UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

Le comunicazioni voce

A.Carini – Elettronica per l'audio e l'acustica

Quantizzazione

- Durante la conversione analogico-digitale i campioni vanno quantizzati.
- Il processo riduce anche la quantità di informazione da memorizzare.
- Ridurre la quantità di bit con cui rappresentiamo i nostri campioni porta diversi benefici: riduce il costo dei dispositivi usati per la loro memorizzazione e la larghezza di banda del canale usato per la loro trasmissione.
- L'operazione è detta *compressione*.
- Il *rapporto di compressione* è il rapporto tra la dimensione originale del segnale digitale e quella del segnale compresso.
- C'è in genere un compromesso tra grado di compressione e la qualità (o l'intelligibilità) del segnale compresso.
- Mediante opportune tecniche è possibile ottenere entrambi i risultati a spese di una maggiore complessità computazionale.

Pulse code modulation (PCM)

- Il formato PCM è quello fornito dalla maggior parte dei convertitori analogico-digitali e quello usato per la rappresentazione del segnale audio nei computer.
- Il suono viene memorizzato nella forma di un vettore di campioni, con ciascun campione memorizzato mediante un valore (in genere a 16 bit).

- La differenza tra segnale originale e segnale quantizzato viene detta errore di quantizzazione.
- Tale errore può essere modellato come un rumore a distribuzione uniforme sovrapposto al segnale.

Pulse code modulation (PCM)

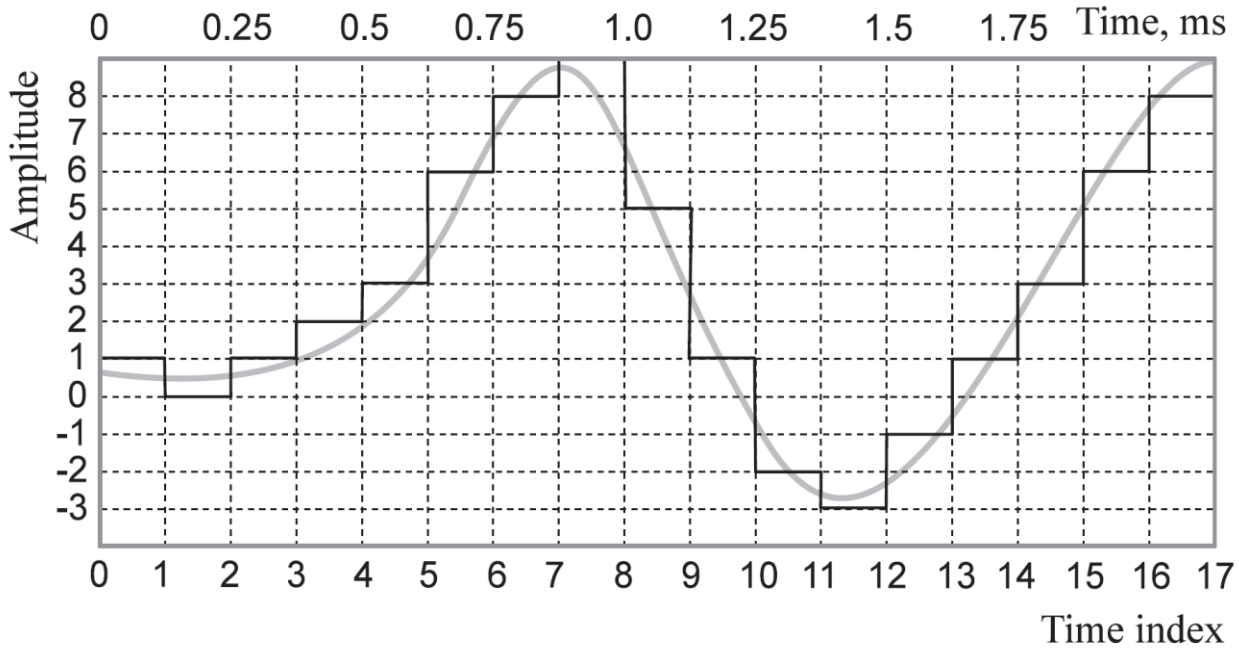


Figure 6.1 Illustration of an audio waveform being pulse code modulated by quantising the amplitude of the signal to the nearest integer at sampling points spaced regularly in time.

Delta modulation

- Il delta si riferisce alla differenza tra un campione e il successivo.
- In questo schema di codifica, la differenza è limitata a +1 o -1.
- Il sistema mantiene un accumulatore che parte da 0. Ad ogni campione, questo accumulatore viene aumentato di 1 o diminuito di 1, a seconda della migliore approssimazione.
- Non è possibile mantenere il valore costante.
- La decisione se aumentare o diminuire viene fatta confrontando il valore dell'accumulatore con l'ampiezza della forma d'onda allo stesso istante.
- Si usano in genere frequenze ben maggiori a quelle usate per l'audio PCM, anche 16-20 volte maggiori.
- L'errore di quantizzazione dipende dal passo di quantizzazione. Se piccolo errore minimo ma sorgono problemi di *slew rate*.

