR_i}} = \frac{SNR_i}{M}$$

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 $\alpha$ è un fattore che dipende dal tipo di modulazione. Sembrerebbe dunque che per migliorare l'$SNR$ complessivo sia sufficiente elevare il livello di trasmissione di tutti gli stadi, in modo da elevare la potenza ricevuta. In realtà la potenza trasmessa non può aumentare a piacere, in quanto intervengono fenomeni di non-linearità.

#### 16.3.4.2 Compromesso tra rumore termico e di intermodulazione

A suo tempo si è osservato come per un segnale modulato, la presenza di un elemento a comportamento non lineare (tipicamente l'amplificatore di potenza del trasmettitore) produce interferenza in banda, la cui potenza dipende con legge cubica dalla potenza del segnale trasmesso. Indicando quindi con $SNR_I = \frac{\mathcal{P}_m}{\mathcal{P}_I}$ il rapporto $SNR$ complessivo del collegamento dovuto a cause di non linearità, *indipendenti dal rumore termico*, osserviamo che questo *diminuisce* all'aumentare della potenza trasmessa da ogni ripetitore. Pertanto, l'$SNR$ complessivo che tiene conto sia del rumore termico che di quello di intermodulazione, e che possiamo ottenere come "il parallelo" di entrambi, ossia

$$SNR = \frac{1}{\frac{1}{SNR_T} + \frac{1}{SNR_I}}$$

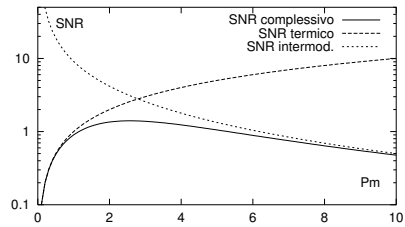
presenta un massimo per un certo valore di potenza trasmessa, ovvero esiste un dimensionamento ottimo in grado di fornire il miglior $SNR$ complessivo.

**Esempio**
La figura a lato mostra l'andamento di

$$SNR = \frac{1}{\frac{1}{SNR_T} + \frac{1}{SNR_I}}$$

dovuto ai due termini

$$SNR_T = \mathcal{P}_m \text{ e } SNR_I = \frac{\mathcal{P}_m}{.1 \cdot \mathcal{P}_m^2 + .01 \cdot \mathcal{P}_m^3}$$




Come si vede, $SNR$ presenta un massimo per $\mathcal{P}_m \simeq 2.5$.

# Capitolo 17

# Teoria dell'informazione e codifica

Sviluppiamo gli argomenti relativi alla codifica di sorgente e di canale, assieme alle basi teoriche che individuano le prestazioni-limite che possono essere conseguite da un sistema di trasmissione dell'informazione.

## 17.1 Codifica di sorgente

Una fonte informativa (o sorgente) può essere per sua natura di tipo discreto, come nel caso di un documento scritto, o di tipo continuo, come nel caso di un segnale analogico, ad esempio audio e video. In base a considerazioni di tipo statistico, la sorgente può essere caratterizzata da una grandezza, l'*Entropia*, che indica il tasso di informazione (in bit/secondo) *intrinseco* per i messaggi prodotti dalla sorgente; d'altra parte, la descrizione in modo nativo dei messaggi prodotti può determinare una *velocità di trasmissione* ben superiore!

Lo scopo della *codifica di sorgente* è quello di individuare rappresentazioni alternative per le informazioni prodotte dalla sorgente, in modo da ridurre la quantità di bit/secondo necessari alla trasmissione a valori quanto più possibile prossimi a quelli indicati dall'Entropia, sfruttando le caratteristiche della sorgente, del processo di codifica, e del destinatario dei messaggi, come

- la particolare distribuzione statistica dei simboli o dei valori emessi dalla sorgente, tale da permettere l'uso di meno bit per rappresentare i simboli *più frequenti* di altri;
- la dipendenza statistica presente tra simboli successivi, ovvero la presenza di un fenomeno di memoria intrinseco della sorgente, tale da rendere possibile entro certi limiti *la predizione* dei valori futuri;
- l'introduzione di un *ritardo di codifica* che permette di analizzare un intero intervallo temporale del messaggio;
- nel caso di segnali multimediali, l'esistenza di *fenomeni percettivi* legati alla fisiologia dell'apparato sensoriale, tali da guidare il codificatore nella scelta delle componenti di segnale da sopprimere, in quanto percettivamente non rilevanti.

Nel caso di sorgenti nativamente discrete, come ad esempio per documenti in formato elettronico, lo scopo della codifica di sorgente è quello di permettere la ricostruzione *integrale* di quanto trasmesso, che dunque in questo caso viene detta *senza perdita di*

*informazione*. Nel caso di sorgenti continue invece, si ottiene una sequenza numerica a seguito di un processo di campionamento e quantizzazione, che determina l'insorgenza di una prima causa di *distorsione* nel messaggio ricostruito, e la velocità binaria può essere ulteriormente ridotta grazie allo sfruttamento dei fenomeni percettivi: in tal caso il processo di codifica viene quindi detto *con perdita di informazione*.

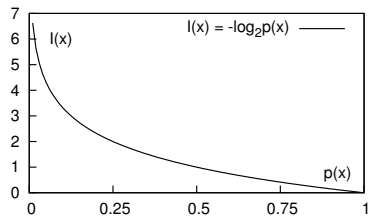
### 17.1.1 Codifica di sorgente discreta

**Sorgente senza memoria** Prendiamo in considerazione una sorgente discreta e stazionaria, che emetta una sequenza $x(n)$ composta di simboli $x_k$ appartenenti ad un alfabeto di cardinalità $L$ (ossia con $k = \{1, 2, \cdots, L\}$), ognuno contraddistinto dalla probabilità di emissione $p_k = Pr(x_k)$. Il termine *senza memoria* si riferisce al fatto che, se indichiamo con $x_h, x_k$ una coppia di simboli emessi uno dopo l'altro (ossia $x(n) = x_h$, $x(n+1) = x_k$), la probabilità del simbolo emesso per secondo non dipende dall'identità di quello(i) emesso precedentemente, ossia $p(x_k/x_h) = p(x_k) = p_k$.

**Misura dell'informazione** Definiamo informazione associata all'osservazione del simbolo $x_k$ il valore[1]

$$I_k = I(x_k) = \log_2 \frac{1}{p_k} = -\log_2 p_k \text{ bit}$$

che rappresenta il grado di incertezza a riguardo del verificarsi un evento, prima che questo si verifichi, ovvero di quanto possiamo ritenerci sorpresi nel venire a conoscenza di evento, di cui riteniamo di conoscere la probabilità. Osserviamo infatti che per come è fatta la funzione logaritmo, a bassi valori di probabilità è associata una informazione elevata. La scelta di usare il logaritmo in base 2 produce i risultati seguenti:




| Prob. $p_k$ | Informazione $-\log_2 p_k$ | Commento |
|---|---|---|
| 1 | 0 | L'evento certo non fornisce informazione |
| 0 | $\infty$ | L'evento impossibile dà informazione infinita |
| $\frac{1}{2}$ | 1 | In caso di scelta binaria (es. testa o croce) occorre una cifra binaria (*bit = binary digit*) per indicare il risultato |

#### 17.1.1.1 Entropia

Come in termodinamica al concetto di entropia si associa il grado di *disordine* in un sistema, così per una sorgente informative l'entropia misura il livello di *casualità* dei simboli emessi. Definiamo quindi *Entropia* (indicata con $H$) di una sorgente discreta $S$, il valore atteso della quantità di informazione apportata dalla conoscenza dei simboli da essa generati

$$H_s = E\{I_k\} = \sum_{k=1}^{L} p_k I_k = \sum_{k=1}^{L} p_k \log_2 \frac{1}{p_k} \text{ bit/simbolo} \tag{17.1}$$

[1] Per calcolare il logaritmo in base 2, si ricordi che $\log_2 \alpha = \frac{\log_{10} \alpha}{\log_{10} 2} \simeq 3.32 \log_{10} \alpha$.

che pesando in probabilità il valore di informazione associato ai diversi simboli, rappresenta il tasso medio di informazione per simbolo delle sequenze osservabili. Come mostrato nell'esercizio che segue, risulta che:

- se i simboli sono *equiprobabili* ($p_k = \frac{1}{L}$ con $\forall k$), la sorgente è *massimamente informativa*, e la sua entropia è la massima possibile per un alfabeto ad $L$ simboli, e pari a $H_{s_{Max}} = \frac{1}{L}\sum_{k=1}^{L} \log_2 L = \log_2 L$ bit/simbolo;
- se i simboli non sono equiprobabili, allora $H_s < \log_2 L$;
- se la sorgente emette sempre e solo lo stesso simbolo, allora $H_s = 0$.

Da queste osservazioni discende l'espressione riassuntiva

$$0 \leq H_s \leq \log_2 L \tag{17.2}$$

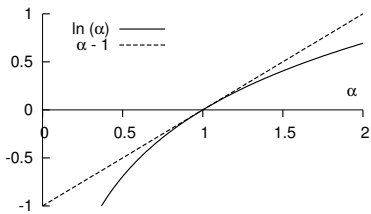
**Esercizio** Per dimostrare formalmente questo risultato, osserviamo innanzitutto che $H_s \geq 0$ in quanto la (17.1) comprende tutti termini positivi o nulli, essendo $\log_2 \alpha \geq 0$ per $\alpha = {}^1/_{p_k} \geq 1$. Quindi, mostriamo che $H_s - \log_2 L \leq 0$. Innanzitutto riscriviamo il primo membro della diseguaglianza come

$$H_s - \log_2 L = \sum_k p_k \log_2 \frac{1}{p_k} - \log_2 L \cdot \sum_k p_k = \sum_k p_k \log_2 \frac{1}{L \cdot p_k}$$

dato che $\sum_k p_k = 1$. Quindi, esprimiamolo in termini dei logaritmi *naturali*, tenendo conto che $\log_2 \alpha = \frac{\ln \alpha}{\ln 2}$, e quindi

$$\sum_k p_k \log_2 \frac{1}{L \cdot p_k} = \frac{1}{\ln 2} \sum_k p_k \ln \frac{1}{L \cdot p_k}$$

A questo punto utilizziamo la relazione $\ln \alpha \leq \alpha - 1$ mostrata in figura, con l'uguaglianza valida solo se $\alpha = 1$, ottenendo così




$$\begin{aligned} H_s - \log_2 L &= \frac{1}{\ln 2} \sum_k p_k \ln \frac{1}{L \cdot p_k} \leq \frac{1}{\ln 2} \sum_k p_k \left( \frac{1}{L \cdot p_k} - 1 \right) = \\ &= \frac{1}{\ln 2} \left( \sum_k \frac{1}{L \cdot} - \sum_k p_k \right) = \frac{1}{\ln 2} (1 - 1) = 0 \end{aligned}$$

con il segno di uguale solo se $\frac{1}{L \cdot p_k} = 1$ ovvero $p_k = \frac{1}{L \cdot}$.

**Entropia di sorgente binaria** Un caso particolare è quello delle sorgenti binarie, che producono uno tra due simboli $\{x_0, x_1\}$ con probabilità rispettivamente $p_0 = p$, $p_1 = q = 1 - p$, che inserite nella formula dell'Entropia, forniscono l'espressione

$$H_b(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \text{ bit/simbolo} \tag{17.3}$$

il cui andamento è mostrato nella figura 17.1, in funzione di $p$.

I due simboli $\{x_0, x_1\}$ possono essere rappresentati dalle 2 cifre binarie $\{0, 1\}$, che in questo caso chiamiamo *binit* (binary digit), per non confonderli con la misura dell'informazione (il bit). Osserviamo quindi che se $p \neq .5$, risulta che $H_b(p) < 1$, ossia la sorgente emette informazione con un tasso inferiore a un bit/simbolo, mentre a prima vista non potremmo usare meno di un binit per rappresentare ogni simbolo binario, introducendo una ridondanza pari a $1 - H_b(p)$ (graficata)[2].

[2]Si presti attenzione sulla differenza: la ridondanza della codifica *di sorgente* indica i binit/simbolo che eccedono il valore dell'entropia, mentre la ridondanza della codifica *di canale* indica il rapporto tra binit di protezione e quelli di effettivamente emessi dalla sorgente, come definito a pag. 5.3.3.1.

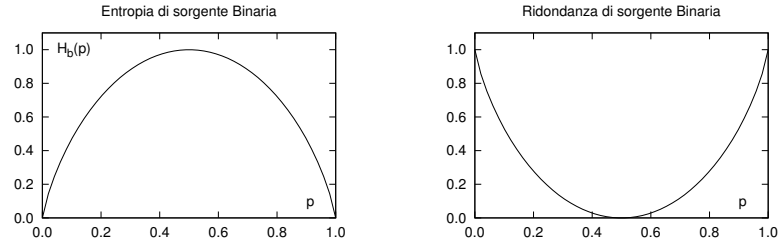



Figura 17.1: Entropia di sorgente binaria, e ridondanza associata

**Esempio** Consideriamo il caso di una sorgente con $p_0 = 0.8$ e $p_1 = 0.2$. L'applicazione della (17.3) fornisce un valore $H_b(0.8) = .8 \log_2 \frac{1}{.8} + .2 \log_2 \frac{1}{.2} = 0.72$ bit/simbolo, minore del valore di 1 bit/simbolo che si sarebbe ottenuto nel caso di equiprobabilità.

**Entropia di sorgente L-aria** L'applicazione della (17.2) al caso di una sorgente che emette simboli *non* equiprobabili ed appartenenti ad un alfabeto di cardinalità $L$, determina per la stessa un valore di Entropia $H_L < \log_2 L$ bit/simbolo: se i simboli sono codificati utilizzando $\lceil \log_2 L \rceil$ binit/simbolo[3], otteniamo una ridondanza pari a $\lceil \log_2 L \rceil - H_L$.

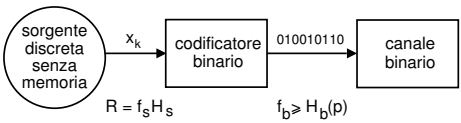
**Esempio** Consideriamo il caso di una sorgente quaternaria con $p_0 = 0.5$, $p_1 = 0.25$, $p_2 = 0.125$, $p_3 = 0.125$. L'applicazione della (17.1) fornisce $H_4 = 1.75$ bit/simbolo, inferiore ai 2 bit/simbolo che si sarebbero ottenuti nel caso di simboli equiprobabili.

#### 17.1.1.2 Tasso informativo e codifica binaria

Abbiamo fin qui preso in esame sorgenti discrete e senza memoria, caratterizzate da una entropia $H_s$ bit/simbolo, ed i cui simboli $x_k$ sono emessi ad una frequenza $f_s$ simboli/secondo, dando luogo ad un flusso informativo di intensità

$$R = f_s \cdot H_s \quad \text{bit/secondo}$$

Volendo trasmettere questa informazione attraverso un canale binario (vedi § 17.2.1), occorre che l'elemento indicato come *codificatore binario di sorgente* faccia corrispondere ad ogni simbolo $x_k$ una univoca sequenza di $N_k$ bit[4], scelti in uno dei modi descritti appresso, producendo una velocità di trasmissione binaria $f_b$. Dal punto di vista del canale, il messaggio è prodotto da una nuova sorgente, i cui simboli binari hanno probabilità $p$ e $1 - p$ non necessariamene pari ad 1/2, e caratterizzata da una entropia $H_b(p) \leq 1$, la cui velocità di trasmissione rispecchia il vincolo




$$f_b \geq H_b(p) = f_s \cdot H_s = R$$

in quanto il codificatore non altera il tasso informativo in transito. Il rapporto

[3] La notazione $\lceil \alpha \rceil$ indica l'intero superiore ad $\alpha$: ad esempio, se $\alpha = 3.7538$, si ha $\lceil \alpha \rceil = 4$.

[4] L'esatta corrispondenza tra i diversi simboli di sorgente e la loro codifca binaria è detta *codifica per blocchi*, discussa al § 17.1.1.4, dove si mostra anche la possibilità di emettere le parole di codice in corrispondenza di più di un simbolo di sorgente.

$\bar{N} = \frac{f_b}{f_s} \geq H_s$ rappresenta il numero *medio* di cifre binarie emesse per ciascun simbolo della sorgente, ed è valutato a partire dalle probabilità $p_k$ dei simboli $x_k$ come $\bar{N} = \sum_k p_k N_k$.

Il primo teorema di Shannon della *codifica di sorgente* afferma che *esiste* un modo di scegliere gli $N_k$ binit associati ai simboli $x_k$ tale che[5]

$$H_s \leq \bar{N} \leq H_s + \epsilon \tag{17.4}$$

con $\epsilon$ piccolo a piacere, e che si annulla in corrispondenza della codifica *ottima*, per la quale risulta $\bar{N} = H_s$. Ma non dice come fare. Ma intanto osserviamo che il rapporto

$$\eta = \frac{H_s}{\bar{N}} = \frac{R}{f_b} \leq 1$$

rappresenta quanti bit di sorgente sono trasportati da ogni binit di codifica, ovvero una misura della efficienza del processo di codifica binaria[6]. D'altra parte, sappiamo già che se i binit emessi a velocità $f_b$ assumono i valori 0 o 1 in modo equiprobabile, allora $H_b\left(\frac{1}{2}\right) = 1$, producendo $f_b = R$ e $\bar{N} = H_s$, e dunque il problema di individuare un codice ottimo sembrerebbe quello di trovare un insieme di *parole di codice* (dette CODEWORD) tali da rendere le cifre binarie equiprobabili, con il vincolo di mantenere il codice *decifrabile*, ovvero tale da rispettare la *regola del prefisso*. Ma andiamo con ordine.

#### 17.1.1.3 Codifica con lunghezza di parola variabile

| Simbolo | Prob. | Codeword | $N_k$ |
|---|---|---|---|
| $x_1$ | .5 | 0 | 1 |
| $x_2$ | .25 | 10 | 2 |
| $x_3$ | .125 | 110 | 3 |
| $x_4$ | .125 | 111 | 3 |

Sperimentiamo immediatamente come il vantaggio di usare codewords con un numero *variabile* di bit risieda nella possibilità di usare le codeword *più lunghe* per rappresentare i simboli *meno probabili*, ed usare pochi bit per i simboli più frequenti. Consideriamo come esempio una sorgente con alfabeto di cardinalità $L = 4$, ai cui simboli competono le probabilità riportate alla seconda colonna della tabella.In questo caso l'Entropia vale

$$\begin{aligned} H_s = \sum_k p_k \log_2 \frac{1}{p_k} &= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{2}{8}\log_2 8 \\ &= \frac{1}{2} + \frac{1}{2} + \frac{2}{8}3 = 1.75 \text{ bit/simbolo} \end{aligned}$$

Se il codificatore di sorgente adotta le corrispondenze mostrate nella terza colonna della tabella, a cui competono le lunghezze in binit riportate quarta colonna, il numero *medio* di binit/simbolo prodotto dalla codifica binaria risulta pari a

$$\overline{N} = E\{L\} = \sum_k N_k p_k = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{2}{8} = 1.75 \text{ binit/simbolo}$$

Osserviamo subito che il risultato ottenuto $H_s = \bar{N}$, ovvero il migliore possibile, non è per nulla scontato, e dipende sia dal tipo particolare delle $p_k$ dell'esempio, tutte potenze negative di due (essendo $0.5 = 2^{-1}$, $0.25 = 2^{-2}$, $0.125 = 2^{-3}$), sia dalla particolare scelta fatta per le codeword adottate.

[5] In effetti la (17.4) sussiste qualora il codificatore non operi indipendentemente su ogni simbolo di sorgente, ma più in generale possa emettere i binit in corrispondenza di sequenze di $x_k$ *via via più lunghe*.

[6] Ad esempio, un valore $\eta$=0.33 indica che ogni binit trasporta solo 1/3 di bit.

**Regola del prefisso** Perché un insieme di codewords possa essere usato come codice di sorgente, queste devono poter essere riconosciute come *distinte* presso il ricevitore, e questo è possibile a patto che nessuna di esse sia *uguale all'inizio* di una codeword più lunga. Si può mostrare che ciò è possibile purché le lunghezze $N_k$ delle codeword soddisfino la *diseguaglianza di Kratf*, espressa come

$$K = \sum_{k=1}^{L} 2^{-N_k} \leq 1 \tag{17.5}$$

**Esempio** Nella tabella sono riportati quattro possibili codici (*A,B,C,D*) per la sorgente quaternaria già discussa, assieme al corrispettivo valore di $\bar{N}$ e $K$.

| Simb. | $p_k$ | *A* | *B* | *C* | *D* |
|---|---|---|---|---|---|
| $x_1$ | .5 | 00 | 0 | 0 | 0 |
| $x_2$ | .25 | 01 | 1 | 01 | 10 |
| $x_3$ | .125 | 10 | 10 | 011 | 110 |
| $x_4$ | .125 | 11 | 11 | 0111 | 111 |
| $\bar{N}$ | | 2.0 | 1.25 | 1.875 | 1.75 |
| $K$ | | 1.0 | 1.5 | 0.9375 | 1.0 |

Il codice *A* corrisponde ad un codificatore particolarmente banale con $N_k = \bar{N}$ per tutti i $k$, dunque la (17.5) diviene $K = L2^{-\bar{N}} \leq 1$ ed è soddisfatta a patto che $\bar{N} \geq \log_2 L$: nel nostro caso essendo $L = 4$ ed $\bar{N} = 2$ si ottiene $\bar{N} = \log_2 L = 2$ e quindi $K = 1$, dunque il codice è decifrabile (anche perché a lunghezza fissa), ma non particolarmente efficiente, in quanto $\frac{H_s}{\bar{N}} \leq \frac{H_s}{\log_2 L}$ e nel nostro caso, anche se vale l'eguaglianza, si ottiene $\frac{H_s}{\bar{N}} = 0,875 < 1$; in questo caso, o in generale quando $H_s < \log_2 L$, si può realizzare una efficienza migliore ricorrendo ad un codice a lunghezza variabile.

Le codeword del codice *B* producono un valore $K = 1.25 > 1$, e dunque rappresentano un codice ambiguo[7]: difatti violano sia la regola del prefisso, che il limite (17.5). Il codice *C* invece è non ambiguo[8], essendo $K < 1$, ma presenta una efficienza $\frac{H_s}{\bar{N}} = 0,9\bar{3} < 1$ e dunque subottimale. Infine, il codice *D* è quello analizzato al precedente paragrafo, ed effettivamente risulta una scelta ottima, dato che $\frac{H_s}{\bar{N}} = 1$ e le sue codeword soddisfano la (17.5).

**Codice ottimo** Si può dimostrare che un codice ottimo, ossia per il quale $\frac{H_s}{\bar{N}}$=1, dove soddisfare la (17.5) con il segno di eguale, e perché questo accada, è necessario che le probabilità di simbolo abbiano valori $p_k = 2^{-N_k}$. In tal caso infatti, risultando $N_k = \log_2 \frac{1}{p_k}$, l'espressione che calcola $\bar{N} = \sum_k p_k N_k$ coincide con quella che fornisce $H_s = \sum_k p_k \log_2 \frac{1}{p_k}$, ovvero ogni simbolo è codificato con una parola di codice lunga tanti binit quanti sono i bit di informazione che trasporta. Per individuare un codice che *si avvicini* a questa proprietà si può realizzare la tecnica di *Huffman* presentata appresso, mentre per *modificare* le $p_k$ si ricorre alla *codifica per blocchi* di simboli.

**Codice di Huffman** Si tratta di un algoritmo che permette di realizzare un codice a lunghezza variabile che soddisfa la regola del prefisso, adotta codeword più lunghe

[7] Ad esempio, la sequenza 10110010 potrebbe essere interpretata come $x_3x_4x_1x_1x_3$ oppure $x_2x_1x_4x_1x_1x_2x_1$ o $x_3x_2x_2x_1x_1x_3$

[8] Nonostante il codice $C$ non soddisfi la regola del prefisso, non è ambiguo in quanto lo zero indica comunque l'inizio di una *nuova* codeword.

ai simboli meno probabili, ed uniforma, per quanto possibile, la probabilità delle cifre binarie. Si basa[9] sulla costruzione di un albero binario, i cui rami sono etichettati con 1 e 0, e può essere descritto come segue:

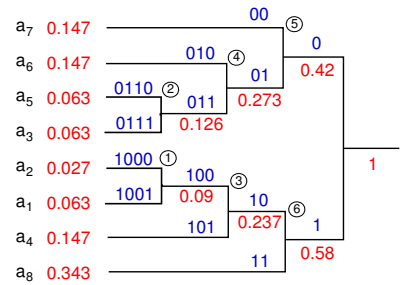
- crea una lista contenente i simboli della sorgente, ordinati per valore di probabilità decrescente, ed associa ad ognuno di essi un nodo-foglia dell'albero;
- finché c'è più di un nodo nella lista:
  - rimuovi dalla lista i due nodi con la probabilità più bassa;
  - crea un nuovo nodo interno all'albero con questi due nodi come figli, e con probabilità pari alla somma delle loro probabilità;
  - aggiungi il nuovo nodo alla lista;
- il nodo rimanente è la radice, e l'albero è completo;
- assegna cifre binarie diverse ad ogni coppia di rami a partire dalla radice, concatenando le quali ottieni le codeword per i simboli sulle foglie

Si può dimostrare che il codice di Huffman generato in questo modo è il migliore possibile nel caso in cui la statistica dei simboli di sorgente sia nota a priori, nel senso che produce una codifica con il minor numero possibile di binit/simbolo medi. La codifica di Huffman è ampiamente utilizzata nel contesto di altri metodi di compressione (metodo DEFLATE di PKZIP) e di codec multimediali (JPEG e MP3), in virtù della sua semplicità, velocità, ed assenza di brevetti.

Ovviamente ci deve essere un accordo a priori tra sorgente e destinatario a riguardo delle corrispondenze tra parole di codice e simboli (o blocchi di simboli) della sorgente. Nel caso in cui ciò non sia vero, oppure nel caso in cui la statistica dei simboli della sorgente sia stimata a partire dal materiale da codificare, occorre inviare all'inizio della comunicazione anche la tabella di corrispondenza, eventualmente in forma a sua volta codificata.

**Esempio** una sorgente con $L = 8$ simboli è caratterizzata dalle probabilità di simbolo riportate in figura, a partire dalle quali si realizza un codice di Huffman mediante la costruzione grafica riportata, in cui le probabilità sono scritte *sotto* i rami ed accanto ai simboli, mentre i binit *sopra*. Dopo aver ordinato i simboli in base alle probabilità, si individuano i due nodi a prob.

| Simbolo | $p_k$ | Codeword |
|---|---|---|
| $a_1$ | 0.027 | 1001 |
| $a_2$ | 0.063 | 1000 |
| $a_3$ | 0.063 | 0111 |
| $a_4$ | 0.147 | 101 |
| $a_5$ | 0.063 | 0110 |
| $a_6$ | 0.147 | 010 |
| $a_7$ | 0.147 | 00 |
| $a_8$ | 0.343 | 11 |




più bassa come $a_1$ e $a_2$, che assommano prob. 0.09; dunque la coppia ora meno probabile è $a_3$ con $a_5$, che cumulano prob. 0.126. Quindi, le due prob. minori divengono quelle di $a_4$ e della coppia $a_1a_2$, per un totale di 0.237, e poi è il turno di raggruppare $a_6$ con $a_3a_5$, per un totale di 0.273. Quest'ultimo nodo viene allora combinato con $a_7$ con prob. complessiva 0.42, quindi è il turno di $a_8$ con $a_1a_2a_4$ che totalizzano 0.58, ed infine questo nodo è collegato con il (5). A questo punto, partendo dalla radice a destra, si assegna un binit pari a 0 o 1 ad ogni coppia di rami, ripetendo l'assegnazione per le coppie discendenti, fino al risultato mostrato.

[9] Vedi `http://en.wikipedia.org/wiki/Huffman_coding`

**Dynamic Huffman coding** Questa variante permette di costruire e modificare l'albero di codifica[10] man mano che i simboli sono trasmessi. In questo modo, se un carattere è già presente nel codebook, viene trasmessa la codeword corrispondente, mentre se non lo è, viene trasmesso il codice del carattere, ed aggiornato il codebook. Lo stesso processo si svolge anche dal lato ricevente, permettendo una codifica in tempo reale, e l'adattamento a condizioni di variabilità nei dati. Ovviamente, il metodo inizia ad essere efficiente solo dopo aver accumulato sufficienti informazioni statistiche.

#### 17.1.1.4 Codifica per blocchi

Riprendiamo la discussione iniziata a pag. 408 relativamente al codice ottimo, notando che data una sorgente con simboli a probabilità $p_k$ è possibile scegliere delle codeword tali che

$$\log_2 \frac{1}{p_k} \leq N_k \leq \log_2 \frac{1}{p_k} + 1 \tag{17.6}$$

che si può mostrare soddisfano la condizione (17.5); moltiplicando ora i membri di (17.6) per $p_k$ e sommando su $k$ otteniamo

$$H_s \leq \bar{N} \leq H_s + 1$$

e quindi avremo $\frac{H_s}{\bar{N}} \simeq 1$ solo se $H_s \gg 1$ oppure se $N_k \simeq \log_2 \frac{1}{p_k}$. Altrimenti, possiamo ricorrere al *trucco* di *trasformare* la nostra sorgente ad $L$ simboli in una equivalente con $L^n$ simboli, ottenuti concatenando $n$ simboli della sorgente originaria. Dato che questi ultimi sono indipendenti, la nuova sorgente esibisce una entropia $H_s^{blocco} = nH_s$, e la regola di codifica (17.6) produce ora il risultato $nH_s \leq n\bar{N} \leq nH_s + 1$ in cui $n\bar{N}$ è il numero medio di binit per blocco. Dividendo per $n$, otteniamo infine

$$H_s \leq \bar{N} \leq H_s + \frac{1}{n} \tag{17.7}$$

che rappresenta una diversa forma del teorema (17.4) con $\epsilon = \frac{1}{n}$, e che permette di ottenere $\bar{N} \to H_s$ se $n \to \infty$, avvicinandosi alle condizioni di codifica ottima per qualsiasi distribuzione delle $p_k$.

Per applicare questo metodo ad un caso pratico, consideriamo una sorgente binaria senza memoria con le $p_k$ mostrate in tabella, e raggruppiamone i simboli a coppie, a cui competono probabilità di emissione ottenute moltiplicando le probabilità originarie[11], e quindi codifichiamo i nuovi simboli quaternari con il codice a lunghezza variabile già esaminato.

| Simbolo | Prob. | Codeword |
|---|---|---|
| $x_1$ | .8 | 1 |
| $x_2$ | .2 | 0 |
| $x_1x_1$ | .64 | 0 |
| $x_1x_2$ | .16 | 10 |
| $x_2x_1$ | .16 | 110 |
| $x_2x_2$ | .04 | 111 |

Mentre il valore dell'entropia della sorgente binaria è ancora quello calcolato nel primo esempio a pag. 17.1.1.1 e pari a $H_b = 0.72$ bit a simbolo, la lunghezza media del nuovo codice ora risulta pari a

$$\overline{N} = 1 \cdot 0.64 + 2 \cdot 0.16 + 3 \cdot 0.16 + 3 \cdot 0.04 = 1.58 \text{ binit}$$

ogni 2 simboli, ossia pari ad una media di 0.79 binit/simbolo, effettivamente più vicina al valore di $H_b = 0.72$.

[10] Presso *Wikipedia* si trova una descrizione dell'algoritmo di VITTER `http://en.wikipedia.org/wiki/Adaptive_Huffman_coding`

[11] operazione resa possibile in conseguenza della indipendenza statistica tra i simboli binari

Come indicato dalla (17.7), realizzando blocchi via via più lunghi è possibile ridurre la velocità media di codifica $\overline{N}$ (in binit/simbolo) rendendola sempre più vicina all'Entropia, ovvero

$$\min\left[\bar{N}\right] = H_s + \varepsilon$$

in cui $\varepsilon \to 0$ se la lunghezza del blocco tende ad infinito. D'altra parte, all'aumentare della dimensione del blocco aumenta di egual misura il *ritardo* che intercorre tra l'emissione di un simbolo e la sua codifica, e di questo va tenuto conto, nel caso sussistano dei vincoli temporali particolarmente stringenti sulla consegna del messaggio.

**Riassumendo** Qualora una sorgente discreta ad $L$ simboli esibisca un valore di entropia inferiore a $\log_2 L$, il flusso informativo dal codificatore binario di sorgente può essere ridotto adottando una codifica a blocchi, e calcolando per i nuovi simboli un opportuno codice di Huffman.

**Esercizio** Si ripeta il calcolo del numero medio di binit/simbolo, adottando lo stesso codice a lunghezza variabile usato finora, per codificare i simboli emessi dalla sorgente binaria Markoviana di primo ordine analizzata all'esempio seguente, e mostrare come in questo caso si riesca ad ottenere una velocità media pari a $0.72$ bit/simbolo. Sperimentare quindi la costruzione di un codice di Huffman basato sul raggruppamento di tre simboli di sorgente, per verificare se ci si riesce ad avvicinare di più al valore limite indicato dalla entropia, come vedremo pari a $0.58$ bit/simbolo.

#### 17.1.1.5 Sorgenti con memoria

Rimuoviamo ora l'ipotesi di indipendenza statistica tra i simboli emessi. In questo caso indichiamo con $\mathbf{x} = \{x(1), x(2), \ldots, x(N)\}$ una sequenza di $N$ di simboli, la cui probabilità congiunta risulta essere

$$p(\mathbf{x}) = p(x_1)\, p(x_2/x_1)\, p(x_3/x_1, x_2) \ldots p(x_N/x_1, x_2, \ldots x_{N-1}) \neq \prod_{k=1}^{N} p(x_k)$$

dato che appunto la dipendenza statistica comporta l'uso delle probabilità condizionali. In questo caso, l'espressione dell'entropia si modifica in

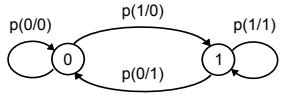
$$H_N = E_{\mathbf{x}}\{I(\mathbf{x})\} = -\frac{1}{N} \sum_{tutti\ gli\ \mathbf{x}\ possibili} \sum \sum \sum \ldots \sum p(\mathbf{x}) \log_2 p(\mathbf{x}) \text{ bit/simbolo}$$

in modo da eseguire la media statistica su tutte le possibili sequenze $\mathbf{x}$ di lunghezza $N$. $H_N$ è indicata come *entropia a blocco*, e si dimostra che al crescere di $N$ il suo valore è non crescente, ossia $H_{N+1} \leq H_N \leq H_{N-1}$, mentre per $N \to \infty$, $H_N$ tende ad un valore $H_\infty \leq H_s$, in cui l'uguaglianza è valida solo per sorgenti senza memoria.

