Politecnico di Torino
Laurea Magistrale in Ingegneria Informatica

appunti di

Elaborazione e trasmissione di informazioni multimediali

Autori principali: Luca Ghio
Docenti: Enrico Masala, Juan Carlos De Martin
Anno accademico: 2014/2015
Versione: 1.0.0.0
Data: 22 dicembre 2015

Ringraziamenti

Oltre agli autori precedentemente citati, quest'opera può includere contributi da opere correlate su WikiAppunti e su Wikibooks, perciò grazie anche a tutti gli utenti che hanno apportato contributi agli appunti *Elaborazione e trasmissione di informazioni multimediali* e al libro *Codifica della voce e dell'audio*.

Informazioni su quest'opera

Quest'opera è pubblicata gratuitamente. Puoi scaricare l'ultima versione del documento PDF, insieme al codice sorgente LaTeX, da qui: http://lucaghio.webege.com/redirs/16

Quest'opera non è stata controllata in alcun modo dai professori e quindi potrebbe contenere degli errori. Se ne trovi uno, sei invitato a correggerlo direttamente tu stesso realizzando un commit nel repository Git pubblico o modificando gli appunti *Elaborazione e trasmissione di informazioni multimediali* su WikiAppunti, oppure alternativamente puoi contattare l'autore principale inviando un messaggio di posta elettronica a artghio@tiscali.it.

Licenza

Indice

Parte I

Codifica della voce e dell'audio

Capitolo 1

Storia[1]

1.1 Telegrafo

Il **telegrafo** permetteva di trasmettere simboli (alfabero Morse) usando la rete elettrica. La banda del telegrafo era pari a circa 150 Hz.

Il telegrafo fu il primo esempio di infrastruttura di telecomunicazione, che richiese un grosso investimento sia statale sia privato.

1.2 Telefono

Il brevetto del **telefono** risale al 1875 a nome di Graham Bell. Il brevetto iniziale era basato su una trasmissione analogica; le prime trasmissioni erano semi-comprensibili, per poi arrivare a comunicazioni sempre più chiare.

L'idea alla base del telefono era la possibilità di mettere in comunicazione tra di loro i diversi uffici e negozi che stavano iniziando a prendere vita proprio in questo periodo.

Uno dei problemi che si ebbe fin dall'inizio fu definire quanta banda fosse necessaria per la comunicazione, in modo da garantire l'intelligibilità (molto importante) e una sufficiente qualità. Empiricamente, e successivamente scientificamente, si scoprì che la banda minima per avere un servizio discreto è pari a circa 3 KHz.

Negli anni successivi la telefonia si espanse per usi personali, portando piano piano alla creazioni di monopoli.

1.3 Era digitale

La rete telefonica è ingegnerizzata per fare bene solamente una cosa: trasmettere voce. La motivazione per la quale la voce è stata il primo elemento che è stato preso in considerazione per la trasmissione digitale è proprio la telefonia.

L'avvento dell'era digitale ha portato con sé molte problematiche: come fare nel campo digitale le stesse operazioni già note nel campo analogico? Questa problematica ha portato alla nascita di una nuova materia, detta "elaborazione numerica dei segnali" (Digital Signal Processing [DSP]).

Il **digitale** è più attraente dell'analogico:

- la trasmissione dei segnali digitali è più resistente al rumore, perché non è soggetta a fonti di rumore analogiche quali polvere, graffi ⇒ maggiori prestazioni (qualità) a parità di potenza (rumore), stesse prestazioni per potenze minori;
- un segnale analogico è più facile da elaborare se viene campionato in un segnale numerico, cioè a tempo discreto (campionamento) e discreto in ampiezza (quantizzazione).

[1] Questo capitolo è basato sugli appunti di Nicola Gallo.

Capitolo 2

Conversione analogico/digitale

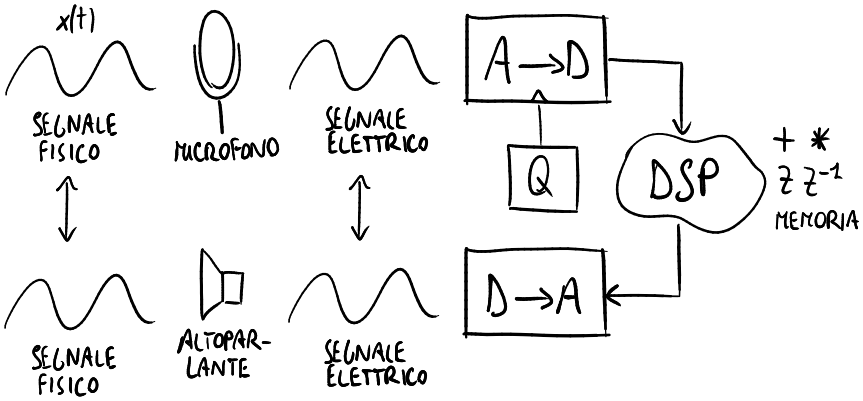


Figura 2.1: Schema a blocchi di un sistema di conversione A/D e D/A.

2.1 Campionamento

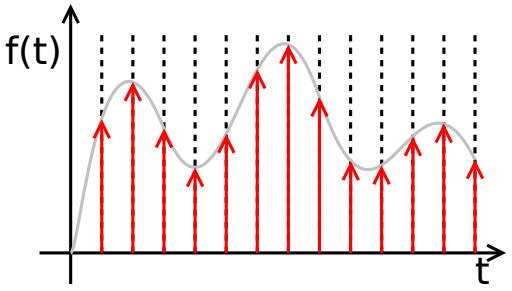


Figura 2.2: Campionamento nel dominio del tempo.[1]

Il **campionamento** di un segnale tempo-continuo $x(t)$ produce il segnale tempo-discreto $x[n]$, che è una sequenza equispaziata di campioni del segnale originario.

Il campionamento consiste nella moltiplicazione del segnale analogico $x(t)$ per un treno di impulsi (delta):

$$x[n] = \sum_n x(t)\,\delta(t - nT)$$

[1] Questa immagine è tratta da Wikimedia Commons (Digital.signal.discret.svg), è stata realizzata da Petr Adámek, dall'utente Rbj e da Walter Dvorak e si trova nel dominio pubblico.

2.1.1 Teorema del campionamento di Shannon

Il **teorema del campionamento di Shannon** definisce come campionare un segnale tempo-continuo senza perdita di informazioni:

> Sotto certe condizioni, un segnale tempo-continuo può essere perfettamente ricostruito a partire dai suoi campioni se la frequenza di campionamento F_c è maggiore del doppio della banda F_0 del segnale:
>
> $$F_c > 2F_0$$

Condizione 1

La banda F_0 del segnale di partenza deve essere limitata.

La maggioranza dei segnali utilizzati in realtà ha banda illimitata: esiste un intervallo al di fuori del quale il segnale è significativamente vicino a zero, ma non è mai identicamente nullo $\Rightarrow$ l'eliminazione delle parti ad alta frequenza porta a un'approssimazione, e il teorema di Shannon non è fisicamente realizzabile.

Condizione 2

Il segnale campionato può essere ricostruito perfettamente se e solo se come filtro interpolatore viene usato il **filtro passa-basso ideale**, con frequenza di taglio pari alla banda F_0, che corrisponde:

- nel dominio del tempo: alla convoluzione con la risposta all'impulso del filtro (ovvero la funzione sinc):
$$x(t) = \sum_n x[n] * \delta\left(t - nT\right)$$
- nel dominio della frequenza: alla moltiplicazione con la funzione di trasferimento del filtro:
$$X(f) = \frac{1}{T} \sum_n X\left(\frac{n}{T}\right) \delta\left(f - \frac{n}{T}\right)$$
  - piatta nella banda del segnale (non distorcente);
  - a pendenza infinita in corrispondenza della frequenza di taglio;
  - nulla al di fuori della banda del segnale.

Anche in questo caso il filtro ideale non è fisicamente realizzabile, e i filtri reali introducono approssimazioni:

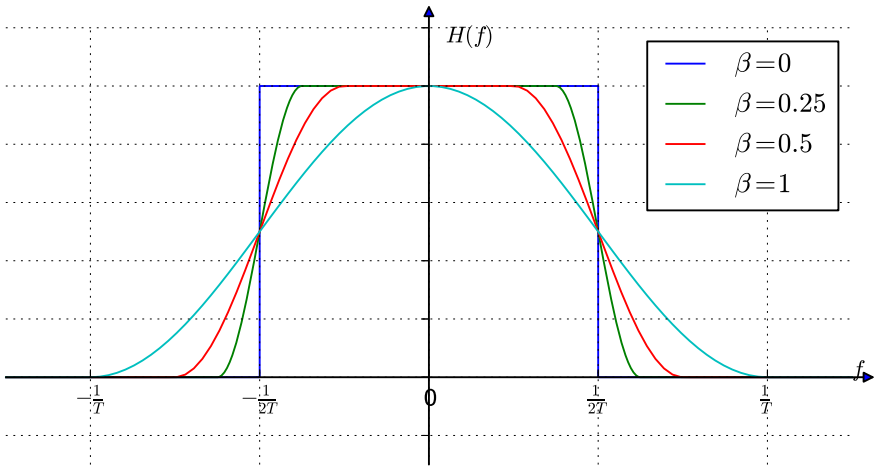


Figura 2.3: Confronto tra il filtro ideale (blu) e alcuni filtri reali.[2]

2.1.2 Diagramma di uguale intensità sonora

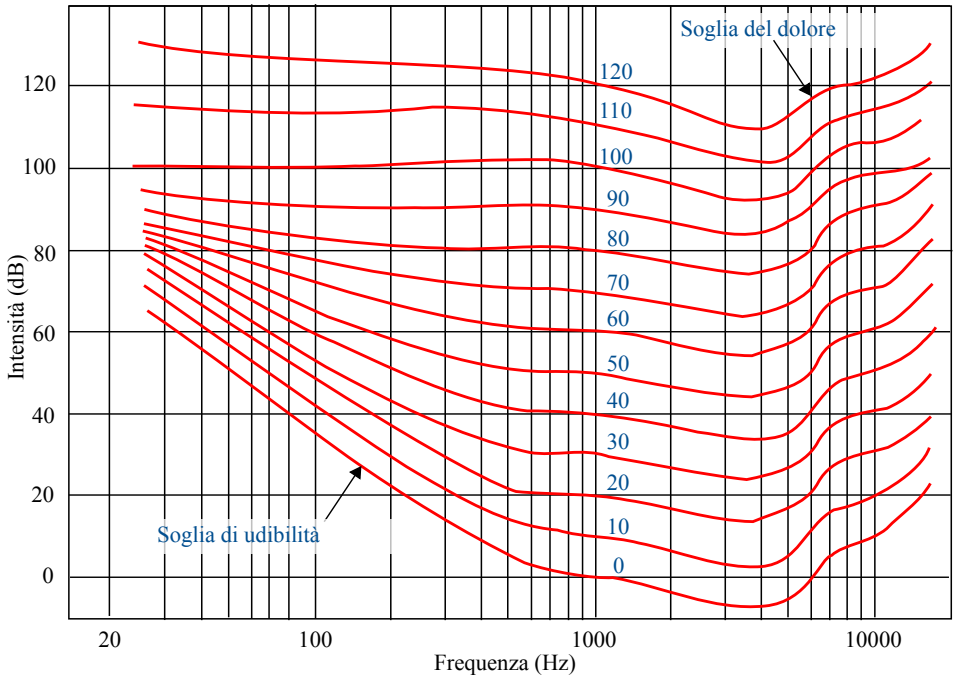


Figura 2.4: Diagramma di uguale intensità sonora.[3]

suono onde trasversali di pressione che si propagano in un mezzo (tipicamente l'aria)

audio l'insieme dei suoni percepibili dal sistema uditivo umano

L'**audio** è caratterizzato da intensità e frequenza.[4]

Intensità (dB)

La misura dell'intensità è il **Sound Pressure Level** (SPL):

$$\mathrm{SPL} = 10 \log_{10} \frac{P}{P_0} \ \mathrm{dB}$$

dove P_0 è la pressione della sinusoide minimamente udibile alla frequenza di riferimento (1 kHz).