Delta modulation

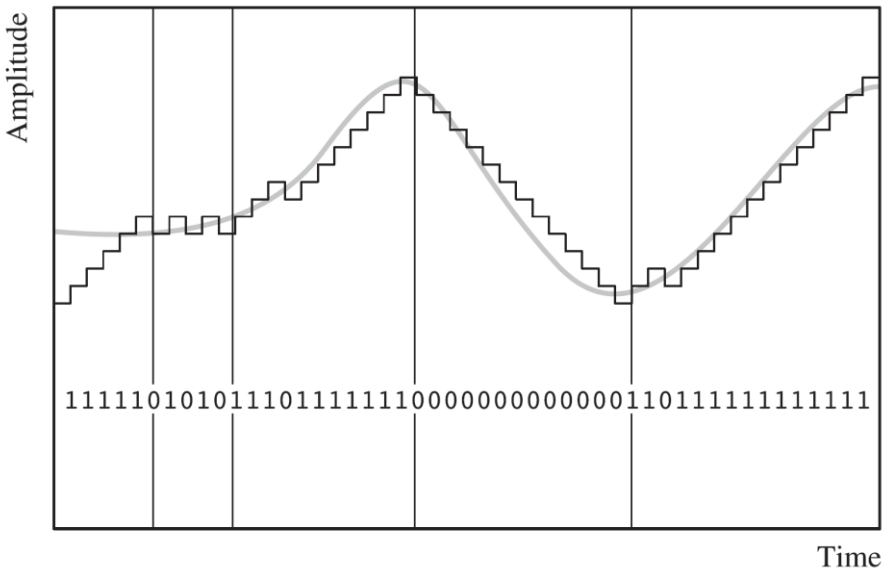


Figure 6.2 Illustration of an audio waveform being represented using delta modulation, showing the encoded vector below the analogue and quantised waveforms. In the encoded vector, a 1 indicates a stepwise increase in amplitude while a 0 indicates a stepwise decrease in amplitude. There is one element in the vector at each sampling instant.

Adaptive delta modulation

- Mantiene i benefici e la semplicità della modulazione delta e supera le limitazioni dello slew rate aumentando l'ampiezza dei gradini sulla base della storia passata del segnale.
- Vengono usati dei livelli di quantizzazione piccoli se il segnale varia lentamente, grandi quando varia più velocemente.
- Ad es. con regole di questo tipo:
  - «Se gli ultimi *n* valori erano uguali allora raddoppia lo step-size, altrimenti dimezzalo».

Adaptive delta modulation

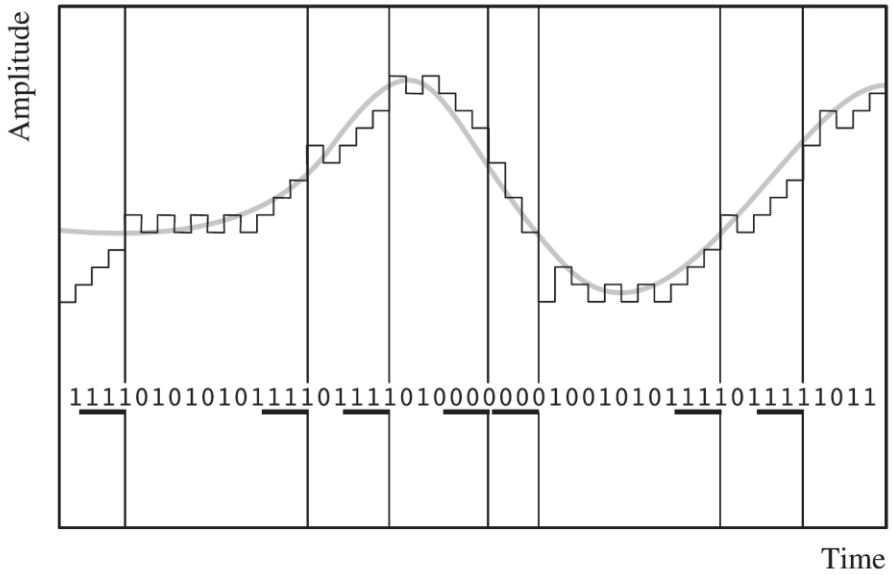


Figure 6.3 Illustration of an audio waveform being represented using adaptive delta modulation, where a 1 indicates a stepwise increase in signal amplitude and a 0 indicates a stepwise decrease in signal amplitude. Repetition of three similar sample values triggers a doubling of stepsize, otherwise the stepsize is halved, until it reaches the minimum (which is the predominant stepsize used throughout).

Adaptive differential pulse code modulation (ADPCM)

- L'ADPCM applica la strategia dell'adaptive delta modulation ai campioni PCM.
- «Differential» indica che viene calcolata la differenza tra i campioni a istanti diversi, «adaptive» indica che lo step-size viene adattato alla storia passata.
- Come nella delta modulation, c'è un accumulatore che parte da zero. Ad ogni nuovo campione si calcola la differenza tra il segnale e il valore accumulato.
- Questa differenza viene quantizzata (e codificata), nonché sommata all'accumulatore.
- Nella codifica DPCM, invece di codificare i singoli campioni si codifica la differenza tra il campione attuale e quello precedente (es.: codifica DECT)
- Nell'ADPCM, viene cambiato il passo di quantizzazione sulla base della storia passata, ed il valore sottratto al segnale è il *valore predetto* dello stesso segnale sulla base della storia passata.