**Sorgente Markoviana** Se oltre ad un certo valore $N_{Max}$ la sequenza $H_N$ non decresce più, allora la sorgente è detta a *memoria finita* o *di Markov* di ordine $N_{Max}$, caratterizzata dal fatto che le probabilità condizionate dipendono solo dagli ultimi $N_{Max}$ simboli emessi.

**Esempio** Analizziamo il caso di una sorgente binaria di Markov del primo ordine, per la quale sono definite le probabilità

$$\begin{array}{lcll} p(0/0) & = & 0.9 & p(1/0) = 0.1 \\ p(0/1) & = & 0.4 & p(1/1) = 0.6 \end{array}$$




ed a cui corrisponde il diagramma di transizione mostrato. In questo caso, l'ultimo simbolo emesso determina *lo stato* della sorgente, condizionando così i valori delle probabilità di emissione di un nuovo simbolo: con i valori dell'esempio, si osserva come la sorgente *preferisca* continuare ad emettere l'ultimo simbolo prodotto, piuttosto che l'altro.

Sotto un certo punto di vista, è come se la sorgente binaria si fosse *sdoppiata*, esibendo due diverse statistiche in base allo stato in cui si trova. Perciò, in questo caso l'entropia di sorgente può essere calcolata applicando la (17.3) ad ognuno dei due stati, ottenendo dei valori di *Entropia condizionata*, che sono poi mediati statisticamente, pesandoli con le probabilità di trovarsi in ognuno degli stati del modello Markoviano. Tornando all'esempio, i valori di entropia condizionata risultano pari a

$$\begin{array}{lcl} H(x/0) & = & -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.47 \\ H(x/1) & = & -0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.97 \end{array}$$

bit/simbolo, mentre il valore della probabiltà di trovarsi in uno dei due stati si ottiene risolvendo il sistema

$$\begin{cases} p(0) & = p(0/0)\, p(0) + p(0/1)\, p(1) \\ 1 & = p(0) + p(1) \end{cases}$$

in cui la prima equazione asserisce che la probabilità di trovarsi in $S_0$ è pari alla somma di quella di esserci già, per quella di emettere ancora zero, più la probabilità di aver emesso uno, ed ora emettere zero. Sostituendo i valori, si ottiene $p(0) = 0.8$ e $p(1) = 0.2$, ossia gli stessi valori dell'esempio binario senza memoria. Ma mentre in quel caso il valore dell'entropia risultava pari a $0.72$ bit/simbolo, ora si ottiene

$$H = p(0)\, H(x/0) + p(1)\, H(x/1) = 0.58 \text{ bit/simbolo}$$

mostrando come la presenza di memoria aumenti la predicibilità delle sequenze emesse dalla sorgente.

**Esercizio** Si ripeta il calcolo dell'entropia per un modello di Markov del primo ordine, caratterizzato dalle probabilità $p(0) = p(1) = 0.5$ e $p(1/0) = p(0/1) = 0.01$, mostrando che in questo caso si ottiene una entropia di 0.08 bit/simbolo.

#### 17.1.1.6 Codifica per sorgenti con memoria

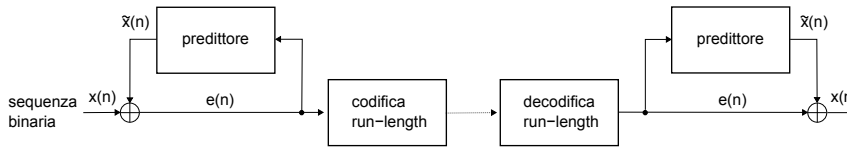
La discussione appena svolta ha evidenziato come nel caso di sorgenti con memoria i valori di entropia si riducano ulteriormente, e conseguentemente anche la velocità di codifica, a patto di accettare un ritardo ancor maggiore legato all'uso di codici a blocchi. A volte però la dimensione dei blocchi da prendere in considerazione può risultare eccessiva, producendo spropositate tabelle di codeword. Inoltre si può ritenere di *non* conoscere la statistica della sorgente, e non si desidera effettuarne una stima e quindi trasmetterla. In questi casi, può essere opportuno adottare tecniche diverse dalle precedenti, come le due riportate appresso.

**Codifica run-length** Prendendo come esempio tipico il caso della trasmissione fax, in cui si ha a che fare con un segnale di immagine in bianco e nero, scansionato per righe, che è assimilabile ad una sorgente binaria che emetta uno zero per il bianco, ed un uno per il nero: per la natura delle immagini scansionate, tipicamente ci saranno lunghe sequenze di uni o di zeri, e dunque si può assumere valido un modello di sorgente Markoviano di primo ordine, con elevate probabilità condizionate di restare nello stesso stato.

Le lunghe sequenze di bit tutti uguali vengono dette *run* (corse), e la codifica *run-length* consiste effettivamente nel trasmettere una parola di codice che indica il numero (*length*) di questi bit uguali. In questo caso quindi, la codeword è di lunghezza fissa (ad esempio $k+1$ binit, il primo dei quali indica se il run è tutto di uni o di zeri), e rappresenta un numero variabile (da 0 a $2^k - 1$) di binit di sorgente. Se ad esempio $k = 6$ binit, questi 6+1 = 7 binit possono codificare fino a 64 bit di immagine: un bel risparmio![12]

**Codifica predittiva** Questa ulteriore tecnica si basa sul fatto che un elevato grado di dipendenza statistica dei messaggi comporta la possibilità di *predire* in qualche modo i simboli a venire, in base all'identità di quelli già emessi. La differenza tra la sequenza predetta $\tilde{x}(n)$ e quella effettiva $x(n)$ è una nuova sequenza indicata come *errore di predizione* $e(n) = \tilde{x}(n) - x(n)$ che, se il predittore *ci azzecca* per la maggior parte del tempo, è quasi tutta nulla.

Nella figura seguente mostriamo l'applicazione della tecnica al caso di sequenze binarie, per le quali l'operazione di differenza è realizzata tramite una




somma modulo due $e(n) = \tilde{x}(n) \oplus x(n)$, in modo che nel caso di predizione corretta, l'errore sia nullo; possiamo verificare l'effettiva invertibilità del processo di predizione binaria confontando lo schema mostrato con quello proposto per la codifica differenziale al § 13.6.1.

Il predittore conserva uno *stato interno* che rappresenta gli ultimi bit di ingresso, in base ai quali determina[13] la stima $\tilde{x}(n)$; l'errore $e(n)$ che è frequentemente nullo, e che viene sottoposto a codifica run-length, è re-inserito anche nel predittore, in modo che questo possa ri-determinare l'ultimo simbolo di ingresso $x(n) = e(n) \oplus \tilde{x}(n)$, ed aggiornare il proprio stato interno. La medesima formula di ricostruzione viene applicata anche in uscita del predittore di ricezione, che condivide con quello di trasmissione la conoscenza dello stato iniziale del segnale trasmesso, in modo da evolvere allo stesso modo.

Notiamo come la codifica run-length non preveda l'esistenza di un accordo a priori tra trasmettitore e ricevitore, a parte il comune stato di partenza ad inizio messaggio, e la medesima struttura del predittore. Per contro, in presenza di errori di trasmissione i due predittori restano disallineati, finchè non si inizia a co-decodificare un nuovo messaggio. Ma lo stesso problema, è comune anche al caso di codifica a lunghezza di parola variabile, ed a quello di Huffman dinamico.

[12] In realtà, nel caso specifico del fax le cose non stanno esattamente in questi termini: infatti, anziché usare una parola di lunghezza fissa di $k$ binit, l'ITU-T ha definito un apposito codebook `http://www.itu.int/rec/T-REC-T.4-199904-S/en` che rappresenta un codice di Huffman a lunghezza variabile, in modo da codificare le run lenght più frequenti con un numero ridotto di bit.

[13] Il lettore più curioso si chiederà a questo punto, come è fatto il predittore. Molto semplicemente, *scommette* sul prossimo simbolo più probabile, in base alla conoscenza di quelli osservati per ultimi, ed ai parametri del modello markoviano: se questo simbolo possiede una probabilità a priori $> 0.5$, allora *la maggioranza* delle volte la predizione sarà corretta, ed il metodo inizia a consentire una riduzione di velocità. Nel caso di sorgenti continue, troveremo invece alcune particolarità aggiuntive.

#### 17.1.1.7 Compressione basata su dizionario

Nella comune accezione del termine, un dizionario è costituito da un *array* di stringhe, popolato con le parole esistenti per un determinato linguaggio. Anzichè operare carattere per carattere, un codificatore di sorgente testuale può ricercare la posizione nel dizionario delle parole del messaggio, e quindi trasmettere l'indice della parola: per un dizionario di 25.000 termini bastano 15 bit di indirizzamento, ottenendo un rapporto di compressione variabile, in funzione della lunghezza della parola codificata.

**Metodo di Lempel-Ziv-Welsh** Per evitare di dover condividere la conoscenza dell'intero dizionario tra sorgente e destinatario, che tra l'altro potrebbe essere assolutamente sovradimensionato rispetto alle caratteristiche dei messaggi da trattare, il metodo LZW prevede che il codificatore generi il dizionario in modo graduale, man mano che analizza il testo, e che il decodificatore sia in grado di replicare questa modalità di generazione. Inoltre, il dizionario non è vincolato a contenere le reali *parole* del messaggio, ma semplicemente ospita le sequenze di caratteri effettivamente osservati, di lunghezza due, tre, quattro...

Operando su di un alfabeto ad $L$ simboli, rappresentabili con $n = \lceil \log_2 L \rceil$ bit, il dizionario iniziale conterrà i simboli di sorgente alle prime $L$ posizioni, e posti liberi nelle restanti $2^n - L - 1$ posizioni[14].

Ogni carattere letto in ingresso viene accodato in una *stringa* ed il risultato confrontato con le stringhe già presenti nel dizionario. Nel caso non si verifichi nessuna corrispondenza, viene aggiunta una nuova voce di dizionario, e quindi viene trasmesso il codice associato alla sua parte iniziale, escludendo cioè il simbolo concatenato per ultimo, e che ha prodotto l'occorrenza della nuova voce.

```
w = NIL;
while (read a char c) do
  if (wc exists in dictionary) then
    w = wc;
  else
    add wc to the dictionary;
    output the code for w;
    w = c;
  endif
done
output the code for w;
```

Nel caso invece in cui la stringa sia già presente (e questo in particolare è vero per la stringa di lunghezza uno corrispondente al primo simbolo analizzato) non si emette nulla, ma si continuano a concatenare simboli fino ad incontrare una stringa mai vista. Presso Wikipedia[15] è presente un esempio di risultato della codifica.

La parte iniziale del testo, ovviamente, ha una alta probabilità di contenere tutte coppie di caratteri mai viste prima, e quindi in questa fase vengono semplicemente emessi i codici associati ai simboli osservati. Con il progredire della scansione, aumenta la probabilità di incontrare stringhe già osservate e sempre più lunghe. Ogni volta che viene esaurito lo spazio residuo per i nuovi simboli, viene aggiunto un bit alla lunghezza della codeword, ovvero viene raddoppiata la dimensione del vocabolario. Man mano che viene analizzato nuovo materiale, aumenta la lunghezza delle stringhe memorizzate nel dizionario, che riflette l'effettiva composizione statistica del documento in esame, ivi compresa la presenza di memoria; allo stesso tempo, la dimensione del dizionario (e la lunghezza delle codeword) resta sempre la minima indispensabile per descrivere il lessico effettivamente in uso. Alla fine del processo, il dizionario ottenuto viene aggiunto *in testa* al file compresso, seguito dalle codeword risultanti dall'algoritmo.

[14] Ad esempio con L=96 simboli si ha n=7 bit, ed un dizionario iniziale con 128 posizioni, di cui 96 occupate e 32 libere.

[15] `http://en.wikipedia.org/wiki/Lempel-Ziv-Welch`

L'algoritmo LZW è usato nel programma di compressione Unix `compress`, per la realizzazione di immagini GIF e TIFF, ed incorporato in altri software, come ad esempio *Adobe Acrobat*.

**Algoritmo Deflate** L'ultimo metodo di compressione senza perdite che esaminiamo è quello che è stato introdotto dal programma PKZIP, e quindi formalizzato nella RFC 1951[16], e tuttora ampiamente utilizzato per le sue ottime prestazioni e l'assenza di brevetti. Usa una variante dell'algoritmo LZW, al cui risultato applica poi una codifica di Huffman. *Deflate* opera su blocchi di dati con dimensione massima 64Kbyte, ognuno dei quali può essere replicato intatto (come nel caso in cui i bit siano già sufficientemente impredicibili), oppure essere compresso con un codice di Huffman statico, oppure ancora dinamico.

Per quanto riguarda la variante di LZW, essa consiste nel *non costruire* esplicitamente il dizionario, ma nell'usare invece *dei puntatori all'indietro* per specificare che una determinata sotto-stringa di ingresso, è in realtà la ripetizione di un'altra già osservata in precedenza. In questo caso, anziché emettere il codice (di Huffman) associato al byte corrente, si emette (il codice di Huffman del) la lunghezza della stringa da copiare, e la distanza (nel passato) della stessa. Quindi in pratica, anzichè usare una codeword di lunghezza fissa per indicizzare gli elementi del dizionario come per LZW, viene usato un puntatore di lunghezza variabile, privilegiando le copie della sottostringa corrente più prossime nel tempo, oppure quelle con un maggior numero di caratteri uguali.

### 17.1.2 Codifica con perdite di sorgente continua

Come verrà mostrato ai § 17.2.3 e 17.2.4, un segnale $r(t)$ di potenza $S$ ricevuto mediante un canale alla cui uscita sia presente un filtro con banda $W$ ed rumore $n(t)$ gaussiano bianco di potenza $N$ non può convogliare un tasso di informazione $R$ più elevato della *capacità di canale* pari a $C = W \log_2 \left(1 + \frac{S}{N}\right)$ bit per secondo. Nel caso discreto e senza perdite trattato finora, $C$ pone semplicemente un limite massimo al tasso di informazione che può essere trasmesso. Viceversa, nel caso di una sorgente continua $x(t)$, abbiamo due strade possibili: intraprendere un processo di campionamento e quantizzazione per produrre un segnale numerico con velocità $R \leq C$, incorrendo così in una distorsione di quantizzazione $D$ tanto minore quanto maggiore è la $R$ consentita, oppure effettuare una trasmissione analogica in cui

- il rumore al ricevitore può essere visto come una distorsione $d(t) = n(t) = x(t) - r(t)$ di potenza $D = N$
- l'entità della distorsione $D$ può essere ridotta aumentando $\frac{S}{N}$, e quindi $C$, permettendo così l'aumento del tasso di informazione $R$ trasferito.

Nel caso continuo pertanto, sia che la sorgente sia quantizzata o che sia trasmessa come segnale analogico, sussiste un legame diretto tra la distorsione ed il tasso informativo, di cui discutiamo ora.

#### 17.1.2.1 Curva velocità-distorsione

La valutazione della dipendenza tra $D$ ed $R$ si avvale di una serie di sviluppi teorici (che vengono omessi) che prendono come caso-tipo quello di una sorgente $x(t)$ gaussiana,

[16] `http://tools.ietf.org/html/rfc1951`

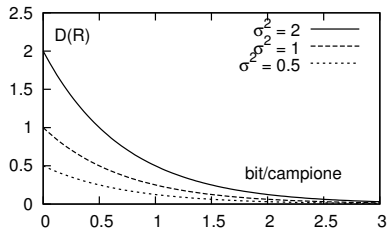
stazionaria, ergodica e bianca, con potenza $\sigma_x^2$. In tal caso si ottiene che la *minima distorsione* conseguibile in corrispondenza di una velocità di trasmissione (*rate*) pari ad $R$, assume un valore pari a

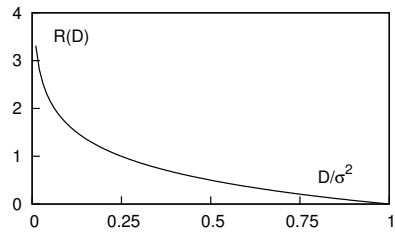
$$D(R)_G = 2^{-2R}\sigma_x^2 \tag{17.8}$$

Questa espressione risulta poi essere *il più grande valore minimo* possibile per una potenza $\sigma_x^2$ assegnata, dato che per sorgenti non gaussiane, oppure gaussiane ma non bianche, si possono ottenere valori inferiori. D'altra parte, invertendo la (17.8) si ottiene la curva $R(D)_G$, che descrive la minima velocità $R$ necessaria per trasmettere i campioni di una sorgente gaussiana con distorsione $D$ e potenza $\sigma_x^2$ assegnate

$$R(D)_G = \begin{cases} -\frac{1}{2}\log_2 \frac{D}{\sigma_x^2} & \text{se} \quad 0 \le D \le \sigma_x^2 \\ 0 & \text{se} \quad D \ge \sigma_x^2 \end{cases}$$

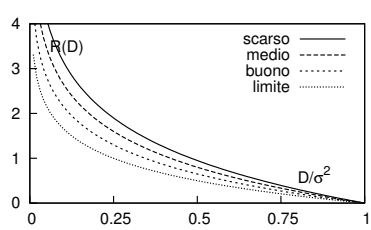
la cui seconda riga può essere interpretata osservando che, se la distorsione è superiore alla potenza di segnale, non occorre trasmettere proprio nulla, dato che tanto il ricevitore può ri-generare un segnale di errore, a partire da un rumore gaussiano di potenza $D$ prodotto in loco.




**Valori limite** Il valore $D(R)_G$ costituisce un *limite superiore* per ciò che riguarda la distorsione ottenibile ad una certa velocità $R$, utile per rapportare le prestazioni del codificatore della nostra sorgente con quelle migliori ottenibili per la sorgente più *difficile*, ossia la gaussiana. Allo stesso tempo, è definito un *limite inferiore* $D(R)_L$ (ossia, la minima distorsione sotto cui non si può scendere per un dato $R$) per sorgenti *non* gaussiane e *senza memoria*, in modo da poter scrivere

$$D(R)_L = 2^{-2R}Q \le D(R) \le D(R)_G = 2^{-2R}\sigma_x^2 \tag{17.9}$$




in cui $Q$ è la *potenza entropica* (17.12) e risulta $Q < \sigma_x^2$ per sorgenti non gaussiane.

In definitiva, per una determinata sorgente per la quale sono disponibili diversi codificatori, potremmo ottenere una famiglia di curve del tipo di quelle mostrate in figura.

#### 17.1.2.2 Entropia di sorgente continua

Nel caso discreto abbiamo apprezzato come l'entropia fornisca un potente strumento per valutare il tasso informativo intrinseco di una sorgente, supportando così la ricerca di metodi di riduzione della ridondanza al fine di avvicinare la velocità di trasmissione alla entropia effettiva. Ci chiediamo allora se anche nel caso continuo possa essere definita una entropia, e come questa possa aiutarci nello stabilire dei limiti prestazionali. Estendendo formalmente al caso continuo la definizione trovata per le sorgenti discrete, si ottiene l'espressione

$$h(X) = E\{-\log_2 p_x(x)\} = -\int p_x(x)\log_2 p_x(x)\,dx \tag{17.10}$$

che è indicata con la $h$ minuscola per distinguerla dal caso discreto, e che viene detta *entropia differenziale* o *relativa* perché il suo valore può risultare positivo, negativo o nullo, in funzione della dinamica della variabile aleatoria $X$.

**Esempio** Se calcoliamo il valore di entropia differenziale per un processo i cui valori sono descritti da una variabile aleatoria a distribuzione uniforme $p_x(x) = \frac{1}{A} rect_A(x)$, otteniamo il risultato $h(X) = -\frac{1}{A}\int_{-\frac{A}{2}}^{\frac{A}{2}} \log_2\left(\frac{1}{A}\right) dx = \log_2 A$ il cui valore effettivo, appunto, dipende dal valore di $A$.

Sebbene inadatta ad esprimere il contenuto informativo *assoluto*[17] di una sorgente continua, l'entropia differenziale può comunque essere utile per confrontare tra loro due sorgenti con uguale varianza $\sigma_x^2$; in particolare, il massimo valore di $h(X)$ per $\sigma_x^2$ assegnata è ottenuto in corrispondenza di un processo gaussiano[18], e risulta pari a

$$h(X)_G = \frac{1}{2}\log_2\left(2\pi e\sigma_x^2\right) > h(X) \tag{17.11}$$

ed è per questo motivo che, a parità di velocità $R$ e di potenza $\sigma_x^2$, il processo gaussiano incorre nella *massima distorsione minima* (17.8). Associata alla definizione di entropia differenziale, sussiste quella di *potenza entropica* $Q$, scritta come

$$Q = \frac{1}{2\pi e} 2^{2h(X)} \tag{17.12}$$

che per sorgenti gaussiane fornisce $Q_G = \sigma_x^2$, mentre per altri tipi di statistiche, si ottiene un valore minore. Applicando questa definizione alla (17.9) osserviamo come il limite inferiore di distorsione $D(R)_L$ si riduce al diminuire di $h(X)$.

[17] In effetti esiste una misura di entropia *assoluta* per sorgenti continue, che però ha la sgradevole caratteristica di risultare sempre infinita. Infatti, approssimando la (17.10) come limite a cui tende una sommatoria, e suddividendo l'escursione dei valori di $x$ in intervalli uguali $\Delta x$, possiamo scrivere

$$\begin{aligned} h_{abs}(x) &= \lim_{\Delta x\to 0}\sum_i p(x_i)\,\Delta x \log_2 \frac{1}{p(x_i)\,\Delta x} \\ &= \lim_{\Delta x\to 0}\sum_i \left[p(x_i)\,\Delta x \log_2 \frac{1}{p(x_i)} + p(x_i)\,\Delta x \log_2 \frac{1}{\Delta x}\right] = h(x) + h_0 \end{aligned}$$

in cui $h(x)$ è proprio la (17.10) mentre $h_0 = -\lim_{\Delta x\to 0}\log_2 \Delta x \int_{-\infty}^{\infty} p(x)\,dx = -\lim_{\Delta x\to 0}\log_2 \Delta x = \infty$. D'altra parte, la differenza tra le entropie assolute di due sorgenti $z$ e $x$ risulta pari a $h_{abs}(z) - h_{abs}(x) = h(z) - h(x) + h_0(z) - h_0(x)$, in cui la seconda differenza tende a $-\log_2 \frac{\Delta z}{\Delta x}$ che, se $z$ ed $x$ hanno la medesima dinamica, risulta pari a zero.

[18] Questo risultato si ottiene massimizzando la (17.10) rispetto a $p(x)$ mediante il metodo dei moltiplicatori di LAGRANGE http://it.wikipedia.org/wiki/Metodo_dei_moltiplicatori_di_Lagrange, in modo da tener conto dei vincoli espressi dalle condizioni $\int p_x(x)\,dx = 1$ e $\int x^2 p_x(x)\,dx = \sigma_x^2$. Notiamo esplicitamente la differenza rispetto al caso continuo, in cui invece la d.d.p. che rende massima l'entropia, è quella uniforme.

#### 17.1.2.3 Sorgenti con memoria

Come per il caso di sorgenti discrete, anche per quelle continue l'esistenza di una dipendenza statistica tra i valori prodotti riduce la quantità di informazione emessa, al punto che a parità di distorsione, questa può essere codificata a velocità ridotta; oppure, a parità di velocità, può essere conseguita una distorsione inferiore. Anche stavolta, la sorgente più *difficile* (ossia a cui compete la *massima distorsione minima*) è quella gaussiana, per la quale risulta che la minima distorsione per $\sigma_x^2$ assegnata può esprimersi come

$$D\left(R\right)_G = 2^{-2R}\gamma_x^2\sigma_x^2 \tag{17.13}$$

in cui $0 \leq \gamma_x^2 \leq 1$ rappresenta una misura di *piattezza spettrale*[19], che vale uno nel caso senza memoria, ovvero di processo *bianco*, e si riduce nel caso di un segnale i cui valori sono correlati tra loro, ed a cui corrisponde una densità spettrale *colorata*.

Nel caso *non gaussiano*, infine, la (17.13) si riscrive sostituendo al posto di $\sigma_x^2$ la potenza entropica $Q$ espressa dalla (17.12), ottenendo valori $D\left(R\right)$ ancora inferiori.

L'applicazione dei principi relativi alla codifica di sorgente al caso specifico dei messaggi multimediali (audio e video) viene trattata al capitolo 18.

## 17.2 Codifica di canale

Come già fatto osservare al § 5.3.3, lo scopo della codifica di canale è quello di ridurre il tasso di errore di una trasmissione numerica, ricorrendo alla aggiunta di ridondanza, ossia trasmettendo più binit di quanti necessari dopo l'attuazione della codifica di sorgente, e dunque (nel caso di una trasmissione *real-time*) occupare più banda dello stretto necessario.

E' quindi naturale chiedersi: fino a che punto si può arrivare, ovvero di quanto si può ridurre la $P_e$, e quanta ridondanza è necessario aggiungere? La risposta fornita in questa sezione è che finché il tasso informativo $R$ si mantiene inferiore al valore di una grandezza $C$ denominata *capacità di canale*, definita ai § 17.2.3 e 17.2.4, il canale può trasportare l'informazione (teoricamente) *senza errori!* Mentre se al contrario $R > C$, non è possibile trovare nessun procedimento in grado di ridurre gli errori - che anzi, divengono praticamente *certi*. Infine per quanto riguarda la ridondanza che occorre aggiungere, pur senza spiegare come fare, la teoria assicura che questa può essere resa *trascurabile!* Ma prima di approfondire l'enunciazione di questi risultati, svolgiamo alcune riflessioni su come

- il processo di decisione svolto dal lato ricevente di un canale numerico può basarsi oltre che sulla conoscenza della statistica di *come avvengano* gli errori, anche su quella che descrive *come sono emessi* i simboli della sorgente;
- il verificarsi di errori costituisce una *perdita di informazione*.

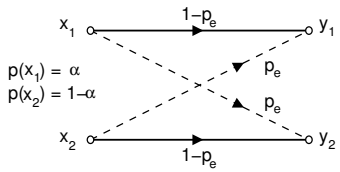
[19] Si può mostrare che $\gamma_x^2$ può essere interpretato come il rapporto tra la media aritmetica e la media geometrica della densità spettrale di potenza $\mathcal{P}_x\left(f\right)$ del processo $x\left(t\right)$: indicando con $S_k = \mathcal{P}_x\left(f_k\right)$, $k = 1, 2, \cdots, N$, i campioni equispaziati della densità spettrale valutati a frequenze positive $f_k$ tra zero e la massima frequenza del processo, si ha

$$\gamma_x^2 = \lim_{N\to\infty} \frac{\left(\prod_{k=1}^{N} S_k\right)^{1/N}}{\frac{1}{N}\sum_{k=1}^{N} S_k}$$

Nel caso di un processo bianco, per il quale i valori $S_k$ sono tutti uguali, le due medie coincidono, e $\gamma_x^2 = 1$. Altrimenti, $\gamma_x^2$ risulta tanto più piccolo quanto più i valori $S_k$ si discostano dal loro valore medio.

### 17.2.1 Canale binario simmetrico e decisore Bayesiano

In figura è mostrato uno schema che rappresenta un canale numerico al cui ingresso si può presentare uno tra due simboli $x_1$ e $x_2$, con probabilità rispettiva $\alpha$ e $1-\alpha$, mentre in uscita si osserva il simbolo $y_1$ oppure $y_2$.




Il canale è detto *simmetrico* perché tali sono le probabilità condizionate *in avanti*: la probabilità di errore

$$p_e = p\,(y_2/x_1) = p\,(y_1/x_2)$$

e la probabilità (complementare) di non-errore

$$p_{ne} = 1 - p_e = p\,(y_1/x_1) = p\,(y_2/x_2)$$

**Rapporto di verosimiglianza** Qualora si osservi in uscita uno dei due valori (ad es. $y_1$), si possono confrontare le probabilità condizionate *in avanti* per le due possibili *ipotesi* che in ingresso sia presente $x_1$ od $x_2$, valutando il rapporto

$$RV_{ML} = \frac{p\,(y_1/x_1)}{p\,(y_1/x_2)} = \frac{p_{ne}}{p_e} \tag{17.14}$$

e quindi si decide per l'ipotesi *piú verosimile* in funzione del valore maggiore o minore di uno per $RV_{ML}$: ad esempio se $p_{ne} > p_e$ (ovvero $p_e < \frac{1}{2}$) si ha $RV_{ML} > 1$, e quindi si decide per $x_1$, mentre nel caso fosse stato ricevuto $y_2$, si sarebbe deciso per $x_2$.

Se disponiamo della conoscenza delle probabilità *a priori* $p\,(x_1)$ e $p\,(x_2)$, ed i due simboli $x_1$ ed $x_2$ non sono equiprobabili[20], possiamo costruire il *rapporto di verosimiglianza* utilizzando le probabilità *a posteriori* $p\,(x_1/y_1)$ e $p\,(x_2/y_1)$, calcolabili applicando il teorema di Bayes (vedi § 7.1.4). Facendo di nuovo il caso di aver ricevuto il simbolo $y_1$, il *rapporto di verosimiglianza* si scrive ora come

$$RV_{MAP} = \frac{p\,(x_1/y_1)}{p\,(x_2/y_1)} = \frac{p\,(y_1/x_1)\,p\,(x_1)}{p\,(y_1)} \cdot \frac{p\,(y_1)}{p\,(y_1/x_2)\,p\,(x_2)} \tag{17.15}$$

$$= \frac{p\,(y_1/x_1)}{p\,(y_1/x_2)}\frac{p\,(x_1)}{p\,(x_2)} \tag{17.16}$$

Anche qui, $RV_{MAP}$ può assumere valore $>, <$ od $=$ ad 1, a seconda di quale delle due probabilità a posteriori sia piú grande, portando rispettivamente la decisione a favore di $x_1$, $x_2$, o l'indifferenza.

**Verifica di ipotesi ML e MAP** La metodologia ora descritta prende il nome di *verifica di ipotesi statistica* e si basa appunto sul confronto di quanto la grandezza osservata sia *verosimile,* compatibilmente con le *ipotesi* possibili. Se $RV$ utilizza solamente le probabilità in avanti (17.14), la decisione si dice *di massima verosimiglianza* (indicata come *ML* o Maximum Likelihood), mentre se si impiegano le probabilità *a posteriori* (17.15), si sta effettuando una *decisione di massima probabilità a posteriori* (indicata come MAP).

[20] In caso contrario (ovvero $x_1$ ed $x_2$ sono equiprobabili) la 17.15 è equivalente alla 17.14. Nei casi in cui *non si conosca* la statistica di sorgente, non si può fare altro che attuare una decisione di massima verosimiglianza.

**Riflessioni** Il meccanismo con cui, nella decisione MAP, le probabilità in avanti si combinano con quelle a priori, può essere analizzato mediante alcune osservazioni: innanzi tutto, $x_1$ potrebbe essere *cosí raro* che, in presenza di una moderata probabilità di errore, il ricevitore potrebbe preferire di decidere sempre $x_2$, attribuendo l'eventuale ricezione di $y_1$ dovuta piú verosimilmente ad un errore del canale, piuttosto che all'effettiva trasmissione di $x_1$. In assenza di canale poi, l'unico rapporto di verosimiglianza possibile sarebbe stato quello tra le probabilità a priori $p(x_1)$ e $p(x_2)$; la ricezione di un simbolo $y_i$ dunque ha portato un miglior livello di informazione, alterando il $RV$, in misura tanto maggiore quanto piú minore è la probabilità di errore.

**Esempio** Verifichiamo i ragionamenti appena svolti riscrivendo per esteso una probabilità a posteriori:

$$\begin{aligned} p(x_1/y_1) &= \frac{p(x_1,y_1)}{p(y_1)} = \frac{p(y_1/x_1)\,p(x_1)}{p(y_1/x_1)\,p(x_1)+p(y_1/x_2)\,p(x_2)} = \\ &= \frac{p_{ne}\cdot p(x_1)}{p_{ne}\cdot p(x_1)+p_e\cdot p(x_2)} \end{aligned}$$

Se $p_e = p_{ne} = \frac{1}{2}$, il canale è *inservibile* e non trasferisce informazione: infatti si ottiene $p(x_1/y_1) = p(x_1)$ in quanto $p(x_1)+p(x_2)=1$. D'altra parte, se $p_e < p_{ne}$, risulta

$$p(x_1/y_1) = \frac{p(x_1)}{p(x_1)+\frac{p_e}{p_{ne}}p(x_2)} > p(x_1)$$

aumentando quindi la probabilità di $x_1$ rispetto alla sua probabilità a priori; se poi la probabilità di errore tende a zero ($p_e \to 0$), osserviamo che $p(x_1/y_1) \to 1$.

### 17.2.2 Informazione mutua media per sorgenti discrete

Abbiamo discusso di come l'entropia permetta di valutare la capacità informativa di una sorgente; estendiamo ora il concetto, per mostrare come *l'informazione condivisa* tra ingresso ed uscita di un canale consenta di determinare anche la quantità di informazione che viene persa a causa degli errori che si sono verificati.

Consideriamo una sorgente discreta che emette simboli $x$ appartenenti ad un alfabeto finito di cardinalità $L$, ossia $x \in \{x_i\}$ con $i = 1, 2, \cdots, L$, a cui è associata la distribuzione $p(x_i)$, ed indichiamo con $y \in \{y_j\}$ (con $j = 1, 2, \cdots, L$) il simbolo ricevuto, in generale diverso da $x$, a causa di errori introdotti dal canale. Conoscendo le densità di probabilità $p(x_i)$, $p(y_j)$, e le probabilità congiunte $p(x_i, y_j)$, possiamo definire la quantità di informazione *in comune* tra $x_i$ e $y_j$, denominata *informazione mutua*, come[21]

$$I(x_i,y_j) = \log_2 \frac{p(x_i,y_j)}{p(x_i)\,p(y_j)} = \log_2 \frac{p(x_i/y_j)}{p(x_i)} = \log_2 \frac{p(y_j/x_i)}{p(y_j)} \quad \text{bit} \tag{17.17}$$

Notiamo che

- se ingresso ed uscita del canale sono indipendenti, si ha $p(x_i,y_j) = p(x_i)\,p(y_j)$, e quindi l'informazione mutua *è nulla;*
- se $p(y_j/x_i) > p(y_j)$, e quindi la conoscenza della emissione di $x_i$ rende la ricezione di $y_j$ più probabile di quanto non lo fosse a priori, allora l'informazione mutua è positiva;
- la definizione di informazione mutua è simmetrica, ovvero $I(x_i,y_j) = I(y_j,x_i)$.