Il suono udibile è compreso tra la **soglia di udibilità** e la **soglia del dolore**:

- 0 dB = soglia di udibilità: suoni al di sotto di questa soglia non sono udibili dal sistema uditivo umano;
- 100 dB = soglia del danno irreversibile: suoni al di sopra di questa soglia possono ridurre la capacità uditiva in maniera permanente;
- 120 dB = soglia del dolore fisico: suoni al di sopra di questa soglia provocano danni fisici al timpano.

[2]Questa immagine è tratta da Wikimedia Commons (Raised-cosine filter.svg), è stata realizzata dall'utente Oli Filth, dall'utente Krishnavedala e da Edgar Bonet, ed è concessa sotto la licenza Creative Commons Attribuzione - Condividi allo stesso modo 3.0 Unported.

[3]Questa immagine è tratta da Wikimedia Commons (FletcherMunson ELC.svg), è stata realizzata da Oarih Ropshkow ed è concessa sotto la licenza Creative Commons Attribuzione - Condividi allo stesso modo 3.0 Unported.

[4]I valori di SPL e di frequenza riportati di seguito sono convenzionali, ma dipendono in realtà da fattori legati alla persona come l'età, la salute, ecc.

Frequenza (Hz)

Il suono udibile è compreso tra 20 Hz e 20 kHz, per un'ampiezza pari a 10 ottave[5]. La curva di udibilità è fortemente non lineare:

- l'intervallo di frequenze tra 1 kHz e 4 kHz comprende i suoni a cui il sistema uditivo è maggiormente sensibile (soglia di udibilità molto bassa);
- a frequenze molto basse o molto alte, possono essere sentiti solo suoni a intensità molto alte (soglia di udibilità molto alta).

2.1.3 Voce

La voce umana naturale è compresa:

- intensità: entro una dinamica ampia 60 dB (dal bisbiglio all'urlo);
- frequenza: nell'intervallo da 20 Hz a 12 kHz.

Tuttavia per la voce trasmessa via telefono si è visto empiricamente che è sufficiente una banda compresa tra 300 e 3400 Hz, detta **banda telefonica**, in modo da garantire:

- l'intelligibilità (indispensabile): capire la sequenza di fonemi che viene pronunciata dall'interlocutore;
- una sufficiente qualità (naturalezza): capire informazioni sul parlatore (come identità, sesso, età...).

La voce in banda telefonica (narrowband voice) deve essere campionata a una frequenza maggiore della minima frequenza di campionamento imposta dal teorema di Shannon $\Rightarrow$ viene campionata alla frequenza di 8 kHz per tenere conto delle non idealità dei filtri.

Oggigiorno nuove tecnologie (ad es. VoIP) rendono possibile la voce a banda larga (wideband):

- larghezza di banda = 50-7000 Hz
- frequenza di campionamento = 16 kHz

2.2 Quantizzazione

La **quantizzazione** permette di trasformare un segnale tempo-discreto $x[n]$ in un segnale digitale (o numerico) $\hat{x}[n]$.

La **zona operativa** (o dinamica, o fondo scala) X_m è l'intervallo di valori che ogni campione può assumere sulla scala reale. Dati N bit:

1. la zona operativa viene suddivisa in $2^N - 1$ intervalli, chiamati **gradini** (o step) **di quantizzazione**;
2. ogni campione viene mappato su uno dei 2^N valori possibili, e in particolare al più vicino (secondo la distanza euclidea).

L'operazione di quantizzazione introduce un errore irreversibile, chiamato **errore** (o rumore) **di quantizzazione** $e\,[n]$, pari alla differenza fra un campione reale $x\,[n]$ e la sua versione quantizzata $\hat{x}\,[n]$:

$$|e\,[n]| = |\hat{x}\,[n] - x\,[n]| \leq \frac{\Delta}{2}$$

[5]Si raddoppia circa 10 volte:

$$20 \to 40 \to 80 \to 160 \to 320 \to 640 \to 1080 \to 2160 \to 4320 \to 8620 \to 17740$$

Per confronto, il sistema visivo si limita a un intervallo di frequenze ampio appena 1 ottava.

dove Δ è l'ampiezza del gradino di quantizzazione. Nel quantizzatore uniforme[6], tutti i gradini di quantizzazione hanno ampiezza costante $\Delta = \frac{X_m}{2^N}$.

Un campione può assumere tipicamente tutti i valori sulla scala reale $\Rightarrow$ la **zona di saturazione** (o overload) comprende i valori al di fuori della zona operativa, in cui l'errore di quantizzazione può essere potenzialmente infinito.

2.2.1 Progetto di un quantizzatore

Numero di bit per campione

Il numero N di bit per campione dipende da:

- ampiezza X_m della zona operativa: a parità di qualità, il numero di livelli necessario cresce con l'ampiezza della zona operativa;
- errore di quantizzazione $e[n]$: a parità di ampiezza della zona operativa, il numero di livelli necessario cresce con la qualità (prestazioni) della quantizzazione.

Valori tipici

- CD audio: 16 bit/campione
- voce telefonica: 12 bit/campione (minore qualità della musica + minore potenza del segnale)
- immagini in scala di grigi: 8 bpp (bit/pixel)
- immagini a colori: 24 bpp

Ampiezza della zona operativa

A parità di numero N di bit, la scelta dell'ampiezza X_m della zona operativa deriva dal compromesso tra:

- zona stretta: più la zona operativa è stretta e i livelli sono fitti, più l'errore di quantizzazione è basso e le prestazioni del quantizzatore sono alte;
- zona ampia: la zona operativa deve includere i valori a probabilità più alta in modo da minimizzare la **probabilità di overload**, ossia la percentuale dei campioni il cui valore cade al di fuori della zona operativa.

Assumendo una distribuzione di probabilità gaussiana, si è visto empiricamente che la scelta di una zona operativa con un'ampiezza X_m pari a 4σ comporta una percentuale di overhead pari allo 0,069% circa.

2.2.2 Rapporto segnale/rumore

La qualità del segnale quantizzato è espressa in termini del **rapporto segnale/rumore** SNR, definito come il rapporto tra la potenza σ_x^2 del segnale non ancora quantizzato $x\,[n]$ e la potenza σ_e^2 dell'errore di quantizzazione $e\,[n]$:

$$\mathrm{SNR} = 10 \log_{10} \frac{\sigma_x^2}{\sigma_e^2} \text{ dB}$$

dove la potenza σ_x^2 di un segnale $x(t)$ avente una funzione densità di probabilità $\mathrm{PDF}_x\,(t)$ è:

$$\sigma_x^2 = \int_{-\infty}^{+\infty} x^2\,(t) \cdot \mathrm{PDF}_x\,(t)\,dt$$

[6] Si rimanda alla sezione 4.1.1.

2.2.3 Quantizzatore ottimo

Un quantizzatore si dice **ottimo** per un certo segnale se la sua distribuzione di livelli è tale che:

- tutti i livelli di quantizzazione vengono utilizzati con pari probabilità, cioè nessun livello è utilizzato più di altri;
- l'energia σ_e^2 dell'errore di quantizzazione $e[n]$ viene minimizzata;
- il rapporto segnale/rumore SNR viene massimizzato.

Il quantizzatore ottimo si ottiene facendo "combaciare" la distribuzione dei livelli e la funzione PDF del segnale. Il **teorema di Max-Lloyd** permette di ricavare la distribuzione ottima di livelli a partire dall'espressione analitica della funzione PDF del segnale.

Il quantizzatore uniforme è un quantizzatore ottimo per segnali distribuiti uniformemente sulla zona operativa, ma i segnali audio tipicamente hanno una distribuzione di probabilità non uniforme.

Capitolo 3

Codifica di sorgente

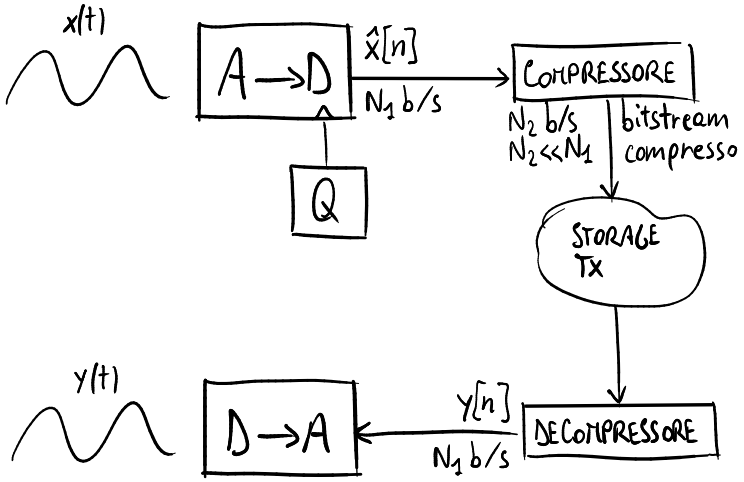


Figura 3.1: Schema a blocchi di un sistema di conversione A/D e D/A con compressore.

La voce umana naturale (fino a 12 kHz) dovrebbe essere rappresentata da un segnale digitale caratterizzato da un bit rate R molto elevato:

$$R = N \times F_c = 16 \text{ bit/campione} \times 24000 \text{ Hz} = 384000 \text{ b/s}$$

dove:

- la frequenza di campionamento F_c è imposta dal teorema di Shannon:

$$F_c > 2 \times 12000 = 24000 \text{ Hz}$$

- come numero di bit N si può assumere quello adottato per i CD audio:

$$N = 16 \text{ bit/campione}$$

Un bit rate troppo elevato può essere un problema importante dal punto di vista di:

- capacità di trasmissione in rete:
  - originariamente i modem potevano trasmettere centinaia/poche migliaia di bit al secondo;
  - oggi è importante risparmiare banda sul backbone dove passano tante telefonate;
- potenza di elaborazione dei microprocessori.

Bit rate tipici di segnali digitali non compressi

- musica non compressa (CD audio):
  - mono:
    $$R = 16 \text{ bit/campione} \times 44100 \text{ Hz} \simeq 715 \text{ kb/s}$$
  - stereo:
    $$R \simeq 1,4 \text{ Mb/s}$$
- musica compressa (MP3):
  $$R \leq 320 \text{ kb/s}$$
- video a risoluzione 640×480:
  $$R = 640 \text{ px} \times 480 \text{ px} \times 25 \text{ fps} \times 24 \text{ bpp} \simeq 184 \text{ Mb/s}$$
- voce in banda telefonica:
  $$R = 12 \text{ bit/campione} \times 8000 \text{ Hz} = 96 \text{ kb/s}$$

3.1 Compressione

	rapporto di compressione	**formati comuni**	**applicazioni**
lossless	da 2 a 3 volte	RAW, FLAC, ZIP	• audiofili che desiderano la massima qualità • immaganizzamento dei "master": si evitano compressioni in cascata • applicazioni particolari: applicazioni legali, ambito medico, astronomia
lossy	da 10 a 100 volte	MP3, JPEG, MPEG-2	• memorizzazione su dischi piccoli • trasmissione su canali con poca banda disponibile

Tabella 3.1: Confronto tra le tecniche di compressione lossless e lossy.