Adaptive differential pulse code modulation (ADPCM)

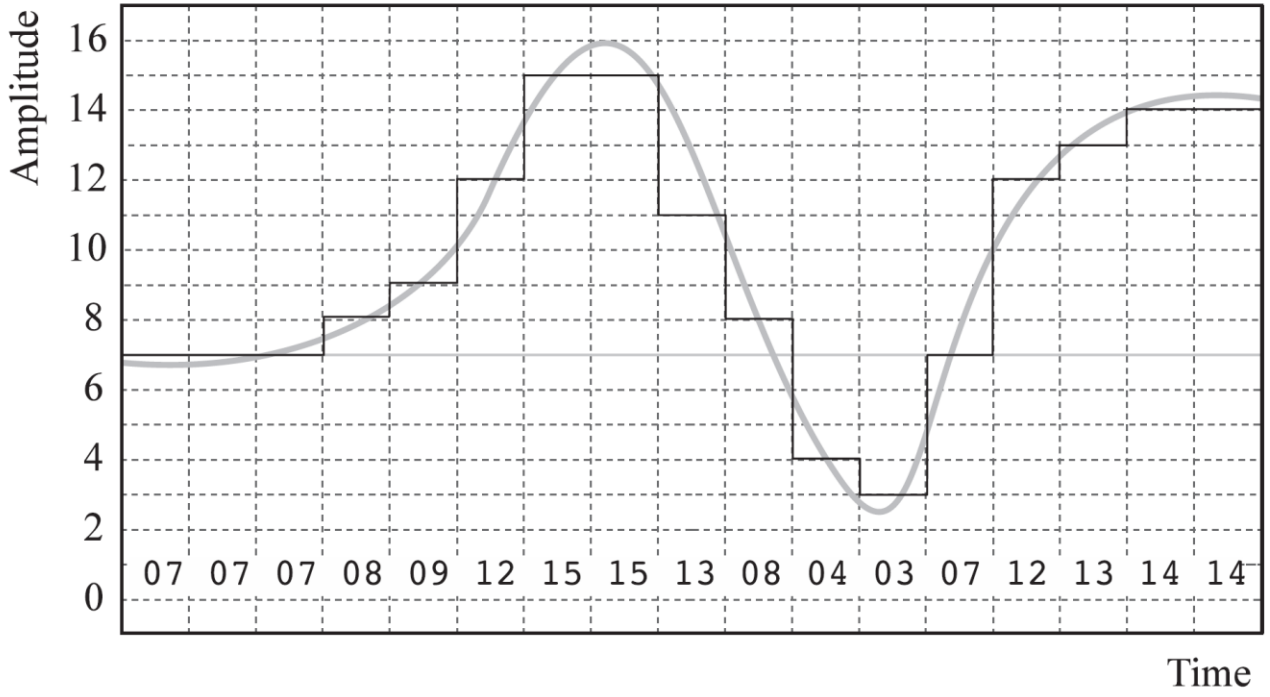


Figure 6.4 Illustration of an audio waveform being quantised to 16 levels of PCM.

Adaptive differential pulse code modulation (ADPCM)

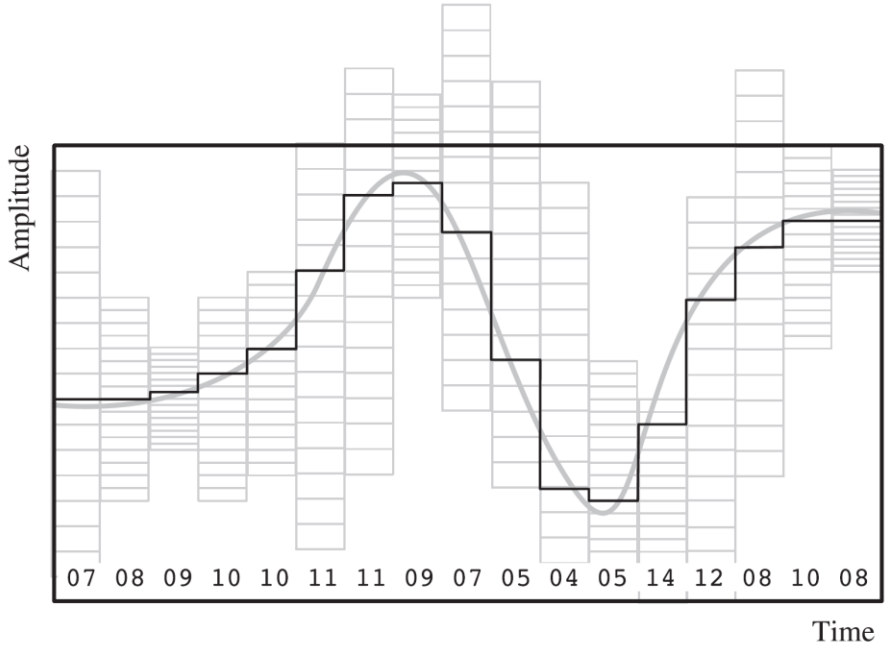


Figure 6.5 Illustration of an audio waveform being quantised to 16 step levels, with the step sizes being adapted based on the previous sample.