[21] Per ottenere (17.17) si ricordi che $p(x_i,y_j) = p(x_i/y_j)\,p(y_j) = p(y_j/x_i)\,p(x_i)$

Per giungere ad una grandezza $I(X,Y)$ che tenga conto del comportamento del canale per qualsiasi simbolo di sorgente e ricevuto, occorre pesare i valori di $I(x_i,y_j)$ con le relative probabilità congiunte, ossia calcolarne il valore atteso rispetto a tutte le possibili coppie $I(X,Y) = E_{X,Y}\{I(x_i,y_j)\}$, e quindi

$$I(X,Y) = \sum_i \sum_j p(x_i,y_j) \log_2 \frac{p(x_i/y_j)}{p(x_i)} \tag{17.18}$$

$$= \sum_i \sum_j p(x_i,y_j) \log_2 \frac{p(y_j/x_i)}{p(y_j)} \tag{17.19}$$

ottenendo così la quantità denominata *informazione mutua media*, misurata in bit/simbolo, e che rappresenta (in media) quanta informazione ogni simbolo ricevuto trasporta a riguardo di quelli trasmessi. In virtù della simmetria di questa definizione, ci accorgiamo[22] che il suo valor medio può essere espresso nelle due forme alternative

$$I(X,Y) = H(X) - H(X/Y) \tag{17.20}$$

$$= H(Y) - H(Y/X) \tag{17.21}$$

in cui l'entropia *condizionale*

$$H(X/Y) = \sum_{i,j} p(x_i,y_j) \log_2 \frac{1}{p(x_i/y_j)} \tag{17.22}$$

prende il nome di *equivocazione*, e rappresenta la quantità media di informazione *persa*, rispetto all'entropia di sorgente $H(X)$, a causa della rumorosità del canale. Nel caso in cui il canale non introduca errori, e quindi $p(x_i/y_j)$ sia pari a 1 se $i = j$ e zero altrimenti, è facile vedere che $H(X/Y)$ è pari a zero, e $I(X,Y) = H(X)$, ossia tutta l'informazione della sorgente si trasferisce a destinazione. D'altra parte

$$H(Y/X) = \sum_{i,j} p(x_i,y_j) \log_2 \frac{1}{p(y_j/x_i)} \tag{17.23}$$

prende il nome di *noise entropy* dato che considera il processo di rumore come se fosse un segnale informativo: infatti, sebbene si potrebbe dire che l'informazione media ricevuta è misurata dalla entropia $H(Y)$ della sequenza di osservazione, una parte di essa $H(Y/X)$ *è falsa*, perché in realtà è introdotta dagli errori.

**Esercizio:** calcolare $I(X,Y)$ per il canale binario simmetrico. Mantenendo la notazione introdotta al § 17.2.1, usiamo la (17.21) per calcolare l'informazione mutua in funzione di $p(x_1) = \alpha$ e $p_e$, e dunque iniziamo con il valutare $H(Y)$ e $H(Y/X)$. Dal punto di vista dell'uscita del canale, i simboli $y_{1,2}$ costituiscono l'alfabeto di una sorgente binaria senza memoria, la

[22] Infatti

$$\sum_i \sum_j p(x_i,y_j) \log_2 \frac{p(x_i/y_j)}{p(x_i)} = \sum_i \sum_j p(x_i,y_j) \left[\log_2 \frac{1}{p(x_i)} - \log_2 \frac{1}{p(x_i/y_j)}\right] =$$

$$= \sum_i \sum_j p(x_i,y_j) \log_2 \frac{1}{p(x_i)} - \sum_i \sum_j p(x_i,y_j) \log_2 \frac{1}{p(x_i/y_j)}$$

e, saturando la sommatoria doppia del primo termine rispetto ad $j$, si ottiene la (17.20). Per la (17.21), il passaggio è del tutto simile.

cui entropia si esprime in termini di $p(y_1)$ mediante la (17.3), ovvero $H(Y) = H_b(p(y_1))$, in cui

$$\begin{aligned} p(y_1) &= p(y_1/x_1)\,p(x_1) + p(y_1/x_2)\,p(x_2) = \\ &= (1-p_e)\,\alpha + p_e\,(1-\alpha) = p_e + \alpha - 2\alpha p_e \end{aligned}$$

e dunque $H(Y) = H_b(p_e + \alpha - 2\alpha p_e)$. Per quanto riguarda la *noise entropy* $H(Y/X)$, sostituendo $p(x_i, y_j) = p(y_j/x_i)\,p(x_i)$ nella (17.23) otteniamo

$$I(Y/X) = \sum_i p(x_i) \left[ \sum_j p(y_j/x_i) \log_2 \frac{1}{p(y_j/x_i)} \right] = H_b(p_e)$$

dato che il termine tra parentesi quadre rappresenta appunto l'entropia di una sorgente binaria con simboli a probabilità $p_e$ e $1 - p_e$. Possiamo quindi ora scrivere l'espressione cercata

$$I(X,Y) = H_b(p_e + \alpha - 2\alpha p_e) - H_b(p_e)$$

che dipende sia dalla probabilità di errore $p_e$, sia dalla statistica dei simboli della sorgente: osserviamo che se $p_e \ll 1$, il canale (quasi) non commette errori e risulta $I(X,Y) \simeq H_b(\alpha) = H(X)$, mentre se $p_e \to \frac{1}{2}$ allora $I(X,Y) \to 0$.

### 17.2.3 Capacità di canale discreto

I risultati ora mostrati, pur permettendo di valutare la perdita di informazione causata dai disturbi, dipendono sia dalle probabilità in avanti $p(y_j/x_i)$ che effettivamente descrivono il comportamento del canale, sia da quelle a priori $p(x_i)$, che invece attengono unicamente al tipo di sorgente in uso. Al contrario, vorremmo trovare una grandezza che esprima esclusivamente l'attitudine (o *capacità*) del canale a trasportare informazione, indipendentemente dalle caratteristiche della sorgente. Questo risultato può essere ottenuto provando a variare la statistica della sorgente in tutti i modi possibili, fino a trovare il valore

$$C_s = \max_{p(x)} I(X,Y) \quad \text{bit/simbolo} \tag{17.24}$$

che definisce la *capacità di canale* come il massimo valore dell'informazione mutua media, ottenuto in corrispondenza della migliore sorgente possibile[23]. Il pedice $_s$ sta per *simbolo*, e serve a distinguere il valore ora definito da quello che esprime il massimo *tasso* di trasferimento dell'informazione espresso in bit/secondo, ottenibile una volta nota la frequenza $f_s$ con cui sono trasmessi i simboli, fornendo per la capacità di canale il nuovo valore[24]

$$C = f_s \cdot C_s \quad \text{bit/secondo} \tag{17.25}$$

L'importanza di questa quantità risiede nel *teorema fondamentale per canali rumorosi* (non dimostrato) già anticipato più volte, che asserisce

Per ogni canale discreto senza memoria di capacità $C$

- esiste una tecnica di codifica che consente la trasmissione di informazione a velocità $R$ e con probabilità di errore per simbolo $p_e$ piccola a piacere, purché risulti $R < C$;

[23] In definitiva, questo modo di ottenere una grandezza rappresentativa del solo canale ricorda un pò la via per la quale si è definita ad es. la potenza *disponibile* di un generatore, al variare di tutti i possibili valori di impedenza di carico.

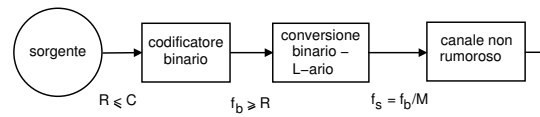
[24] Osserviamo l'invarianza di (17.25) rispetto alla modifica del numero di livelli: se $M$ bit sono raggrupati per generare simboli ad $L = 2^M$ livelli, come noto $f_s$ si riduce di $M$ volte, mentre $C_s$ aumenta della stessa quantità, dato che ogni simbolo trasporta ora $M$ bit anziché uno.

- se è accettabile una probabilità di errore $p_e$, si può raggiungere (con la miglior codifica possibile) una velocità $R(p_e) = \frac{C}{1-H_b(p_e)} > C$ in cui $H_b(p_e)$ è l'entropia di una sorgente binaria (17.3);
- per ogni valore di $p_e$, non è possibile trasmettere informazione a velocità maggiore di $R(p_e)$

Il teorema non suggerisce come individuare la tecnica di codifica, né fa distinzioni tra codifica di sorgente e di canale, ma indica le prestazioni limite ottenibili mediante la migliore tecnica possibile, in grado di ridurre a piacere la $p_e$ purché $R < C$, mettendoci al tempo stesso in guardia a non tentare operazioni impossibili. Da questo punto di vista, le prestazioni conseguibili adottando le tecniche di codifica note possono essere valutate confrontandole con quelle *ideali* predette dal teorema. Inoltre, dato che la capacità di canale è definita come massimo valore di $I(X,Y)$ per la migliore $p(x)$, qualora la statistica dei messaggi prodotti dal codificatore di sorgente differisca da quella ottima per il canale, l'informazione mutua media risulterà ridotta, così come la massima velocità $R$.

Illustriamo l'applicazione di questi risultati con un paio di esempi:

**Canale $L-$ario non rumoroso** Consideriamo il caso mostrato alla figura seguente, che rappresenta un canale che trasporta *senza errori* simboli con $L = 2^M$ livelli: in tal caso l'equivocazione $H(Y/X)$ è nulla, e la (17.20) permette di scrivere $I(X,Y) = H(X)$, che è massima se $P(x_i) = 1/L$ per tutti gli $i$, risultando così $C_s = \log_2 L = M$ bit/simbolo, e $C = f_s \cdot C_s = f_s \cdot M$ bit/secondo.




I simboli ad $L$ livelli sono ottenuti a partire da $M$ bit prodotti da una codifica binaria a velocità $f_b$, risultando $f_b \geq R = H_x$ in funzione della ottimalità o meno del codificatore; pertanto, risulta $R \leq f_b = f_s M = C$ con l'uguaglianza valida nel caso in cui il codificatore riesca a rimuovere tutta la ridondanza dei messaggi della sorgente, conseguendo in tal caso il massimo trasferimento di informazione.

Al contrario, volendo realizzare una velocità $R > C$, il codificatore di sorgente dovrebbe produrre codeword con lunghezze tali da violare la disuguaglianza di Kraft (17.5)[25], e quindi la regola del prefisso non sarebbe rispettata, causando in definitiva errori di decodifica anche in assenza di rumore!

**Capacità del canale simmetrico binario** Esaminiamo l'effetto della presenza di rumore per questo caso già studiato, e per il quale abbiamo valutato che l'espressione dell'informazione mutua media risulta

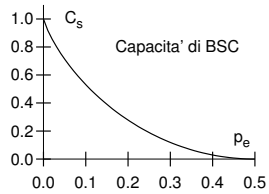
$$I(X,Y) = H_b(p_e + \alpha - 2\alpha p_e) - H_b(p_e)$$

in cui $H_b(p_e)$ dipende solo dalla probabilità di errore, mentre il termine $H_b(p_e + \alpha - 2\alpha p_e)$ dipende anche dalla statistica di sorgente, e risulta massimizzato e pari ad 1 se $p_e + \alpha - 2\alpha p_e = \frac{1}{2}$, come risulta per qualunque $p_e$ se $\alpha = \frac{1}{2}$, ossia per simboli equiprobabili. Pertanto, la capacità in questo caso è

$$C_s = 1 - H_b(p_e)$$

[25] 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

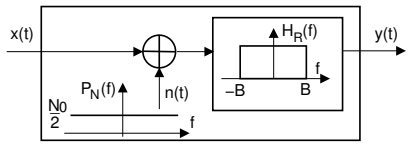
ed il suo andamento è rappresentato in figura[26], evidenziando che $C_s \simeq 1$ se $p_e \ll 1$, ma decade rapidamente a zero se $p_e \to 0.5$.




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, e di preferire tra queste le tecniche che vi riescono mantenendo al minimo la quantità dei bit aggiuntivi. Infatti spesso il canale impone una velocità di trasmissione, parte della quale è impiegata per trasmettere i soli binit di protezione e non l'informazione della sorgente, riducendo di fatto il tasso $R$ effettivamente trasmesso.

### 17.2.4 Capacità per canali continui

Come noto, un canale numerico è in realtà una astrazione che ingloba internamente un codificatore di linea o *modem* che, a partire da una sequenza numerica, produce un segnale trasmissibile su di un canale analogico, che a sua volta può essere caratterizzato da un valore di capacità, espresso nei termini dei parametri che descrivono la trasmissione analogica soggiacente. Una situazione tipica è quella rappresentata in figura, in cui al segnale ricevuto è sommato un rumore $n(t)$ gaussiano, bianco e a media nulla, mentre il filtro di ricezione $H_R(f)$ impone una limitazione di banda $2B$, in modo che la potenza di rumore in ingresso al decisore vale $P_n = \sigma_n^2 = N_0 B$. Una tale descrizione viene indicata come *canale* AWGN (*additive white gaussian noise*) *limitato in banda*.




Indicando ora con $p(x)$, $p(y)$, $p(x/y)$, $p(y/x)$ le densità di probabilità marginali e condizionali che descrivono *un campione* dei processi di ingresso $x(t)$ ed uscita $y(t)$, entrambi limitati in banda $\pm B$, l'applicazione formale della (17.18) al caso continuo porta a scrivere l'espressione dell'informazione mutua media come

$$I(X,Y) = \int \int_{-\infty}^{\infty} p_{XY}(x,y) \log_2 \frac{p_Y(y/x)}{p_Y(y)} dxdy \quad \text{bit/campione} \tag{17.26}$$

che è una misura *assoluta*[27] del trasferimento di informazione per campione di uscita. Il massimo valore di (17.26) al variare di $p_X(x)$ consente anche questa volta di definire la capacità di canale per campione $C_s$=$\max_{p(x)} I(X,Y)$; in virtù della limitazione di banda, i campioni prelevati ad una frequenza di campionamento $f_c = 2B$ risultano indipendenti tra loro (vedi § 9.2.3), cosicchè la capacità di canale risulta definita come

$$C = 2B \cdot \max_{p(x)} \{I(X,Y)\} \quad \text{bit/secondo} \tag{17.27}$$

Riscrivendo la (17.26) nella forma

$$I(X,Y) = h(Y) - h(Y/X) \tag{17.28}$$

[26] Sono mostrati solo i valori per $0 \leq p_e \leq 0.5$ dato che successivamente l'andamento di $C_s$ si riflette in modo speculare.

[27] Per il fatto di avere una ddp di $y$ sia a numeratore che a denominatore del logaritmo, la (17.26) non soffre dei problemi discussi alla nota 17

si ottiene una espressione analoga alla (17.21) ma i cui termini sono ora da intendersi come entropia differenziale, definita in (17.10). Osserviamo ora che il termine di *noise entropy* $h(Y/X) = \int\int_{-\infty}^{\infty} p_X(x)p_Y(y/x)\log_2 \frac{1}{p_X(y/x)}dxdy$ dipende esclusivamente dal rumore additivo, in quanto $y(t) = x(t) + n(t)$ e quindi $p_Y(y/x) = p_Y(x+n) = p_N(y-x)$[28]; pertanto

$$h(Y/X) = \int_{-\infty}^{\infty} p_N(n)\log_2 \frac{1}{p_N(n)}dn = \frac{1}{2}\log_2\left(2\pi e\sigma_n^2\right)$$

come risulta per l'entropia differenziale di sorgenti gaussiane (17.11). Pertanto, nella (17.28) ora il termine che deve essere massimizzato rispetto a $p(x)$ è solo il primo $h(Y)$, 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 $\sigma_x^2$ la potenza di quest'ultimo, ed in virtù della indipendenza statistica tra $x(t)$ e $n(t)$, risulta $\sigma_y^2 = \sigma_x^2 + \sigma_n^2$, e quindi

$$h(Y) = \frac{1}{2}\log_2\left[2\pi e\left(\sigma_x^2 + \sigma_n^2\right)\right]$$

cosicchè la (17.27) si riscrive come

$$\begin{aligned} C &= 2B\cdot\left\{\frac{1}{2}\log_2\left[2\pi e\left(\sigma_x^2+\sigma_n^2\right)\right] - \frac{1}{2}\log_2\left(2\pi e\sigma_n^2\right)\right\} = \\ &= B\cdot\log_2\frac{\sigma_x^2+\sigma_n^2}{\sigma_n^2} = B\cdot\log_2\left(1+\frac{P_x}{P_n}\right) \quad \text{bit/secondo} \end{aligned}$$

che è proprio il risultato tanto spesso citato, che prende il nome di *legge di Hartley-Shannon* 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 trasmesso $P_s$, riscriviamo l'espressione della capacità nella sua forma più nota:

$$C = B\cdot\log_2\left(1+\frac{P_s}{N_0 B}\right) \quad \text{bit/secondo} \tag{17.29}$$

che, associata al teorema fondamentale della codifica espresso al § 17.2.3, stabilisce il massimo tasso informativo tramissibile senza errori su di un canale AWGN limitato in banda come $R < B\cdot\log_2(1 + {}^{P_s}/_{N_0 B})$. Discutiamo ora delle conseguenze di questo risultato.

**Sistema di comunicazione ideale** Una volta noto il massimo tasso di informazione $R < C$ che il canale può trasportare senza errori, come fare per evitare, appunto, questi ultimi? Il metodo suggerito da Shannon, anziché introdurre ridondanza come avviene per le tecniche di codifica di canale classiche, effettua invece la trasmissione semplicemente ripartendo l'informazione in blocchi codificati mediante simboli di durata elevata. In pratica, si tratta di realizzare una sorta di *trasmissione multilivello* (vedi § 5.1.2.4) come mostrato alla figura 17.2 dove l'informazione generata ad una

[28] Osserviamo infatti che $p_Y(y/x)$ altro non è che la gaussiana del rumore, a cui si somma un valor medio fornito dal campione di $x$, e quindi $p_Y(y/x) = p_N(n) + x$; pertanto $h(Y/X)$ si riduce all'entropia differenziale di un processo gaussiano, che non dipende dal valor medio, ma solo dall'andamento di $p_N(n)$

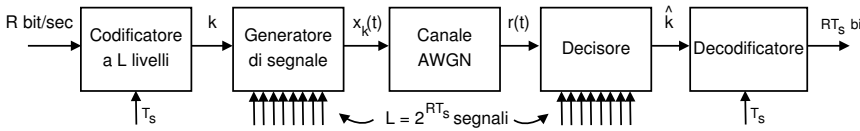



Figura 17.2: Schema ideale di codifica di canale ad errore asintoticamente nullo

velocità $R$ bit/secondo viene trasmessa mediante simboli emessi con periodo $T_s$ secondi, ognuno dei quali deve quindi convogliare una quantità di informazione pari a $M = RT_s$ bit, e dunque occorrono $L = 2^M$ diversi simboli.

Nella dimostrazione di Shannon ogni simbolo, anziché essere rappresentato da un livello costante, 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 fornisce l'errore $\varepsilon_k$ 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[29] 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!

In realtà, uno schema di trasmissione numerica che approssima piuttosto bene questo ideale appena discusso esiste veramente, ed è quello esposto al § 13.1.3 e indicato come FSK ortogonale. Infatti, il grafico delle sue prestazioni a pag. 300 mostra come, aumentando $L$, lo stesso valore di $\frac{E_b}{N_0}$ permette di conseguire valori di $P_e$ via via più piccoli.

**Minima energia per bit** Lo stesso grafico consente anche di verificare come allo stesso tempo il valore di $\frac{E_b}{N_0}$ necessario a conseguire una ben determinata $P_e$ diviene sempre più piccolo all'aumentare di $L$, anche se non può ridursi a meno di un valore limite, ossia deve comunque risultare[30]

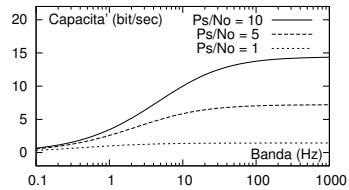
$$\frac{E_b}{N_0} \geq \ln 2 = 0,693 \quad \text{ovvero} \quad \left.\frac{E_b}{N_0}\right|_{dB} \geq -1.6\,\text{dB} \tag{17.30}$$

D'altra parte nell'FSK l'aumento di $L$ comporta l'aumento, oltre che di $T_s$, anche della banda occupata per la trasmissione, e questo ci dà lo spunto per le osservazioni che seguono.

[29] 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 $\varepsilon_k$ 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 per $E\{\varepsilon_k\}$ si ha $\varepsilon_k^{(h)} = \frac{1}{T_s}\int_0^{T_s}(x_h(t) + n(t) - x_k(t))^2\,dt \to \sigma_n^2 + 2\sigma_x^2$, 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.

[30] La (17.30) si ottiene considerando che se la capacità di canale per $B \to \infty$ fornita dalla (17.31) vale $C_\infty = \frac{P_s}{N_0 \ln 2}$, e se deve risultare $R \leq C$, allora $\ln 2 = \frac{P_s}{N_0 C_\infty} \leq \frac{P_s}{N_0 R} = \frac{E_b}{N_0}$.

**Compromesso banda-potenza e capacità massima** Il valore limite (17.30) trae origine da una conseguenza della (17.29) già fatta notare a pag. 145, ovvero la possibilità di risparmiare potenza aumentando l'occupazione di banda (o viceversa), dato che a ciò corrisponde un aumento di $C$, che però *non può oltrepassare* un valore massimo. Infatti, se nella (17.29) si aumenta $B$, aumenta anche la potenza di rumore, e l'effetto finale è che per un canale con *banda infinita* non si ottiene una capacità infinita, bensì il valore[31]




$$C_\infty = \lim_{B\to\infty} B \cdot \log_2\left(1 + \frac{P_s}{N_0 B}\right) = \frac{P_s}{N_0 \ln 2} \simeq 1.44\frac{P_s}{N_0} \tag{17.31}$$

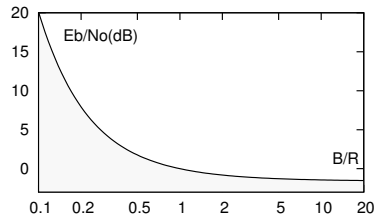
che individua anche il limite *assoluto* al massimo tasso informativo $R$ trasmissibile. La figura precedente mostra l'andamento effettivo della (17.29) in funzione di $B$, per alcuni valori di $\frac{P_s}{N_0}$ di esempio.

**Limite inferiore per $\frac{E_b}{N_0}$** Una volta assegnato il tasso informativo $R < C$ della sorgente e la banda $B$ del canale, partendo dalla (17.29) si può ottenere[32] una relazione che esprime l'$\frac{E_b}{N_0}$ necessario a conseguire una trasmissione senza errori (nel caso ideale):

$$\frac{E_b}{N_0} > \frac{B}{R}\left(2^{\frac{R}{B}} - 1\right) \tag{17.32}$$

e che, espressa in dB, è graficata nella figura a lato, in cui l'area grigia indica i valori di $\frac{E_b}{N_0}$ *vietati*, ossia per i quali è impossibile ottenere una trasmissione senza errori.

Notiamo innanzitutto che, mentre per $\frac{B}{R} = 1$ il sistema ideale richiede un valore di $\frac{E_b}{N_0}$ pari ad almeno 0 dB, questo si riduce nel caso in cui la trasmissione occupi una banda maggiore del tasso informativo $R$, fino a raggiungere (già per valori $B > 10R$) il




Compromesso banda-potenza per un sistema ideale

[31] La (17.31) si ottiene riscrivendo la (17.29) nella forma

$$C = \frac{P_s}{N_0 \frac{P_s}{N_0 B}} \cdot \frac{\ln\left(1 + \frac{P_s}{N_0 B}\right)}{\ln 2} = \frac{P_s}{N_0 \ln 2} \cdot \frac{\ln(1+\lambda)}{\lambda}$$

in cui ln è il logaritmo *naturale* in base $e$, e si è posto $\frac{P_s}{N_0 B} = \lambda$. Ricordando ora lo sviluppo di Maclaurin $f(x) = f(0) + \sum_{n=1}^{\infty}\left(\frac{\partial^n f(x)}{\partial x^n}\Big|_{x=0} \cdot \frac{x^n}{n!}\right)$ e che $\frac{d}{dx}\ln x = \frac{1}{x}$, il termine $\ln(1+\lambda)$ può essere espanso in serie di potenze come $\ln(1+\lambda) = \lambda - \frac{1}{2}\lambda^2 + \frac{1}{3}\lambda^3 + \cdots$; 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 (17.31).

[32] Riscrivendo la (17.29) come $2^{\frac{C}{B}} - 1 = \frac{P_s}{N_0 B}$, moltiplicando ambo i membri per $\frac{B}{R}$, e semplificando il risultato, si ottiene $\frac{B}{R}\left(2^{\frac{C}{B}} - 1\right) = \frac{P_s}{N_0 R}$. L'uguaglianza individua la circostanza limite in cui $R = C$, mentre se nell'esponente di 2 a primo membro sostituiamo $C$ con $R$, e $R < C$, il primo membro diviene più piccolo, e pertanto $\frac{B}{R}\left(2^{\frac{R}{B}} - 1\right) < \frac{P_s}{N_0 R}$. Infine, notiamo che $\frac{P_s}{N_0 R} = \frac{E_b}{N_0}$, da cui il risultato mostrato (17.32).

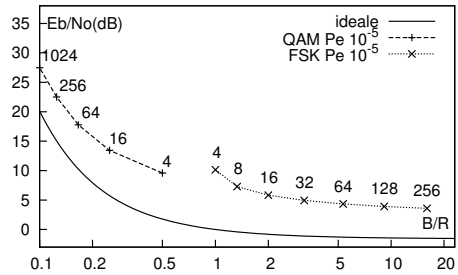



Figura 17.3: Prestazioni di QAM ed FSK confrontate con quelle ideali

limite (17.30) di -1.6 dB. D'altra parte, qualora la trasmissione impegni una banda inferiore ad $R$, il valore di $\frac{E_b}{N_0}$ necessario aumenta in modo piuttosto brusco.

**Prestazioni di sistemi di comunicazione reali** La verifica dei comportamenti appena evidenziati può essere svolta confrontando le prestazioni ideali (17.32) con quelle ottenibili adottando le tecniche di modulazione numerica già discusse, e per le quali si riesce a ridurre la banda occupata come nel caso della trasmissione multivello[33], oppure la si aumenta, come nel caso dell'FSK.

La figura 17.3 riporta i valori di $\frac{E_b}{N_0}$ *vs* $\frac{B}{R}$ per le tecniche di modulazione numerica QAM (§ 13.3.1) e FSK ortogonale (pag. 298): a partire dai rispettivi andamenti della $P_e$ in funzione di $\frac{E_b}{N_0}$ ed $L$, si sono ricavati i valori di $\frac{E_b}{N_0}$ necessari ad ottenere una $P_e$ pari a $10^{-5}$ per diversi valori di $L$, e questi sono stati riportati nel grafico assieme alla banda occupata, valutata come segue.

Considerando di adottare per il QAM un impulso di Nyquist a banda minima, la banda occupata risulta pari a $B_{QAM} = \frac{f_b}{\log_2 L}$, e pertanto

$$\left.\frac{B}{R}\right|_{QAM} = \frac{1}{\log_2 L}$$

mentre come riportato a pag. 300, per l'FSK ortogonale si ha $B_{FSK} \simeq \frac{f_b}{2}\frac{L}{\log_2 L}$, e dunque

$$\left.\frac{B}{R}\right|_{FSK} = \frac{L}{2\log_2 L}$$

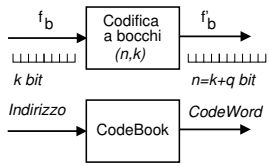
Possiamo osservare come per le due tecniche di trasmissione l'andamento dei valori di $\frac{E_b}{N_0}$ in funzione di $\frac{B}{R}$ ricalchi abbastanza fedelmente quello ideale, a parte una perdita di efficienza, che si riduce per $L$ crescente.

## 17.3 Codici di canale

Dopo aver esposto i risultati che la teoria dell'informazione fornisce a riguardo delle migliori prestazioni ottenibili, proseguiamo il discorso iniziato al § 5.3.3.1 e relativo a come aggiungere ridondanza ad un flusso binario a velocità $f_b$ da trasmettere su di un canale numerico, in modo realizzare una protezione FEC capace di ridurre la probabilità di errore sul bit $P_e$ in ricezione.

[33] Vedi ad es. il caso di banda base al § 7.5.5 o quello del QAM al § 13.3.1.

Riprendiamo la notazione introdotta a pag. 82 per i codici a blocchi, in cui per ogni $k$ bit in ingresso, sono prodotte *codeword* con lunghezza $n = k+q > k$, avendone aggiunti $q$ *di protezione* in funzione dei primi $k$, ed avendo denominato tale modo di procedere *un codice* $(n, k)$, la cui efficienza è misurata dal *tasso di codifica* (CODE RATE)




$$R_c = \frac{k}{n} < 1$$

che rappresenta la frazione di bit informativi sul totale di quelli trasmessi. La nuova velocità di trasmissione vale pertanto

$$f_b^{'} = \frac{f_b}{R_c}$$

La teoria esposta al § 17.2.3 afferma che la probabilità di errore può essere resa piccola a piacere purché $f_b = R < C$, ma se questo limite è applicato alla effettiva velocità di segnalazione $f_b^{'}$, allora il massimo tasso di trasferimento dell'informazione si riduce a $f_b = R_c f_b^{'} < R_c C$. Inoltre, aumentando la velocità di segnalazione (essendo $f_b^{'} > f_b$) diminuisce di pari misura il rapporto $E_b/N_0$, e conseguentemente peggiora anche la probabilità di errore del decisore, in modo che la capacità correttiva del codice deve essere tale da compensare anche quest'altro fattore. In definitiva, vorremmo trovare codificatori per cui $R_c$ sia il più possibile vicino ad uno.

Al tempo stesso, la capacità di correzione del codice è direttamente legata alla *distanza di Hamming* $d_H$ definita a pag. 83 come il minimo numero di bit diversi tra due parole di codice, sussistendo le relazioni

- per detettare almeno $l$ errori per codeword occorre $d_H \geq l + 1$
- per correggere almeno $t$ errori per codeword occorre $d_H \geq 2t + 1$

Un codice è tanto più *potente* quanti più errori è in grado di correggere, e dunque deve possedere $d_H$ elevato. In un codice a blocchi $(n, k)$ i $k$ bit del messaggio originale assumono tutte le configurazioni possibili, e quindi contribuiscono alla distanza tra codeword per un solo bit; per ottenere $d_H > 1$ occorre pertanto sfruttare gli $n - k = q$ bit di protezione, portando a scrivere

$$d_H \leq n - k + 1 = q + 1$$

che evidenzia la relazione tra $d_H$ e la quantità di bit aggiunti $q$. Purtroppo l'uguaglianza sussiste solo per i codici a ripetizione, discussi a pag. 83, che adottando una dimensione di blocco in ingresso $k = 1$ hanno un tasso di codifica $R_c = k/n = 1/n$ molto inefficiente. Mostriamo allora delle soluzioni che consentono di ottenere un adeguato potere di detezione senza per questo aumentare di molto la velocità di trasmissione del flusso codificato.

### 17.3.1 Codici lineari a blocchi

Le proprietà di questa classe di codici di canale possono essere meglio analizzate interpretando l'insieme delle possibili codeword da un punto di vista algebrico, che ci porta ad adottare una notazione matriciale idonea a descrivere la classe di codici *di Hamming*, mentre per i codici ciclici e di *Reed-Solomon* interverrà una notazione polinomiale.

**Distanza $d_H$ per codici lineari** Scegliamo di rappresentare una codeword arbitraria $X$ di un codice a blocchi $(n, k)$ mediante un vettore ad elementi binari

$$\mathbf{X} = (x_1 \quad x_2 \quad \cdots \quad x_n)$$

che può assumere solo $2^k$ diversi valori tra i $2^n$ possibili, mentre le ricezione di una delle rimanenti $2^n - 2^k$ combinazioni segnala la presenza di almeno un errore. Le $2^k$ codeword costituiscono uno *spazio lineare* se comprendono la codeword nulla, e se la somma di due vettori è anch'essa una parola di codice. La somma tra due codeword è definita in base alla matematica binaria modulo due, espressa mediante l'operatore di OR esclusivo $\oplus$ come

$$\mathbf{X} + \mathbf{Y} = (x_1 \oplus y_1 \quad x_2 \oplus y_2 \quad \cdots \quad x_n \oplus y_n)$$

Indicando ora come *peso* $w(\mathbf{Z})$ di un vettore $\mathbf{Z}$ il numero di suoi elementi non zero, una conseguenza della linearità è la possibilità di valutare la distanza di Hamming tra parole di codice come il *minimo peso* tra tutte le codeword non zero[34], ossia

$$d_H = \min_{\mathbf{X} \neq 0} [w(\mathbf{X})]$$

**Rappresentazione matriciale per codici sistematici** Benché la ripartizione netta degli $n$ bit delle codeword in una prima parte contenente i $k$ bit da proteggere seguiti dai $q = n - k$ bit di protezione è stata finora data per scontata, questa non è per nulla una circostanza ineludibile, ma se si verifica, il codice risultante viene detto *sistematico*. In tal caso possiamo scrivere le codeword come

$$\mathbf{X} = (m_1 \quad m_2 \quad \cdots \quad m_k \quad c_1 \quad c_2 \quad \cdots \quad c_q)$$

ovvero in forma partizionata $\mathbf{X} = (\mathbf{M} \mid \mathbf{C})$ che permette di rappresentare $\mathbf{X}$ a partire dai bit da proteggere $\mathbf{M}$ e dalla definizione di una matrice *generatrice* $k \times n$ con struttura generale $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$ in cui $\mathbf{I}_k$ è una matrice identità $k \times k$ e $\mathbf{P}$ è una sotto-matrice di elementi binari $k \times n$, potendo così scrivere $\mathbf{X} = \mathbf{M} \cdot \mathbf{G}$ ovvero

$$\begin{bmatrix} m_1 \cdots m_k & c_1 \cdots c_q \end{bmatrix} = \begin{bmatrix} m_1 \cdots m_k \end{bmatrix} \cdot \begin{bmatrix} 1 & \cdots & 0 & p_{11} & \cdots & p_{1q} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & p_{k1} & \cdots & p_{kq} \end{bmatrix}$$

in modo che $\mathbf{P}$ produca[35] i $q$ bit di protezione come $\mathbf{C} = \mathbf{M} \cdot \mathbf{P}$. Dato che il valore della (generica) $j$-esima componente di $\mathbf{C}$ si calcola come

$$c_j = m_1 p_{1j} \oplus m_2 p_{2j} \oplus \cdots \oplus m_k p_{kj}$$

osserviamo che ciascuna colonna di $\mathbf{P}$ individua un sotto-insieme di componenti di $\mathbf{M}$ su cui calcolare una *somma di parità*, ed è per questo che il sotto-blocco di matrice è rappresentato dalla lettera $\mathbf{P}$. Ma non è ancora stato definito nulla che ci possa aiutare a scegliere i coefficienti $p_{ij}$ allo scopo di ottenere i valori $d_H$ e $R_c$ desiderati: il codice di Hamming ci fornisce una possibile soluzione.