Il segnale digitale in uscita dal convertitore analogico/digitale, prima di essere trasmesso o archiviato, attraversa tipicamente due blocchi:

- **codifica di sorgente**: si occupa di codificare il segnale digitale in maniera compatta, al fine di ridurne il bit rate, tramite tecniche di compressione che conoscono il contenuto del segnale (ad es. MP3);
- **codifica di canale**: si occupa di codificare i bit da trasmettere in maniera robusta tramite tecniche di protezione (rilevamento e recupero) dagli errori che lavorano su sequenze di bit indipendentemente dal contenuto del segnale (ad es. codice di Hamming).

3.1.1 Classificazione delle tecniche di compressione

- **lossless** (senza perdite) (es. RAW, FLAC, ZIP): il segnale decompresso $y[n]$ è identico al segnale originario $x[n] \Rightarrow$ i due segnali $y[n]$ e $x[n]$ sicuramente sono percettivamente indistinguibili indipendentemente dall'ascoltatore;
- **lossy** (con perdita) (es. MP3, JPEG, MPEG-2): il segnale decompresso $y[n]$ è diverso dal segnale originario $x[n] \Rightarrow$ i due segnali $y[n]$ e $x[n]$:
  - devono essere almeno percettivamente simili secondo la statistica (cioè percepiti come simili dalla maggior parte delle persone);
  - se il bit rate è sufficientemente elevato, possono essere percepiti come indistinguibili.

3.2 Caratteristiche dei codificatori multimediali

- bit rate: sezione 3.2.1
- complessità: sezione 3.2.2
- ritardo: sezione 3.2.3
- robustezza: sezione 3.2.4
- qualità: sezione 3.2.5

3.2.1 Bit rate

- **constant bit rate** (CBR): il bit rate è costante $\Rightarrow$ la codifica è più semplice da implementare, ma meno efficiente;
- **variable bit rate** (VBR): il bit rate è variabile nel tempo:
  - source-driven: il bit rate cambia in base alla variabilità del segnale (es. telefonata: voce/silenzio);
  - network-driven: il bit rate cambia in base alla banda disponibile del canale (es. YouTube).

3.2.2 Complessità

La complessità di un codificatore può essere quantificata in base a:

- numero di operazioni della CPU al secondo, espresso in Million Instructions Per Second (MIPS) (unità di misura indipendente dalla CPU);
- quantità di memoria RAM occupata durante l'esecuzione dell'algoritmo;
- quantità di memoria ROM richiesta per memorizzare il codice (ad es. tabelle di quantizzazione).

3.2.3 Ritardo

Il ritardo di un codificatore è dato dalla somma di due componenti:

- ritardo computazionale: dipende dalle prestazioni della tecnologia (ad es. potenza della CPU) in uso, e può essere ridotto con hardware più potente;
- ritardo algoritmico: dipende dall'algoritmo per come è stato progettato, e non può essere ridotto se non modificando l'algoritmo (ad es. ritardo di bufferizzazione per algoritmi che lavorano su segmenti di campioni).

3.2.4 Robustezza

La robustezza può essere intesa in due modi:

- robustezza ai segnali di ingresso: quanto un codificatore tarato per una specifica categoria di segnale è sensibile a un segnale diverso da quello atteso, ad esempio:
  - musica invece di voce a un codificatore vocale;
  - voce + rumore di fondo (es. finestrino aperto);
- robustezza agli errori di trasmissione: anche la codifica di sorgente si preoccupa di proteggere il bitstream compresso dagli errori comprimendo in modo robusto:
  - robustezza agli errori su singoli bit;
  - robustezza alle perdite di pacchetti (frame): ad esempio il codificatore ILBC di Skype fu progettato per Internet (rete IP con perdite).

3.2.5 Qualità

L'elaborazione dei segnali multimediali guarda non tanto al rapporto segnale/rumore SNR, quanto alla percezione del segnale multimediale da parte dell'essere umano.

percezione l'effetto di un segnale tenuto conto di tutto il processo di elaborazione umana (cervello incluso)

psico-acustica in ambito acustico, la disciplina che studia la percezione uditiva

La valutazione delle prestazioni di un algoritmo di compressione si basa principalmente sulla **valutazione percettiva**: si chiede a un numero statisticamente significativo di persone di valutare il risultato della decodifica.

Voce

La voce decodificata deve essere caratterizzata da:

- l'intelligibilità (indispensabile): capire la sequenza di fonemi che viene pronunciata dall'interlocutore;
- una sufficiente qualità (naturalezza): capire informazioni sul parlatore (come identità, sesso, età...).

Come misurare la naturalezza?

1. ad ogni ascoltatore viene fatta ascoltare la voce decodificata e viene chiesto di esprimere un giudizio soggettivo da 1 (pessimo) a 5 (eccellente);
2. si calcola il **Mean Opinion Score** (MOS) che è la media di tutti i voti.

Il valore 4 del MOS corrisponde alla **toll** (= pedaggio) **quality**, ovvero alla qualità percepita come uguale a quella della telefonia analogica tradizionale e che la telefonia digitale deve avere come minimo perché la gente sia disposta ad acquistare il servizio. Esempi di codificatori che raggiungono la toll quality sono: PCM lineare, G.711, ADPCM G.726.

Per ottenere un risultato significativo dal punto di vista statistico, cioè affetto da un margine di errore sufficientemente piccolo ($\approx 0,1$ MOS), occorre che:

- il campione di ascoltatori sia il più possibile rappresentativo del pubblico degli utenti di telefonia:
  - il numero di ascoltatori deve essere significativo ($\approx 40 \div 50$);
  - il campione di ascoltatori deve essere il più possibile eterogeneo (ad es. età, sesso...);

- il campione di stimoli sia il più possibile rappresentativo del segnale:
  - gli ascoltatori devono essere sottoposti a un insieme statisticamente rappresentativo di tutti gli stimoli possibili (ad es. voce maschile/femminile, lingue diverse...) ($\approx$ centinaia).

Audio

Il campione di ascoltatori è composto da esperti (golden ears) in grado di compiere un'analisi critica.

L'obiettivo è raggiungere la **trasparenza** (indistinguibilità) **percettiva**: l'ascoltatore non riesce a distinguere l'originale dal risultato della decodifica $\Rightarrow$ il campione di ascoltatori dà un voto statisticamente casuale (50%).

Misure oggettive

I test soggettivi:

- sono costosi e lenti perché coinvolgono un gran numero di persone;
- non sono completamente affidabili: sono basati sulla statistica.

Recentemente sono nate delle misure oggettive:

- SNR e derivati: sono tecniche rudimentali e poco utili, in quanto due segnali diversi possono suonare simili;
- tecniche oggettive di tipo percettivo (derivate dalla psico-acustica): sono in grado di simulare come il segnale verrà percepito dall'orecchio umano, ma sono usati solo per misure relative (cioè per confronti tra due algoritmi):
  - voce: ITU PESQ (Perceptual Evaluation of Speech Quality);
  - audio: ITU PEAQ (Perceptual Evaluation of Audio Quality).

Capitolo 4

Tecniche PCM

Le **tecniche di quantizzazione PCM** (Pulse Code Modulation) si basano su:

- codifica campione-per-campione: lavorano su un campione alla volta, e per ogni campione $x[n]$ in ingresso producono un campione quantizzato $\hat{x}[n]$ in uscita;
- codifica di forma d'onda: l'obiettivo è produrre una forma d'onda geometricamente simile all'originale $\Rightarrow$ la forma d'onda risultante sarà anche percettivamente simile.

Le tecniche PCM per la codifica della voce in banda telefonica possono essere suddivise in:

- statiche: una volta che l'algoritmo è stato progettato, esso non cambia nel tempo:
  - senza memoria (o stateless): ogni campione è quantizzato indipendentemente dagli altri campioni: sezione 4.1
  - differenziali o predittive: la quantizzazione di ogni campione sfrutta anche informazioni dagli altri campioni nel passato e/o nel futuro: sezione 4.4
- adattative: l'algoritmo si adatta al segnale corrente stimato: sezione 4.3

Caratteristiche delle tecniche PCM

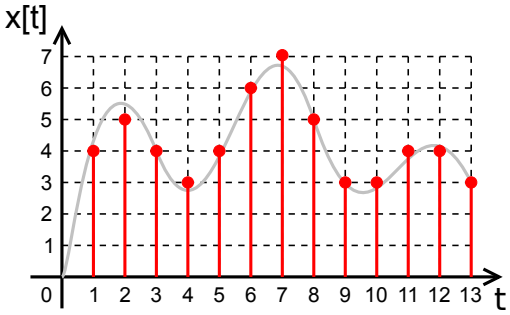
+ robustezza ai segnali di ingresso: poiché l'algoritmo non fa assunzioni sul tipo di segnale, esso continua a funzionare dando buone prestazioni se il tipo di segnale fornito in input non è voce;
+ complessità: è quasi nulla, al massimo pari a 1 MIPS;
+ ritardo: è basso;
− bit rate: le tecniche PCM non riescono a garantire la toll quality con un bit rate al di sotto di 32 kb/s (4 bit/campione) $\Rightarrow$ è un bit rate medio-alto, e può essere troppo alto per specifiche applicazioni (ad es. telefonia satellitare).

4.1 Tecniche PCM senza memoria

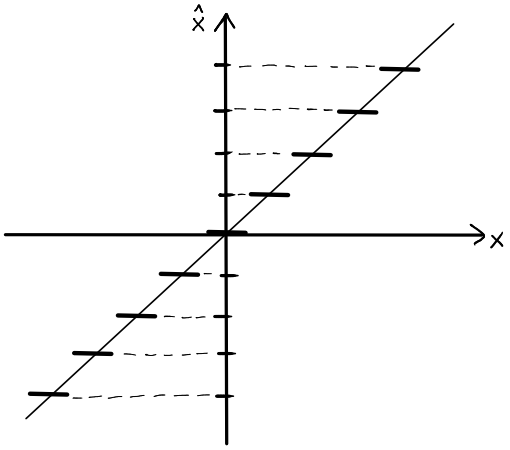
4.1.1 Quantizzatore uniforme: PCM lineare

Il **quantizzatore uniforme** è caratterizzato da una distribuzione dei livelli uniforme: la zona operativa X_m è suddivisa in $2^N - 1$ gradini di quantizzazione di ampiezza costante $\Delta = \frac{X_m}{2^N}$.

[1] Questa immagine è tratta da Wikimedia Commons (Digital.signal.discret.svg), è stata realizzata da Petr Adámek e da Walter Dvorak e si trova nel dominio pubblico.



(a) Esempio di quantizzazione uniforme.[1]



(b) Caratteristica ingresso/uscita di un quantizzatore uniforme.

La potenza σ_e^2 dell'errore di quantizzazione $e\,[n]$, avente una funzione densità di probabilità $\mathrm{PDF}_e\,(t)$ uniforme, è:

$$\sigma_e^2 = \int_{-\frac{\Delta}{2}}^{+\frac{\Delta}{2}} e^2\,(t) \cdot \ \mathrm{PDF}_e\,(t)\,dt = \frac{\Delta^2}{12}$$

Il rapporto segnale/rumore SNR è lineare nel numero di bit $6N$:

$$\mathrm{SNR} = 10 \log_{10} \frac{\sigma_x^2}{\sigma_e^2}\ \mathrm{dB} = K + \alpha \frac{X_m}{\sigma_x} + 6N$$

$\Rightarrow$ il rapporto segnale/rumore SNR migliora di 6 dB per ogni bit in più utilizzato.