Adaptive differential pulse code modulation (ADPCM)

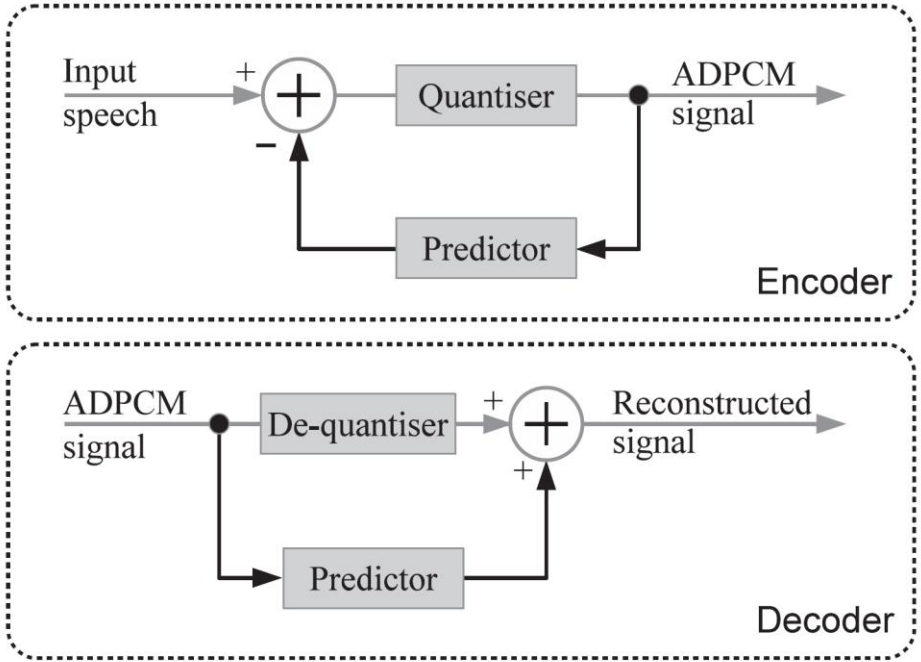
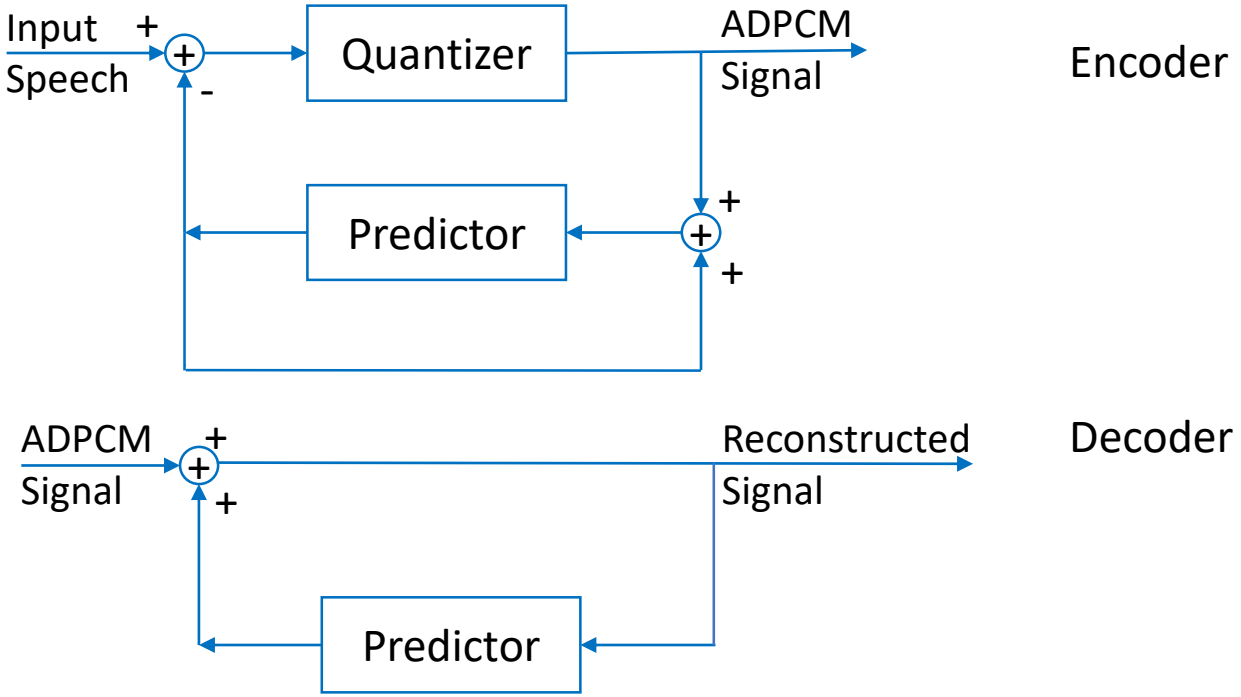


Figure 6.6 A block diagram of an ADPCM encoder (top) and decoder (bottom), showing the predictor in both, fed with identical information (i.e. the ADPCM signal). Both the predictor and the quantiser would usually be adaptive.

Adaptive differential pulse code modulation (ADPCM)



Sub-band ADPCM (SB- ADPCM)

- Introdotta con lo standard ITU G.722.
- Include due stadi di codifica ADPCM separati. Ciascuna unità opera su metà del range di frequenze (da 0--4 kHz e da 4--8 kHz).
- Alla banda bassa viene dedicato 4 volte il numero di bit della banda alta perché è la più importante per la fedeltà del suono.

Name	Description
G.711	8 kHz sampling A-law and μ-law compression
G.721	32 kbits/s ADPCM standard (replaced by G.726)
G.723	24 and 40 kbits/s ADPCM (replaced by G.726)
G.722	64 kbits/s SB-ADPCM sampled at 16 kHz
G.726	24, 32 and 40 kbits/s ADPCM sampled at 8 kHz

Codificatori parametrici

- Sfruttano la conoscenza di come viene prodotta e udita la voce per parametrizzare il segnale vocale in dipendenza al suo contenuto.
- I «parametri» sono valori scelti per rappresentare gli aspetti più importanti della voce, che vanno trasmessi da codificatore a decodificatore, dove sono usati per ricreare una forma d'onda simile ma diversa.
- Il processo che scompone la voce in questi parametri è detto *analisi*, mentre la ricostruzione del decodificatore è detta *sintesi*.