[34] Infatti dalla definizione di somma otteniamo che la distanza tra due codeword $\mathbf{X}$ e $\mathbf{Y}$ è pari al peso della codeword $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$: infatti $\mathbf{Z}$ presenterà componenti $z_j = 1$ solo in corrispondenza di elementi $x_j \neq y_j$. Ma per la linearità anche $\mathbf{Z}$ appartiene al codebook, e dunque la ricerca su tutte le coppie si trasforma in una ricerca su tutte le codeword.

[35] Sono valide le normali regole di moltiplicazione tra matrici, tranne per l'accortezza di usare la somma modulo due anziché quella convenzionale.

### 17.3.2 Codice di Hamming

E' un codice a blocchi $(n, k)$ sistematico con $q \geq 3$ bit di controllo, e per il quale si definisce

$$n = 2^q - 1 \quad \text{e} \quad k = n - q$$

per cui il tasso di codifica vale

$$R_c = \frac{k}{n} = \frac{n-q}{n} = 1 - \frac{q}{2^q - 1}$$

| $q$ | $n$ | $k$ | $R_c$ |
|---|---|---|---|
| 3 | 7 | 4 | 0.57 |
| 4 | 15 | 11 | 0.73 |
| 5 | 31 | 26 | 0.84 |
| 6 | 63 | 57 | 0.9 |
| 7 | 127 | 120 | 0.94 |

che aumenta con il crescere di $q$, come mostrato in tabella. Le codeword si individuano ponendo *le righe* della sottomatrice $P$ pari a tutte le parole di $q$ bit con due o più uni, in qualsiasi ordine. Ma la cosa *più simpatica*, è che per un codebook siffatto si ottiene $d_H = 3$, indipendentemente dalla scelta di $q$.

**Esempio:** codice di Hamming $(7, 4)$. Consideriamo un codice di Hamming con $q = 3$, e quindi $n = 2^3 - 1 = 7$ e $k = 7 - 3 = 4$. La matrice generatrice è quindi pari a

$$G = \left[\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{array}\right]$$

a cui corrispondono le seguenti $2^4 = 16$ codeword, confermando che $d_H = 3$.

| **M** | **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 |

| **M** | **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 |

**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 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 $\hat{\mathbf{Y}}$ più vicina, ossia quella per la quale il peso $w\left(\mathbf{Y} \oplus \hat{\mathbf{Y}}\right)$ è minimo.

**Correzione basata sulla sindrome** Un metodo che non richiede una ricerca esaustiva si basa invece sul calcolo della cosiddetta *sindrome*, ottenuta mediante moltiplicazione del vettore $\mathbf{Y}$ ricevuto per una matrice $n \times q$ di *controllo parità* $\mathbf{H}$, definita come $\mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I}_q \end{bmatrix}$ in cui $\mathbf{P}$ è la stessa matrice di parità utilizzata nella matrice generatrice $\mathbf{G}$, e $\mathbf{I}_q$ è una matrice identità di dimensioni $q \times q$. La matrice $\mathbf{H}$ esibisce la simpatica proprietà che, se moltiplicata per una qualunque codeword valida, fornisce un vettore *nullo*, ossia

$$\mathbf{X} \cdot \mathbf{H} = (0 \quad 0 \quad \cdots \quad 0)$$

Al contrario, se moltiplicata per un vettore $\mathbf{Y}$ non appartenente al codebook, fornisce un vettore detto *sindrome* $\mathbf{S} = \mathbf{Y} \cdot \mathbf{H}$ *non nullo*, e quindi il suo calcolo permette la *rivelazione* (nei limiti consentiti da $d_H$) dell'occorrenza di errori.

**Esempio** considerando di nuovo il caso di $q = 3$, la corrispondente matrice di controllo parità $\mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I}_q \end{bmatrix}$ è mostrata a lato. E' facile verificare che per tutte le possibili codeword $\mathbf{X}$ (ad es. $\mathbf{X} = [0100111]$) si ottiene $\mathbf{X} \cdot \mathbf{H} = [0000000]$. 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 $\mathbf{S} = \mathbf{Y} \cdot \mathbf{H} = [100]$.

$$\mathbf{H} = \left[\begin{array}{ccc} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]$$

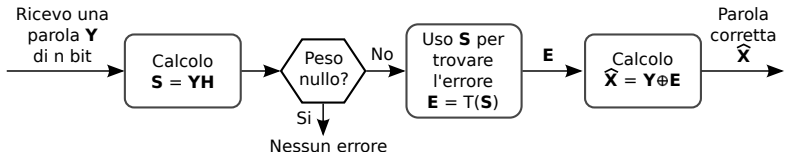
Per quanto riguarda la *correzione*, se scriviamo il vettore ricevuto come $\mathbf{Y} = \mathbf{X} \oplus \mathbf{E}$, dove $\mathbf{E}$ è un vettore di $n$ bit le cui componenti sono diverse da zero in corrispondenza dei bit errati di $\mathbf{Y}$, il calcolo della sindrome risulta

$$\mathbf{S} = \mathbf{Y} \cdot \mathbf{H} = (\mathbf{X} \oplus \mathbf{E}) \cdot \mathbf{H} = \mathbf{X} \cdot \mathbf{H} \oplus \mathbf{E} \cdot \mathbf{H} = \mathbf{E} \cdot \mathbf{H}$$

visto che come detto sopra, la sindrome delle codeword è nulla. Dato però che la sindrome ha dimensione di $q$ componenti, tutte le $2^n$ possibili sequenze di errore $\mathbf{E}$ danno luogo a sole $2^q$ diverse sindromi, e quindi la conoscenza della sindrome non consente di risalire direttamente a $\mathbf{E}$. Ma riprendendo i risultati esposti a pag. 78, risulta che la probabilità $P(i, n)$ che si siano verificati $i$ errori su $n$ bit decresce al crescere di $i$, e pertanto il vettore $\mathbf{E}$ 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:

$$\hat{\mathbf{E}} = \underset{\mathbf{E}:\mathbf{E}\cdot\mathbf{H}=\mathbf{S}}{\operatorname{argmin}} \{w(\mathbf{E})\}$$

Anziché effettuare questo calcolo ad ogni codeword ricevuta, si può *precalcolare*, per ogni possibile $\mathbf{E}$, la relativa sindrome, e compilare una tabella $T()$ in cui memorizzare per ogni possibile $\mathbf{S}$, il vettore $\mathbf{E}$ di peso minore che la produce. La stessa tabella può quindi essere consultata usando come chiave di ricerca la sindrome $\mathbf{S} \neq 0$ calcolata per ogni $\mathbf{Y}$ ricevuto, ottenendo il vettore di errore $\mathbf{E} = T(\mathbf{S})$ più probabile, e che se sommato ad $\mathbf{Y}$, permette la correzione degli errori. La figura che segue riepiloga i passi necessari alla correzione mediante sindrome.




**Limiti** A questo punto va spesa una parola di cautela, perché se si sono verificati più errori di quanti $d_H$ permetta correggere, è inutile (anzi dannoso) cercare di eseguire la correzione, perché il numero di errori complessivo potrebbe essere ancora più elevato. Ad esempio considerando un codice di Hamming caratterizzato da $d_H = 3$, esistono solo $n$ vettori $\mathbf{E}$ di peso unitario, a cui corrispondono sindromi esattamente pari alle $n = 2^q - 1$ righe di $\mathbf{H}$. Supponiamo ora che $\mathbf{E}$ contenga due errori: la sua moltiplicazione per la sindrome produce comunque una delle $2^q$ possibili sindromi e (se è risultato $\mathbf{S} \neq 0$) il tentativo di correzione produrrà un vettore $\hat{\mathbf{X}}$ contenente comunque uno o tre errori.

### 17.3.3 Codici convoluzionali

Si tratta di una classe di codificatori di canale che, a differenza dei codici a blocchi, produce una sequenza binaria i cui valori dipendono da gruppi di bit di ingresso *temporalmente sovrapposti*, in analogia all'integrale di convoluzione che fornisce valori che dipendono da *intervalli* dell'ingresso, pesati dai valori della risposta impulsiva.Un generico codice convoluzionale è indicato con la notazione $CC\,(n,k,K)$, che lo descrive capace di generare gruppi di $n$ bit di uscita (sequenza $\{b\}$) in base alla conoscenza di $K$ simboli di ingresso (sequenza $\{a\}$), ognuno composto da $k$ bit.

La memoria dei $K$ ingressi è usualmente rappresentata mediante un registro a scorrimento che ospita gli ultimi $K \cdot k$ bit di ingresso, dove per ogni nuovo simbolo $a_j$ che entra da sinistra, i precedenti scorrono a destra, ed il più "vecchio" viene dimenticato. Ognuno degli $n$ bit di uscita $b_j\,(i)$, $i = 1, 2, ..., n$ è ottenuto a partire dai $k \cdot K$ bit di memoria, eseguendo una somma modulo 2 tra alcuni di essi[36]. Resta valido quindi il concetto di *coding rate* $R_c = \frac{k}{n}$ che rappresenta *quanti bit di informazione* sono presenti per ogni bit di uscita dal codificatore.

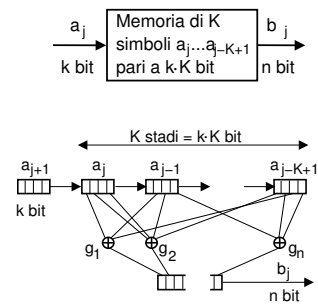



**Diagramma di transizione** Una volta definito il meccanismo di calcolo di $b_j$ a partire da $a_j$, il codificatore può essere descritto mediante un *diagramma di transizione*, costituito da $2^{(K-1)k}$ *stati* $S$ associati a tutte le possibili combinazioni di bit degli ultimi $K-1$ simboli di ingresso; ad ogni stato competono quindi $2^k$ transizioni, una per ogni possibile nuovo $a_j$ di ingresso, ed ognuna delle quali termina sul nuovo stato che si è determinato. Ad ogni transizione, è infine associato in modo univoco il gruppo di $n$ bit $b_j\,(a_j, S_j)$ da produrre in uscita[37].

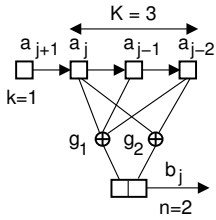
**Esempio: CC(2,1,3)** Per fissare le idee, definiamo un $CC\,(n,k,K) = CC\,(2,1,3)$ con $g_1 = \left[\begin{array}{ccc} 1 & 1 & 1 \end{array}\right]$ e $g_2 = \left[\begin{array}{ccc} 1 & 0 & 1 \end{array}\right]$, caratterizzato da un coding rate $\frac{1}{2}$, e mostrato nella parte sinistra della figura 17.4. Per questo, costruiamo il diagramma di transizione mostrato al centro, in cui le transizioni tra stati sono tratteggiate o piene, in corrispondenza dei valori $a_j$ pari a zero o ad uno, e sono etichettate con la coppia di bit in uscita $b_j$, calcolabile mediante i vettori generatori $g_{1,2}$, come risulta dalla tabella della verità mostrata a destra:

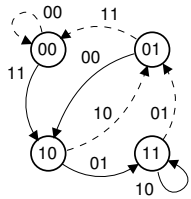
Come si può notare, ogni stato ha *solo due* transizioni, e quindi solo *due valori* di uscita (la metà dei 4 possibili con due bit); inoltre, questi valori differiscono in *entrambi i bit*[38].

[36] Gli $n$ modi di scegliere quali dei $k \cdot K$ bit sommare, per ottenere ognuno degli $n$ bit di uscita, sono determinati mediante $n$ *vettori generatori* $g_i$, $i = 1, 2, ..., n$, di lunghezza $k \cdot K$, contenenti una serie di cifre binarie zero od uno, a seconda che l'$i$-*esimo* sommatore modulo due sia connesso (o meno) al corrispondente bit di memoria.

[37] Lo stesso valore di $b_i$ potrebbe essere prodotto da più di una delle $2^{k \cdot K}$ diverse memorie del codificatore.

[38] *La dipendenza di* $b_j$ *da* $(a_j, S_j)$ *è legata alla scelta dei generatori* $g_i$. *Nel caso in cui un valore* $b_j\,(i)$ *sia sempre uguale ad uno dei* $k$ *bit di* $a_j$, *il codice è detto* sistematico. *La scelta dei* $g_i$ *può essere effettuata via computer, per individuare il gruppo che determina le migliori prestazioni.*




| $a_j$ | $S_j$ | $b_j\,(1)$ | $b_j\,(2)$ |
|---|---|---|---|
| 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 |

Figura 17.4: Architettura del codice convoluzionale $CC\,(2,1,3)$, diagramma di transizione e tabella della verità.

Qualora si ponga in ingresso la sequenza $\{a\} = \{...010100\}$([39]), è possibile osservare che la sequenza di stati risulta $\{S\} = \{..., 01, 10, 01, 10, 00\}$, mentre quella di uscita è $\{b\} = \{..., 10, 00, 10, 11\}$, semplicemente seguendo il diagramma di transizione, a partire dallo stato iniziale 00. D'altra parte, è possibile anche il procedimento opposto: conoscendo $\{b\}$, si può risalire ad $\{a\}$, percorrendo di nuovo le transizioni etichettate con i simboli $b_j$. In definitiva, osserviamo come ad ogni coppia $(\{a\}, \{b\})$ sia biunivocamente associata una sequenza di stati $\{S\}$.

**Diagramma a traliccio** Per meglio visualizzare le possibili sequenze di stati, costruiamo il *diagramma a traliccio* (TRELLIS) del codificatore, mostrato a sinistra di fig. 17.5 riportando sulle colonne i possibili stati attraversati ai diversi istanti $j$, collegando i *nodi* del traliccio con transizioni piene o tratteggiate nei casi in cui siano relative ad ingressi pari ad 1 o 0, e riportando sulle transizioni stesse i valori di uscita. Con riferimento alla sequenza *codificata* $\{b\}$ riportata in basso, l'effettiva successione di stati $\{S\}$ è rappresentata dalle linee più spesse.

[39] Per comodità di rappresentazione, il bit più a destra nella sequenza $\{a\}$ è il primo in ordine di tempo ad entrare nel codificatore.

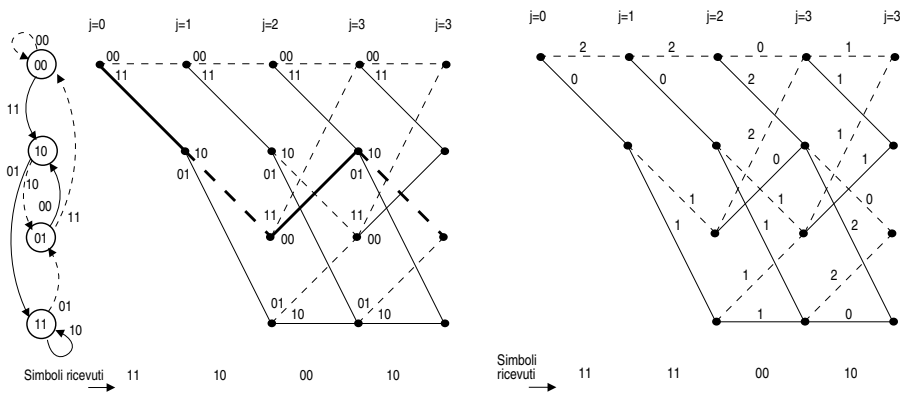
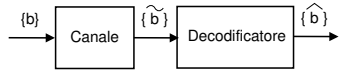



Figura 17.5: Diagramma a traliccio e costi $d_H$ per la sequenza ricevuta

**Decodifica di Viterbi** Consideriamo ora il caso in cui la sequenza codificata $\{b\}$ venga trasmessa su di un canale, e che sia ricevuta *con errori*.



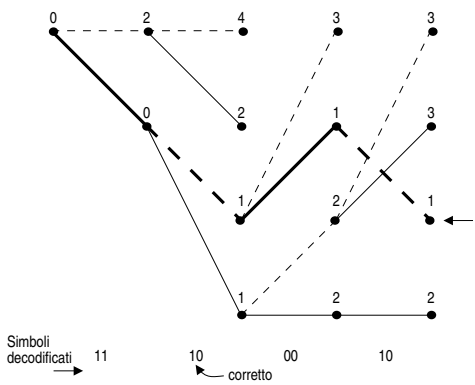

In generale, non sarà più possibile rintracciare una sequenza di stati tale da produrre *esattamente* la sequenza ricevuta $\left\{\tilde{b}\right\}$, ed il problema diviene quello di individuare la sequenza di stati $\{S\}$ tale da produrre una $\left\{\hat{b}\right\}$ la più vicina a $\left\{\tilde{b}\right\}$. Allo scopo di misurare questa differenza, utilizziamo la *distanza di Hamming* $d_H\left(\widehat{b},\tilde{b}\right)$ pari al numero di bit diversi tra $\left\{\tilde{b}\right\}$ e le possibili $\left\{\hat{b}\right\}$([40]), e etichettiamo gli archi del traliccio con le $d_H\left(\widehat{b}_j,\tilde{b}_j\right)$ tra il simbolo da emettere $\widehat{b}_j$ e quello osservato in ricezione $\tilde{b}_j$ (vedi lato destro di fig. 17.5). In tal modo, per ogni particolare sequenza di stati $\{S\}$ è possibile determinare un *costo* pari alla somma delle $d_H\left(\widehat{b}_j,\tilde{b}_j\right)$ relative alle transizioni attraversate dal traliccio[41]. Pertanto, la sequenza $\left\{\hat{b}\right\}$ più vicina a $\left\{\tilde{b}\right\}$, ossia tale che

$$d_H\left(\widehat{b},\tilde{b}\right)=\min_{\{S\}}\left\{\sum_j d_H\left(b_j,\tilde{b}_j\right)\right\}$$

può essere individuata come quella associata alla sequenza di stati $\left\{\hat{S}\right\}$ *di minimo costo*[42].

Dato che da ogni stato si dipartono $2^k$ archi, ad ogni istante il numero di percorsi alternativi aumenta di un fattore $2^k$, crescendo molto velocemente all'aumentare di $j$. L'enumerazione *completa* dei percorsi può essere però evitata, notando che quando due percorsi con costi diversi si incontrano in uno stesso nodo, quello di costo *maggiore* sicuramente *non è* la parte iniziale del percorso di minimo costo, e quindi può essere eliminato.

Questa filosofia si applica al caso in questione con riferimento alla figura a lato, che mostra come il calcolo dei *costi parziali* avvenga *per colonne* da sinistra a destra, scrivendo sopra ad ogni nodo il costo del *miglior* percorso che lo raggiunge. Ad ogni colonna sono scartati i percorsi che si incontrano con uno migliore, cosicchè il numero di percorsi *sopravvissuti* è sempre pari al numero di stati $2^k$. All'estremità destra della figura, una freccia indica la minima $d_H\left(\widehat{b},\tilde{b}\right)$, asso-




[40] Questo caso viene indicato con il termine *hard-decision decoding* in quanto il ricevitore *ha già* operato una decisione (quantizzazione) rispetto a $\tilde{b}$. Al contrario, se i valori ricevuti sono passati *come sono* al decodificatore di Viterbi, questo può correttamente valutare le probabilità $p\left(\tilde{b}/\widehat{b}\right)$ ed operare in modalità *soft decoding*, conseguendo prestazioni migliori.

[41] Ad esempio, con riferimento alla fig. 17.5, la $\{S\}=\{00,10,11,01,10\}$ ha un *costo* pari a 3.

[42] Qualora la distanza tra $\tilde{b}_j$ ed un possibile $\widehat{b}_j$ sia espressa come probabilità condizionata $p\left(\tilde{b}_j/\widehat{b}_j\right)$, il processo di decodifica è detto *di massima verosimiglianza*.

ciata al percorso a tratto spesso, e che
permette di individuare la $\left\{\tilde{b}\right\}$, che come si vede *è quella esatta.*
Tralasciamo ora di approfondire la teoria che consente l'analisi dettagliata dell'algoritmo, e ci limitiamo alle seguenti

**Riflessioni**

- l'esempio fornito si mostra in grado di correggere un errore pur impiegando un coding rate pari ad $\frac{1}{2}$, migliore 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 $\tilde{b}$ *ricevuta*;
- si verifica errore (cioè $\left\{\hat{b}\right\} \neq \{b\}$) se $d_H\left(\widehat{b}, \tilde{b}\right)$ è *minore* di $d_H\left(b, \tilde{b}\right)$;
- le capacità di correzione del codice migliorano aumentando la $d_H$ tra le possibili sequenze $\{b\}$[43];
- la $d_H$ tra diverse $\{b\}$ aumenta con $\frac{K}{k}$, in quanto la matrice di transizione tra stati diviene più sparsa, ed i valori di $\{b\}$ sono più interdipendenti;
- se il miglioramento di cui sopra è ottenuto aumentando K, ciò equivale ad estendere nel tempo la memoria del codificatore, ma senza per questo alterare il tasso di codifica $R_c = \frac{k}{n}$.

[43] La minima distanza tra le sequenze codificate è indicata come $d_{min}$, e può essere trovata come la $d_H$ tra una $\{b^0\}$ tutta nulla ($\{b^0\} = \{...000000000\}$) e quella con il minor numero di uni, che si diparte e ritorna (nel traliccio) dallo/allo stato 00.

# Capitolo 18

# Codifica di sorgente multimediale

Affrontiamo ora alcuni aspetti applicativi delle teorie finora esposte al presente capito, come la *codifica audio*, di *immagine*, e *video*.

## 18.1 Codifica audio

Al § 7.4 abbiamo svolto una valutazione approssimata della distorsione introdotta dal processo di quantizzazione di segnale audio, ricavando che l'utilizzo di $M$ bit/campione si traduce in $SNR_q\left(M\right)|_{dB} \simeq 6 \cdot M$ dB. Quindi, al § 7.6.1 si è mostrato come adottando una caratteristica di quantizzazione logaritmica anziché lineare, ci si può adattare meglio alla effettiva densità di probabilità del segnale vocale, rendendo inoltre $SNR_q$ relativamente poco sensibile alla sua effettiva dinamica, dando luogo alla cosidetta codifica PCM con *legge A* o *legge* $\mu$, standardizzata nel 1988 da ITU-T come G.711[1]. Mentre questa costituisce un formato universale di scambio permettendo la compatibilità tra dispositivi e tecnologie, nel seguito sono state sviluppate diverse tecniche alternative[2], capaci di offrire la stessa (o migliore) qualità di ascolto con velocità di trasmissione contenute, non solo per segnali vocali in banda telefonica, ma anche per segnali a banda larga, musicali, e multicanale, di cui tentiamo ora una sommaria rassegna.

### 18.1.1 Codifica di forma d'onda

Questa classe di codificatori opera esclusivamente nel dominio del tempo, operando campione per campione, e ottiene una qualità comparabile o superiore a quella del PCM sfruttando le caratteristiche di memoria presenti nel segnale, e/o adattando alcuni parametri di funzionamento alle caratterististiche tempo varianti del segnale.

#### 18.1.1.1 DPCM o PCM Differenziale

La prima variazione rispetto al PCM è stata quella di applicare il principio della codifica predittiva (pag. 413) usando come predittore semplicemente il precedente campione di ingresso. Il corrispondente schema di eleborazione è mostrato in fig. 18.1, ed il suo

[1] `http://www.itu.int/rec/T-REC-G.711/e`

[2] Una raccolta di riferimenti a risorse relative a codec audio orientati alle applicazioni multimedali può essere trovata presso `http://labtel.ing.uniroma1.it/codecs`

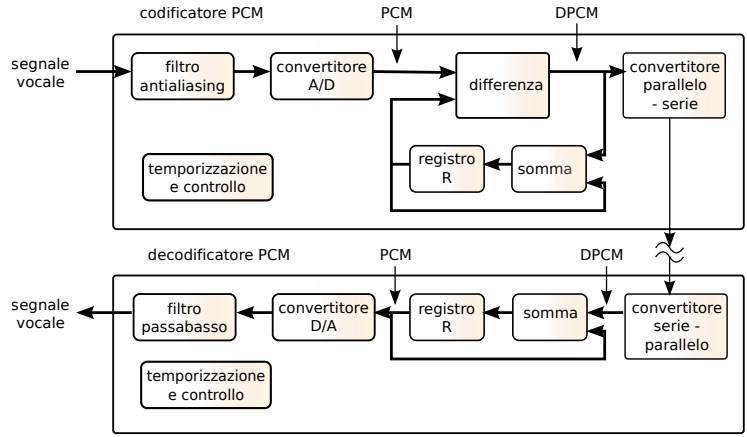
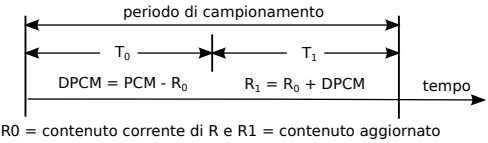



Figura 18.1: Codec audio Differential PCM o DPCM

funzionamento è suddiviso in due fasi: nella prima il codificatore sottrae il campione precedente (all'inizio nullo) all'attuale, e nella seconda questa differenza è risommata al valore di differenza precedente (all'inizio nullo) in modo da ri-calcolare il valore attuale, e salvarlo nel registro di memoria. Il segnale differenza è caratterizzato da valori di ampiezza ridotti rispetto all'originale, e può essere codificato con 7 bit/campione, producendo ora una velocità per la qualità telefonica di 56 kbps. Il decodificatore si limita quindi a sommare alla differenza ricevuta il valore ricostruito del campione precedente, ed effettuare l'operazione di restituzione analogica. Osserviamo che il codificatore calcola il valore precedente mediante un circuito identico a quello presente al decodificatore, e per questo l'operazione è perfettamente invertibile.




### 18.1.1.2 ADPCM o DPCM Adattivo

Questo metodo differisce dal precedente per due aspetti: da un lato il processo di predizione tiene conto di più di un campione passato e non di uno solo come nel DPCM, come descritto in fig. 18.2 in cui è mostrato un predittore del terzo ordine che in pratica consiste in un filtro trasversale i cui coefficienti sono fissati in base alle caratteristiche statistiche medie del segnale vocale. Il secondo aspetto è che ora il quantizzatore *modifica nel tempo* la propria dinamica di azione (da cui il termine *adattativo*) in base ad una stima della dinamica del segnale.

Nel lato sinistro della fig. 18.3 è mostrata una caratteristica di quantizzazione uniforme operante su di una dinamica di ingresso $\phi_x \hat{\sigma}_x$, con $\phi_x > 1$ scelto in modo da rendere trascurabile la probabilità che un valore di ingresso troppo elevato determini la saturazione del quantizzatore. Utilizzando una stima a breve termine della varianza $\hat{\sigma}_x^2$ calcolata sugli ultimi campioni di segnale (a media nulla), ossia ad es. calcolando $\hat{\sigma}_x^2(n) = \frac{1}{N}\sum_{i=1}^{N} x^2(n-i)$, si possono rendere gli intervalli di decisione $\Delta$ piccoli nelle fasi di segnale piccolo, in modo da mantenere l'SNR costante anche per segnali con ampizze molto variabili. Inoltre, è possibile *omettere* la trasmissione della stima di

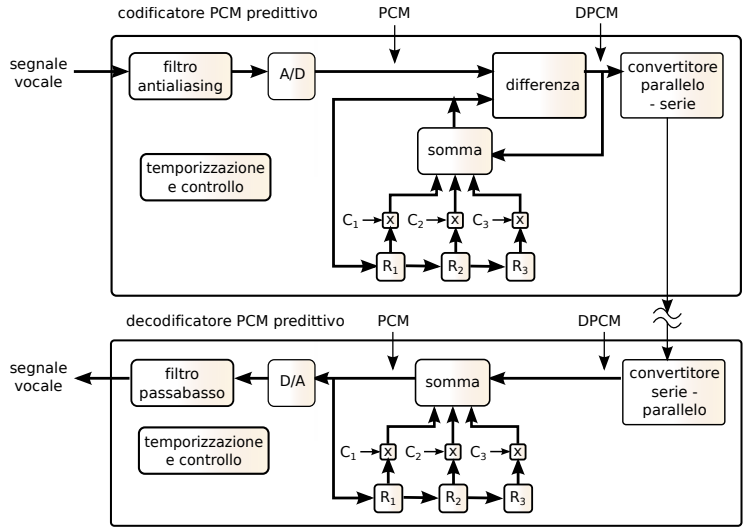



Figura 18.2: Codec DPCM con predittore a tre coefficienti costanti

varianza se quest'ultima è calcolata in modalità *backward*, ossia a partire dai valori $y(n) = Q[x(n)]$, dato che la stessa operazione è eseguibile in modo indipendente anche dal lato del decodificatore. Infine, la stima della varianza è ulteriormente semplificata se realizzata mediante una formula recursiva, ossia

$$\hat{\sigma}_x^2(n) = \alpha\hat{\sigma}_x^2(n-1) + (1-\alpha)\,y^2(n)$$

il cui risultato è mostrato in fig. 18.4, dove la linea tratteggiata rappresenta il valore istantaneo di $y^2(n)$, mentre quella continua mostra i valori di $\hat{\sigma}_x^2(n)$ ottenuti in modo

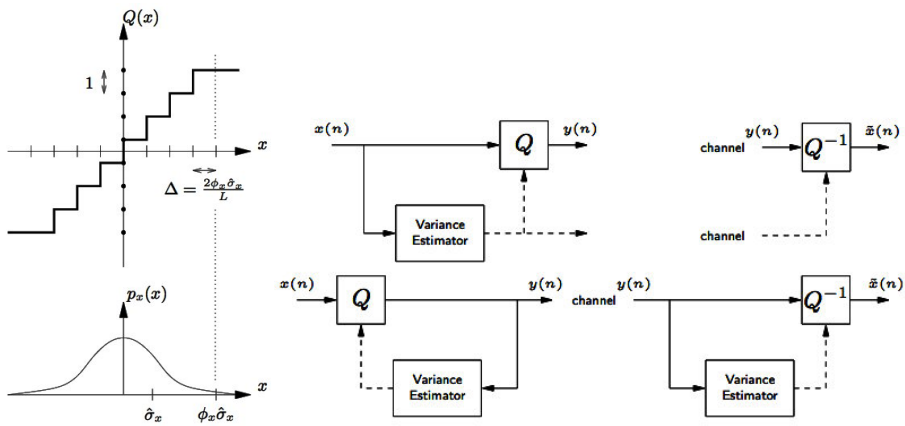



Figura 18.3: Dinamica di quantizzazione e sua stima diretta o *backward* - tratto da http://cnx.org/content/m32074/latest/

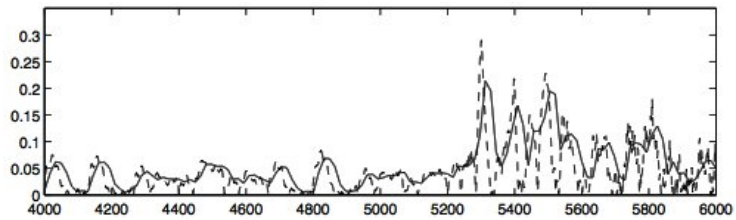



Figura 18.4: Stima recursiva backward della varianza, confrontata con i valori di $y^2(n)$, per $\alpha = 0.9$

recursivo. Infine, la fig. 18.5 mostra i due estremi del codec ADPCM, che rimangono sincronizzati anche nel caso di saturazione del quantizzatore adattativo.

Il miglioramento della qualità ottenibile ha portato a ridurre il numero di bit (e di livelli) del quantizzatore a 5, 4, 3, 2 bit/campione, a cui corrispondono velocità di codifica di 40, 32, 24, 16 kbps. Questi sono i valori a cui si riferisce lo standard ITU-T G.721, successivamente confluito nel G.726.

#### 18.1.1.3 Codica per sottobande

Anche la raccomandazione G.722 è basata sulla codifica ADPCM, ma applicata ad un segnale con una banda audio più larga, riproducendo correttamente frequenze fino a 7 KHz. Ciò avviene dopo aver suddiviso le componenti frequenziali del segnale in due sottobande, come mostrato in fig. 18.6, mediante una coppia di filtri passa-basso e passa-alto con comune frequenza di taglio di 3.5 KHz. Il canale relativo alla semi banda superiore è quindi campionato a frequenza di 16 kHz, mentre l'altro è praticamente equivalente al segnale telefonico preso in esame fino ad ora. Per entrambi i canali è applicata la codifica ADPCM, ma le velocità dei due sono impostate in modo differente, dando più importanza alla componente di bassa frequenza, percettivamente più rilevante: ad esempio, si può scegliere di assegnare 16 kbps alle alte frequenze e 48 alle basse, ottenendo un totale di 64 kbps per una qualità risultante migliore del G.711, in quanto ora si opera su di un segnale a larga banda, con risultati idonei ad applicazioni come la videoconferenza.

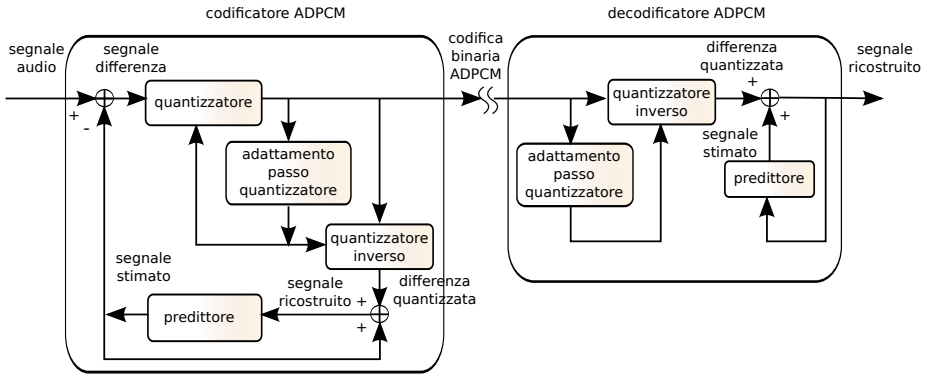



Figura 18.5: Architettura di un codec ADPCM

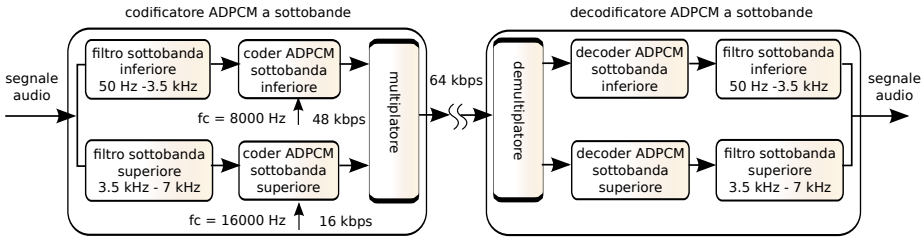



Figura 18.6: Architettura di un codec ADPCM a sottobande

Lo stesso schema di codifica per sottobande più ADPCM è proposto anche dallo standard G.726, ma applicato ad un segnale a qualità telefonica, offrendo le velocità di 40, 32, 24 e 16 kbps.