La codifica **PCM lineare** è basata su un quantizzatore uniforme a 4096 livelli:

- frequenza di campionamento: (imposta dal teorema di Shannon)

$$F_c = 8000\ \mathrm{Hz}\ = 125\ \mu\mathrm{s} > 2 \times 3400\ \mathrm{Hz}$$

- numero di bit:

$$N = 12\ \mathrm{bit/campione} \Rightarrow N_Q = 2^N = 4096\ \mathrm{livelli}$$

- bit rate:

$$R = 12\ \mathrm{bit/campione} \times 8000\ \mathrm{Hz} = 96\ \mathrm{kb/s}$$

4.1.2 Quantizzatore ottimo: PCM logaritmico (log PCM)

Il quantizzatore uniforme è un quantizzatore ottimo[2] per segnali distribuiti uniformemente sulla zona operativa, ma i segnali audio naturali hanno una distribuzione di probabilità non uniforme. In particolare, la voce ha una funzione distribuzione di probabilità PDF gaussiana fortemente concentrata intorno al valor medio $\Rightarrow$ a parità di qualità, è possibile risparmiare bit utilizzando un quantizzatore avente una distribuzione dei livelli non uniforme:

- intorno all'intensità media il segnale è più probabile $\Rightarrow$ servono livelli più fitti;
- alle basse e alle alte intensità il segnale è meno probabile $\Rightarrow$ i livelli possono essere più radi.

[2] Si veda la sezione 2.2.3.

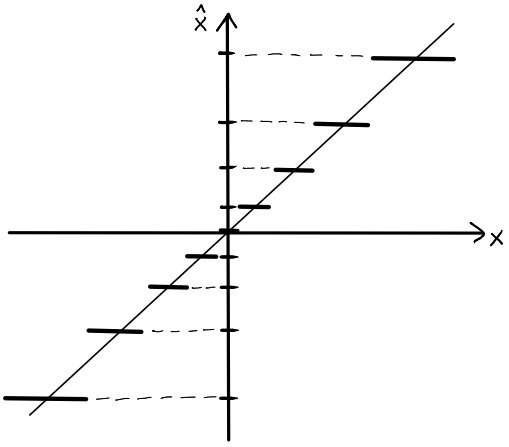


Figura 4.2: Caratteristica ingresso/uscita di un quantizzatore ottimo per la voce.

Poiché l'orecchio umano è sensibile in modo para-logaritmico, il **quantizzatore ottimo** per la voce ha una distribuzione dei livelli simil-logaritmica:

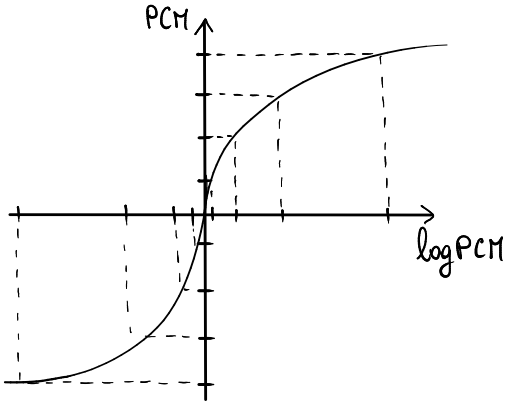


Figura 4.3: I livelli del quantizzatore uniforme sono mappati ai livelli del quantizzatore ottimo secondo una distribuzione simil-logaritmica.

- intorno all'intensità media, i livelli del quantizzatore uniforme sono mappati a tanti livelli vicini tra loro del quantizzatore ottimo;
- alle basse e alle alte intensità, i livelli del quantizzatore uniforme sono mappati a pochi livelli lontani tra loro del quantizzatore ottimo.

Standard ITU G.711

Lo standard **G.711**, sviluppato da International Telecommunication Union (ITU), usa una codifica **PCM logaritmica** (log PCM) basata su un quantizzatore ottimo a 256 livelli:

- numero di bit: i livelli di quantizzazione sono in minor numero ma sono meglio distribuiti secondo le caratteristiche del segnale vocale:

$$N = 8 \text{ bit/campione} \Rightarrow N_Q = 2^N = 256 \text{ livelli}$$

- bit rate: lo standard G.711 raggiunge un bit rate più basso rispetto al PCM lineare pur mantenendone le stesse prestazioni:

$$R = 8 \text{ bit/campione} \times 8000 \text{ Hz} = 64 \text{ kb/s}$$

Applicazioni

- il primo standard per la telefonia digitale, chiamato ISDN

4.1.3 Ulteriori evoluzioni

- telefonia cellulare (GSM, 3G...): il bit rate arriva a 13 kb/s, anche se con la tecnologia di oggi si potrebbe arrivare a circa 6 kb/s;
- applicazioni militari (es. telefoni criptati) e civili (es. telefoni satellitari): il bit rate scende addirittura a 1 kb/s, ma la voce, seppur intelligibile, non è tanto naturale.

4.2 Tecniche PCM differenziali o predittive

Le tecniche PCM senza memoria sono adatte per la codifica del rumore bianco: ogni bit vale 0 o 1 con probabilità 50% $\Rightarrow$ dato un qualunque campione, nessun campione nel passato o nel futuro può fornire informazioni sul campione corrente, perché i campioni sono tutti completamente scorrelati tra loro. Nei segnali audio naturali invece esistono molte correlazioni tra un campione e l'altro, che possono essere sfruttate per comprimere di più.

4.2.1 Quantizzatore differenziale: PCM differenziale (DPCM)

L'idea delle tecniche differenziali è quella di codificare e trasmettere non il campione del segnale originario, con tutta la sua ampia dinamica possibile di valori, ma solo la differenza, detta **segnale differenziale**, tra ogni campione e uno o più dei suoi campioni precedenti.

Se i campioni sono sufficientemente in media correlati tra loro, il segnale differenziale ha una dinamica X_m molto inferiore e una distribuzione gaussiana più stretta ($\sigma_x \gg \sigma_d$) rispetto al segnale originario $\Rightarrow$ servono meno livelli di quantizzazione per raggiungere le stesse prestazioni.

Quantizzatore differenziale del 1° ordine

La differenza $d[n]$ codificata e trasmessa è calcolata tra il campione corrente $x[n]$ e il campione precedente $x[n-1]$:

$$d\left[n\right] = x\left[n\right] - x\left[n-1\right]$$

Il **coefficiente di correlazione** ρ dice quanto due campioni consecutivi sono correlati tra loro:[3]

$$\rho = \frac{E\left[x\left[n\right]x\left[n-1\right]\right]}{E\left[x^2\left[n-1\right]\right]}, \quad 0 \leq \rho \leq 1$$

- se il campione $x\left[n\right]$ è uguale al campione precedente $x\left[n-1\right]$, la correlazione ρ è pari a 1:

$$x\left[n\right] = x\left[n-1\right] \Rightarrow \rho = \frac{E\left[x^2\left[n-1\right]\right]}{E\left[x^2\left[n-1\right]\right]} = 1$$

- se il campione $x\left[n\right]$ è completamente differente rispetto al campione precedente $x\left[n-1\right]$, la correlazione ρ è pari a 0.

Il coefficiente di correlazione ρ è il valore ottimo che minimizza l'energia del segnale differenza $d[n]$:

$$\min \sigma_d^2 \Leftrightarrow d\left[n\right] = x\left[n\right] - \rho x\left[n-1\right]$$

[3] $E\left[X\right]$ è la funzione di valore atteso della variabile casuale X.

Dimostrazione Dato il segnale differenziale:

$$d\left[n\right] = x\left[n\right] - \alpha x\left[n-1\right]$$

si vuole trovare il valore ottimo α che ne minimizza l'energia:

$$\sigma_d^2 = E\left[d^2\left[n\right]\right] = E\left[\left(x\left[n\right] - \alpha x\left[n-1\right]\right)^2\right] =$$

$$= E\left[x^2\left[n\right]\right] + \alpha^2 E\left[x^2\left[n-1\right]\right] - 2\alpha E\left[x\left[n\right] \cdot x\left[n-1\right]\right] \Rightarrow$$

$$\Rightarrow \frac{\partial \sigma_d^2}{\partial \alpha} = 0;\ 2\alpha E\left[x^2\left[n-1\right]\right] - 2E\left[x\left[n\right]x\left[n-1\right]\right] = 0;\ \alpha = \frac{2E\left[x\left[n\right]x\left[n-1\right]\right]}{2E\left[x^2\left[n-1\right]\right]} = \rho$$

Il quantizzatore differenziale funziona molto bene con la voce telefonica grazie al fatto che statisticamente è un segnale fortemente correlato:

$$\rho \simeq 0,9 \Rightarrow d[n] = x[n] - 0,9x[n-1]$$

Processo di codifica e decodifica

1. il codificatore calcola il segnale differenziale $d[n]$ tra il campione corrente $x[n]$ e il campione precedente $x[n-1]$:
$$d\left[n\right] = x\left[n\right] - \rho x\left[n-1\right]$$
2. il codificatore invia al decodificatore la versione quantizzata $\hat{d}[n]$ del segnale differenziale;
3. il decodificatore riceve il segnale differenziale quantizzato $\hat{d}[n]$ e ricostruisce il campione corrente $\hat{x}[n]$:
$$\hat{x}[n] = \hat{d}[n] + \rho\hat{x}[n-1]$$

Quantizzatore differenziale di ordine N

La differenza $d[n]$ è calcolata tra il campione corrente $x[n]$ e la combinazione lineare degli N campioni precedenti:

$$d[n] = x[n] - f\left(x[n-1], x[n-2], \ldots, x[n-N]\right) = x[n] - \sum_{i=1}^{N} \alpha_i x[n-i]$$

L'ordine N deve essere scelto dal compromesso tra:

- prestazioni di compressione: più l'ordine è alto, più informazioni da campioni passati vengono prese per il campione corrente;
- prestazioni di calcolo: all'aumentare dell'ordine aumentano:
  - la memoria necessaria per bufferizzare gli N campioni passati;
  - la complessità di calcolo.

Per la voce telefonica, la **correlazione di breve termine** (= relativa ai campioni adiacenti) è concentrata in media entro 8÷12 campioni ⇒ per la codifica della voce in banda telefonica è sufficiente il **quantizzatore differenziale di ordine 10**: il campione corrente viene codificato prendendo informazioni fino a 10 campioni (equivalenti a 1,2 ms) nel passato.

I valori ottimi dei parametri α_i possono essere calcolati risolvendo un sistema di N derivate parziali in modo analogo al caso del 1° ordine:

$$\begin{cases} \dfrac{\partial \sigma_d^2}{\partial \alpha_1} = 0 \\ \vdots \\ \dfrac{\partial \sigma_d^2}{\partial \alpha_N} = 0 \end{cases}$$

4.2.2 Codifica predittiva: Linear Predictive Coding (LPC)

Un approccio alternativo alla codifica differenziale è la **codifica predittiva**, che affronta un problema di predizione: data la serie storica dei valori passati, è possibile fare una predizione del campione $x[n]$ a partire dai campioni passati?