Codificatori parametrici

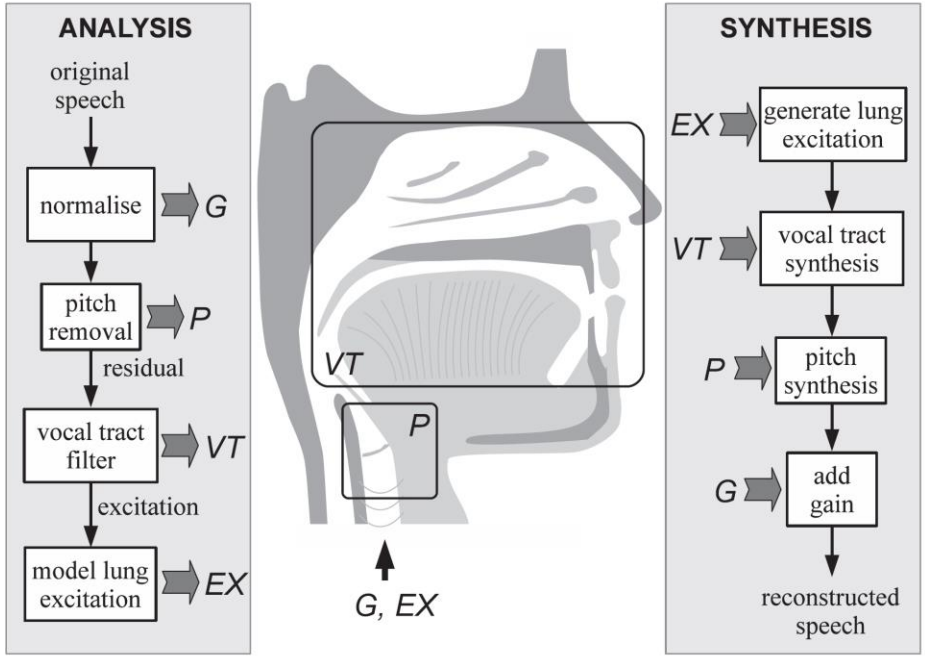


Figure 6.7 Parameterisation of the speech signal into components based loosely upon the human speech production system to convey gain (G), pitch (P), vocal tract filter (VT) and lung excitation (EX) information.

Predizione lineare – Linear predictive coding (LPC)

- E’ presente in gran parte dei codificatori parametrici, sin dal 1976.
- Si basa sulle caratteristiche di quasi-stazionarietà della voce sul breve periodo (~30 ms) dovuta al comportamento dei muscoli umani.
- Mediante un filtro predittore, stimato su piccole finestre di 20-30ms, si ricava un modello dell’azione del tratto vocale sul segnale proveniente dai polmoni.
- L’azione della glottide, che genera i picchi del pitch, è spesso più rapida dei 30 ms. Attraverso una opportuna elaborazione, si rimuove l’effetto del pitch ottenendo un segnale a minore energia detto *residuo*.
- La quasi-stazionarietà della voce su 30ms implica che, con un frequenza di campionamento di 8kHz, 240 campioni avranno una statistica simile.
- Rimosso il pitch, questi campioni possono essere parametrizzati con un insieme limitato di parametri: 8-10 coefficienti del filtro di predizione più una stima dell’eccitazione dei polmoni.

Filtro predittore

- Minimizza secondo qualche criterio l'errore:

$$e(n) = x(n) - p(n) = x(n) - \sum_{p=1}^{P}(-a[p])x(n-p)$$

- E' un filtro FIR con funzione di trasferimento *A(z)* data da

$$E(z) = A(z)X(z) = \left(1 + \sum_{p=1}^{P} a[p]z^{-p}\right) X(z)$$

- Dato il segnale *e(n)* possiamo ricostruire il segnale *x(n)* con un filtro IIR:

$$X(z) = \frac{E(z)}{A(z)} = \frac{1}{1 + \sum_{p=1}^{P} a[p]z^{-p}} E(z)$$

$$y(n) = e(n) - \sum_{p=1}^{P} a[p]\, y(n-p)$$

Predizione lineare – Linear predictive coding (LPC)

- I coefficienti del filtro di predizione lineare definiscono un filtro IIR che opportunamente eccitato ricostruisce le caratteristiche del segnale originale.
- Il segnale prodotto non sarà identico all'originale se visto nel dominio del tempo, ma il suo spettro d'ampiezza sarà approssimativamente lo stesso e l'effetto sull'ascoltatore sarà lo stesso.
- Nei moderni algoritmi di codifica, la tecnica LPC viene in genere abbinata a ulteriori tecniche
  - per la codifica dei coefficienti del filtro,
  - l'estrazione e l'elaborazione del pitch,
  - per la rappresentazione dell'eccitazione dei polmoni.