### 18.1.2 Codifica basata su modello

I metodi discussi fin qui non tengono in particolare conto la natura del segnale da codificare; restringendo invece il campo al solo caso di segnale vocale, le conoscenze relative alla sua particolare modalità di produzione possono essere usate per ridurre le informazioni da trasmettere, costituite ora dai parametri che caratterizzano un suo modello di generazione. Essendo questo il dominio delle scienze linguistiche e fonetiche, svolgiamo una piccola digressione in tal senso.

**Fisiologia dell'apparato fonatorio** Può essere descritta (vedi fig. 18.7) mediante il cosiddetto *modello a tubi*, in cui il tratto compreso tra le corde vocali e le labbra (nonché il *tratto nasale*) è idealizzato come una concatenazione di tubi di diversa sezione, la cui effettiva conformazione varia con continuità, fonema per fonema, in base alla posizione della lingua e delle labbra. Nei suoni vocalici l'aria che attraversa le corde vocali ne determina la chiusura periodica, dando origine ad un *segnale di eccitazione* periodico; la differenza di area delle diverse sezioni provoca un *disadattamento di impedenza acustica*[3] e la conseguente formazione di onde riflesse (vedi fig. 18.8),

[3]Si applica in pratica la stessa teoria valida per le linee elettriche, in cui al posto di tensione e corrente, ora si considerano rispettivamente pressione $p$ e velocità $u$

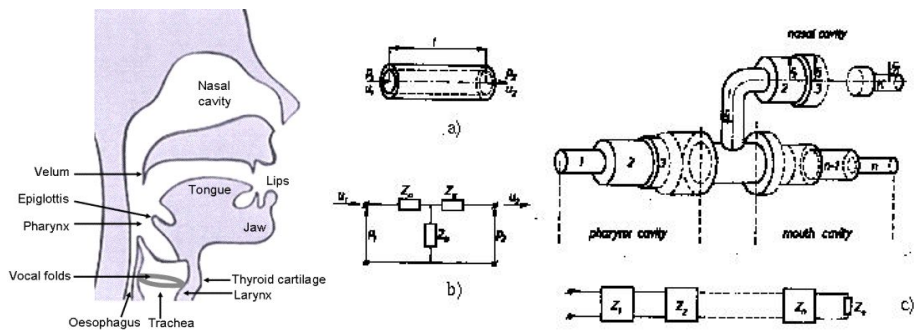



Figura 18.7: Tratto vocale (a ds.) e relativo modello a tubi (a sin.)

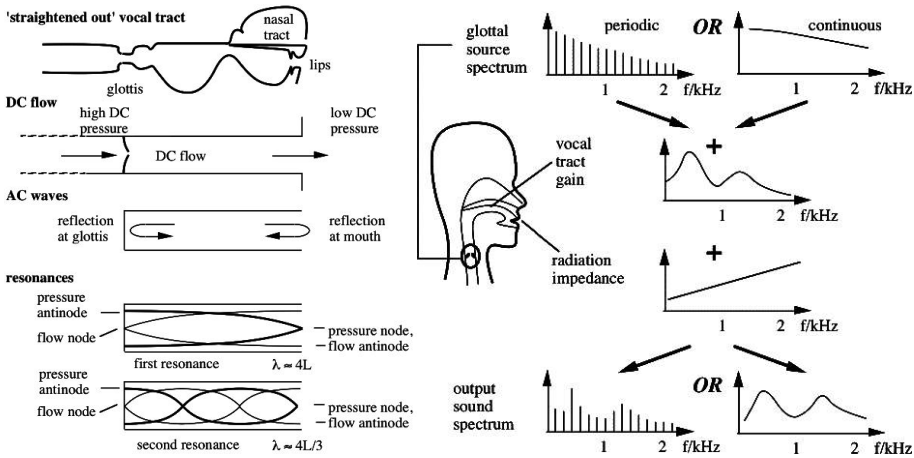



Figura 18.8: Natura delle risonanze del tratto vocale e loro effetto filtrante sull'onda glottale

che per lunghezze d'onda in relazione intera con la lunghezza del tratto vocale, determinano fenomeni di *onde stazionarie*, ovvero di *risonanze* (che in fonetica sono dette *formanti*), la cui disposizione in frequenza modifica lo spettro dell'onda glottale, determinando così il timbro corrispondente ai suoni della lingua; a questo fenomeno è infine associata una enfasi delle alte frequenze dipendente da un effetto derivata prodotto dalle labbra. Quindi, si assume il modello valido anche per i suoni *fricativi*, prodotti anziché mediante le corde vocali, mediante una occlusione che causa una turbolenza nel flusso d'aria.

**Codifica a predizione lineare - LPC** Il modello di produzione vocale illustrato consente di riformulare il processo di codifica nel suddividere il segnale vocale in segmenti (di durata da 10 a 30 msec) indicati come *finestre di analisi* che comprendono un numero fisso di campioni, durante i quali il segnale può essere considerato stazionario[4], e condurre una *analisi* (o stima) dei parametri del modello, che consistono in

- il tipo di eccitazione (periodica o caotica), la sua frequenza fondamentale (o *pitch*) se periodica, e la sua intensità,
- il valore e la banda delle frequenze di risonanza

e trasmettere poi questi valori, da usare in un processo di sintesi, come illustrato dalla fig. 18.9. In particolare, il modello del tratto vocale si presta[5] alla definizione di un *predittore lineare* di ordine $p$, ovvero a predire un campione futuro come combinazione lineare di campioni passati. Indicando con $\hat{y}_n$ il valore predetto per il campione $y_n$, scriviamo quindi $\hat{y}_n = \sum_{i=1}^{p} a_i y_{n-i}$ a cui corrisponde un *errore di predizione*

[4] Alla frequenza di campionamento di 8 KHz una finestra di 10 msec contiene 80 campioni; d'altra parte la durata di una sillaba può estendersi da 10-15 msec per le vocali *ridotte*, fino a più di 100 msec per quelle accentate.

[5] La congruenza del modello con la predizione lineare discende dal fatto che poi il filtro di sintesi risulta essere di tipo *recursivo* ovvero *a soli poli*, in accordo appunto con il modello basato sulle risonanze.

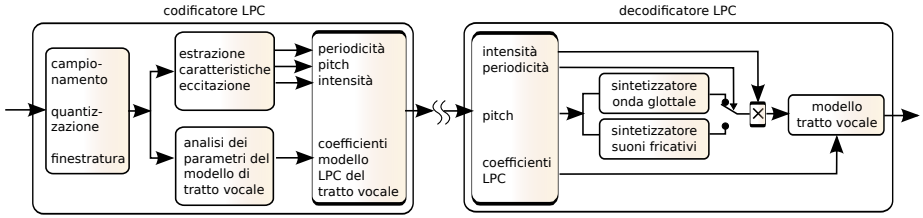



Figura 18.9: Schema di codificatore e decodificatore LPC

$$e_n = y_n - \hat{y}_n = y_n - \sum_{i=1}^{p} a_i y_{n-i} \tag{18.1}$$

I coefficienti $a_i$ si ottengono ora come quelli che rendono minimo il valore atteso di $e_n^2 = \left(y_n - \sum_{i=1}^{p} a_i y_{n-i}\right)^2$ (ovvero, l'energia dell'errore), e quindi cercando quei valori di $a_j$ per cui le derivate parziali $\frac{\partial}{\partial a_j} E\left\{e_n^2\right\}$ si annullano. Scriviamo dunque

$$\frac{\partial}{\partial a_j} E\left\{\left(y_n - \sum_{i=1}^{p} a_i y_{n-i}\right)^2\right\} = 2E\left\{\left(y_n - \sum_{i=1}^{p} a_i y_{n-i}\right) y_{n-j}\right\} = 0$$

ovvero

$$E\left\{y_n y_{n-j}\right\} = \sum_{i=1}^{p} a_i E\left\{y_{n-i} y_{n-j}\right\} \tag{18.2}$$

I valori attesi mostrati sono quindi stimati[6] a partire dai campioni temporali presenti nella finestra di analisi, ovvero

$$\begin{aligned} E\left\{y_{n-i} y_{n-j}\right\} &= R_{yy}(|i-j|) \qquad \text{e ponendo } k = |i-j| \\ &= R_{yy}(k) = \sum_{n=1}^{N-k} y_n y_{n+k} \end{aligned} \tag{18.3}$$

dove l'estremo superiore della sommatoria varia in modo da includere solo i campioni effettivamente presenti nella finestra di analisi[7]. La (18.3) permette di riscrivere (18.2) come

$$R_{yy}(j) = \sum_{i=1}^{p} a_i R_{yy}(|i-j|)$$

che valutata per $j = 1, \cdots, p$ individua un sistema di $p$ equazioni[8] in $p$ incognite

$$\begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} = \begin{bmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \tag{18.4}$$

[6] Sottintendendo una ipotesi di stazionarietà ed ergodicità non vera, ma molto comoda per arrivare ad un risultato.

[7] La (18.3) è effettivamente una stima della *autocorrelazione* del segnale a durata limitata che ricade nella finestra di segnale, mentre l'inclusione nella sommatoria di un numero di termini pari al numero di campioni disponibili porta ad un diverso tipo di risultato, detto *metodo della covarianza*, ed un diverso modo di risolvere il sistema (18.4).

[8] dette di *Yule-Walker*

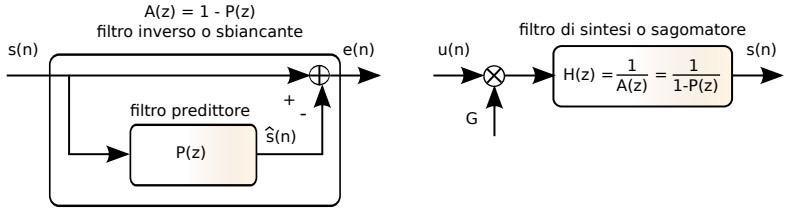



Figura 18.10: Filtro predittore, filtro inverso associato, e filtro di sintesi LPC

che può essere risolto nei termini dei coefficienti $a_i$ mediante metodi particolarmente efficienti[9].

Omettendo una serie di passaggi e considerazioni, evidenziamo che gli $a_i$, una volta determinati, possono essere effettivamente utilizzati nella (18.1), ed il filtro trasversale così ottenuto viene indicato come *predittore*, ed associato ad un polinomio[10] $P(z) = \sum_{i=1}^{p} a_i z^{-i}$, in modo che facendo passare i campioni di segnale $y_n$ attraverso tale filtro, si ottengono i campioni *predetti*; inoltre, il filtro FIR con trasformata zeta $A(z) = 1 - P(z)$ può essere usato per ottenere i campioni dell'errore di predizione (o residuo) $e_n$ come espresso dalla (18.1), e mostrato nel lato sinistro della fig. 18.10. Indicando ora con $G \cdot u_n$ una codifica del residuo $e_n$, il segnale di partenza può essere (quasi) ri-ottenuto come mostrato nella parte destra della fig. 18.10, ossia facendo passare $u_n$ attraverso il filtro IIR $H(z) = \frac{1}{A(z)} = \frac{1}{1-P(z)}$.

Dato che, in base a considerazioni che non svolgiamo, $e_n$ risulta possedere una densità spettrale *bianca*, $|H(z)|^2$ (calcolato per $z = e^{i\omega}$) rappresenta una vera e propria *stima spettrale* del segnale di partenza, come mostrato in fig. 18.11 per diversi valori di $p$, mostrando che per suoni vocalici si ottengono risultati accettabili già per valori tra 8 e 14, mentre per le fricative l'ordine può essere ancora inferiore.

[9] In base alle assunzioni adottate, $R_{yy}(j)$ risulta una funzione pari dell'indice $j$, e la corrispondente matrice dei coefficienti viene detta di *Toeplix*, consentendone l'inversione mediante il metodo di *Levinson-Durbin*, che presenta una complessità $O(n^2)$ anziché $O(n^3)$, come sarebbe necessario per invertire la matrice dei coefficienti.

[10] Una breve analisi della relazione tra DFT e trasformata *zeta* è svolta al § 4.2.1.

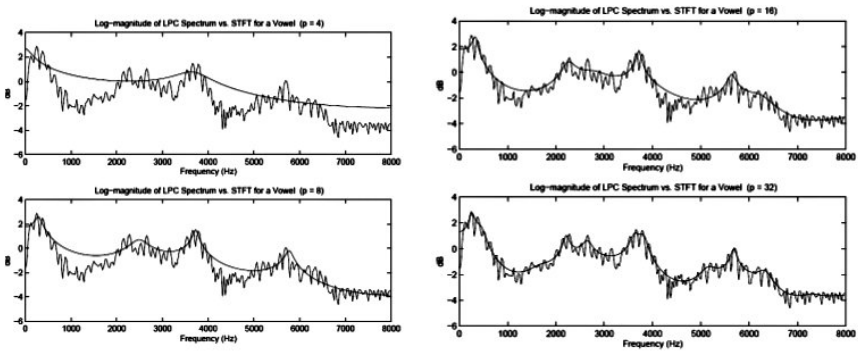



Figura 18.11: Approssimazione spettrale LPC per diversi ordini di predizione


Figura 18.12: Schema di codifica vocale ABS - *Analysis by Synthesis*

Effettivamente, per i suoni sordi si ottengono buoni risultati di sintesi usando come eccitazione un vero e proprio rumore bianco; invece, per i suoni sonori l'uso di forme d'onda impulsive con periodo pari al pitch stimato, sebbene capaci di produrre un bit rate riducibile fino a 2.4 kbps, non fornisce risultati particolarmente utilizzabili, producendo un voce piuttosto robotica. Per questo motivo, si sono sviluppate le tecniche seguenti.

**Predizione lineare ad eccitazione residuale - RELP** In effetti lo schema di sintesi riportato in fig. 18.9 è fin troppo semplificato, epossono essere ottenuti risultati migliori se, dopo aver svolto l'analisi spettrale LPC, il residuo di predizione relativo alla finestra di analisi viene calcolato veramente, e quindi su questo operata una codifica di forma d'onda[11], come avviene per i codificatori RELP (*Residual Excited LP*).

**Analysis by synthesis - ABS** Anziché *calcolare* il residuo di predizione, codificarlo, e trasmetterlo direttamente, la tecnica di *analisi per sintesi* adotta una tecnica *ad anello chiuso*, cercando di trovare quale segnale di eccitazione[12] fornire al filtro di sintesi in modo che il risultato sia quanto più possibile simile al segnale originale (vedi fig. 18.12); quindi, i parametri del filtro di sintesi e della eccitazione sono trasmessi al decoder. La funzione di minimizzazione opera dunque una vera e propria *ricerca tra i possibili segnali* di eccitazione

**Filtraggio percettivo** Sempre in fig. 18.12 si mostra come il processo di minimizzazione prende in considerazione un segnale di errore ottenuto filtrando l'errore effettivo mediante un filtro di *pesatura percettiva*, il cui andamento frequenziale è tendenzialmente *reciproco* rispetto a quello stimato del segnale[13] (vedi fig. 18.13), in modo da attenuare la rilevanza dell'errore di predizione nelle regioni dove c'è più segnale[14] ed esaltarla invece nelle regioni con meno segnale, sfruttando così il fenomeno percettivo

[11] In questo modo si evita di dover operare una esplicita decisione *sonoro/sordo*, visto che in realtà le due fonti di eccitazione posso essere presenti contemporaneamente, come per i cosiddetti suoni *affricati*.

[12] Generato per tentativi, oppure da sceglire in un dizionario di sequenze di eccitazione già codificate.

[13] Il filtro di pesatura percettiva si ottiene a partire dagli stessi coefficienti di predizione $a_i$ che descrivono l'andamento spettrale della finestra di segnale, definendo la sua trasformata zeta come $W(z) = \frac{A(z/\alpha_1)}{A(z/\alpha_2)} = \frac{H(z/\alpha_2)}{H(z/\alpha_1)}$ in cui, se $\alpha_{1,2}$ sono numeri reali, i poli di $W(z)$ si trovano alle stesse frequenze di quelli di $H(z)$ ma con raggio $\alpha_2$ volte maggiore, così come gli zeri di $W(z)$ hanno modulo $\alpha_1$ volte maggiore. Scegliendo $0 < \alpha_{1,2} < 1$ e $\alpha_1 > \alpha_2$ per la $W(z)$ si ottiene l'effetto desiderato, e mostrato in fig. 18.13

[14] La procedura di minimizzazione determina una eccitazione tale da rendere bianco il residuo al suo ingresso; dato però che questo ha subito il filtraggio da parte di $W(z)$, significa che le frequenze da questo depresse sono in realtà enfatizzate per il segnale di errore reale.

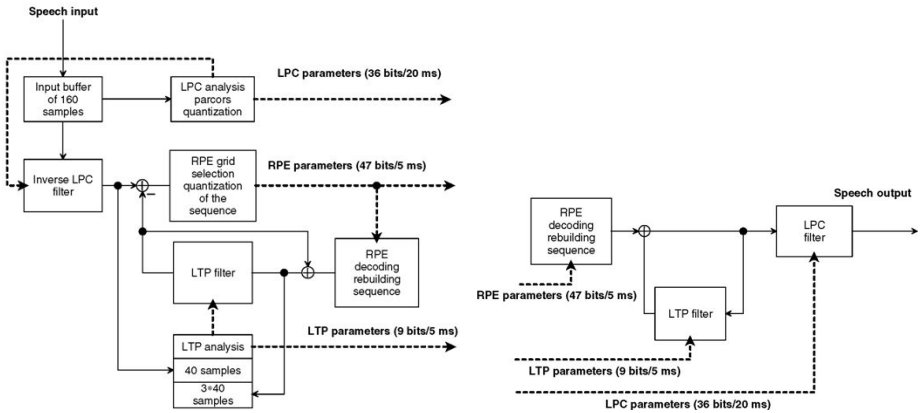



Figura 18.14: Codifica e decodifica GSM 6.10 *full rate* o RPE-LTP

noto come *mascheramento uditivo* (vedi più avanti). Anche se per questa via l'energia totale del rumore è maggiore, l'effetto soggettivo è migliore.

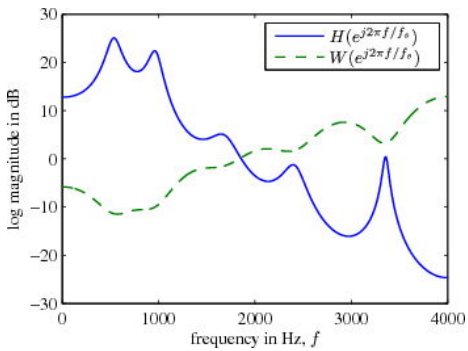



Figura 18.13: Spettro LPC vocalico e relativo filtro di pesatura percettiva dell'errore di predizione

**Multi pulse linear prediction - MPLP** Lo schema operativo suggerito dalla tecnica ABS è stato inizialmente realizzato cercando di *costruire* la sequenza di eccitazione ottima (ossia in grado di minimizzare l'errore pesato percettivamente) come una sequenza di pochi impulsi sparsi, decidendone uno alla volta: pertanto viene inizialmente trovata l'ampiezza e la posizione *ottime* per un primo impulso, poi per un secondo (con il primo fisso), e così via, fino al numero di impulsi desiderati, tipicamente 4-5 ogni 5 msec, ottenuti suddividendo una finestra di 20 msec in quattro sotto-trame, ognuna con 40 campioni, se $f_c = 8000$ Hz.

**Regular pulse excitation with long-term prediction - RPE-LTP o GSM 6.10** Il metodo MPLP presentava una complessità proibitiva, ma ha dato luogo alla versione semplificata RPE-LTP usata inizialmente nella telefonia GSM per fornire una velocità di 13 kbps. In questo caso dopo aver determinato la posizione del primo impulso nella sottofinestra, ne sono piazzati altri 9 ad intervalli regolari (un campione si e tre no), e quindi l'ottimizzazione riguarda solo i valori delle ampiezze.

Rispetto allo schema di fig. 18.12 viene aggiunto un *predittore a lungo termine* o LTP, utilizzato per rimuovere dal segnale di eccitazione l'eventuale periodicità caratteristica dei suoni vocalici, e stimato a partire da sotto-finestre consecutive (vedi fig. 18.14). Il filtro LTP in essenza consiste in un semplice ritardo pari al periodo di pitch (e dunque $\gg p$), ed il predittore LTP relativo (vedi lo schema di decodifica) *ripropone* in uscita una copia ritardata ed attenuata dell'uscita stessa. Il codificatore

GSM pertanto determina ritardo e attenuazione dell'LTP in base all'analisi del residuo di predizione LPC[15], e lo usa per reintrodurre la componente periodica nella sequenza RPE di cui si sta valutando l'idoneità. Una volta che al residuo LPC viene sottratta la componente predicibile per tramite del LTP, ciò che rimane risulta effettivamente assimilabile ad un rumore, ed è indicato anche come *processo di innovazione*.

**Quantizzazione vettoriale** Dato che la codifica del segnale di eccitazione è la parte che impegna più bit da trasmettere, l'evoluzione successiva della codifica vocale fa uso di questa tecnica per evitare la trasmissione di tutti i suoi valori, che sono invece visti come *un vettore* e quindi rappresentati per mezzo di *una unica codeword* che identifica un elemento scelto in un *codebook* ottenuto mediante un processo noto come *quantizzazione vettoriale*.

Una distribuzione di vettori può essere partizionata in più regioni di decisione come quelle mostrate nell'esempio di fig. 18.15, in modo che un nuovo vettore possa essere classificato[16] come contenuto in una di esse, e venire quindi rappresentato dal *centroide* (i punti rossi) associato alla regione. I centroidi ed i confini di decisione sono determinati mediante un procedimento iterativo tale da minimizzare l'errore quadratico medio di rappresentazione[17].

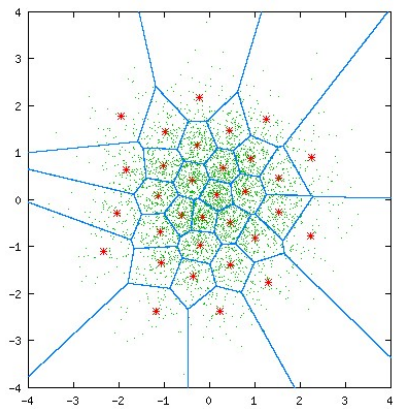



Figura 18.15: Regioni di decisione e centroidi per un quantizzatore vettoriale bidimensionale

I coefficienti dei vettori associati ai centroidi sono quindi memorizzati in un dizionario (o *codebook*) noto sia al codificatore che al decodificatore, in modo che ogni vettore può essere rappresentato, anziché da tutti i suoi coefficienti, dal solo indice della *codeword* del centroide più vicino: al solito, utilizzando $M$ bit per rappresentare l'indice, il codebook sarà formato da $2^M$ diverse codeword. Oltre al codebook utilizzato per rappresentare i possibili vettori di innovazione, la codifica del segnale vocale si può avvantaggiare anche di un secondo codebook usato per approssimare il vettore dei possibili coefficienti spettrali.

**Predizione lineare ad eccitazione codificata - CELP** La fig. 18.16 mostra lo schema realizzativo di un codificatore CELP, in cui sono evidenziati il filtro di predizione a lungo termine ed il filtro LPC, stimati in modalità *ad anello aperto*, ed il *filtro percettivo* che fa in modo che la densità spettrale dell'errore di predizione sia concentrata nelle regioni dove è presente segnale.

Per ogni codeword di eccitazione selezionata dal codebook, ed il guadagno associato, viene calcolata l'energia dell'errore ottenuto, ed il risultato confrontato con quello

[15] In effetti, mentre i coefficienti spettrali (denominati *parcor* in questo caso) sono determinati a partire dall'analisi dell'intera finestra di 20 msec, l'eccitazione RPE ed i parametri LTP sono ottenuti a partire da *sottofinestre* di 40 campioni, pari a 5 msec.

[16] Per questa classificazione, così come per poter definire l'insieme dei centroidi, occorre che sia definita una funzione di *distanza* tra vettori.

[17] Vedi `http://www.data-compression.com/vq.html`, ma anche la nota 53 a pag. 150.

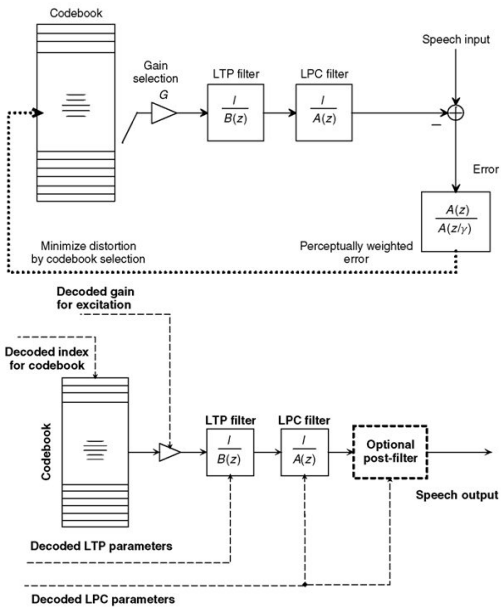



Figura 18.16: Codificatore e decodificatore CELP

ottenibile mediante le altre codeword, finché non si trova la codeword che minimizza l'errore. Ovviamente questo modo di procedere è estremamente oneroso, ma si sono trovati metodi di ricerca più efficienti adottando tecniche di costruzione del codebook come combinazione di sequenze elementari, dando luogo alla famiglia dei codificatori *algebrici* o ACELP.

D'altra parte, anche l'identificazione del LTP può essere ricondotta ad una ricerca ad anello chiuso, svolta ora nell'ambito di un *codebook adattivo*, costruito a partire dalla precedente sequenza di eccitazione ottima, replicata in forma traslata di un campione alla volta, come illustrato in fig. 18.17, che mostra appunto l'uso della eccitazione per la trama precedente per popolare il codebook adattativo: da questo viene quindi individuata la codeword $I_a$ ed il guadagno $G_a$ ottimi, e quindi individuata la codeword di innovazione $I_s$ e $G_s$ ottimi, riferiti ad un codebook detto *stocastico* perché costituito da sequenze pseudo casuali.

Infine, viene mostrato come anche i coefficienti spettrali LPC sono trasmessi mediante una codeword derivata da un processo di quantizzazione vettoriale. Possiamo elencare i seguenti standard che adottano una tecnica di questo tipo:

- Federal Standard 1016 (4800-16000 bit/s) CELP
- ITU-T 8-kbit/s G.729 CS-ACELP (*conjugate-structure algebraic* CELP);
- dual-rate multimedia ITU-T G.723.1 a 5.3 kbit/s con ACELP e 6.3 kbit/s con MP-MLQ (*multi-pulse maximum likelihood quantization*);
- ITU-T low-delay CELP 16-kbit/s G.728 - usa finestre di analisi multo brevi e una predizione lineare all'indietro per conseguire un ritardo di 2 msec;
- ETSI enhanced full-rate EFR-GSM e half-rate HR-GSM, con velocità di 12.2 e 5.6 kbps, così come i codec AMR (*adaptive multirate*) e WB-AMR, con velocità da 7.95 a 4.75 kbps;
- Speex[18] - un insieme di codecs open source esenti da brevetti e liberamente utilizzabili, con velocità (a banda stretta) da 5,95 a 24,6 kbps, e da 5.75 a 42,4 kbps per segnali con banda di 16 kHz

### 18.1.3 Codifica psicoacustica

Mentre la codifica di forma d'onda (§ 18.1.1) non fa assunzioni a riguardo della natura del segnale, i metodi esposti al § 18.1.2 sono tutti fortemente orientati a rappresentare segnali vocali. D'altra parte, il gruppo di lavoro MPEG di ISO si è dedicato ad individuare metodi di codifica idonei alla trasmissione di segnali multimediali di natura qualsiasi, come ad esempio brani musicali. Inoltre, i vincoli relativi al basso ritardo

[18] vedi `http://en.wikipedia.org/wiki/Speex`

necessario ad assicurare un buon grado di interattività vengono meno, e si possono dunque intrapredendere elaborazioni più complesse, e che richiedono un tempo maggiore. Infine, vengono trascurati rigidi vincoli sulla velocità risultante, accettando invece che questa *vari* nel tempo in funzione del tipo di segnale da rappresentare.

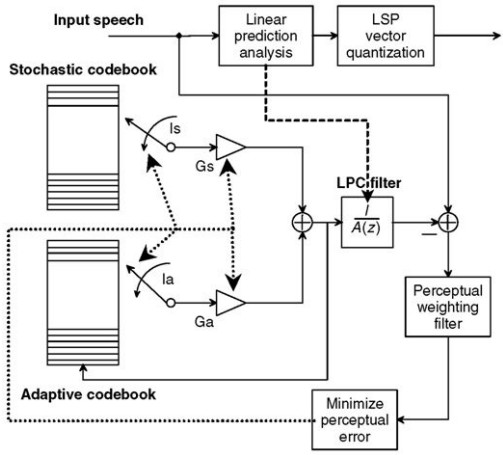



Figura 18.17: Codificatore CELP con codebook adattivo per la predizione a lungo termine

Come vedremo tra breve, per queste tecniche si fa di nuovo uso di una codifica per sottobande, introdotta nella discussione dell'ADPCM, tenendo però anche conto di caratteristiche molto importanti della percezione sonora, il cui sfruttamento è già stato illustrato nella discussione del filtro di pesatura percettiva, ma che ora hanno un impatto ancora maggiore sulla realizzazione del codificatore. I codificatori che fanno uso di queste caratteristiche sono l'MPEG *layer* 3 o MP3, il *Dolby* AC, e l'*advanced audio coding* o AAC.

**Sensibilità uditiva e mascheramento in frequenza** La fig. 18.18a mostra la curva di sensibilità del sistema uditivo, ovvero il livello di intensità minimo perché possa essere percepito un suono: come si vede, questo è molto variabile con la frequenza, per cui anche se il suono B (sinusoide o tono puro) ha la stessa intensità di A non può essere udito, mentre invece A si. Ma ad una analisi più approfondita, si scopre che la presenza di un suono in una determinata regione di frequenza ha l'effetto di modificare la curva di sensibilità per le frequenze vicine, di fatto *mascherando* suoni a frequenze vicini che altrimenti avrebbero superato la soglia di sensibilità, come mostrato in fig. 18.18b: la presenza del suono B rende A non più udibile.

In realtà, l'estensione in frequenza per cui si verifica l'effetto di mascheramento dipende sia dalla frequenza del tono mascherante (come mostrato in fig. 18.19 ottenute con toni a 1, 4 ed 8 kHz) che dalla sua intensità. In particolare, la banda delle frequenze mascherate viene detta *banda critica* ed ha una estensione differente

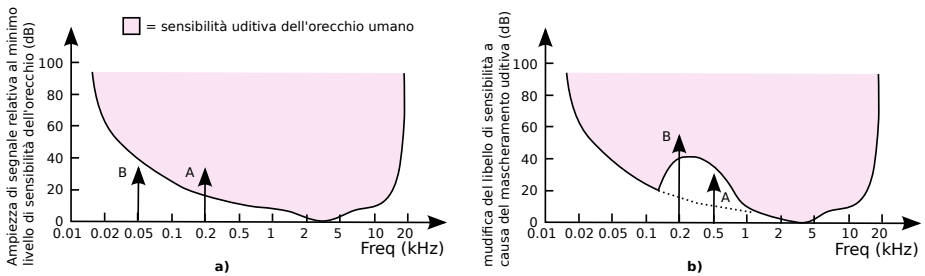



Figura 18.18: a) sensibilità uditiva alle diverse frequenze; b) mascheramento uditivo

alle diverse frequenze: si trova che sotto i 500 Hz la banda critica ha una estensione di circa 100 Hz, mentre a frequenze superiori aumenta (circa) linearmente per multipli di 100 Hz. Ad esempio, un segnale ad 1 KHz (2x500) produce una banda critica di 200 Hz (2x100), mentre a 5 kHz (10x500) questa vale circa 1 kHz (10x100).

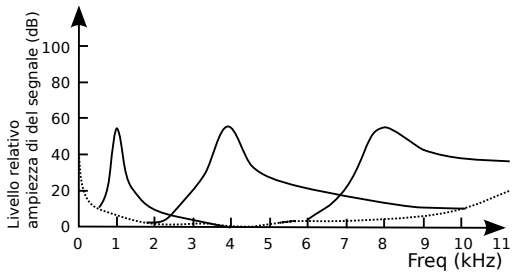



Figura 18.19: Variazione della banda critica in funzione della frequenza

**Mascheramento temporale** Il secondo effetto percettivo riguarda ancora una modifica alle curve di sensibilità, stavolta in modo *non selettivo* in frequenza, ma che coinvolge tutte le frequenze: si verifica infatti che dopo aver udito un suono forte, per il tempo necessario all'estinzione del suono e che tipicamente dura qualche decina di millisecondi (vedi fig. 18.20), l'orecchio non è più in grado di percepire suoni con intensità minore a quello che si sta estinguendo.

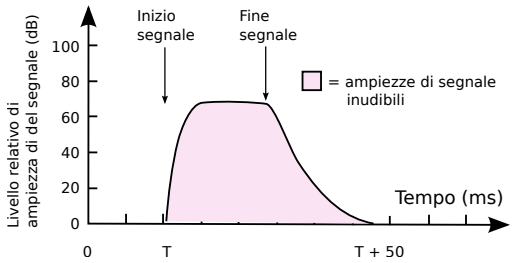



Figura 18.20: Mascheramento temporale

**MPEG layer 3** Il gruppo di lavoro MPEG di ISO ha definito uno standard di codifica audio basato su tre livelli di complessità (e potere di compressione) crescente, ed il terzo (o MP3) è quello di gran lunga più popolare, anche grazie alla diffusione che ha avuto via Internet.