L'idea delle tecniche predittive è quella di codificare e trasmettere l'**errore di predizione** $e[n]$, cioè la differenza tra il valore effettivo del campione corrente $x[n]$ e il valore predetto $\tilde{x}[n]$:

$$e[n] = x[n] - \tilde{x}[n]$$

- codifica predittiva di ordine 1: la predizione $\tilde{x}[n]$ del campione corrente è basata solo sull'ultimo campione $x[n-1]$:

  $$\tilde{x}[n] = f\left(x[n-1]\right) = \alpha x[n-1]$$

  Se α è il coefficiente di correlazione ρ tra il campione predetto $\tilde{x}[n]$ e il campione effettivo $x[n]$, l'errore di predizione $e[n]$ è minimizzato e la codifica è ottima;

- codifica predittiva di ordine N: la predizione $\tilde{x}[n]$ del campione corrente è basata sulla combinazione lineare degli ultimi N campioni:

  $$\tilde{x}[n] = f\left(x[n-1], x[n-2], \ldots, x[n-N]\right) = \sum_{i=1}^{N} \alpha_i x[n-i]$$

  Se i parametri α_i sono i **coefficienti di predizione lineare**, l'errore di predizione $e[n]$ è minimizzato e la codifica è ottima.

Processo di codifica e decodifica La codifica predittiva funziona grazie al fatto che, dato che il decodificatore ha a disposizione una serie storica simile a quella a disposizione del codificatore, le predizioni svolte da entrambi indipendentemente l'uno dall'altro saranno simili:

1. il codificatore calcola il **valore predetto** del campione corrente a partire dagli ultimi N campioni:

   $$\tilde{x}[n] = f\left(x[n-1], x[n-2], \ldots, x[n-N]\right) =$$

   - ordine 1:

     $$= \rho x[n-1]$$

   - ordine N:

     $$= \sum_{i=1}^{N} \alpha_i x[n-i]$$

2. il codificatore calcola l'**errore di predizione** $e[n]$ confrontando il valore predetto $\tilde{x}[n]$ con il valore effettivo $x[n]$:

   $$e[n] = x[n] - \tilde{x}[n] =$$

   - ordine 1:

     $$= x[n] - \rho x[n-1]$$

- ordine N:

$$= x[n] - \sum_{i=1}^{N} \alpha_i x[n-i]$$

3. il codificatore invia al decodificatore la versione quantizzata $\hat{e}[n]$ dell'errore di predizione;
4. anche il decodificatore calcola il valore predetto per il campione corrente a partire dagli ultimi N campioni ricostruiti:

$$\tilde{x}[n] = f\left(\hat{x}[n-1], \hat{x}[n-2], \ldots, \hat{x}[n-N]\right) =$$

- ordine 1:

$$= \rho \hat{x}[n-1]$$

- ordine N:

$$= \sum_{i=1}^{N} \alpha_i \hat{x}[n-i]$$

5. il decodificatore riceve l'errore di predizione quantizzato $\hat{e}[n]$ e ricostruisce il campione corrente $\hat{x}[n]$:

$$\hat{x}[n] = \hat{e}[n] + \tilde{x}[n] =$$

- ordine 1:

$$= \hat{e}[n] + \rho \hat{x}[n-1]$$

- ordine N:

$$= \hat{e}[n] + \sum_{i=1}^{N} \alpha_i \hat{x}[n-i]$$

4.3 Tecniche PCM adattative: adaptive PCM (APCM)

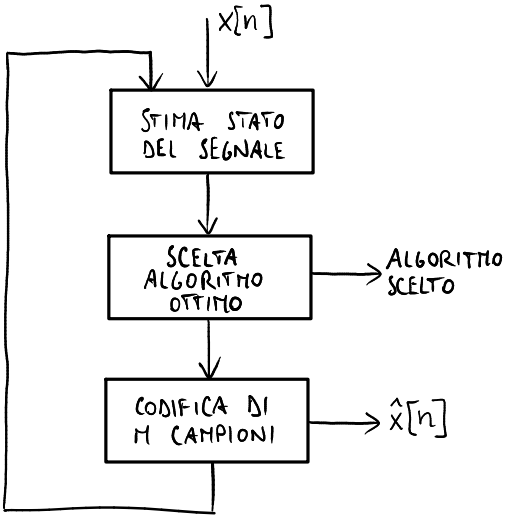


Figura 4.4: Schema a blocchi dell'algoritmo usato dalle tecniche APCM.

Le tecniche PCM statiche sono progettate in base alle caratteristiche statistiche di lungo termine del segnale (valor medio μ, varianza σ, funzione PDF...) $\Rightarrow$ sono adatte per segnali stazionari le cui caratteristiche non dipendono dal tempo. I segnali audio naturali tuttavia sono fortemente non stazionari.

L'idea delle tecniche adattative è quella di usare un algoritmo in grado di adattarsi al segnale corrente stimato nel tempo, con l'obiettivo di risparmiare bit quando il segnale è meno complesso da codificare.

Algoritmo

1. stima dello stato del segnale: si determina lo **stato** del segnale (ad es. rumore o voce) all'interno di una finestra ampia M campioni centrata in n_0;
2. scelta dell'algoritmo ottimo: si sceglie quale algoritmo di codifica è il più adatto al segnale corrente stimato entro la finestra corrente.
   L'algoritmo di codifica scelto deve essere mandato direttamente al ricevitore, cosicché il ricevitore sappia in che modo è stato codificato il segnale. I bit necessari per comunicare queste informazioni al ricevitore sono detti **bit di overhead** perché sono inviati insieme ai campioni quantizzati del segnale e quindi pesano sul bit rate complessivo;
3. codifica di M campioni: si applica l'algoritmo di codifica scelto sulla sequenza di M campioni compresa nella finestra corrente, e i campioni quantizzati sono mandati al ricevitore;
4. si ritorna al passo 1 avanzando la finestra alla sequenza di M campioni successivi.

Il numero M di campioni su cui viene applicato l'algoritmo di codifica scelto è un compromesso tra:

- prestazioni di compressione dei bit che trasportano informazioni multimediali: un adattamento molto frequente permette di seguire fedelmente l'evoluzione del segnale nel tempo e stimare lo stato in modo meno grezzo;
- limitazione dei bit di overhead: occorre contenere il bit rate complessivo evitando di inviare troppi bit di overhead.

Siccome il segnale vocale varia approssimativamente da 50 a 100 volte al secondo, è sufficiente aggiornare la scelta dell'algoritmo ottimo:

- ogni 20 ms:
$$\frac{1 \text{ s}}{50 \text{ volte/s}} = 20 \text{ ms}$$
- ogni $M = 160$ campioni:
$$\frac{20 \text{ ms}}{125 \ \mu\text{s/campione}} = 160 \text{ campioni}$$

Vantaggi/svantaggi

\+ prestazioni di compressione;

\− complessità di calcolo: occorre stimare lo stato del segnale 50 volte al secondo (per la voce);

\− overhead: i bit di overhead, essendo inviati insieme ai campioni quantizzati del segnale, pesano sul bit rate complessivo;

\− robustezza: a volte è difficile stimare lo stato del segnale (ad es. voce con rumore di fondo).

4.3.1 Energy-tracking APCM

La codifica **energy-tracking APCM** è basata su un quantizzatore uniforme con fondo scala variabile nel tempo al fine di adattarsi ai cambiamenti nel tempo dell'energia del segnale:

- il fondo scala si riduce quando il segnale ha meno energia;
- il fondo scala si allarga quando il segnale ha più energia.

Riducendo del fondo scala quando possibile, si possono ottenere due risultati:

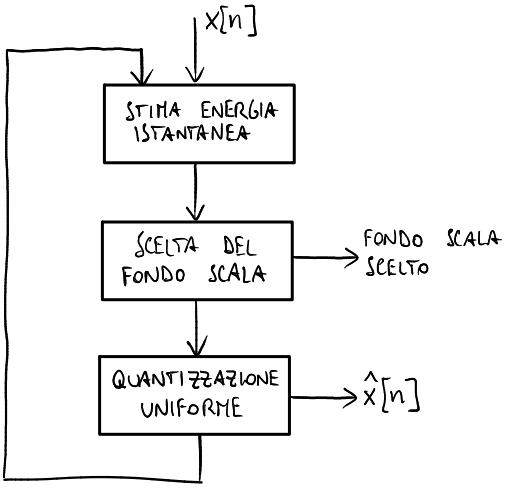


Figura 4.5: Schema a blocchi dell'algoritmo usato dalla codifica energy-tracking APCM.

- aumento del rapporto segnale/rumore SNR a parità di bit rate: viene ridotta l'ampiezza Δ del gradino di quantizzazione, e quindi l'errore di quantizzazione $e[n]$, mantenendo costante il numero N_Q di livelli di quantizzazione;
- riduzione del bit rate a parità di rapporto segnale/rumore SNR: viene ridotto il numero N_Q di livelli di quantizzazione, mantenendo costante l'ampiezza Δ del gradino di quantizzazione.

Algoritmo

1. stima dell'energia istantanea: si misura l'energia locale istantanea del segnale $x[n]$ all'interno della finestra corrente:

$$\mathrm{E}\left[n_0\right] = \sum_{i=n_0-\frac{M}{2}}^{n_0+\frac{M}{2}} x^2[i]$$

2. scelta del fondo scala: si calcola il fondo scala più adatto per la finestra corrente (ad es. tramite la regola euristica del 4σ[4]), e si invia come overhead al ricevitore il fondo scala scelto;
3. quantizzazione uniforme di M campioni con il fondo scala scelto;
4. si ritorna al passo 1.

4.4 Tecniche ADPCM

Le tecniche ADPCM introducono nelle tecniche DPCM l'adattività ai cambiamenti nel tempo dell'energia del segnale differenziale:

- DPCM: i valori ottimi dei parametri α_i sono calcolati una volta in fase di progetto, in modo da minimizzare globalmente l'energia σ_d^2 del segnale differenziale:

$$d[n] = x[n] - \sum_{i=1}^{N} \alpha_i x[n-i], \quad -\infty < n < +\infty$$

[4]Si veda la sezione 2.2.1.

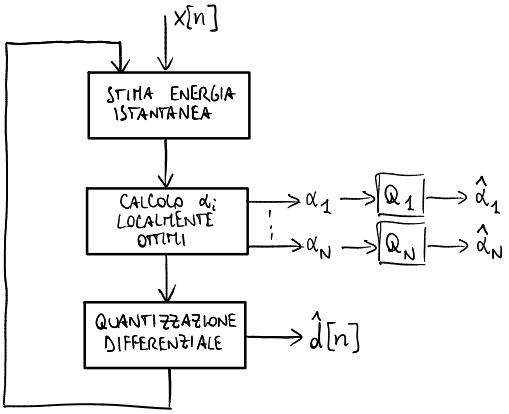


Figura 4.6: Schema a blocchi dell'algoritmo usato dalle tecniche ADPCM.

- ADPCM: i valori ottimi dei parametri α_i sono calcolati di volta in volta per la finestra corrente di M campioni, in modo da minimizzare localmente l'energia istantanea $\mathrm{E}[n_0]$ del segnale differenziale:

$$d[n] = x[n] - \sum_{i=1}^{N} \alpha_i x[n-i], \quad n_0 - \frac{M}{2} < n < n_0 + \frac{M}{2}$$

Algoritmo

1. stima dell'energia istantanea: si misura l'energia istantanea del segnale differenziale $d[n]$ all'interno della finestra corrente:

$$\mathrm{E}[n_0] = \sum_{i=n_0-\frac{M}{2}}^{n_0+\frac{M}{2}} d^2[i]$$

2. calcolo dei valori localmente ottimi dei parametri α_i: si risolve il sistema di N derivate parziali ($N = 10$ per la voce), e si inviano come overhead al ricevitore i valori ottimi calcolati e quantizzati $\hat{\alpha}_i$ (il ricevitore dovrà compiere un'operazione di inversione della matrice);

3. quantizzazione differenziale di ordine N di M campioni con i parametri α_i calcolati, e il segnale differenziale quantizzato $\hat{d}[n]$ è mandato al ricevitore;

4. si ritorna al passo 1.