Il filtro LPC

- I coefficienti LPC che modellano un segmento di voce possono essere usati in due modi diversi:
  - In un filtro di *sintesi* per aggiungere le caratteristiche del tratto vocale a un vettore di campioni di eccitazione.
  - In un filtro di *analisi* per rimuovere tali caratteristiche.
- Dati i coefficienti di un filtro predittore di ordine *P, a[1], a[2], …, a[P],* il filtro di sintesi LPC è un filtro IIR a soli poli

$$y(n) = x(n) - \sum_{p=1}^{P} a[p]y(n-p)$$

  - *x(n)* segnale di ingresso che rappresenta l'eccitazione dei polmoni,
  - *y(n)* segnale di uscita con le caratteristiche del tratto vocale aggiunte.

Il filtro LPC

- In Matlab possiamo usare la funzione `lpc()`

```
seg=speech(1:160);
wseg=seg.*hamming(160);
a=lpc(wseg,10);
```

- Da cui otterremo qualcosa di simile a:

```
a=[1;-1.6187;2.3179;-2.9555;2.8862;-2.5331;
2.2299;-1.3271;0.9886;-0.6126;0.2354];
```

Il filtro LPC

- I coefficienti LPC definiscono il denominatore di un filtro IIR a soli poli:

$$H(z) = \frac{1}{A(z)} = 1 \Big/ \left\{ 1 + \sum_{i=1}^{P} a_i z^{-i} \right\}.$$

- Il filtro di analisi è un filtro FIR (a soli zeri) con funzione di trasferimento *A(z)*.

Il filtro LPC

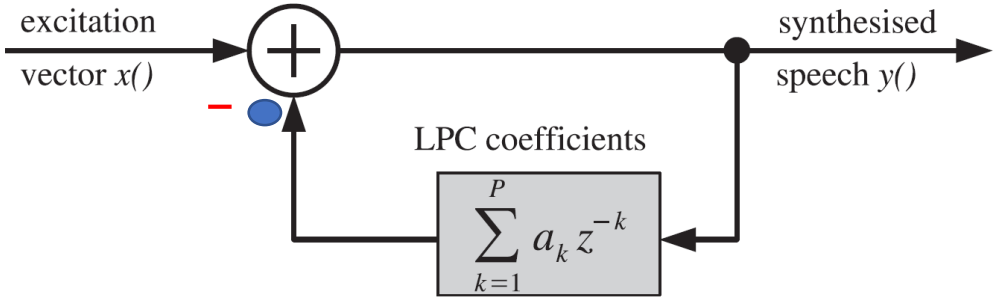


Figure 6.8 Use of LPC coefficients in a synthesis filter.

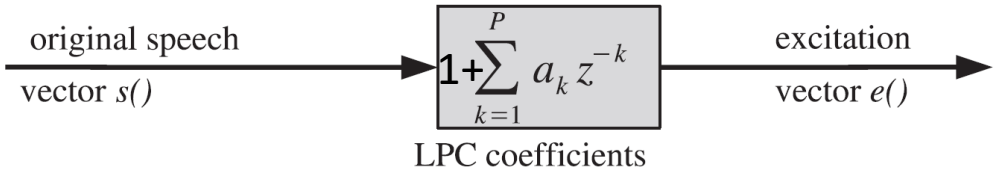
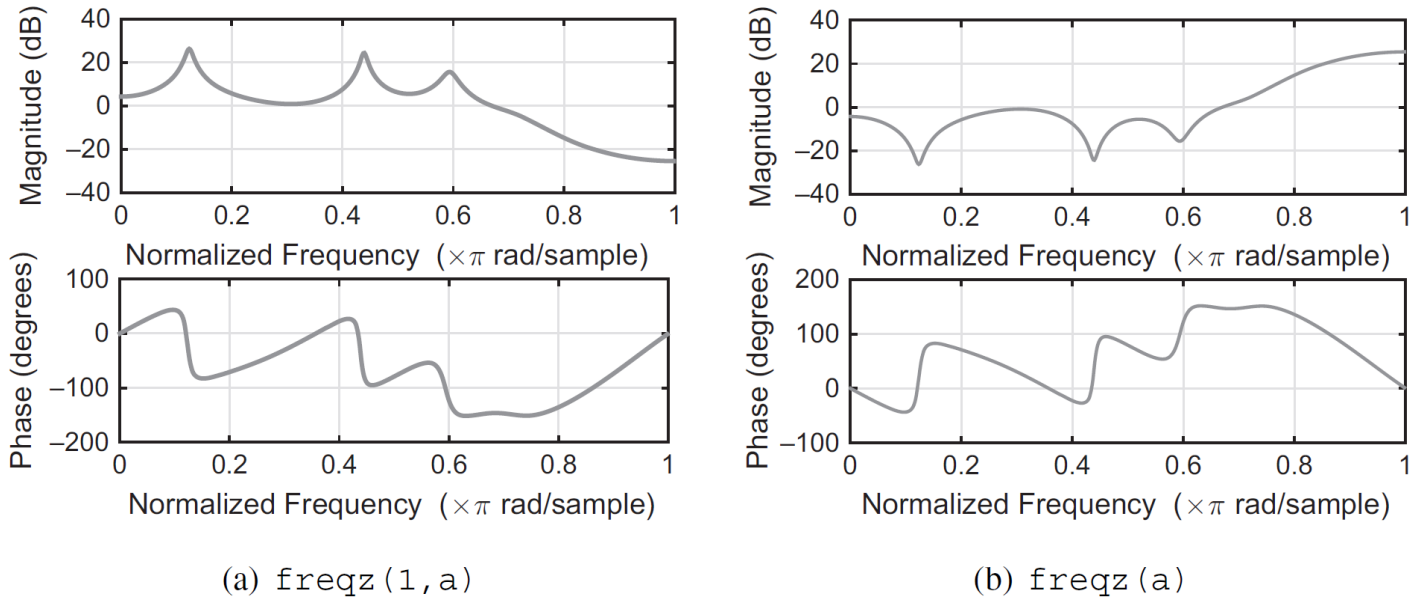


Figure 6.9 Use of LPC coefficients in an analysis filter.