Lo schema di funzionamento di principio è mostrato in fig. 18.21: il segnale campionato in ingresso (a 32, 44.1 o 48 kHz) transita attraverso un PCM *encoder* che esegue un filtraggio[19] in 32 sottobande di eguale ampiezza, le cui uscite sono campionate a frequenza $1/32$ di quella di ingresso. Ogni 384 campioni di ingresso (pari a 12 msec se $f_c = 32$ kHz) sono quindi prodotti $384/32 = 12$ campioni per ogni sotto-banda, e per ognuna di esse è individuato il valore del campione più grande, che contribuisce sia ad impostare la dinamica del quantizzatore per quella banda, sia come parametro per il modello psicoacustico.

Il modello psicoacustico riceve le informazioni prodotte da un banco di filtri di analisi realizzati mediante una MDCT,[20] che produce una stima spettrale con risoluzione maggiore di quella del primo banco di filtri, su cui basare le valutazioni di mascheramento uditivo, che a loro volta determinano per ogni sottobanda l'indicazione di un *signal to mask ratio* ($SMR$), che a sua volta determina *quanti bit utilizzare* (e quindi

[19] Eseguito mediante un banco di *filtri polifase*, vedi ad es. `http://en.wikipedia.org/wiki/Polyphase_quadrature_filter` o `http://cnx.org/content/m32148/latest/`. Le uscite dei filtri polifase, anche se campionate a frequenza inferiore a quella di Nyquist, sono esenti da aliasing, che viene cancellato dall'effetto delle altre sottobande.

[20] Vedi `http://en.wikipedia.org/wiki/Modified_discrete_cosine_transform`

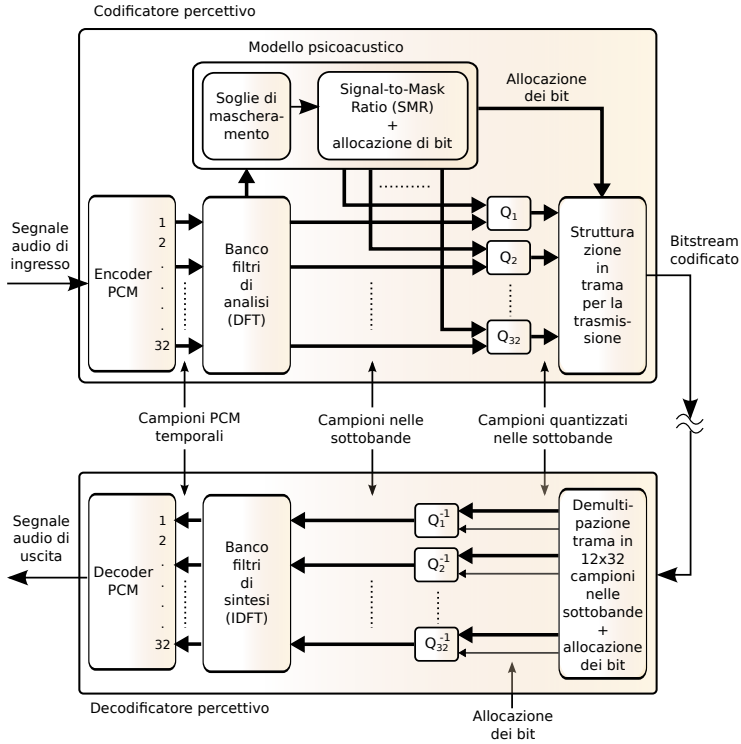



Figura 18.21: Codec percettivo MPEG

quanti livelli) per la quantizzazione $Q$ dei campioni relativi alle singole sottobande. Quelle contraddistinte da una maggiore sensibilità (ovvero nelle quali si percepiscono anche suoni deboli) saranno quantizzate con più accuratezza, e quindi con più bit e meno rumore; mentre le sottobande caratterizzate da una sensibilità inferiore possono essere quantizzate con meno bit, almeno finché l'$SNR$ di quantizzazione si mantiene superiore all'$SMR$, dato che in tal caso il rumore è mascherato, e dunque non viene udito. Quindi, i 12 campioni delle 32 sottobande sono quantizzati tenendo conto sia della dinamica effettiva, che del numero di livelli in cui suddividere la dinamica. Infine, viene prodotta una struttura di trama che contiene, oltre ai campioni, anche le informazioni sulla effettiva allocazione dei bit.

Ad una futura edizione, un trattazione più approfondita.

**Riferimenti** Si citano dei riferimenti essenziali sulla codifica audio, da cui sono anche tratte alcune illustrazioni

- Introduction to Digital Speech Processing, L. R. Rabiner and R. W. Schafer, `http://www.nowpublishers.com/product.aspx?product=SIG&doi=2000000001&section=xstart`
- Beyond VoIP Protocols: Understanding Voice Technology And Networking, O. Hersent, J.P. Petit, D. Gurle `http://what-when-how.com/category/voip-protocols/`
- `http://www.data-compression.com/index.shtml`

| | *banda* | *linee* | *fps* | *aspetto* | *colonne* | *righe* | *colore* |
|---|---|---|---|---|---|---|---|
| PAL | 6 MHz | 625 | 25 int | 4:3 | | | |
| NTSC | 5 MHz | 525 | 30 int | 4:3 | | | |
| HDTV | | 1080 | | 4:3 | 1440 | 1152 | |
| | | | | 16:9 | 1920 | 1152 | |
| SVGA | | | | 4:3 | 1024 | 768 | |
| 4:2:2 | | 625/525 | 50/60 non int | 4:3 | 720 | 576/480 | 360 × 576/480 |
| 4:2:0 | | 625/525 | 25/30 int | 4:3 | 720 | 576/480 | 360 × 288/240 |
| VGA | | | | 4:3 | 640 | 480 | |
| SIF | | 625/525 | 25/30 non int | 4:3 | 360 | 288/240 | 180 × 144/120 |
| CIF | | | 30 non int | 4:3 | 360 | 288 | 180 × 144 |
| QCIF | | | 15:7.5 non int | 4:3 | 180 | 144 | 90 × 72 |

Tabella 18.1: Griglia dei parametri corrispondenti ai formati video

- Voice Acoustics: an introduction - University of New South Wales, J. Wolfe, M. Garnier, J. Smith `http://www.phys.unsw.edu.au/jw/voice.html`
- Let's build an MP3-decoder! - Björn Edström `http://blog.bjrn.se/2008/10/lets-build-mp3-decoder.html`

## 18.2 Codifica di immagine

Un segnale di immagine può essere di tipo *vettoriale*[21], come nel caso di un disegno prodotto da un *plotter*, e rappresentato mediante un linguaggio descrittivo che codifica le operazioni grafiche necessarie alla sua realizzazione; al contrario, un segnale di immagine è detto di tipo *bitmap*, o *raster* (griglia, reticolo), quando è il risultato di un campionamento spaziale, come nel caso di una foto digitale, di un fax, o del risultato di un processo di scansionamento elettronico. Mentre le immagini vettoriali sono pienamente scalabili e ridimensionabili senza perdita di definizione, quelle bitmap sono ottimizzate per essere riprodotte nelle loro dimensioni originali, avendo già operato un processo di distorsione tale da sfruttare al più possibile le caratteristiche di predicibilità e di sensibilità percettiva.

### 18.2.1 Dimensioni

Per quanto riguarda le immagini bitmap, queste sono definite nei termini di una matrice di elementi di immagine o PIXEL (*picture elements*)[22], che sono l'equivalente bidimensionale dei campioni estratti da un segnale unidimensionale. Per ogni pixel è definito un valore associato alla intensità con la quale deve essere riprodotto: nel caso di immagini a colori, sono necessari tre valori di intensità, per cui una immagine è in realtà descritta da tre matrici, come approfondiamo di seguito.

[21] Esempi di formati per la grafica vettoriale sono SVG, EPS, PDF, e VRML.

[22] Per alcuni anni, si è usato come sinonimo anche il termine PEL `http://www.foveon.com/files/ABriefHistoryofPixel2.pdf`.

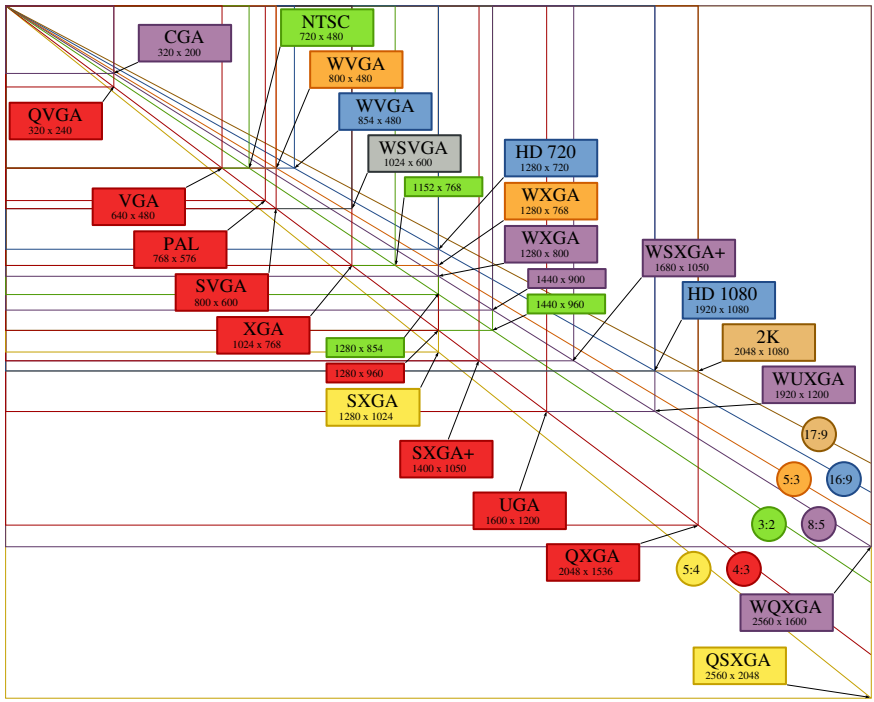



Figura 18.22: Risoluzioni standard o modalità video digitale

Sebbene le dimensioni della matrice di pixel possano essere qualunque, nel corso del tempo si sono affermate una serie di valori di riferimento, associati ad altrettante serie di sigle, legate al tipo di dispositivo che deve poi riprodurre l'immagine, ma anche a quello da cui l'immagine viene acquisita; la tabella 18.1 riassume tali corrispondenze.

Ad esempio, la risoluzione VGA (640 x 480) trae origine dai parametri dello standard NTSC della televisione analogica, i cui quadri sono composti da una serie di 525 linee, di cui solo 480 visibili: volendo mantenere una risoluzione orizzontale pari a quella verticale, con un rapporto d'aspetto di 4:3, ogni linea deve essere campionata su 480/3x4 = 640 punti. Prima ancora dell'uso broadcast della TV digitale, la raccomandazione BT 601[23] stabilisce le regole per la conversione tra standard video differenti, mediante l'uso di una comune frequenza di campionamento del segnale video a 13.5 MHz, individuando così nei 52 $\mu$sec ([24]) di una linea 52 x $10^{-6}$x 13.5 $^{6}$ = 702 campioni per linea, a cui si aggiungono 9 campioni neri in testa ed in coda per ottenere 720 campioni per linea; per un segnale a 525 linee si ottiene quindi la matrice 720 x 480 del formato 4:2:2, che approfondiremo tra breve.

Le matrici più grandi di 1024 x 768 sono spesso descritte in termini di *Megapixel* (es 1600 x 1200 = 1,9 Mpixel), spesso usati per confrontare la risoluzione (ma non necessariamente la qualità) degli attuali mezzi di cattura fotografica digitale; inoltre, i

[23] Il sito di ITU-R `http://www.itu.int/ITU-R/index.asp?category=information&link=rec-601&lang=en` non consente l'accesso pubblico alla raccomandazione. Un approfondimento può essere svolto presso Wikipedia `http://it.wikipedia.org/wiki/BT.601`.

[24] Vedi fig. 11.3 a pag. 281.

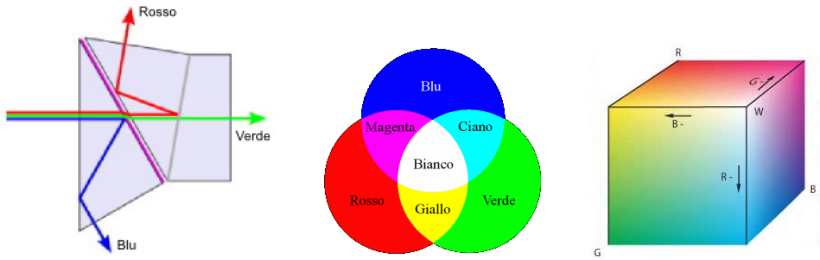



Figura 18.23: Prisma dicroico, sintesi cromatica additiva, cubo dei colori

*grandi formati* traggono origine anche dalla tecnologia delle schede video per computer da un lato, e da quella della televisione ad alta definizione da un altro, come riassunto nella figura 18.22[25].

Il formato SIF (*source intermediate format*) è ottenuto a partire dal 4:2:2, conservando la metà dei pixel sia in verticale che in orizzontale, e trascurando la metà dei quadri di immagine; il suo uso è orientato alla memorizzazione, e quindi usa una scansione non interallacciata. Il formato CIF (*common intermediate format*) è simile al SIF, tranne per aver perso il riferimento al numero di linee analogiche da cui deriva; il suo uso è orientato ai sistemi di videoconferenza, e da questo sono definiti formati a maggior risoluzione, come il 4CIF ed il 16CIF, equivalenti al 4:2:2 ed all'HDTV. Il formato QCIF (*quarter* CIF) è orientato alla videotelefonia, dimezzando ancora sia la risoluzione spaziale che quella temporale. Da questo è a sua volta derivato il formato SUB-QCIF (o S-QCIF) di 128 x 96 pixel, orientato a collegamenti lenti come quelli via modem.

### 18.2.2 Spazio dei colori

I dispositivi di acquisizione e riproduzione di immagini a colori operano su tre diverse matrici di pixel, che rappresentano i tre colori di base della *sintesi additiva*, ossia *rosso*, *verde*, e *blu*, o RGB (dalle iniziali inglesi *Red*, *Green* e *Blue*). In figura 18.23 viene mostrato il principio di funzionamento di un *prisma dicroico*, che devia le tre componenti di colore verso tre diversi dispositivi di acquisizione. Variando quindi la proporzione con cui si sommano gli stimoli dei tre colori, si ottiene, oltre al bianco, anche qualunque altro colore. Sebbene dalle figure riportate sembra che il bianco risulti dal contributo in parti uguali delle tre componenti RGB, in realtà la scala di grigi della immagine *monocromatica* corrispondente si ottiene calcolando un segnale $Y$ di *luminanza* secondo la formula

$$Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \tag{18.5}$$

che è quella usata per modulare il segnale video analogico[26]. Come già discusso, in tale ambito la componente di colore viene trasmessa utilizzando due altri segnali, $C_b$ o *crominanza blu* e $C_r$ o *crominanza rossa*, secondo la formula

$$C_b = B - Y \quad \text{e} \quad C_r = R - Y \tag{18.6}$$

[25] La figura è tratta da Wikipedia, dove possono essere approfonditi gli altri aspetti legati a queste risoluzioni video `http://it.wikipedia.org/wiki/Risoluzioni_standard`.

[26] Vedi nota 39 a pag. 282.

Disponendo dei segnali $Y$, $C_b$ e $C_r$, si possono riottenere i valori RGB inserendo la (18.5) nelle (18.6), e risolvendo il sistema di tre equazioni in tre incognite risultante.

**Segnale video composito** Al § 11.4.4 abbiamo descritto come nel segnale televisivo analogico la componente di colore sia trasmessa assieme alla luminanza, su di una diversa portante, con modulazione in fase e quadratura. In realtà, per diversi motivi, le componenti trasmesse non sono direttamente quelle individuate dalle (18.6), ma piuttosto componenti denominate $U$, $V$ oppure $I$, $Q$, e così definite:

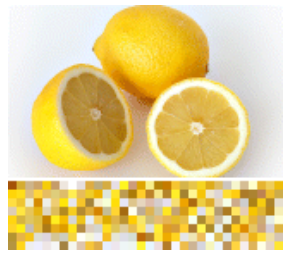
$$\begin{array}{lrcl@{\qquad}lrcl} PAL: & U & = & 0.493 \cdot C_b & NTSC: & I & = & 0.74 \cdot C_r - 0.27 \cdot C_b \\ & V & = & 0.877 \cdot C_r & & Q & = & 0.48 \cdot C_r + 0.41 \cdot C_b \end{array}$$

Pertanto, in funzione delle diverse modalità di rappresentazione, un segnale video a colori può essere descritto indifferentemente da una delle seguenti quattro terne di segnali: RGB, $YC_rC_b$, YUV, YIQ.

Una descrizione alternativa dello spazio di colore è fornita dai parametri di *Tinta*, *Saturazione* e *Chiarezza*, ovvero HUE, SATURATION e LIGHTNESS, o HSL: si tratta di attributi più legati alla descrizione percettiva che non alle tecnologie della riproduzione dell'immagine. Mentre la tinta descrive una famiglia di colori (es tutti i rossi), la saturazione ne indica il grado di purezza, ossia la presenza congiunta di altre tonalità; la chiarezza, infine, denota la luminosità del colore, rispetto ad un punto bianco. La terna HSL viene a volte usata per descrivere un colore nell'ambito di programmi di *computer graphic*, mediante i quali è fornito anche l'equivalente RGB.

**Profondità di colore** Dato che l'occhio umano non distingue più di 250 tinte diverse, e di 100 livelli di saturazione, si ritiene che utilizzare 8 bit per ogni componente dello spazio di colore RGB sia più che sufficiente. Con 8x3=24 bit per pixel (bpp) si possono infatti rappresentare $2^{24} - 1$ diversi colori, ovvero più di 16 milioni, molti dei quali indistinguibili ad occhio nudo. Modalità più spinte di quella a 24 bpp (detta *truecolor*) adottano 10, 12, 16 bit/componente, o rappresentazioni in virgola mobile, e sebbene non migliorino la qualità visiva, possono comunque essere usate in contesti professionali, per non perdere precisione nelle operazioni di editing ripetuto. Al contrario, profondità inferiori sono comunemente usate per risparmiare memoria, come nel caso di 15 bpp, che usa 5 bit per componente, o 16 bpp, che usa 6 bit per il verde, offrendo 65.536 colori diversi.

**Palette** Nel caso si decida di adottare profondità molto ridotte, come 8 bpp, si preferisce ricorrere ad una modalità detta a *colore indicizzato*: l'insieme dei colori presenti nell'immagine viene *quantizzato*[27] in un insieme ridotto, i cui valori a 24 bpp sono memorizzati in una tavolozza (la *palette* detta anche *colour look-up table* o CLUT), che viene quindi utilizzata come un dizionario. La figura a lato mostra una immagine di esempio, assieme alla palette dei colori che usa. In questo modo, per ogni pixel dell'immagine è ora sufficiente specificare l'indice della palette dove è memorizzata la rappresentazione a 24 bpp del colore più prossimo.



[27] Per una breve introduzione alla *quantizzazione cromatica*, può essere consultata Wikipedia http://en.wikipedia.org/wiki/Color_quantization

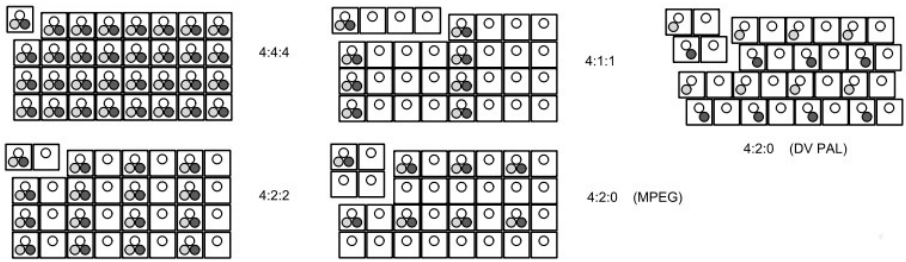



Figura 18.24: Sottocampionamento delle componenti di colore

**Esempio** Consideriamo una immagine in formato VGA rappresentata mediante una palette di 256 elementi da 24 bit: ognuno dei 640x480 = 307.200 pixel può quindi assumere uno tra 256 diversi colori, scelti tra $2^{24} = 16$ milioni. La dimensione di memoria occupata si ottiene considerando che per ogni pixel occorrono 8 bit per l'indice nella palette, e che la palette stessa ha dimensioni 256x24 = 6144 bit = 768 byte, e quindi in totale 307.968 byte.

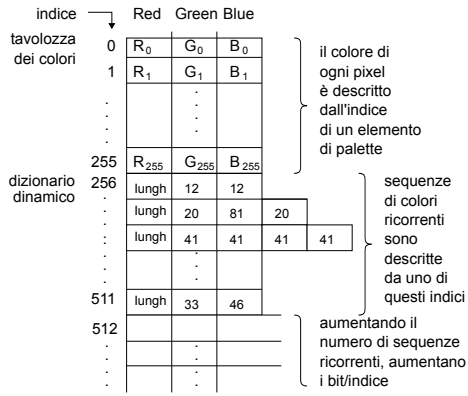
**Sottocampionamento del colore** Nella tabella riportata a pag. 452 è presente la colonna *colore*, che mostra come la dimensione riservata alle matrici di pixel che codificano le informazioni di crominanza sia ridotta di metà rispetto a quella della luminanza. Questo fatto trae origine da due buoni motivi: il primo è che l'acutezza visiva dell'occhio umano per ciò che riguarda le variazioni cromatiche è ridotta rispetto a quella relativa alle variazioni di luminosità; il secondo è che il segnale di crominanza presente nel segnale video composito occupa una banda circa metà di quella del segnale di luminanza. Pertanto, le componenti di luminanza sono generalmente campionate con una risoluzione spaziale inferiore a quella del segnale di luminanza. Il tipo di sottocampionamento spaziale adottato per le componenti di crominanza è generalmente caratterizzato da quattro numeri, in accordo allo schema seguente:

- **4:4:4** - Non si effettua sottocampionamento, e le tre componenti hanno lo stesso numero di campioni. Applicato principalmente a segnali RGB trattati in studio di produzione.
- **4:2.2** - Questo schema si applica tipicamente alle rappresentazioni $YC_bC_r$, memorizzando per ogni 4 campioni di luminanza, 2 campioni della componente $C_b$ e 2 della componente $C_r$, ed è utilizzato in ambito professionale e broadcast.
- **4:1:1** - In questo caso ogni quattro campioni di luminanza su una riga, ne viene preso uno per $C_b$ ed uno per $C_r$. E' lo schema usato nello standard DV NTSC.
- **4:2:0** - Ogni 4 campioni di luminanza, ne vengono salvati uno per $C_b$ ed uno per $C_r$ come per il caso 4:1:1, ma ora la crominanza è campionata su righe alterne. In particolare, la versione utilizzata per l'MPEG-1 campiona assieme entrambi i segnali di crominanza, una riga si ed una no, mentre quella usata con il DV PAL li campiona a righe alternate, e prevede una riproduzione in modalità interallacciata.

### 18.2.3 Formato GIF

Il *Graphics Interchange Format* è un formato ad 8 bpp definito da *CompuServe* nel 1987[28] e da allora ha continuato ad essere molto popolare. Usa una *palette* con cui rappresentare 256 colori scelti tra 16 milioni, e quindi comprime l'immagine mediante l'algoritmo LZW, individuando sequenze ricorrenti dei valori di colore. Un singolo file può contenere più immagini (ognuna con la sua palette) in modo da realizzare brevi animazioni. Il numero ridotto di colori rende il formato poco idoneo alla riproduzione di fotografie, ma più che adatto ad immagini più semplici, come ad es. un logo di pagina web. Per rappresentare i colori assenti dalla palette, il codificatore può ricorrere ad una operazione di *dithering,* alternando colori che, osservati da lontano, ricreano l'effetto della tonalità mancante.

Il metodo di compressione è illustrato con l'ausilio della figura che segue, e adotta come anticipato l'algoritmo LZW[29], il cui dizionario è inizialmente composto dalla palette, o meglio dai 256 valori ad 8 bit che indicizzano la terna RGB a 24 bit nella palette. Quando si incontra una sequenza di codici di colore già osservata, viene aggiunta una riga al dizionario, ed il valore dell'indice corrispondente viene usato per rappresentare tutta la sottosequenza; eventualmente, il numero di bit usati per indicare le righe del dizionario viene aumentato di uno. Per disegnare le sequenze di pixel rappresentate da indici inclusi nella sezione dinamica della tabella, occorre dunque individuare prima le rispettive terne RGB nella tavolozza.




**PNG** Dato che la compressione LZW era stata brevettata, venne sviluppata una codifica alternativa, denominata *Portable Network Graphics.* Al giorno d'oggi i brevetti relativi al formato GIF sono tutti scaduti, ed il formato PNG è stato standardizzato nella RFC 2083[30]. Come per GIF, anche PNG è di tipo *lossless* (senza perdite), ossia individua una compressione invertibile, capace di replicare in modo identico l'immagine di partenza, ovviamente senza considerare il processo di quantizzazione che porta alla generazione della palette. Oltre alla modalità di colore indicizzato, PNG offre anche una modalità *truecolor* a 24 o 32 bpp, e per questo può correttamente rappresentare anche materiale fotografico, al punto da consigliare l'uso di PNG (anzichè JPEG) nel caso si prevedano successive operazioni di editing dell'immagine.

Per quanto riguarda la compressione, PNG fa uso dell'algoritmo *deflate*, preceduto da un passaggio di compressione differenziale, in cui al valore che rappresenta il colore di un pixel viene sottratto il valore predetto a partire dai pixel adiacenti: in tal modo l'algoritmo *deflate* riesce a conseguire rapporti di compressione più elevati, riuscendo quasi sempre a battere le prestazioni di GIF.

[28] Il documento di specifica può essere trovato presso W3C: http://www.w3.org/Graphics/GIF/spec-gif89a.txt

[29] Vedi § 17.1.1.7.

[30] Reperibile presso il sito di IETF: `http://tools.ietf.org/html/rfc2083`

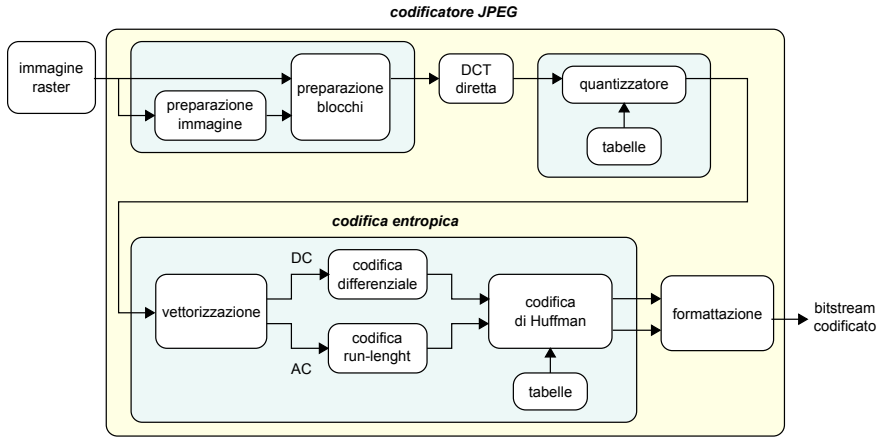



Figura 18.25: Stadi di elaborazione nella compressione jpeg

### 18.2.4 Codifica JPEG

Il *Joint Photographic Experts Group* è un comitato congiunto ISO/ITU che ha definito lo standard internazionale per la compressione di immagini ISO 10918-1[31], particolarmente adatto alla codifica di immagini fotografiche. Descriviamo di seguito il funzionamento della modalità operativa detta *baseline*, o *lossy sequential mode*, che è quella che offre il migliore grado di compressione, e che prevede cinque stadi di elaborazione, mostrati alla fig. 18.25: preparazione dei blocchi, Discrete Cosine Transform (DCT), quantizzazione, codifica entropica, e formattazione.

**Preparazione dell'immagine e dei blocchi** L'immagine *raster* di partenza è formata da una o più matrici bidimensionali di valori (scala di grigi, oppure a colori indicizzati, o RGB, $YC_rC_b$, YUV, ...), eventualmente di dimensioni differenti (come nel caso $YC_rC_b$). Sebbene sia possibile elaborare direttamente una rappresentazione RGB, le migliori prestazioni si ottengono nello spazio $YC_rC_b$ con sotto-campionamento spaziale 4:2:2 o (meglio) 4:2:0, e dunque il primo passo è quello di convertire l'immagine in questa modalità di rappresentazione.

Ogni matrice viene quindi suddivisa in *blocchi* della dimensione di 8x8 pixel[32], ognuno dei quali è elaborato in sequenza in modo indipendente dagli altri.

**DCT diretta** Prima di procedere, la matrice Y (oppure le tre matrici R, G e B) che contiene valori ad 8 bit tutti positivi, viene normalizzata sottraendo ad ogni pixel il valore 128, in modo da ottenere valori tra -128 e 127. Quindi, per ogni blocco di 8x8 pixel, i cui valori indichiamo con $p(x, y)$, viene calcolata una nuova matrice di 8x8 valori $D(i, j)$ ottenuti come coefficienti di una *trasformata coseno discreta* (DCT)

[31] Scaricabile presso il W3C: `http://www.w3.org/Graphics/JPEG/itu-t81.pdf`

[32] Notiamo incidentalmente come le dimensioni definite nella tabella di pag 452 siano multipli interi di 8. Se questo non è il caso, i blocchi ai bordi destro ed inferiore vengono riempiti con pixel scelti in modo da minimizzare le distorsioni risultanti.

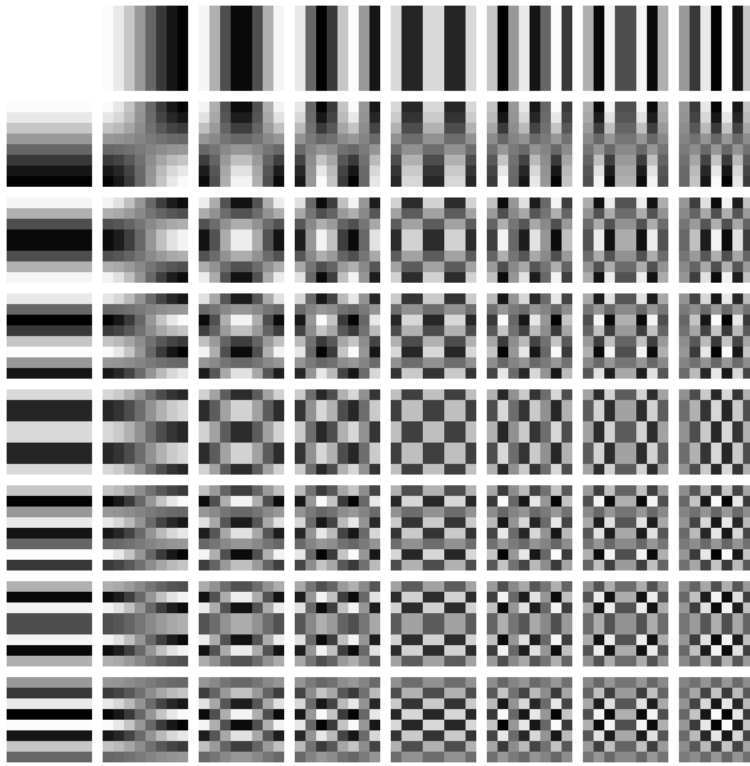


Figura 18.26: Rappresentazione grafica delle superfici DCT

bidimensionale (vedi §4.2.2):

$$D(i,j) = \frac{1}{4} c_i c_j \sum_{x=0}^{7} \sum_{y=0}^{7} p(x,y) \cos \frac{(2x+1)\, i\pi}{16} \cos \frac{(2y+1)\, j\pi}{16}$$

in cui $c_i$ e $c_j$ sono ognuno pari a $1/\sqrt{2}$ con indice $i$ o $j$ pari a zero, oppure $c_i = c_j = 1$ negli altri casi, mentre gli indici $i$ e $j$ variano tra zero e sette. Tralasciando di approfondire le relazioni esistenti tra DCT e DFT[33], consideriamo invece come i coefficienti $D(i,j)$ così ottenuti permettano la ricostruzione della matrice originaria nei termini di una somma pesata delle superfici rappresentate (per mezzo di una scala di grigi) nel diagramma riportato alla figura 18.26, mediante l'applicazione della DCT *inversa*

$$p(x,y) = \frac{1}{4} \sum_{i=0}^{7} \sum_{j=0}^{7} c_i c_j D(i,j) \cos \frac{(2x+1)\, i\pi}{16} \cos \frac{(2y+1)\, j\pi}{16}$$

Ma se fosse tutto qui, non avremmo realizzato la funzione di compressione! Questa è infatti realizzata dalle elaborazioni successive, a partire dalla rappresentazione in

[33] Potremmo tentare comunque di estendere le considerazioni svolte al § 4.2.2 al caso bidimensionale...

termini di blocchi DCT, di cui ora approfondiamo il significato. Osserviamo quindi che ognuna delle superfici elementari rappresentate in fig. 18.26 è legata ad una coppia $i, j$ associata ad un coefficiente della DCT calcolata, in modo che tale coefficiente esprime il contenuto di frequenze spaziali descritto da quella particolare funzione della base. Per questo l'elemento $(i, j) = (0, 0)$ in alto a sinistra, ad andamento costante, è indicato come *coefficiente* DC, o componente continua, dato che essendo calcolato come somma di tutti i pixel, riflette un valore che è legato alla intensità media dell'intero blocco. I coefficienti legati alle funzioni della prima riga rappresentano contenuti di frequenza spaziale orizzontale, con un periodo via via minore spostandosi verso il margine destro, mentre quelli della prima colonna, frequenze verticali. I coefficienti localizzati all'interno della matrice esprimono contenuti di frequenze spaziali in entrambe le direzioni, con valori di frequenza tanto più elevati, quanto più ci si sposta verso l'angolo in basso a destra. Pertanto, i coefficienti descritti da indici diversi da $(0, 0)$ sono indicati come *coefficienti* AC.

L'esperienza pratica mostra come quasi sempre i coefficienti $D(i, j)$ presentino nella regione in alto a sinistra valori ben più elevati di quelli riscontrabili in basso a destra, come conseguenza della predominanza dei blocchi posti in corrispondenza ad aree dell'immagine quasi costanti, rispetto a quelli associati alla presenza di contorni netti e particolari dettagliati.