4.4.1 Quantizzazione dei parametri α_i

Quantizzatore uniforme I valori ottimi dei parametri α_i calcolati per la finestra di trasmissione corrente sono numeri reali ⇒ oltre ai campioni del segnale stesso, occorre quantizzare anche questi valori per poterli mandare al ricevitore in modo digitale ⇒ occorre progettare un quantizzatore uniforme per ognuno dei 10 parametri α_i:

1. creazione di un database: si raccoglie un numero statisticamente significativo di valori del parametro α_i a partire da un campione rappresentativo di segnali vocali;

2. caratterizzazione statistica: si costruisce la funzione densità di probabilità PDF del parametro α_i, ricavandone le caratteristiche statistiche (per una gaussiana: la media μ e la varianza σ);

3. scelta del fondo scala X_m, ad esempio tramite la regola euristica del 4σ[5];

4. scelta del numero N_Q di livelli:

   - se è noto a priori il rapporto segnale/rumore SNR desiderato, è facile ricavare il numero di livelli per mezzo della formula:

$$\text{SNR} = 10 \log_{10} \frac{\sigma_x^2}{\sigma_e^2} \text{ dB} = K + \alpha \frac{X_m}{\sigma_x} + 6N_Q$$

   - nel caso di segnali multimediali, si usano tanti livelli quanti bastano per ottenere una quantizzazione percettivamente trasparente: la voce ricostruita usando il parametro quantizzato α_i è percettivamente indistinguibile dalla voce ricostruita usando il parametro non quantizzato $\hat{\alpha}_i$.

Quantizzatore ottimo La distribuzione di probabilità di ognuno dei parametri α_i però è fortemente concentrata intorno al valor medio $\Rightarrow$ occorre progettare un quantizzatore ottimo, con distribuzione di livelli non uniforme, per ognuno di questi parametri.

Una volta progettato il quantizzatore ottimo, esso è in grado di quantizzare ogni parametro α_i su $3 \div 4$ bit $\Rightarrow$ i 10 parametri quantizzati $\hat{\alpha}_i$ richiedono complessivamente circa 40 bit (sarebbe richiesto circa il doppio dei bit con il quantizzatore uniforme) $\Rightarrow$ essendo inviati 50 volte al secondo (ogni 20 ms), generano un overhead di 2000 b/s: le prestazioni di compressione devono apportare un miglioramento tale da giustificare questo notevole overhead.

4.4.2 Standard ITU G.726

Lo standard **ITU G.726**, grazie a una codifica molto complessa che è derivata dalla tecnica ADPCM, riesce a dimezzare il bit rate del precedente standard, l'ITU G.711, mantenendo la stessa qualità (toll quality):

$$R = 4 \text{ bit/campione} \times 8000 \text{ Hz} = 32 \text{ kb/s}$$

al prezzo di una complessità molto alta pari a 1 MIPS.

Applicazioni

- cordless
- ambito spaziale

[5] Si veda la sezione 2.2.1.

Capitolo 5

Tecniche parametriche

Lo sviluppo delle **tecniche parametriche** iniziò negli anni '70 con la necessità da parte del dipartimento della difesa statunitense di creare una rete di telefonia digitale criptata da usare in ambito militare su cui trasmettere la voce attraverso i modem dell'epoca, che avevano velocità di trasmissione (2400 b/s) assai più basse dei bit rate richiesti dalle tecniche PCM (ogni campione richiederebbe di essere codificato in frazioni di bit!).

Anche la nascente telefonia cellulare digitale a partire dall'inizio degli anni '80 cerca una codifica per la voce digitale a bit rate più bassi:

- qualità: toll quality
- bit rate: $\approx$ 12 kb/s

Le tecniche parametriche descrivono il segnale vocale inteso come contenuto, non come forma d'onda: viene creato un modello di produzione del sistema fonatorio umano, chiamato **modello di produzione** del segnale vocale, e vengono trasmessi solo i parametri di questo modello.

5.1 Codifica predittiva lineare: LPC-10

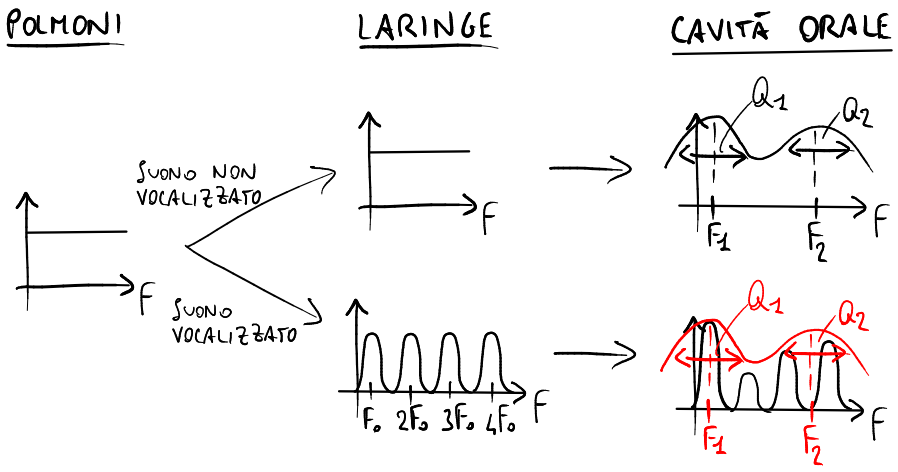


Figura 5.1: Produzione del segnale vocale attraverso il sistema fonatorio umano.

Lo standard **LPC-10** (Linear Predictive Coding), sviluppato negli anni '70, fu il primo standard NATO utilizzato per la telefonia digitale criptata per usi militari e diplomatici:

+ bit rate: solo 650 parametri al secondo (13 parametri ogni 20 ms) vengono trasmessi (contro gli 8000 campioni al secondo delle tecniche PCM) ⇒ il bit rate si abbassa notevolmente:

$$R = 60 \text{ b} \times 50 \text{ volte/s} = 3000 \text{ b/s} \simeq 2400 \text{ b/s}$$

+ intelligibilità: è garantita;

− naturalezza: MOS $\approx$ 2,7: è medio-bassa e al di sotto della toll quality, ma comunque sufficiente per applicazioni militari e satellitari.

5.1.1 Polmoni

Dai polmoni esce un flusso d'aria turbolento, che a livello di segnale produce rumore bianco (avente uno spettro piatto).

Parametri

- **gain** (dB, 7 bit): indica l'intensità del rumore bianco

5.1.2 Laringe

La laringe può produrre due tipi di suoni:

- suoni non vocalizzati (ad es. consonanti "s" e "f", bisbiglio): la laringe è aperta ⇒ il flusso d'aria turbolento in arrivo dai polmoni non viene modificato, e continua a produrre rumore bianco;
- suoni vocalizzati (ad es. vocali): la laringe si apre e si chiude periodicamente producendo una sequenza di sbuffi (segnale glottale) ⇒ le membrane ostruiscono il passaggio dell'aria, e viene prodotto un suono dallo spettro armonico.

Parametri

- **flag** (booleano, 1 bit): specifica se il suono è vocalizzato o non vocalizzato
- **frequenza fondamentale** (Hz, 7 bit): nel caso vocalizzato, lo spettro armonico del segnale presenta dei picchi periodici in corrispondenza dei multipli della frequenza fondamentale F_0 (quindi a F_0, a $2F_0$, ecc.)

Frequenza fondamentale

La **frequenza** (o pitch) **fondamentale** F_0 corrisponde all'altezza (nel senso musicale) della voce:

- è una funzione a tratti: esiste solo quando il suono è vocalizzato;
- è variabile nel tempo: la prosodia studia la variabilità della frequenza fondamentale in un discorso.

La funzione fondamentale è inutile ai fini dell'intelligibilità (un suono non vocalizzato è comunque intelligibile), ma fornisce informazioni importanti per la naturalezza:

- componenti linguistiche:
  - segnali linguistici (ad es. intonazione delle domande);
  - arricchimento semantico della frase (ad es. evidenziare parole o fare parentesi);
- componenti personali:
  - identità del parlatore (sesso, età...): è legata alla frequenza fondamentale media (valori tipici: 100 Hz per gli uomini, 200 Hz per le donne);
  - accento e inflessione dialettale: danno un'idea della provenienza geografica;
  - stato d'animo e di salute del parlatore.

5.1.3 Cavità orale

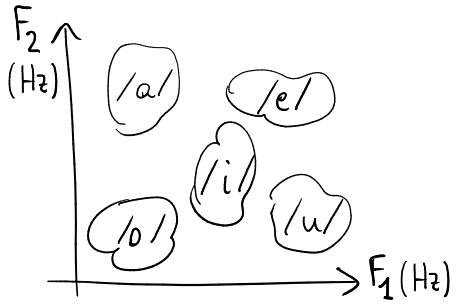


*Figura 5.2: Grafico dei fonemi: i suoni riconosciuti come lo stesso fonema per una certa lingua sono detti **suoni allofoni**.*

Il tratto vocale è una cavità complicata: la voce si propaga sia attraverso la cavità nasale sia attraverso quella orale. Per semplicità, si considera la propagazione della voce solamente attraverso la cavità orale.

La cavità orale è la componente responsabile dell'intelligibilità attraverso l'emissione di fonemi che verranno poi interpretati dal sistema uditivo umano: il sistema uditivo identifica i fonemi grazie all'analisi dello spettro del segnale, e in particolare delle risonanze (effettua una trasformata di Fourier del suono).

La cavità orale può essere vista come un sistema tempo-variante composto da un "tubo" a sezione variabile a causa del diverso posizionamento degli articolatori, in primo luogo della lingua che "sposta" le risonanze, e questi articolatori, svariate volte al secondo, enfatizzano alcune frequenze nello spettro del segnale che attraversa la cavità provocando la nascita di **risonanze** tempo-varianti (quindi non periodiche).

È detta **formante** la frequenza che viene maggiormente enfatizzata per formare una risonanza. Per la modellizzazione della cavità orale occorre individuare le formanti:

- suoni non vocalizzati: le risonanze agiscono su un segnale piatto, producendo un segnale con dei picchi in frequenza ⇒ quei picchi sono proprio le formanti;
- suoni vocalizzati: il segnale in arrivo dalla laringe ha già dei picchi dello spettro periodici, che vengono aumentati in corrispondenza delle risonanze e abbassati tra di esse ⇒ le formanti corrispondono ai picchi dell'**inviluppo** dello spettro, cioè i punti dove è concentrata l'energia.

Il sistema uditivo identifica le prime due o tre formanti per identificare i fonemi pronunciati ⇒ sono sufficienti le prime due formanti F_1 e F_2 ai fini dell'intelligibilità (anche se con 3 formanti l'intelligibilità è migliore).

Parametri ideali

- le prime due **formanti** F_1 e F_2
- le **ampiezze** Q_1 e Q_2 delle prime due formanti

Filtro di sintesi

Assumendo che il sistema fonatorio cambi 50 volte al secondo, ogni 20 ms sarebbe necessario determinare l'inviluppo dallo spettro del segnale e ricavare le formanti ⇒ la complessità di calcolo sarebbe eccessiva.

Un approccio alternativo è modellare la cavità orale come un filtro di ordine 4, chiamato **filtro di sintesi** o **filtro di predizione lineare**, la cui risposta in frequenza ha la "forma" di un generico inviluppo con due picchi.

Questa forma è descritta nel dominio della frequenza da un polinomio complesso che dipende da 4 **coefficienti di predizione lineare** α_i, i quali costituiscono i parametri da trasmettere al ricevitore: ogni 20 ms si calcolano i coefficienti di predizione lineare ottimi, tali da minimizzare localmente la distanza (nel senso euclideo) tra lo spettro del segnale e la risposta in frequenza del filtro (il calcolo è analogo alle tecniche ADPCM[1], ma nel dominio della frequenza anziché nel dominio del tempo).