Il filtro LPC

`freqz(1, a);`	`freqz(a);`



(a) `freqz(1,a)` (b) `freqz(a)`

Figure 6.10 Comparison of (a) `freqz(1,a)` and (b) `freqz(a)`.

Il filtro LPC

```
Fs=8000;  %sample rate
N=100;    %frequency resolution
[H, F] = freqz(1,a,N);
%Plot magnitude with a log scale on the y-axis
semilogy(0:Fs/(N-1):Fs,abs(H));
[y,n]=max(abs(H));
PeakF=(n-1)*Fs/(N-1);
```

Il filtro LPC

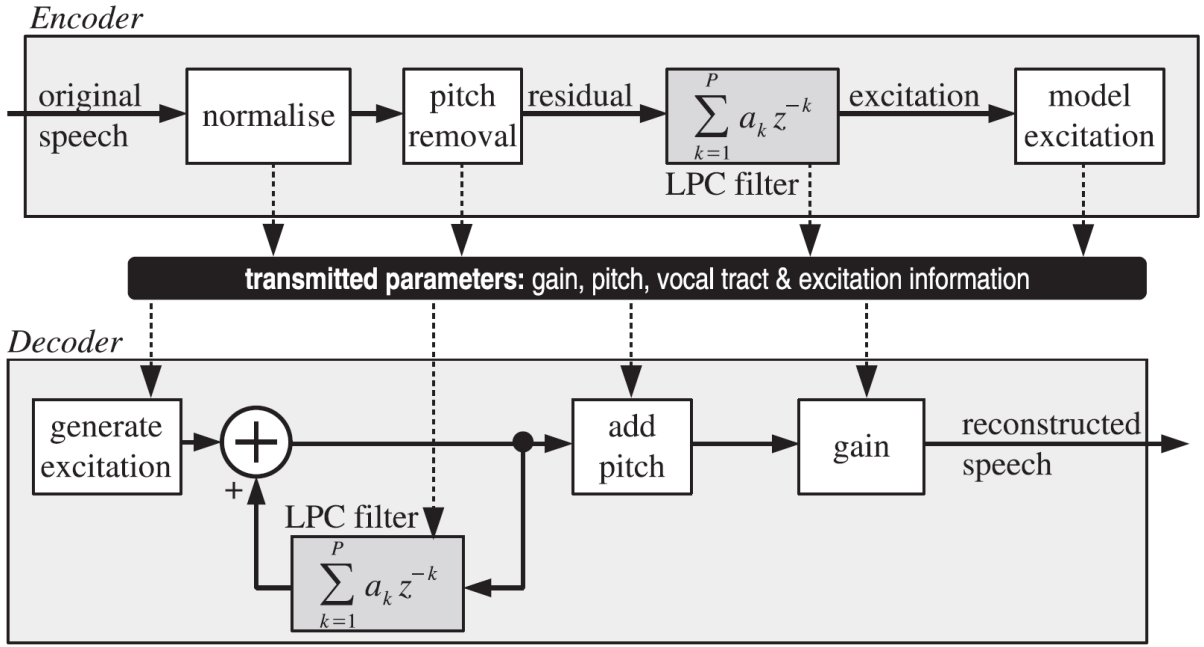


Figure 6.11 The role of LPC analysis and synthesis filters in many speech coding algorithms to convey parameterised vocal tract information from encoder to decoder.

Problemi di stabilità del filtro LPC

- Il processo di quantizzazione dei coefficienti LPC a un ridotto numero di cifre purtroppo produce spesso filtri di sintesi instabili nel decodificatore.
- Questo produce dei segnali la cui ampiezza aumenta incontrollatamente e all'ascolto udiremo dei fastidiosi *click*.
- Per questo motivo, i coefficienti LPC non vengono quantizzati direttamente ma sono convertiti in forme alternative, come i *coefficienti di riflessione*.
- I *log area ratio, LAR,* sono usati nel GSM e corrispondono a una trasformazione logaritmica dei coefficienti di riflessione.
- Le migliori prestazioni sono date dai *line spectral pair* LSP, che pure sono determinati dai coefficienti di riflessione.

Preenfasi del segnale vocale

- Il filtro LPC rappresenta bene lo spettro del tratto vocale alle basse frequenze, ma in modo peggiore alle alte frequenze a causa del roll-off spettrale (dovuto al passaggio della voce dalle labbra all'esterno).
- Possiamo contrastare l'effetto di questo roll-off mediante una pre-enfasi del segnale vocale, mediante un filtro con f.d.t.:

$$(1 - \alpha z^{-1})$$

$$\alpha = 15/16 = 0.9375$$

$$s'(n) = s(n) - 0.9375 \times s(n-1).$$

Preenfasi del segnale vocale

```
% Create emphasis/de-emphasis filter coeffs
h=[1, -0.9375];
% Apply the emphasis filter
es=filter(h, 1, s);
% Apply the de-emphasis filter
ds=filter(1, h, es);
```

Coefficienti di riflessione

- Sono i coefficienti di una diversa realizzazione del filtro LPC (una realizzazione a traliccio).
- In questa realizzazione, il tratto vocale viene modellato con una serie di tubi di uguale lunghezza ma diametro diverso.
- I coefficienti di riflessione quantificano l'energia riflessa da ciascuna giunzione del sistema.
- Sono talvolta chiamati *Partial Correlation (PARCOR) coefficients* dal metodo usato per ottenerli.
- La conversione da PARCOR a LPC è semplice e in genere i coefficienti LPC vengono ottenuti mediante l'analisi PARCOR.