**Quantizzazione** Questo passo della elaborazione JPEG mira a sfruttare il fenomeno percettivo della ridotta sensibilità dell'occhio umano alle frequenze spaziali più elevate, ovvero la capacità di *filtrare percettivamente* le componenti di errore corrispondenti ai dettagli più minuti. Per questo, il processo di quantizzazione è orientato a ridurre, ed eventualmente sopprimere, le componenti di immagine legate alle frequenze spaziali più elevate, introducendo di fatto *una soglia* sotto la quale si stabilisce di non trasmettere quelle informazioni che tanto non sarebbero percepibili. A questo scopo, ogni coefficiente $D(i, j)$ viene diviso per un coefficiente $Q(i, j)$ dipendente da $(i, j)$, ed il risultato viene arrotondato:

$$B(i, j) = \text{round}\left(\frac{D(i, j)}{Q(i, j)}\right)$$

Il risultato corrisponde ad un processo di quantizzazione, perché quando in ricezione il processo viene invertito (ri-moltiplicando il coefficiente per la stessa quantità), viene persa la precisione legata all'arrotondamento, e pari alla metà del coefficiente di divisione. La scelta dei $Q(i, j)$ è fatta in modo tale da utilizzare valori più elevati per gli indici $(i, j)$ più elevati, in modo da ottenere due risultati: ridurre le componenti ad alta variabilità *spaziale* dell'immagine, e poter usare meno bit per codificare questi valori (più piccoli). Inoltre, molti dei coefficienti con $(i, j)$ elevato, già piccoli di per sè, quando divisi per un coefficiente di quantizzazione più elevato, non *sopravvivono* all'operazione di arrotondamento, in modo che tipicamente la parte in basso a destra della matrice $B(i, j)$ sarà tutta pari a zero, facilitando il compito della codifica run-lengh dello stadio successivo.

**Esempio** La figura 18.27 mostra un esempio di matrice di coefficienti DCT, assieme alla tabella di quantizzazione, ed al risultato dell'operazione. Notiamo come il valore dei coefficienti

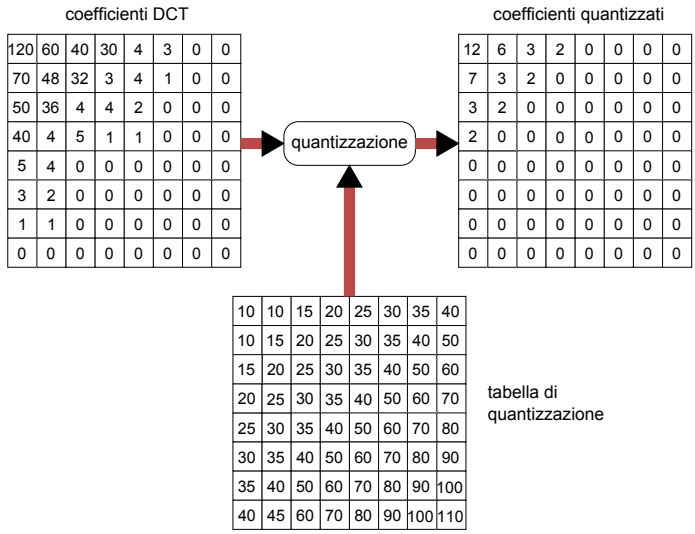


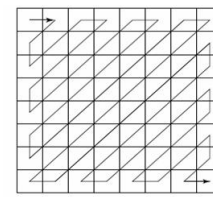

Figura 18.27: Processo di quantizzazione dei coefficienti DCT

di quantizzazione aumenti allontanandosi dal coefficiente DC, e come nella matrice dei coefficienti quantizzati siano *sopravvissuti* solo i coefficienti relativi alle frequenze spaziali più basse.

Sebbene esistano delle tabelle di quantizzazione predefinite, i valori effettivi possono essere variati in base ad un compromesso tra qualità che si intende conseguire e fattore di compressione; tali valori vengono poi acclusi assieme al bitstream codificato durante la fase di formattazione, in modo che il processo di quantizzazione possa essere invertito in fase di riproduzione dell'immagine.

**Codifica entropica** Questo passo è un processo senza perdita, nel senso che non aggiunge altre distorsioni oltre a quelle introdotte dal passo di quantizzazione, ma è essenziale ai fini della compressione, e sfrutta le caratteristiche statistiche del risultato delle elaborazioni precedenti. Come posto in evidenza nello schema di pag. 458, la codifica entropica adotta due diverse procedure per i coefficienti DC e AC, che in entrambi i casi culminano con uno stadio di codifica a lunghezza variabile mediante codici di Huffman.

**Vettorizzazione** Le matrici 8x8 relative ai blocchi di elaborazione visti fin qui vengono ora trasformate in sequenze lineari da un processo di scansione a *zig zag* delle stesse, come mostrato dalla figura seguente.



La sequenza così ottenuta presenta il coefficiente DC in testa, a cui fanno seguito i rimanenti 63 coefficienti AC, ordinati in base al massimo valore di frequenza spaziale che rappresentano. Se applichiamo la scansione zig-zag ai valori riportati nell'esempio

precedente, otteniamo come risultato la sequenza

12 6 7 3 3 3 2 2 2 2 0 0 0 ..... 0 0

**Codifica differenziale** I blocchi adiacenti generalmente possiedono coefficienti DC molto simili tra loro, in virtù dell'omogeneità di ampie zone dell'immagine (pensiamo ad un porzione di cielo). Per questo motivo, anziché codificarli in modo indipendente, i singoli coefficienti DC di blocchi consecutivi vengono sottratti l'uno all'altro, e viene codificata solo la loro differenza. Ad esempio, se una sequenza di coefficienti DC risultasse pari a 12 13 11 11 10 ..., il risultato di questo processo di codifica differenziale darebbe luogo alla sequenza 12 1 −2 0 −1 ... (infatti, il valore *precedente* al primo coefficiente si assume pari a zero). Dato che differenze in valore assoluto piccole sono relativamente più frequenti di differenze grandi, si è scelto di adottare per queste una codifica a lunghezza di parola variabile, realizzata prima descrivendo ogni valore di differenza mediante la coppia (sss, valore), in cui sss rappresenta il numero di bit necessario per rappresentare il valore, e quindi concatenando una codeword di Huffman corrispondente ad sss, al codice binario che rappresenta il valore.

**Esempio** Per chiarire le idee, mostriamo le corrispondenze citate mediante due tabelle, che poi applichiamo all'esempio precedente.

| *differenza* | *N. di bit sss* | *valore codificato* | |
|---|---|---|---|
| 0 | 0 | | |
| -1, 1 | 1 | 1=1 | -1=0 |
| -3, -2, 2, 3 | 2 | 2=10 | -2=01 |
| | | 3=11 | -3=00 |
| -7...-4, 4...7 | 3 | 4=100 | -4=011 |
| | | 5=101 | -5=010 |
| | | 6=110 | -6=001 |
| | | 7=111 | -7=000 |
| -15...-8, 8...15 | 4 | 8=1000, | -8=0111 |
| ⋮ | | ⋮ | |

| *sss* | codeword di Huffman |
|---|---|
| 0 | 010 |
| 1 | 011 |
| 2 | 100 |
| 3 | 00 |
| 4 | 101 |
| 5 | 110 |
| 6 | 1110 |
| 7 | 11110 |
| ⋮ | ⋮ |
| 11 | 111111110 |

Tornando dunque al nostro esempio della sequenza differenziale 12 1 −2 0 −1 ..., in termini di coppie (sss, valore) questa diviene (4, 12), (1, 1), (2, -2), (0, 0), (1, -1),... e quindi, sostituendo ad sss il relativo codice di Huffman preso dalla seconda colonna della seconda tabella, ed ai valori la loro rappresentazione indicata dalla terza colonna della prima tabella, otteniamo la sequenza di bit 101 1100, 011 1, 100 01, 010, 011 0,... in cui si sono mantenute le virgole per chiarezza. In definitiva, abbiamo usato un totale di 23 bit per rappresentare 5 differenze, che ne avrebbero richiesti 45 se codificate con 9 bit.

**Codifica run-lenght** Viene applicata alla sequenza di coefficienti AC che è il risultato dello *zig-zag scan*. In base all'effetto congiunto delle caratteristiche dei coefficienti della DCT, e del processo di quantizzazione, la sequenza degli AC in uscita dal vettorizzatore presenta lunghe sequenze di zeri, consentendo di conseguire buoni rapporti di compressione mediante l'uso di una codifica *run-lenght*, realizzata scrivendo gli AC come una sequenza di coppie *(skip,* ACN*)*, in cui *skip* rappresenta il numero di zeri nel

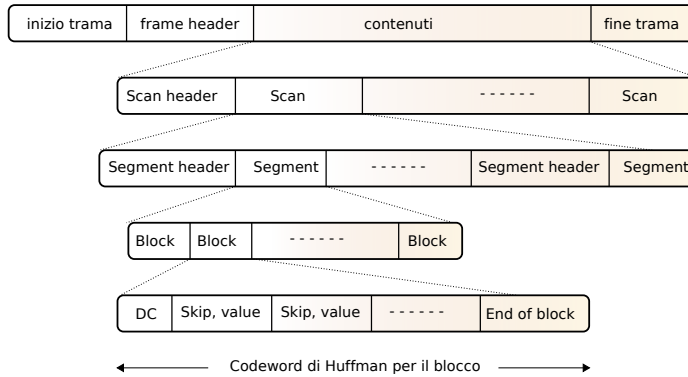



Figura 18.28: Formato del bitstream per la codifica JPEG

run, e ACN è il coefficiente AC non nullo che viene dopo la sequenza di zeri. Quindi, il campo ACN viene espresso a sua volta nella forma *sss, valore*, come indicato dalla prima tabella riportata nell'ultimo esempio. Infine, la coppia *skip, sss* viene rappresentata con una codeword di Huffman individuata in un nuovo codebook appositamente definito.

**Esempio** Applicando la codifica run-length alla sequenza dei coefficienti AC individuati nell'esempio di vettorizzazione, ossia alla sequenza 6 7 3 3 3 2 2 2 2 0 0 0 ... 0 0, si ottiene una sequenza di coppie *(skip,* ACN*)*, pari a *(0,6), (0,7), (0,3), (0,3), (0,3), (0,2), (0,2), (0,2), (0,2) (0,0)* in cui l'ultima coppia *(0,0)* indica la fine del blocco, che in fase di decodifica viene quindi ricostruito riempiendolo di zeri. Anziché usare questa, proseguiamo adottando una diversa sequenza di coppie *(skip,* ACN*)*, pari a *(0,6), (0,7), (3,3), (0,-1), (0,0)*[34]: sostituendo ai termini ACN di questa, la coppia *sss, valore*, e codificando quindi il termine *valore* come indicato nella prima tabella dell'esempio precedente, si ottiene *(0, 3, 110), (0, 3, 111), (3, 2, 11), (0, 1, 0), (0,0)*. Il *bitstream* finale viene quindi realizzato sostituendo alle attuali coppie *skip, sss*, le rispettive codeword individuate alla colonna *Run/Size* della tabella a pagina 150 e segg. delle specifiche ITU-T T.81 `http://www.digicamsoft.com/itu/itu-t81-154.html`, ottenendo *(100, 110), (100, 111), (111110111, 11), (00, 0), (1010)*, e producendo così un totale di 30 bit per rappresentare i 63 coefficienti AC.

**Formattazione** Lo standard JPEG definisce, oltre alla sequenza di operazioni indicata, anche il formato di trama con il quale devono essere memorizzato il bitstream finale. La struttura risultante è gerarchica, e mostrata alla figura 18.28. Al livello superiore troviamo un *frame header* che contiene le dimensioni complessive dell'immagine, il numero ed il tipo di componenti usate (CLUT, RGB, $YC_bC_r$, etc), ed il formato di campionamento (4:2:2, 4:2:0, etc.). Al secondo livello, troviamo uno o più *Scan*, ognuno preceduto da una intestazione in cui viene riportata l'identità del componente (R, G, B, o Y, $C_b$, $C_r$), il numero di bit usato per rappresentare ogni coefficiente di DCT, e la tabella di quantizzazione usata per quella componente. Ogni Scan è composta da uno o più *segmenti*, preceduti da un'ulteriore intestazione, che contiene il codebook di Huffman usato per rappresentare i valori dei blocchi del segmento, nel caso non siano

[34] La nuova sequenza di coppie corrisponde ad una sequenza di coefficienti AC pari a 6 7 0 0 0 3 −1 0 0 ..... 0

Figura 18.29: Due fotogrammi consecutivi, e la differenza tra i rispettivi valori di luminanza

stati usati quelli standard. Infine, nel segmento trovano posto le sequenze di blocchi dell'immagine, così come risultano dopo lo stadio di codifica entropica.

## 18.3 Codifica video

In accordo al metodo di realizzazione dei segnali video analogici, in cui i singoli quadri sono codificati indipendentemente gli uni dagli altri, la codifica video digitale può essere realizzata semplicemente applicando tecniche di codifica di immagine (come JPEG) ad ognuno dei quadri che costituiscono la sequenza video: questo tipo di approccio prende il nome di *moving* JPEG o MJPEG.

D'altra parte, i quadri relativi ad istanti temporali vicini sono spesso molto simili tra loro, anche se quanto siano simili, e per quanto tempo, dipende dal tipo di filmato. La presenza di *memoria* nella sorgente determina quindi la possibilità di ridurre il tasso informativo prodotto dalla codifica ricorrendo a tecniche predittive, tentando quindi di *stimare il movimento* presente in quadri contigui, e trasmettere solo l'informazione necessaria a *compensare* l'errore di predizione. Nella parte destra di fig. 18.29 è mostrata l'immagine differenza $\Delta Y$ tra la componente di luminanza di due quadri consecutivi, consentendo di apprezzarne la relativa semplicità. Considerando poi che alcune regioni si sono mosse più di altre, il quadro da codificare è scomposto in sottoimmagini, per ognuna delle quali si ha un diverso spostamento, e viene calcolata una specifica differenza rispetto alla sotto-immagine precedente (e spostata); questa tecnica prende il nome di *compensazione del movimento*.

**Tipo di quadro** Come abbiamo fatto notare al § 17.1.1.6, le tecniche di codifica predittiva sono particolarmente sensibili agli errori di trasmissione, che possono causare una perdita di sincronismo tra i predittori di trasmissione e ricezione, e quindi l'impossibilità di ricostruire la restante parte di segnale. Per questo, nella codifica video sono presenti dei quadri *di riferimento* in corrispondenza biunivoca con un unico quadro di partenza, detti *intracoded frames* o **I**-frames, che permettono al ricevitore di ri-partire da una condizione nota. Tra due quadri **I** sono poi presenti un certo numero di quadri **P** (*predicted*) come in fig. 18.30a, oltre che quadri **B** (*bidirectional*) come in fig. 18.30b, e che corrispondono rispettivamente alla codifica della compensazione del movimento calcolato a partire da un unico quadro precedente, o da una coppia di quadri passato e futuro.

I quadri **I** sono codificati mediante l'algoritmo JPEG, usando lo stesso coefficiente di quantizzazione per tutti i pixel delle DCT, conseguendo un rapporto di compressione relativamente basso, e sono inseriti a cadenza fissa con un periodo $N$ tipicamente compreso tra 3 e 12: la sequenza di quadri compresi tra due quadri **I** è detta *group of pictures* o GOP. Come mostrato in figura 18.30, la codifica di quadri **P** può dipendere

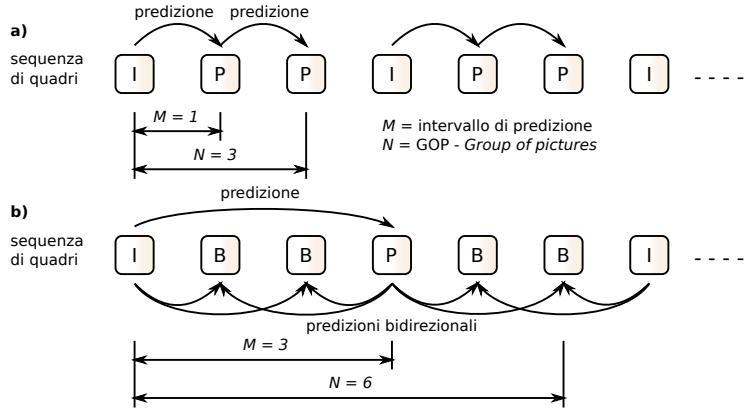



Figura 18.30: Esempi di sequenze di quadri con **a)** solo quadri di tipo I e P; **b)** quadri di tipo I, P e B

dal quadro **I** immediantamente precedente, o dalla ricostruzione di un precedente quadro **P**, ottenendo un fattore di compressione maggiore che per i quadri **I**; la distanza temporale tra **P** e l'originale **I** è detta *intervallo di predizione*, indicato con $M$.

Per realizzare la compensazione del movimento, ogni regione del nuovo quadro è confrontata con regioni *limitrofe* del quadro precedente, riducendo così la complessità di ricerca. Nel caso dei quadri **B** la ricerca delle regioni simili è invece svolta rispetto ai quadri **I** (o **P**) situati sia nel passato che al futuro, migliorando la precisione della stima di movimento, e conseguendo rapporti di compressione ancora maggiori, a patto di subire un aumento del ritardo di codifica, legato al dover attendere un quadro futuro.

Allo scopo di ridurre il ritardo di decodifica, la sequenza di quadri viene trasmessa con un ordine diverso da quello dei quadri orginali, consentendo ai quadri **B** di essere riprodotti non appena ricevuti, e non dopo la ricezione del quadro *futuro* da cui dipendono. Pertanto, se la sequenza originale è ad esempio

**IBBPBBPBBIBBP...**

questa verrà trasmessa nell'ordine

**IPBBPBBIBBPBB...**

**Stima del movimento e compensazione** Specifichiamo innanzitutto cosa intendere con il termine *regione* prima usato per definire il dominio dell'operazione di confronto necessaria alla stima di movimento. Come mostrato in fig. 18.31a, considerando uno suddivisione in componenti $Y$ $C_b$ $C_r$ ed un sottocampionamento 4:1:1, il quadro originale è suddiviso in $N$ righe e $M$ colonne di *macroblocchi* di 16x16 pixel, ed ogni macroblocco è rappresentato da sei blocchi 8x8 pixel, di cui quattro blocchi per la luminanza, più due blocchi per le componenti di crominanza; ogni blocco 8x8 corrisponde quindi ad un equivalente numero di coefficienti DCT, ed è individuato all'interno del quadro, in base al suo indirizzo di riga e colonna.

Nella codifica dei quadri **P**, ogni macroblocco $M_T$ del quadro corrente (*target*) è confrontato pixel per pixel con il corrispondente macroblocco $M_R$ del quadro *di*

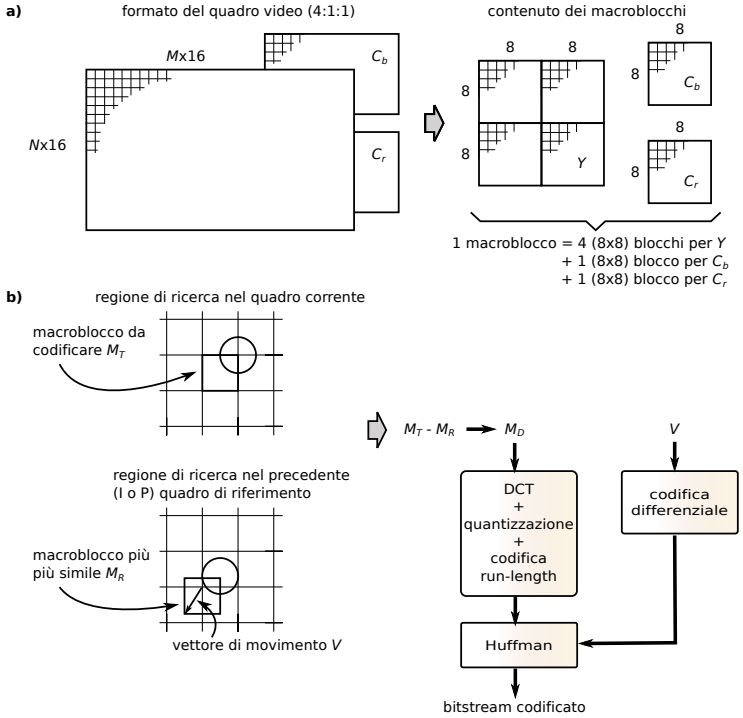



Figura 18.31: Codifica di un quadro P: a) struttura del macroblocco; b) procedura di codifica

*riferimento*, e nel caso sia riscontrata una sufficiente similitudine[35] complessiva, viene trasmesso solo l'indirizzo del blocco. Altrimenti, il confronto viene ripetuto per tutti i possibili spostamenti del macroblocco target nell'ambito dei macroblocchi contigui[36], e qualora sia individuata una buona corrispondenza, il macroblocco viene codificato dal *vettore di movimento* $V$ e dall'*errore di predizione* $M_D$. Con riferimento alla fig. 18.31b in cui l'immagine è simboleggiata da un cerchio, $V$ rappresenta lo spostamento da applicare a $M_T$ per portarlo a coincidere al meglio con il quadro precedente, ed è codificato come una coppia $(x, y)$ corrispondente ad una *risoluzione di un pixel*. Al contrario $M_D$ è composto dalle tre matrici $(Y\ C_b\ C_r)$ dei valori differenza tra quelli di $M_T$ spostato di $V$, ed $M_R$. I valori di $V$ e di $M_D$ relativi ai diversi macroblocchi di un quadro seguono poi due diversi percorsi di codifica, come specificato appresso.

Nel caso in cui la regione di ricerca sia estesa, i valori $V$ possono risultare relativamente grandi; d'altra parte è probabile che macroblocchi vicini esibiscano vettori di spostamento molto simili tra loro. Per questi motivi, la sequenza dei $V$ calcolati per macroblocchi contigui viene prima sottoposta ad un processo di codifica differenziale, e quindi i valori di differenza sono rappresentati da codeword a lunghezza variabile di Huffman. D'altra parte, le tre matrici differenza sono invece sottoposte alla stessa

[35] Il confronto è svolto considerando i soli valori di luminanza, e la similitudine valutata come media tra i valori assoluti delle differenze di luminanza.

[36] l'effettiva estensione dell'area di ricerca non è oggetto di standardizzazione, mentre lo è la rappresentazione del risultato della ricerca.

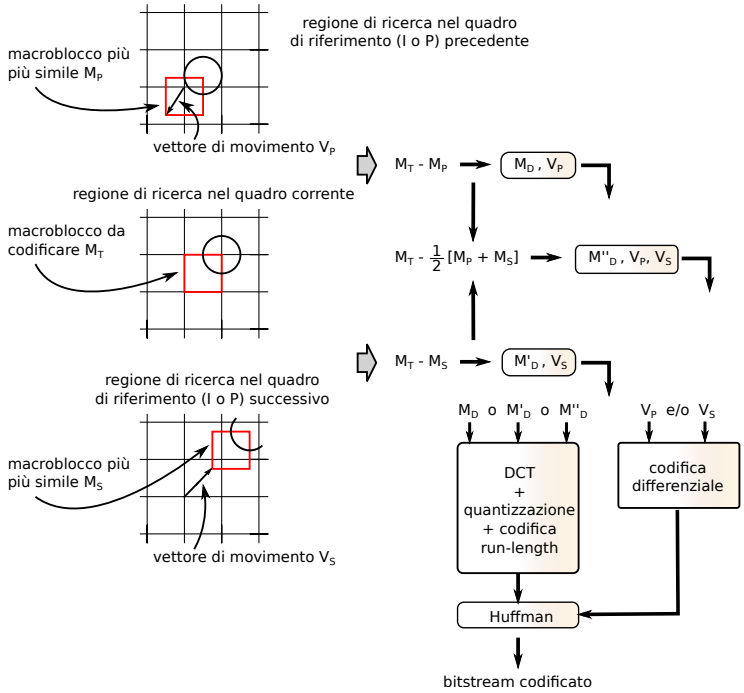



Figura 18.32: Procedura di codifica dei quadri B

sequenza di operazioni dei quadri **I** (DCT, quantizzazione, codifica entropica), conseguendo però un fattore di compressione più elevato, essendo il macroblocco differenza con valori quasi tutti molto piccoli.

Nel caso in cui la stima di movimento fallisca[37] (o a causa di una estensione di ricerca insufficiente, oppure per un reale cambio di scena), il macroblocco è codificato in modo indipendente come avviene per i quadri **I**.

I macroblocchi dei quadri B (vedi fig. 18.32) sono invece confrontati sia con il *precedente* quadro $M_P$ che con il *successivo* $M_S$, ottenendo due possibili insiemi di matrici differenza $M_D$ e $M'_D$ ed associati vettori $V_P$ e $V_S$; viene inoltre calcolato un ulteriore insieme $M''_D$ come differenza tra $M_R$ e la media dei marcoblocchi (spostati) di riferimento, e determinato infine quale delle tre possibilità fornisca il minimo errore di predizione. In base a questa scelta, si individua quale macroblocco differenza codificare, assieme ai rispettivi vettori di movimento. Nel caso prevalga la predizione basata sulla media tra macroblocchi di riferimento, il vettore di movimento complessivo può determinare un potere di risoluzione a livello di *sub-pixel*.

**Questioni realizzative** La fig. 18.33 riassume la sequenza di operazioni applicate alle tre tipologie di quadro **I**, **P** e **B**. Mentre nel primo caso queste seguono lo schema previsto dalla codifica JPEG, i quadri **P** meritano qualche commento: allo scopo di ali-

[37] Viene decretato il fallimento quando anche la migliore compensazione di movimento possibile non determina una riduzione della quantità di bit, rispetto ad una codifica JPEG.

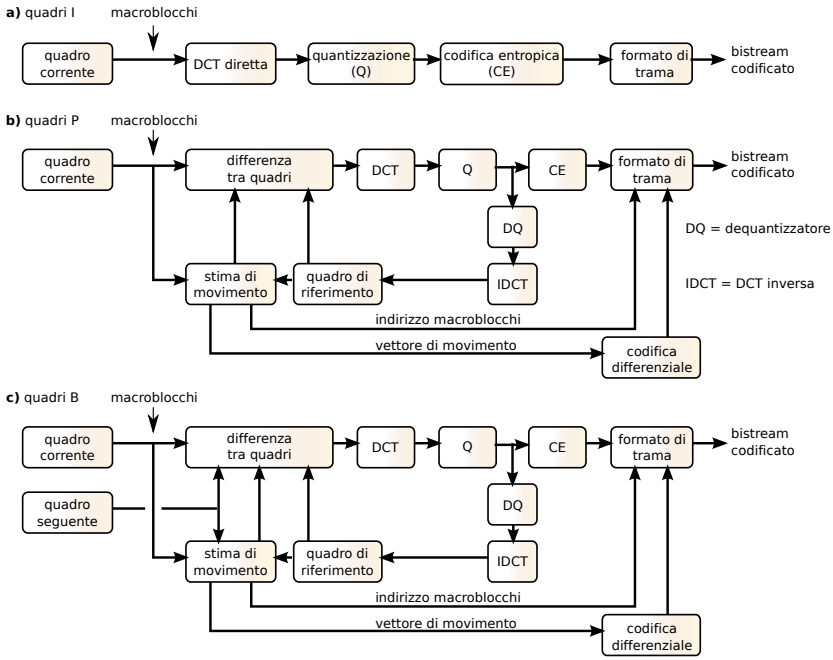



Figura 18.33: Stadi di elaborazione nella codifica di: a) quadri I; b) quadri P; c) quadri B

mentare correttamente il componente di stima di movimento, il codificatore mantiene memoria del quadro di riferimento, all'inizio posto pari ad un quadro **I**, e quindi sostituito da una copia dell'ultimo quadro **P**, ottenuto risommando il quadro differenza al precedente quadro di riferimento. Lo stesso schema di calcolo è svolto nel caso di quadri **B**, tenendo ora conto anche del quadro successivo.

Rimarchiamo ora il fatto che, in funzione dell'esito del processo di stima di movimento, esistono tre diverse possibilità di rappresentazione per ogni macroblocco dei quadri **P** e **B**:

- se non vi è movimento, viene trasmessa solo la sua posizione;
- se vi è movimento e si trova un riferimento abbastanza simile, sono trasmessi il vettore di movimento e le matrici differenza;
- se non si è trovato un riferimento abbastanza simile, viene effettuata una codifica *inter* come per il caso dei quadri **I**.

Ciò determina l'esigenza di disporre di formato di trama di dimensione (e velocità) variabile, come realizzato nell'esempio mostrato in fig. 18.34, in cui ad ogni macroblocco è associato un tipo (**I**, **P** o **B**), il suo indirizzo nell'ambito del quadro, il coefficiente di quantizzazione relativo ai termini della DCT, ed il vettore di movimento (se presente). Quindi si dichiara l'identità dei blocchi presenti (che potrebbero essere assenti in caso di immagini statiche), e per questi viene infine prodotta la sequenza di informazioni previste dalla codifica JPEG.

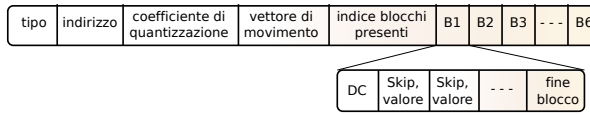



Figura 18.34: Esempio di formato trama per i macroblocchi di un bitstream video

## 18.3.1 Standard video

### 18.3.1.1 H.261

E' lo standard di codifica video definito da ITU-T a fine anni '80 per le applicazioni di videotelefonia su ISDN, ed anche se oggi tecnicamente superato, resta comunque un valido sistema di riferimento che consente la retro-compatibilità tra apparati[38]. Il suo principale limite è la necessità di produrre una velocità ridotta e vincolata a multipli di 64 kbps.

La scelta del formato di immagine è limitata a quanto mostrato in tabella[39], mentre la scansione è non interallacciata e la velocità di rinfresco di 30 quadri/secondo per CIF oppure 15 o 7.5 per QCIF. Sono usati solo quadri di tipo **I** e **P**, con un GOP di 4 (ossia 3 **P** ogni **I**), e sono usate le procedure descritte alla sezione precedente per rappresentare ogni quadro nei termini di macroblocchi composti da 16x16 pixel (4 blocchi di 8x8) di luminanza e 2 blocchi 8x8 per ogni componente di colore $C_b$, $C_r$.

| *Formato* | $Y$ | $C_b$, $C_r$ |
|---|---|---|
| CIF | 352 x 288 | 176 x 144 |
| QCIF | 176 x 144 | 88 x 72 |

Ogni macroblocco segue la tipica formattazione mostrata in fig. 18.35a; tre file di 11 macroblocchi sono poi raggruppati in una nuova struttura sintattica detta GOB (*Group of (macro)Blocks*), che si articola in un contenuto (fig. 18.35b) ed una intestazione (fig. 18.35c), in cui troviamo un *codice di inizio* scelto in modo da non poter essere presente nella sequenza di codici di Huffman che seguono, e che permette la risincronizzazione nel caso di GOB *mancanti* (vedi appresso), in modo da poter tornare a riprodurre un quadro in corrispondenza del primo GOB disponibile. L'intero quadro è quindi realizzato con il formato di fig. 18.35d), in cui compare un *codice di inizio quadro*, un *riferimento temporale* necessario alla sincronizzazione con la traccia audio, e l'indicazione del tipo di quadro (**I** o **P**); a cui segue la sequenza dei GOB, in numero di 3 oppure 12 a seconda se il quadro rappresenti una immagine QCIF o CIF, in modo da permettere l'interoperabilità tra formati come mostrato in fig. 18.35e.

**Controllo di velocità** Dato che la codifica video produce una velocità di trasmissione variabile, questa può eccedere la capacità del canale a disposizione, ed un modo *drastico* per risolvere il problema è di scartare alcuni GOB. Il campo *Group number* dell'intestazione dei GOB permette quindi di collocare il nuovo GOB anche in mancanza dei suoi predecessori.

Un approccio più articolato è quello mostrato dalla figura 18.36a, che ripercorre le tappe già discusse e relative al calcolo del vettore di movimento ed alla codifica

[38] Vedi ad es. `http://www0.cs.ucl.ac.uk/teaching/GZ05/08-h261.pdf` (una presentazione di *Mark Handley*), o la trattazione su `http://en.wikipedia.org/wiki/H.261`.

[39] Il *Common Intermediate Format* è stato pensato per facilitare la compatibilità con PAL e NTSC; il *Quarter*-CIF ha una superficie di 1/4. Sono poi stati anche definiti il 4CIF e 16CIF, oltre che il SIF (352 x 240) che interopera con flussi MPEG.

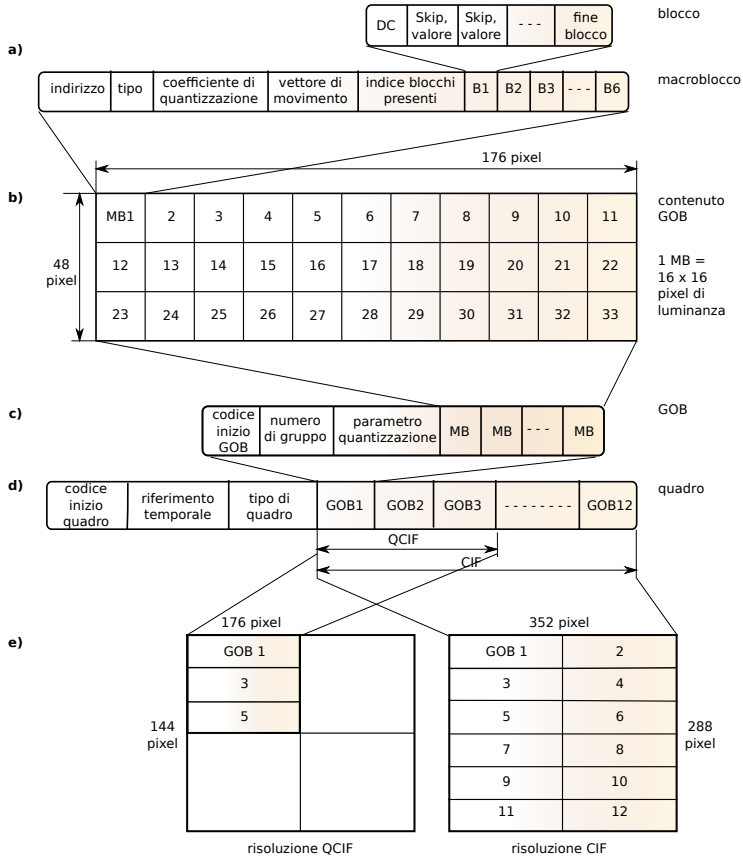



Figura 18.35: Formato codifica H.261: a) macroblocco; b) costruzione e c) formato di un GOB; d) trama di quadro; c) interoperabilità tra formati

degli errori di predizione, ma pone in evidenza il campo di intervento di un componente di *controllo quantizzazione*, che variando l'entità dei coefficienti di quantizzazione della DCT, permette di ridurre e/o aumentare la velocità di codifica complessiva. In particolare, il controllo di quantizzazione opera in base alle stato di riempimento del *buffer* FIFO[40] mostrato in fig. 18.36b, alimentato dal risultato del processo di codifica e formattazione video, e da cui sono prelevati i dati da inviare a velocità costante. Nel caso in cui la velocità media di codifica ecceda quella disponibile, l'aumento della occupazione del buffer determina l'aumento del coefficiente di quantizzazione, e quindi una riduzione di qualità ma anche della velocità media di codifica; ovviamente, anche l'inverso è possibile, ossia un miglioramento di qualità mediante riduzione del coefficiente di quantizzazione, nel caso in cui la scena sia statica, e la codifica produca un basso bit rate che consente alla FIFO di svuotarsi.