In pratica lo standard LPC-10 preferisce un filtro di ordine 10 ($8 \div 12$), che equivale a 5 risonanze, invece di un filtro di ordine 4 (2 risonanze) perché:

- è meglio considerare le prime tre formanti per una migliore intelligibilità;
- la cavità nasale introduce degli zeri nella risposta in frequenza $\Rightarrow$ è necessario un ordine superiore per bilanciare il fatto che la cavità nasale è stata trascurata $\Rightarrow$ la voce è più naturale.

Parametri

- 10 **coefficienti di predizione lineare** α_i ($\sim$40-45 bit)

5.1.4 Codifica e decodifica

Organo	Parametri	Numero di bit (ogni 20 ms)
polmoni	gain	7 bit
laringe	flag vocalizzato/non vocalizzato	1 bit
	frequenza fondamentale (se vocalizzato)	7 bit
cavità orale	10 coefficienti di predizione lineare	45 bit
	totale:	60 bit

Tabella 5.1: Riepilogo dei parametri del modello di produzione del segnale vocale.

Il codificatore è una batteria di algoritmi di stima, seguiti dai corrispondenti quantizzatori:

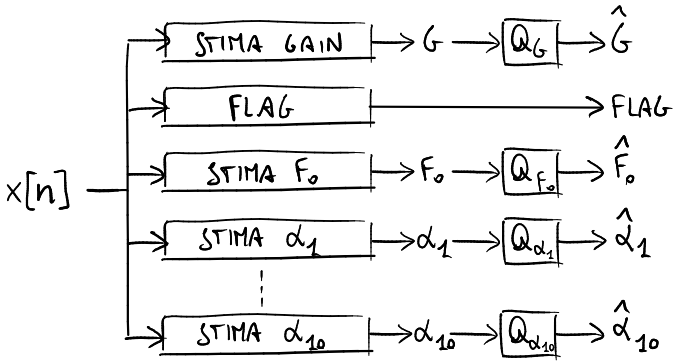


Figura 5.3: Diagramma a blocchi del codificatore LPC-10.

Il decodificatore produce la **voce sintetica** usando il modello di produzione del segnale vocale:

- suono vocalizzato: viene generato del rumore bianco sintetico, formato da una sequenza di numeri casuali;
- suono non vocalizzato: una sequenza di impulsi approssima molto bene la sequenza di sbuffi delle membrane della laringe.

[1] Si veda la sezione 4.4.

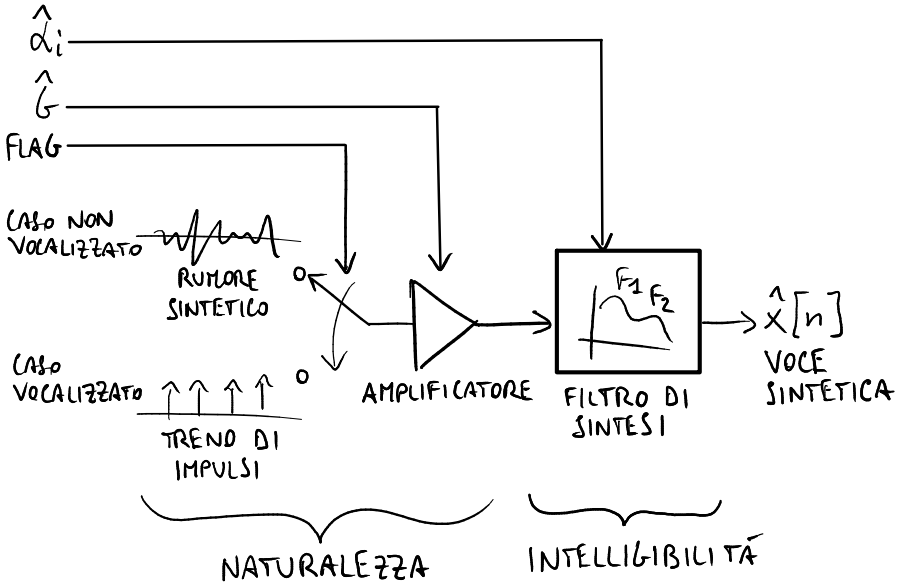


Figura 5.4: Diagramma a blocchi del decodificatore LPC-10.

5.1.5 Limiti LPC-10

La naturalezza della codifica LPC-10 è medio-bassa (MOS $\approx 2{,}7$); per migliorare la naturalezza non è sufficiente aumentare il bit rate a disposizione, ma occorre modificare l'architettura di base.

Il modello di produzione usato dalla codifica LPC-10 presenta alcuni limiti:

- bontà del modello: determina il limite superiore della qualità: i parametri possono anche essere perfetti, ma non basta se il modello è scadente:
  - \+ polmoni: la modellizzazione funziona bene perché sono semplici da modellare;
  - − laringe: il flag vocalizzato/non vocalizzato è un'assunzione troppo semplicistica:
    - l'intero segmento da 20 ms è forzato essere o tutto vocalizzato o tutto non vocalizzato;
    - il suono può essere una via di mezzo tra vocalizzato e non vocalizzato (ad es. voce roca);
  - \+ cavità orale: la modellizzazione funziona molto bene, considerando che è stato scelto un filtro di ordine 10;
- affidabilità algoritmi di stima (con segnali naturali):
  - \+ gain: esistono algoritmi stabili;
  - \+ coefficienti di predizione lineare: esistono algoritmi stabili;
  - − frequenza fondamentale: è difficile da stimare fedelmente con sufficiente affidabilità (ad es. musica di sottofondo, finestrino aperto...);
- \+ prestazioni dei quantizzatori: la quantizzazione è già percettivamente trasparente.

5.2 MELP

Il **MELP** (Mixed Excitation Linear Prediction) fu il nuovo standard della NATO pubblicato a metà degli anni '90.

Il principale limite di LPC-10 era la modellizzazione della laringe: il suono può essere una via di mezzo tra vocalizzato e non vocalizzato (ad es. voce roca). Il MELP migliora la modellizzazione della laringe:

1. lo spettro del segnale vocale viene suddiviso sottobanda per sottobanda;
2. a ogni sottobanda viene assegnato un numero reale che identifica il tipo di componente sonora:
   - 1 = vocalizzato (periodico)
   - 0,5 = una via di mezzo tra vocalizzato e non vocalizzato
   - 0 = non vocalizzato (rumore)

Caratteristiche

- bit rate: rimane lo stesso di LPC-10:

$$R = 2400 \text{ b/s}$$

- naturalezza: rimane al di sotto della toll quality, ancora troppo bassa per la telefonia commerciale:

$$\text{MOS} \approx 3,2$$

Capitolo 6

Tecniche CELP

Le codifiche parametriche ad anello aperto non sono in grado di raggiungere la toll quality: per migliorare la naturalezza non è sufficiente aumentare il bit rate a disposizione, ma occorre modificare l'architettura di base.

Negli anni '80 si studia come combinare insieme le tecniche PCM e le tecniche parametriche:

- codifiche PCM: seguono la forma d'onda ⇒ sono robuste ai segnali in ingresso;
- codifiche parametriche: il filtro di sintesi garantisce l'intelligibilità con pochi bit.

6.1 Quantizzazione vettoriale

Negli anni '70 viene teorizzata la **quantizzazione vettoriale** (VQ):

- il quantizzatore scalare di ordine N trasforma 1 numero reale in uno dei 2^N indici, ciascuno lungo N bit;
- il quantizzatore vettoriale di ordine N trasforma un vettore di N numeri reali in uno dei 2^N indici, ciascuno lungo N bit.

Quantizzatore vettoriale di ordine 3

- un vettore contiene le tre coordinate di un punto nello spazio
- ognuno dei $2^N = 8$ possibili indici è associato a un punto nello spazio
- dato un punto nello spazio in ingresso, il quantizzatore lo trasforma nell'indice associato al punto più vicino al punto dato

Vantaggi/svantaggi

\+ prestazioni: è stato dimostrato che, sotto certe condizioni, un quantizzatore vettoriale di ordine N produce un rapporto segnale/rumore SNR migliore di N quantizzatori scalari di ordine N in parallelo;

− complessità: cresce esponenzialmente con la dimensionalità N, poiché occorre calcolare la distanza euclidea in ogni dimensione.

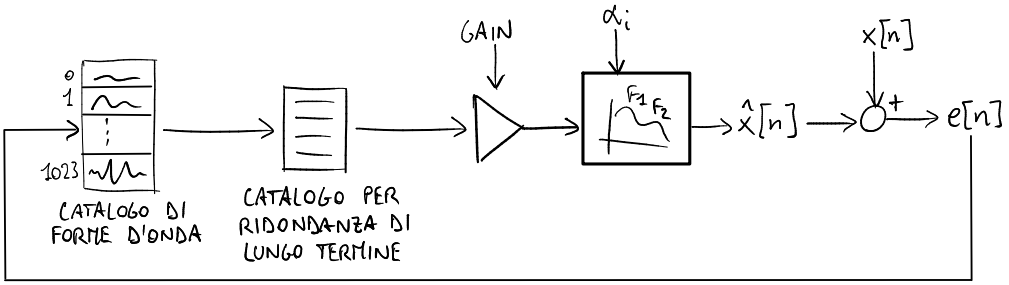


Figura 6.1: Algoritmo ad anello chiuso dell'analisi per sintesi delle tecniche CELP.

6.2 Analisi per sintesi

A metà anni '80 nasce l'idea di usare la quantizzazione vettoriale per fare codifica di forma d'onda a un bit rate più basso delle tecniche PCM: l'idea di base è quella di quantizzare il segnale in ingresso trasformando ogni sequenza di 40 campioni in un indice associato a una forma d'onda.

Il modello della laringe di LPC-10 viene sostituito da due cataloghi contenenti forme d'onda, ciascuna di durata 5 ms (40 campioni):

- **catalogo di forme d'onda**: contiene 1024 forme d'onda predefinite, ciascuna associata a un indice da 10 bit, che rappresentano la componente rumorosa;
- **catalogo per ridondanza di lungo termine**: contiene delle forme d'onda che rappresentano la componente periodica (viene sommata alla componente rumorosa quando il segnale ha delle componenti vocalizzate).

Algoritmo ad anello chiuso (per forza bruta) Per ogni sequenza $x[n]$ di 40 campioni (5 ms) del segnale in ingresso:

1. per ogni forma d'onda nel catalogo:
   (a) si produce la voce sintetica $\hat{x}[n]$ attraverso il filtro di sintesi a partire dalla forma d'onda corrente;
   (b) si calcola l'errore $e[n]$ tra la voce sintetica e la sequenza in ingresso;
2. si seleziona la forma d'onda nel catalogo che genera l'errore minimo, cioè che genera la sequenza di campioni più vicina alla sequenza in ingresso;
3. si invia al decodificatore l'indice associato alla forma d'onda selezionata (insieme al gain G e ai coefficienti di predizione lineare α_i).

Vantaggi/svantaggi

\+ bit rate: abbatte il limite delle tecniche PCM, pur seguendo la forma d'onda (la forma d'onda ricostruita è simile alla forma d'onda originale):

$$R \sim 8 \div 10 \text{ kb/s}$$

\+ naturalezza: la tecnica di analisi per sintesi raggiunge la toll quality;

− complessità: è astronomica ($\div$20 MIPS) poiché occorre calcolare 1024 errori ogni 5 ms.

6.3 Standard CELP

A partire dalla fine degli anni '80, tutti gli standard a toll quality appartengono alla famiglia **CELP** (Codebook-Excited LP).

Prima del 3G, esistevano diversi standard per la telefonia cellulare digitale commerciale:

- USA: erano in competizione tra loro gli standard CDMA e TDMA;
- Europa: è stato imposto un unico standard chiamato GSM.

	Nome	**Anno**	**Bit rate**	**Qualità**
GSM-FR (Full Rate)	RPE-LTP	anni '80	13 kb/s	MOS $\approx$ 3,5[1]
GSM-EFR (Enhanced Full Rate)	ACELP (Algebric)	1995	12,2 kb/s	toll quality (MOS $\approx$ 4)
GSM-AMR (Adaptive Multi-Rate)		$\sim$2000	variabile tra 4,8 e 12,2 kb/s	toll quality

Tabella 6.1: Standard GSM.

	Nome	**Anno**	**Bit rate**	**Applicazioni**
G.728	LD-CELP (Low-Delay)	1995	16 kb/s	poco usato
G.729	CS-ACELP	1998	8 kb/s	VoIP
G.4K	mai standardizzato (per motivi non tecnologici di brevetto)		4 kb/s	

Tabella 6.2: Standard ITU.

L'arrivo del 4G ha permesso di passare dalla voce in banda telefonica alla **voce a larga banda** (wideband):

	voce in banda telefonica	**voce a larga banda**
frequenza di campionamento F_c	8000 Hz	16000 Hz
larghezza di banda B	300-3400 Hz	50-7000 Hz

Tabella 6.3: Standard ITU.

La ricerca si è progressivamente spenta.

6.3.1 GSM-FR

Il **GSM-FR** fu il primo standard europeo per la telefonia cellulare digitale, quasi a toll quality, e venne pubblicato prima che lo sviluppo della tecnica CELP fosse stato portato a termine.

Un canale GSM full rate è ampio complessivamente 22,8 kb/s, ed è suddiviso in bit di voce e bit di correzione degli errori (Forward Error Correction [FEC]), usati dal decodificatore per rilevare e correggere gli errori causati dalla trasmissione su un canale radio (non è possibile reinviare il frame come nel TCP/IP).

L'ampiezza della suddivisione è stata stabilita prendendo come riferimento il canale tipico (Carriage/Interference [C/I] = 7 dB):

[1] Un po' al di sotto della toll quality a causa di un piccolo fruscio tuttavia considerato all'epoca accettabile dagli utenti dei telefoni cellulari.

- FEC: 9,8 kb/s
- voce: 13 kb/s suddivisi in:
  - bit molto importanti: sono protetti dalla FEC, e se è rilevato un errore residuo l'intero frame viene buttato (ripetendo il precedente);
  - bit meno importanti: sono protetti dalla FEC, ma in caso di errori non butto il frame;
  - bit poco importanti: non sono protetti dalla FEC.

6.3.2 GSM-AMR

Il GSM-FR impone una suddivisione fissa tra voce e FEC, che è adatta al caso tipico ma non è adatta a situazioni particolari:

- nei pressi della cella telefonica: i bit di protezione degli errori non servono;
- al limite del raggio d'azione della cella telefonica: il codice di correzione degli errori esce dalla sua zona operativa e provoca ancora più errori in cascata.

Il **GSM-AMR** è un algoritmo a bit rate variabile di tipo network-driven, che permette di costruire un sistema voce/FEC adattativo alle condizioni istantanee del canale.

La suddivisione tra voce e FEC non è fissa, ma si può spostare per adattarsi alla potenza di ricezione del segnale come compromesso tra qualità della voce ed efficacia della FEC:

- quando il segnale è buono, il bit rate è più alto ma il segnale è meno protetto;
- quando il segnale è cattivo, il bit rate si abbassa ma il segnale è molto protetto.

Sono previsti 8 bit rate possibili:

- qualità minima: voce 4,8 kb/s, FEC 18 kb/s;
- qualità massima: voce 12,2 kb/s, FEC 9,8 kb/s (come il GSM-FR).

È meglio una voce codificata peggio ma ben protetta, o una voce codificata meglio ma soggetta a più errori?

- segnale buono: conviene migliorare la qualità di partenza del segnale vocale, anche se ci potrà essere qualche errore;
- segnale cattivo: conviene proteggere molto il segnale vocale e codificare una voce di minore qualità, piuttosto che avere una voce originariamente di alta qualità ma "bombardata" da errori.

Capitolo 7

Codifica dell'audio

audio l'insieme dei suoni percepibili dal sistema uditivo umano

I segnali audio naturali si estendono:

- intensità: tra la soglia di udibilità e la soglia del dolore;
- frequenza: tra 20 Hz a 20000 Hz.

Come codificare un generico segnale audio naturale?

− codifica parametrica: non è pensabile un modello di produzione per tutti i suoni naturali possibili;

- codifica di forma d'onda: le tecniche PCM lavorano campione per campione:
  + bassissimo ritardo
  + bassa complessità
  + elevata qualità
  − alto bit rate

\+ codifica percettiva: grazie a risultati di psicoacustica è possibile produrre forme d'onda differenti, ma percettivamente simili (o anche uguali).

7.1 Codifica percettiva

7.1.1 Fenomeno del mascheramento simultaneo in frequenza

Tra i tanti fenomeni di psicoacustica, spicca (ai fini della compressione) una famiglia di fenomeni, i **fenomeni di mascheramento**: alcune componenti di frequenza di un segnale complesso vengono mascherate (= sono presenti, ma non vengono percepite) da altre componenti di frequenza.

Il sistema uditivo è modellabile come un banco di filtri con 32 bande diseguali, chiamate **bande critiche**: ogni sensore (cilia) è stimolato dagli impulsi sonori entro la sua banda critica.

Data una sinusoide di frequenza F_0 e intensità I_0, si definisce **curva** (o campana) **di mascheramento** la curva al di sotto della quale altre sinusoidi non vengono percepite dal sistema uditivo umano. Questo fenomeno è dovuto al fatto che il sensore non riesce a catturare tutti gli impulsi all'interno della stessa banda critica, ma alcuni segnali, chiamati **segnali mascherati**, possono essere "nascosti" da un **segnale mascherante**: il segnale mascherato dovrebbe essere catturato dallo stesso sensore del segnale mascherante, ma in realtà il sensore, essendo già sovraeccitato dall'impulso mascherante, non riesce a discriminare questa informazione aggiuntiva.

Gli studi di psicoacustica hanno definito le curve di mascheramento al variare della frequenza F_0 e dell'intensità I_0: le curve di mascheramento sono strette ($\sim$100 Hz) alle basse frequenze, e diventano più ampie al crescere della frequenza a causa della sensibilità para-logaritmica del sistema uditivo.

7.1.2 Compressione

Come sfruttare il fenomeno ai fini della compressione? L'idea di base è quella di comprimere mantenendo il rumore (prodotto dal quantizzatore) sotto le curve di mascheramento impedendo quindi che sia percepito dall'orecchio umano.

Ipotizzando la sovrapposizione degli effetti, cioè la linearità del sistema:

1. si costruisce la **curva globale di mascheramento** come la somma delle curve di mascheramento delle singole sinusoidi;

2. si applica una tecnica PCM quantizzando in modo da produrre un rumore che non superi la curva globale di mascheramento.

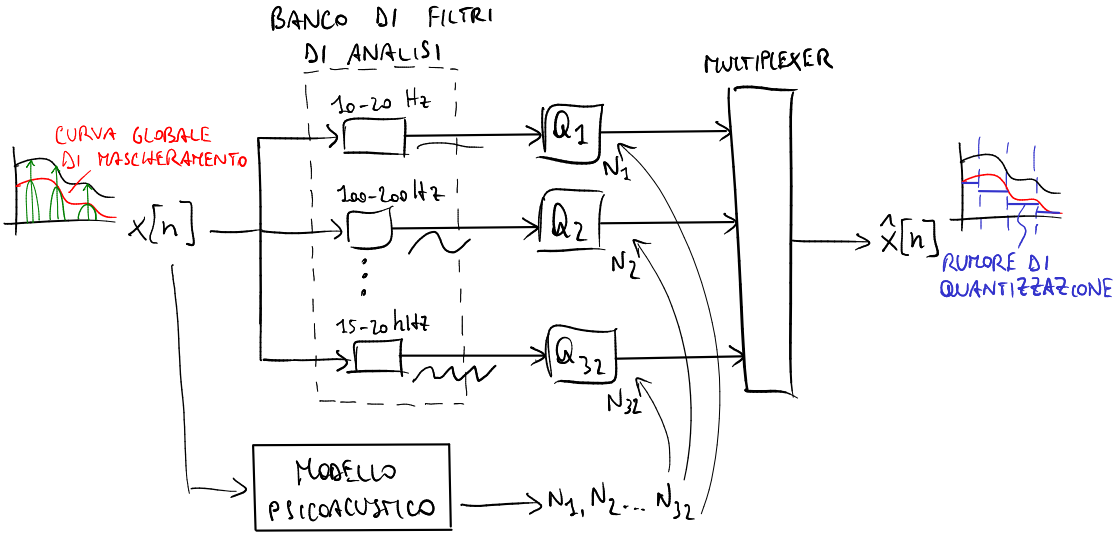


Figura 7.1: Codificatore audio con quantizzatore a tratti.

Tipicamente il rumore prodotto da un quantizzatore è a intensità costante a tutte le frequenze (per questo motivo è detto **pavimento di rumore** o noise floor), ma la curva globale di mascheramento può essere assai variabile ⇒ il pavimento di rumore non può superare la minima intensità della curva globale di mascheramento, anche alle frequenze dove potrebbe essere più alto. Un'ottimizzazione per ovviare a questo problema è approssimare la curva di mascheramento a gradini massimizzando il rumore a tratti:

1. si suddivide lo spettro del segnale in sottobande (idealmente le 32 bande critiche) tramite un **banco di filtri di analisi**;

2. in ogni sottobanda, si quantizza in modo indipendente attraverso un quantizzatore uniforme con il numero di bit definito dal **modello psicoacustico**:

   - se i quantizzatori uniformi possono usare sempre i numeri di bit indicati dal modello psicoacustico, cioè i numeri di bit minimi per mantenere il rumore al di sotto della curva globale di mascheramento, allora il segnale ricostruito sarà percettivamente trasparente perché il rumore iniettato è sempre non udibile;
   - se i quantizzatori non possono usare i numeri di bit del modello psicoacustico perché l'utente specifica un bit rate inferiore, il rumore supera la curva globale di mascheramento diventando udibile e il suono sarà percepito come distorto.

7.2 Standard MPEG

Negli anni ‘80, venne costituito il gruppo ISO Moving Picture Experts Group per lo sviluppo di standard multimediali digitali.

7.2.1 MPEG-1

Il primo standard, MPEG-1, venne sviluppato con l’obiettivo di rimpiazzare il VHS, e rese possibile codificare un flusso audio e video all’interno di un CD audio (bit rate: 1,4 Mb/s):

- parte di audio: 128 kb/s
- parte di video: 1 Mb/s

MPEG-1 Audio usa la codifica percettiva con filtri di uguale banda per semplicità di calcolo, e offre tre layer:

- layer I: complessità minima, rapporto di compressione minimo
- layer II
- layer III (MP3): complessità massima, rapporto di compressione massimo (fattore 10), bit rate fino a 320 kb/s

7.2.2 MPEG-2

MPEG-2 riuscì ad aumentare la qualità del video grazie all’impiego di un supporto più grande, il DVD (bit rate: 6 Mb/s).