Nel caso di un aumento improvviso di velocità, come anticipato si possono addirit-

[40] *First in First out,* è la disciplna di coda del primo arrivato primo servito, opposta a LIFO *Last In First Out*, realizzata come uno *stack*.

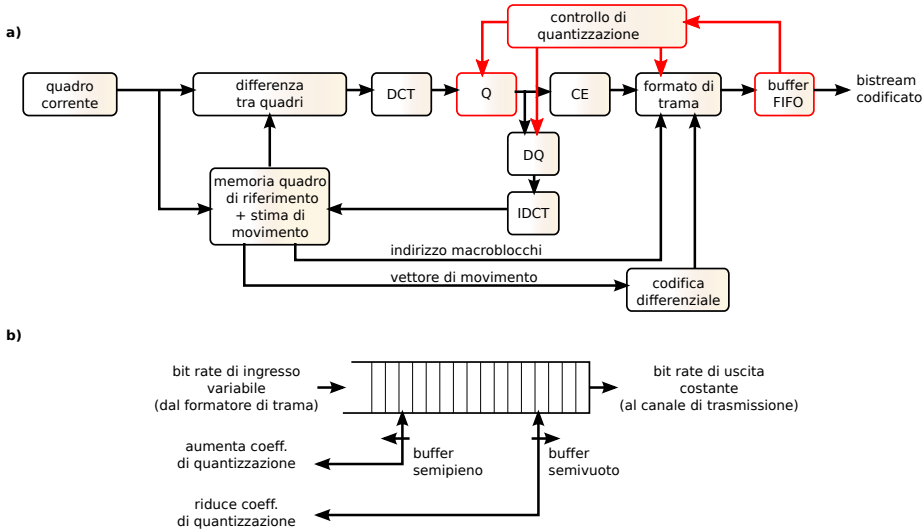



Figura 18.36: Principi della codifica H.261: a) schema del codificatore b) funzionamento del buffer FIFO

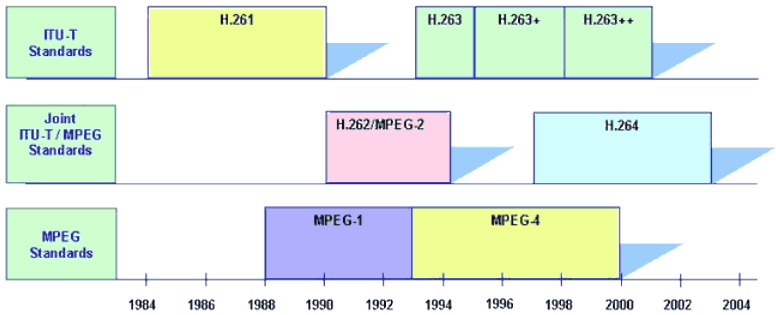
tura *scartare* alcuni GOB, mentre per i successivi si adottano coefficienti di quantizzazione ridotti, comunicati anche al lato ricevente per mezzo dell'apposito campo della intestazione GOB, come mostrato in fig. 18.35c.

#### 18.3.1.2 H.263

Anche questo definito da ITU-T a partire dal 1995, nasce per risolvere i problemi di bassa qualità dell'H.261 a velocità molto ridotte, come quelle offerte dai collegamenti modem *dial-up* precedenti all'introduzione dell'ADSL, ovvero per migliorare la gestione delle possibili condizioni di errore sia sul canale *dial-up* che *wireless*. Le specifiche originarie si sono in seguito arricchite[41] di estensioni, favorendo l'adozione del codec da parte di altre applicazioni (inclusi i filmati di *youtube*), ed aggiungendo il supporto oltre che ai formati nativi CIF e QCIF, anche a S-QCIF, 4CIF, 16CIF, SIF e 4SIF. A partire dal 2003 si è formato un gruppo di lavoro congiunto tra ITU-T VCEG (*Video Coding Expert Group*) e ISO/IEC MPEG (*Moving Pictures Experts Group*), che segue la definizione del suo successore, l'H.264 detto anche AVC (Advanced Video Coding) o MPEG-4 *part 10*, determinando l'arresto dello sviluppo di H.263, che resta comunque (assieme all'H.261) supportato da un gran numero di applicazioni multimediali. Sebbene la struttura generale del codificatore e del bitstream ricalchi quella vista per l'H.261, sono state introdotte alcune novità significative, che tentiamo di elencare appresso.

**Tipi di quadro** In H.263 sono usati, oltre ai quadri di tipo **I** e **P**, anche quelli bidirezionali **B**, consentendo di ottenere fattori di compressione maggiori, a parità di qualità percepita.

[41] Nel 1998 viene rilasciato l'H.263v2, noto anche come H.263+ o H.263 1998, e nel 2000 è emesso l'H.263v3 noto anche come H.263++ o H.263 2000; inoltre l'MPEG-4 Part 2 è compatibile con l'H.263, in quanto un bitstream H.263 di base viene correttamente riprodotto da un decodificatore MPEG-4.




**Slice** I GOB sono ridefiniti come singole *strisce* di macroblocchi, quindi ad esempio per i formati CIF e QCIF un GOB è ora formato da 11 macroblocchi in fila, anziché 33 come avveniva per l'H.261.

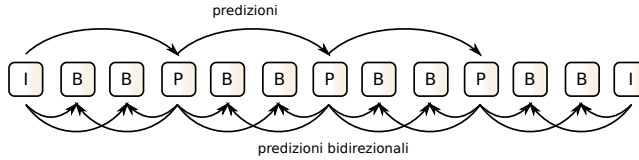
**Vettori di movimento estesi** La stima di movimento dell'H.261 si arresta in corrispondenza dei bordi del quadro, per cui anche se un oggetto è solo parzialmente uscito di scena, il macroblocco corrispondente viene codificato in modalità *intra*. Al contrario H.263 permette di estendere la ricerca anche a vettori di spostamento che cadono al difuori del quadro, alla ricerca di una corrispondenza parziale, consentendo al contempo maggiore efficienza e minor distorsione.

**Predizione avanzata** Anziché determinare il vettore di movimento in base al confronto di un intero macroblocco, i 4 blocchi 8x8 che lo costituiscono sono confrontati in modo indipendente con il quadro di riferimento, permettendo una migliore compensazione del movimento anche per l'immagine di oggetti che non solo traslano, ma si deformano. In definitiva, sono prodotti 4 diversi vettori di movimento per ogni macroblocco.

**Resistenza agli errori** La presenza di un errore nella ricezione[42] di un GOB, oltre ad impedire la corretta riproduzione dello stesso, ostacola la riproduzione anche dei quadri successivi che dipendeno dai pixel presenti nel GOB, e peggio ancora l'errore finisce per estendersi anche ad altri GOB, in virtù degli effetti dell'errore sulla ricostruzione dei macroblocchi *predetti* in presenza di movimento.

Per ridurre l'estensione temporale dell'effetto dell'errore, e non dover attendere fino alla ricezione del successivo quadro **I**, si può usare il canale di ritorno presente nei collegamenti punto-punto, consentendo al decodificatore di inviare dei NACK che notificano al mittente la coppia (quadro, GOB) per la quale si è rilevato un errore. Il codificatore è quindi in grado di valutare esso stesso le conseguenze sui quadri successivi, e può provvedere a fornire una codifica *intra* per tutti i blocchi che necessitano di essere rapidamente risincronizzati.

[42] Qualcuno potrebbe aver notato che nella definizione degli standard fin qui discussi, non sono previsti controlli di tipo *checksum* nel bitstream prodotto. D'altra parte essendo le informazioni codificate di natura auto-sincronizzante, la presenza di errori determina presto presso il ricevitore una condizione di disallineamento, e la decodifica di valori non previsti, come ad esempio la ricezione di vettori di movimento o coefficienti DCT fuori dinamica, o codeword di Huffman non valide, od un numero eccessivo di coefficienti. Per tale via, il ricevitore diviene in grado di accorgersi dell'errore che si è verificato.




Esempio di sequenza di quadri MPEG-1

#### 18.3.1.3 MPEG-1

Il *Moving Pictures Expert Group* di ISO emette una serie di standard ognuno orientato ad un particolare dominio applicativo di segnali multimediali, come

- MPEG-1 adotta un formato SIF di 352x288 pixel inteso per la *memorizzazione* audio-video a qualità VHS su CDROM, a velocità fino a 1.5 Mbps;
- MPEG-2 è orientato alla *memorizzazione e trasmissione* audio-video secondo quattro livelli di risoluzione, per ognuno dei quali diversi profili individuano tecniche alternative di codifica;
- MPEG-4 è stato inizialmente concepito per applicazioni simili a quelle dell'H.263, ma il suo uso si è successivamente esteso ad un'ampia gamma di applicazioni Internet.

Anche MPEG-1 adotta tecniche del tutto simili a quelle dell'H.261, con una scansione dell'immagine progressiva ed un sottocampionamento delle componenti di colore 4:1:1, una frequenza di quadro di 25 Hz, l'adozione di quadri di tipo **I**, **P** e **B**, la rappresentazione dei quadri in termini di marcoblocchi composti da 16x16 pixel di luminanza, più due blocchi 8x8 per ciascuna componente di colore. Le principali differenze sono che

- possono essere inseriti riferimenti temporali *all'interno* di un quadro, permettendo al decodificatore di sincronizzarsi più rapidamente. L'intervallo tra due marche temporali è chiamato *slice* e comprende una sequenza orizzontale di macroblocchi, tipicamente che copre una intera riga[43], o meno, ma non di più;
- l'uso dei quadri di tipo B aumenta la distanza temporale tra i quadri di tipo P ed il loro riferimento, e quindi determina una maggiore distanza coperta dalle porzioni di immagine in movimento, cosicché l'ampiezza della finestra di ricerca adottata dal componente di detezione di movimento è stata estesa.

La figura 18.37a illustra la *struttura gerarchica* del bitstream risultante, secondo il quale l'intero filmato (*sequenza*) è costituito da una successione di GOP, ed ogni GOP da una sequenza di quadri, ognuno costituito da una successione di *slice* che comprendono ognuno 22 macroblocchi, ognuno con 6 blocchi. La sezione b di fig. 18.37 entra più nel dettaglio del formato del bitstream.

#### 18.3.1.4 MPEG-2

Allo scopo di poter usare questo stesso standard per diversi contesti applicativi, sono stati definiti i quattro *livelli* qualitativi mostrati in tabella, e per ogni

[43] ad es., 22 macroblocchi in risoluzione CIF

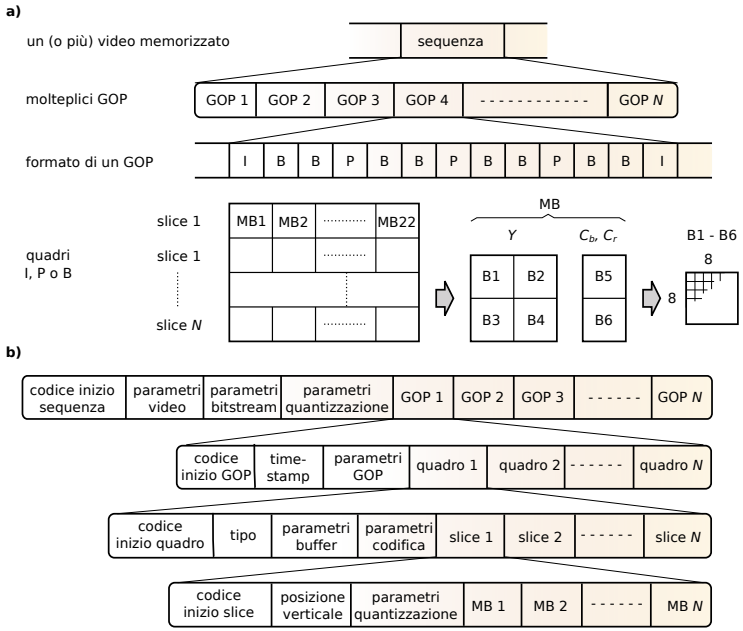



Figura 18.37: Struttura del bitstream MPEG-1: a) composizione; b) formato gerarchico

livello sono quindi definiti cinque profili (*simple, main, spatial resolution, quantization accuracy, high*) in modo da permettere lo sviluppo di nuove tecnologie. Il livello *Low* è compatibile con MPEG-1. Affrontiamo ora la descrizione di ciò che è offerto dal profilo *Main* al livello *Main* (MP@ML).

| Livello MPEG-2 | formato | bit rate (Mbps) | applicazione |
|---|---|---|---|
| Low | SIF | $< 1.5$ | registrazione qualità VHS |
| Main | 4:2:0 | $< 15$ | DVB - MP@ML |
| | 4:2:2 | $< 20$ | |
| High 1440 | 4:2:0 | $< 60$ | HDTV 4/3 |
| | 4:2:2 | $< 80$ | |
| High | 4:2:0 | $< 80$ | HDTV 16/9 |
| | 4:2:2 | $< 100$ | |

**MP@ML** L'obiettivo è la diffusione televisiva DVB, con scansione interallacciata, risoluzione 720x576 a 25 quadri/secondo (PAL), sottocampionamento 4:2:0 per una velocità risultante tra i 4 ed i 15 Mbps. La principale differenza rispetto all'MPEG-1 è legata alla modalità di scansione *interallacciata*, in modo che come mostrato in fig. 18.38, ogni quadro è costituito da due sottoquadri (o *campi*) con le righe rispettivamente dispari e pari, ponendo la questione: come comporre i blocchi da 8x8 pixel su cui eseguire la DCT? Sono possibili due alternative:

- la *modalità campo* (fig. 18.39a) in cui i 16x16 pixel di un macroblocco sono ripartiti tenendo assieme prima le sole righe dispari del primo campo, e quindi le sole righe pari del secondo campo, oppure
- la *modalità quadro* (fig. 18.39b) in cui si usa la stessa suddivisione già vista per il caso non interallacciato, mescolando i due campi in ognuno dei blocchi.

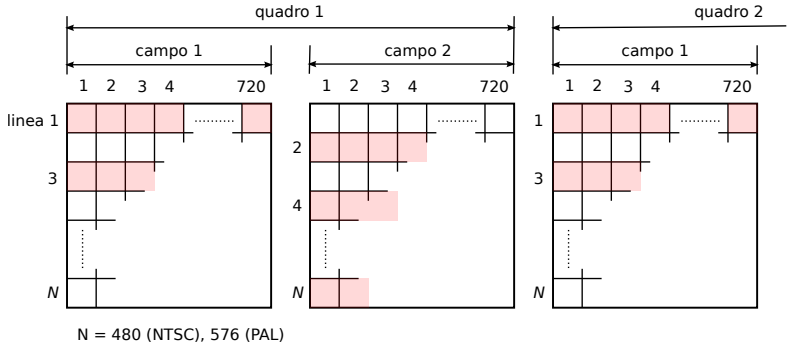



Figura 18.38: Effetto della scansione interallacciata in MPEG-2

La scelta migliore su quale tra le due modalità adottare dipende dal tipo di scena che si sta rappresentando. Se è presente molto movimento, è meglio adottare la modalità campo: essendo infatti i pixel di uno stesso blocco collezionati in un tempo pari a metà dell'intervallo di quadro (mentre nella seconda metà si collezionano i pixel della seconda serie di blocchi), si ottiene un *fotogramma meno mosso*; viceversa in presenza di una scena con poco movimento, può essere adottata la modalità quadro.

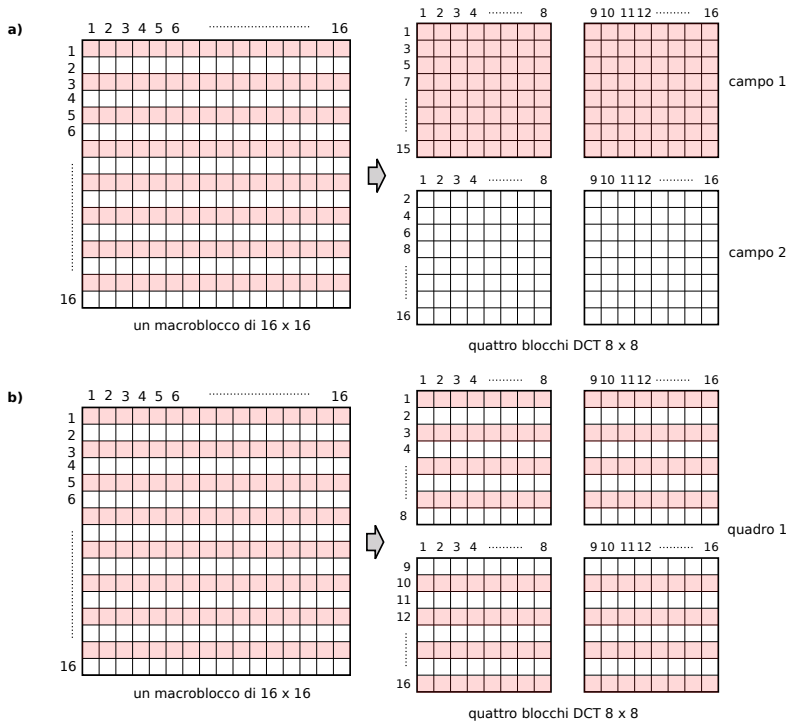



Figura 18.39: Composizione dei blocchi DCT per i quadri I di MPEG-2: a) modalità campo; b) modalità quadro

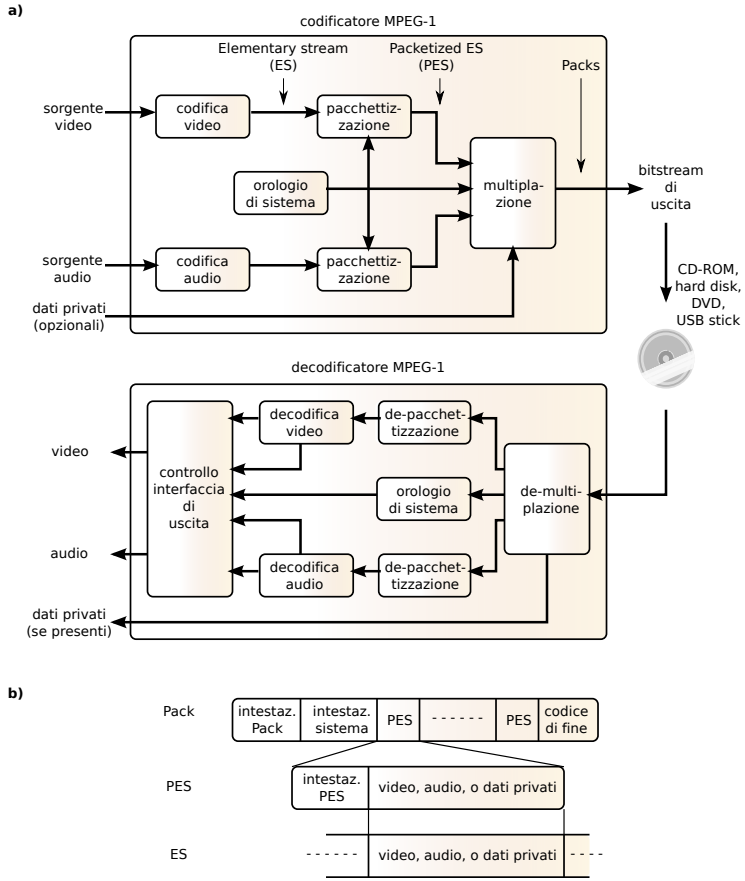



Figura 18.40: Generazione di un contenuto multimediale MPEG-1: a) co-decodificatore; b) formato del bistream di uscita

Per quanto riguarda la stima di movimento, sono ora previste tre possibilità: la *modalità campo* prevede che i campi dispari usino come riferimento i campi pari del quadro precedente, ed i campi pari quelli dispari dello stesso quadro: in tal modo l'intevallo temporale su cui è valutato il movimento è metà del periodo di quadro. Nella *modalità quadro* invece, i campi pari e dispari usano come riferimento i rispettivi campi pari e dispari del quadro precedente, ed è più idoneo nel caso di movimenti lenti. Il meglio di entrambi i modi si ottiene con la *modalità mista*, in cui sono attuati entrambi gli approcci, e viene scelto per la trasmissione quello in grado di dar luogo alla distorsione minore.

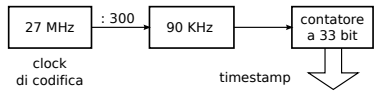
**HDTV** Sono definiti tre standard, ATV, DVB e MUSE, rispettivamente per il Nord America, l'Europa ed il Giappone, a cui si aggiunge la specifica HDTV di ITU-R relativa a studi di produzione e scambio internazionale, e che definisce un rapporto di aspetto 16/9 con 1920 colonne per 1152 righe (di cui solo 1080 visibili), con scansione inte-

rallacciata e sottocampionamento 4:2:2. ATV include le specifiche di ITU-R, oltre che un formato ridotto sempre con aspetto 16/9 ma risoluzione 1280x720, e che adotta la codifica video MPEG-2 MP@ML e quella audio AC-3. Il DVB è basato su di un rapporto di aspetto 4/3 con 1440x1152 pixel (di cui 1080 visibili), pari cioè al doppio della risoluzione PAL di 720x576. La codifica video è MPEG-2 SSP@H1440 (*Spatially Scaleable Profile at High 1440*), simile all'MP@ML, mentre la codifica audio è MPEG audio layer 2.

### 18.3.2 Contenitori

I flussi binari prodotti dai diversi codificatori (audio, video) prendono il nome di *Elementay Stream* (ES), e possono essere multiplati assieme per produrre un nuovo formato idoneo alla registrazione di un contenuto multimediale completo, eventualmente arricchito da un flusso di *dati privati*, come mostrato in fig. 18.40a per il caso di MPEG-1.

Prima di effettuare la multiplazione, gli ES sono suddivisi in *pacchetti* di dimensione variabile denominati *Packetized* ES (PES) in cui trova posto un *payload* contenente[44] il risultato della codifica (ad es. un intero quadro), preceduto da una *intestazione* contenente un codice univoco di inizio PES ed un codice che individua il tipo di payload del pacchetto (audio, video o dati). L'intestazione PES può inoltre contenere un riferimento temporale necessario alla sincronizzazione audio-video, con risoluzione 33 bit, prodotto dal clock a 90 kHz descritto in fig. 18.40 come *orologio di sistema*, e ottenuto a partire da un oscillatore a 27 MHz come mostrato alla figura seguente.




Nel momento in cui un ES è pacchettizzato, viene inserito un *presentation timestamp* (PTS) che ne individua l'istante di riproduzione; per i flussi video è inserito anche[45] un *decode timestamp* (DTS) perché come anticipato a pag. 465 l'ordine di trasmissione (e quindi di decodifica) può differire dall'ordinamento naturale.

I PES derivanti da diversi ES possono essere quindi inseriti in una struttura di trama detta *Pack*, che può infine essere memorizzata ai fini di una successiva riproduzione.

#### 18.3.2.1 Transport Stream

Nel caso di trasmissione dei contenuti mediante un mezzo broadcast, tipicamente il singolo *Program Stream* PS (equivalente a quello prima indicato come *Pack*) viene ulteriormente multiplato assieme ad altri, in modo da realizzare un *Transport Stream* (TS), come mostrato in fig. 18.41. In particolare, la parte b) della figura mostra come i PES siano ora suddivisi in segmenti di lunghezza fissa e pari a 184 byte, intestati con 4 byte, producendo una struttura di trama con pacchetti di 188 byte[46]; l'ultimo pacchetto del TS che origina da uno stesso PES è riempito con uni fino a raggiungere i 188 byte. L'intestazione contiene, oltre ad un byte di inizio con pattern unico, un *Packet IDentification code* (PID) di 13 bit, che identifica il PES a cui appartiene il pacchetto, permettendo al decoder di recuperare il programma a cui è interessato.

[44] Per un approfondimento, vedi `http://en.wikipedia.org/wiki/Packetized_elementary_stream`

[45] In realtà PTS e DTS non sono inseriti in tutti i pacchetti, ma una volta ogni tanto (con intervalli fino a 700 msec per i PS e 100 msec per i TS): il decoder rigenera infatti localmente il clock, ed i timestamp ricevuti servono a mantenerlo al passo con quello trasmesso.

[46] in realtà le intestazioni dei pacchetti del TS possono essere estese e contenere più di 4 byte: in questo caso, la dimensione del paylod si riduce, in modo che il totale sia ancora 188.

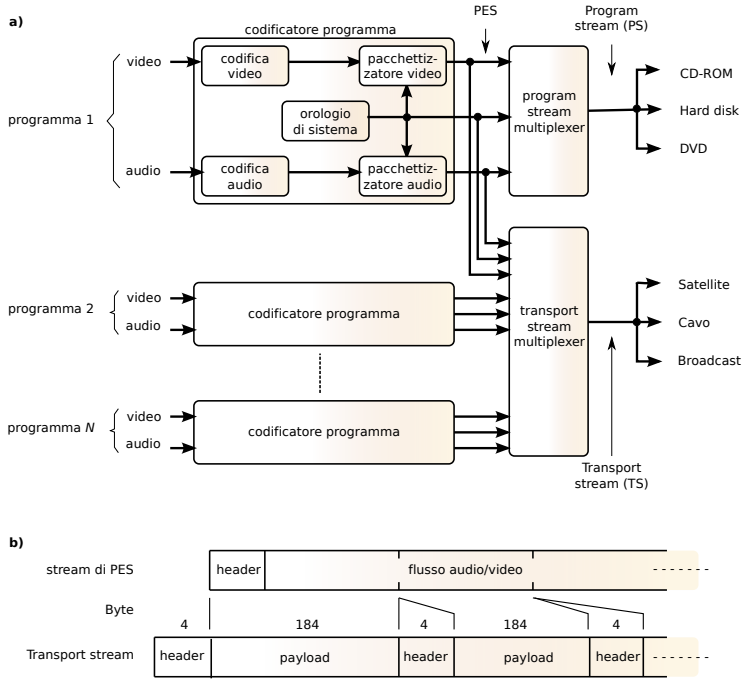



Figura 18.41: Multiplazione di programmi: a) generazione dei PS e TS; b) formato del Transport Stream

Alcuni PID sono riservati, come il PID 8191 che indica assenza di payload[47], ed il PID 0, che annuncia l'inserimento nel payload della *Program Association Table* (PAT), la cui ricezione permette al decoder di conoscere quali PID sono utilizzati per individuare i diversi PES (audio, video) di uno stesso programma[48]. Questo avviene per mezzo delle *Program Map Table* (PMT) rappresentate in fig. 18.42: ogni riga della PAT individua infatti un nuovo PID, alla ricezione del quale, viene estratta dal payload la PMT che descrive i PID dei flussi che compongono il programma. Pertanto, quando uno spettatore seleziona un programma, il decodificatore cerca nella PAT il PID della PMT associata, e quindi inizia a prelevare i pacchetti intestati con ognuo dei PID trovati nella PMT, per riscostruire i relativi PES, sincronizzarli, ed iniziare la riproduzione.

[47] Un payload vuoto è in realtà comunque riempito di 184 bytes inutili, e viene inserito da parte del multiplatore che realizza il TS per mantenere una riserva di banda che consenta di assecondare le fluttazioni di velocità dei tributari.

[48] Altri PID riservati sono l'uno, che annuncia la presenza di una *Conditional Access Table* (CAT) contenente i parametri crittografici per visualizzare contenuti a pagamento, ed il PID 18, che annuncia la presenza della *Network Information Table* (NIT), che descrive altri TS disponibili. Per approfondimenti, vedi `http://en.wikipedia.org/wiki/Program-specific_information`.

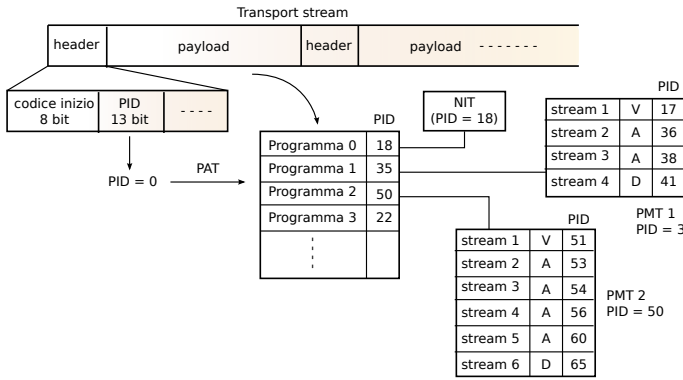



Figura 18.42: Estrazione della PAT e delle PMT dal Transport Stream

# Bibliografia


**A.A. V.V.,** *Wikipedia, L'enciclopedia libera e collaborativa*, `http://it.wikipedia.org`

**S. Barbarossa, T. Bucciarelli,** *Teoria dei Segnali*

**S. Benedetto, E. Biglieri,** *Teoria della Probabilità e Variabili casuali*, Quaderni di Elettronica, 1980 Boringhieri

**C. A. Bentivoglio, A. Caldarelli,** *Tecniche e tecnologie multimediali*, 2007 EUM edizioni università di macerata

**J. Bellamy,** *Digital Telephony*, 1991 John Wiley and Sons, New York

**G. Cancellieri,** *Telecomunicazioni*, 2000 Pitagora editrice - Bologna

**A. B. Carlson,** *Communication Systems*, $3^{rd}$ Edition, 1986 Mc Graw Hill

**F. Cuomo,** *Telematica*, 2001 http://net.infocom.uniroma1.it/tlem/lucidi/lucidi.php3

**R. Cusani, T. Inzerilli,** *Teoria dell'Informazione e Codici*, Ed. Ingegneria 2000, 2007 Roma

**M. Decina, A. Roveri,** *Code e Traffico nelle Reti di Comunicazione*, 1991 Editrice La Goliardica - Roma

**M. Decina, A. Roveri,** *Introduzione alle Reti Telefoniche Analogiche e Numeriche*, 1989 Editrice La Goliardica - Roma

**G. Fedele,** *Complementi ed applicazioni di Teoria dei Segnali*, Ed. Ingegneria 2000, 1996 Roma

**M. G. Di Benedetto, P. Mandarini,** *Comunicazioni Elettriche*, 2000 Editrice La Goliardica - Roma

**R. L. Freeman,** *Telecommunication System Enginnering*, $2^{nd}$ Edition, John Wiley & Sons

**F. Halsall,** *Multimedia Communications*, 2001 Pearson Education Limited

**C. W. Helstrom,** *Probability and Stocastic Processes for Engineers*, $2^{nd}$, 1991 Macmillaman Publishing Company

**N. S. Jayant, P. Noll,** *Digital Coding of Waveforms*, 1984 Prentice-Hall, NJ

**M. Listanti, A. Roveri,** *Comunicazioni Dati*, Appunti

**P. Mandarini,** *Teoria dei Segnali*, 1979 Editrice La Goliardica - Roma

**A. V. Oppenheim, R. W. Shafer,** *Digital Signal processing*, 1975 Prentice Hall, NJ

**A. Papoulis,** *Probability, Random variables, and Stochastic Processes*, 1991 McGraw-Hill Int.Eds.

**B. Peroni,** *Comunicazioni Elettriche*, Ed. Scientifiche Siderea, 1973 Roma

**G. M. Poscetti,** *Elementi di teoria dell'informazione*, Ed. Ingegneria 200, 1996 Roma

**U. Reimers,** *Digital Video Broadcasting*, Springer-Verlag Berlin Heidelberg 2001

**A. Roveri,** *Reti di telecomunicazione*, Appunti

**M. Schwartz,** *Information Transmission, Modulation, and Noise*, $4^{th}$ 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 $6^{th}$ 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 Comunivcation Systems*, 1986 Mc Graw Hill

**J. Watkinson,** *The MPEG handbook*, 2001 Focal Press

La stesura di questa opera inizia nel 2001, come materiale di supporto didattico per un corso di Telecomunicazioni per la Laurea in Ingegneria Elettrica presso l'Università di Roma La Sapienza, con l'intento di affrontare in modo coerente temi generalmente sviluppati in corsi separati. Ad oggi, si è accumulato materiale su *segnali, sistemi, probabilità, campionamento, quantizzazione, elaborazione, modulazione analogica e numerica, multiplazione, commutazione, traffico, trasmissione dati, reti, trasferimenti energetici, dimensionamenti, collegamenti radio, via satellite, in mobilità, in cavo ed su fibra ottica, rumore, teoria dell'informazione, compressione dati, codifica di immagine, codifica audio e video, capacità di canale e codici correttori.*

Di anno in anno la didattica svolta in aula, per corsi diversi, ha dato modo di sviluppare nuovo materiale e riorganizzare l'esistente migliorandolo, arricchito di innumerevoli illustrazioni di schemi e curve. I diversi argomenti si intrecciano nel testo, a volte ritardandone alcuni, o anticipandone altri, ma sempre tutti legati da un reticolo di rimandi che consente di sviluppare percorsi di lettura alternativi. Fin dall'inizio si è cercato di mantenere un equilibrio tra rigore analitico, speculazioni concettuali, ed esempi concreti di realizzazioni operative, sviluppando la trattazione su livelli paralleli, mediante l'uso di numerose note ed appendici, sede di approfondimento.

Il testo è rimasto fin dall'inizio disponibile per il download presso *http://infocom.uniroma1.it/alef/wiki/Didattica.LibroTLC*, totalizzando circa 2000 scaricamenti/anno, pemettendo il feedback diretto da parte dei lettori, ed incoraggiando così il continuo perfezionamento ed ampliamento dell'opera.

*Alessandro Falaschi è nato nel 1959 a Roma, e si laurea con lode nel 1983 in Ingegneria Elettronica presso l'Università di Roma La Sapienza, dove in seguito consegue il Dottorato di Ricerca. Diviene quindi ricercatore presso l'Università di Perugia, e dopo essere tornato al Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni (DIET) di Roma, attualmente insegna presso il polo didattico di Latina.*

*Si è occupato di elaborazione del segnale vocale e di trasmissioni numeriche, ed i suoi interessi attuali riguardano le applicazioni telematiche di comunicazione multimediale, come VoIP e WebRTC, le video conferenze di gruppo, e il live streaming Internet; inoltre, segue gli sviluppi dei sistemi peer to peer e delle VANET.*