Coefficienti di riflessione

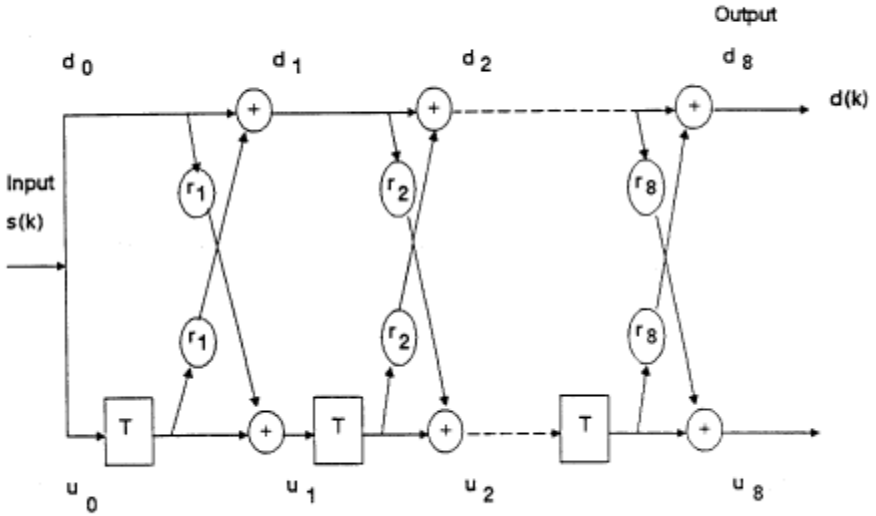


Fig 3.3. Short term analysis filter

Coefficienti di riflessione

Cambia il segno dei coefficienti!

- Consideriamo un segmento vocale di *n* campioni.
- Supponiamo che ciascun campione possa essere stimato dalla combinazione di *P* campioni precedenti:

$$x'[n] = a_1 x[n-1] + a_2 x[n-2] + a_3 x[n-3] + \cdots + a_P x[n-P].$$

- Errore di predizione:

$$e[n] = x[n] - x'[n].$$

- I coefficienti possono essere ottenuti minimizzando il Mean Square Error (MSE) sul segmento

$$E = \sum_n e^2[n] = \sum_n \left\{ x[n] - \sum_{k=1}^{P} a_k x[n-k] \right\}^2.$$

Coefficienti di riflessione

- Imporremo:

$$\frac{\delta E}{\delta a_j} = -2 \sum_n x[n-j] \left\{ x[n] - \sum_{k=1}^{P} a_k x[n-k] \right\} = 0.$$

- Da cui

$$\sum_{k=1}^{P} a_k \sum_n x[n-j]x[n-k] = \sum_n x[n]x[n-j],$$

- Ci sono diversi metodi per risolvere questo problema. I principali sono:
  - *Covariance method* – fissa l'intervallo su cui viene calcolato il MSE
  - *Autocorrelation method* – considera una somma infinita nel MSE, assumendo il segnale abbia energia finita (ottenuta con funz. finestra)
- Il primo metodo è più accurato per i piccoli frame ma spesso porta a instabilità.
- Il secondo metodo produce un risultato sempre stabile.

Coefficienti di riflessione

- Se il MSE viene calcolato con una somma infinita

$$\sum_{n=-\infty}^{\infty} x[n-j]x[n-k] \equiv \sum_{n=-\infty}^{\infty} x[n-j+1]x[n-k+1] \equiv \sum_{n=-\infty}^{\infty} x[n]x[n+j-k].$$

- Il sistema viene riformulato nella seguente forma

$$\sum_{k=1}^{P} a_k \sum_{n=-\infty}^{\infty} x[n]x[n+j-k] = \sum_{n=-\infty}^{\infty} x[n]x[n-j].$$

$$R(k) = \sum_{n=-\infty}^{\infty} x[n]x[n+k].$$

Coefficienti di riflessione

- Otteniamo il sistema

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(P-1) \\ R(1) & R(0) & R(1) & \dots & R(P-2) \\ R(2) & R(1) & R(0) & \dots & R(P-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(P-1) & R(P-2) & R(P-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(P) \end{bmatrix}.$$

- Con la matrice di autocorrelazione che è Toeplitz.
- Il sistema può essere risolto in vari modi
  - Invertendo la matrice e ricavando i coefficienti LPC
  - Usando il *metodo di Levinson-Durbin* che calcola i coefficienti di riflessione (di una realizzazione a traliccio), da cui possiamo ottenere i coefficienti LPC con una ricorsione.

Il modello a tubi

- Il modello basato sui coefficienti di riflessione è ispirato al fatto che il tratto vocale può essere interpretato come una serie di tubi di uguale lunghezza ma diametro diverso interconnessi.
- L'insieme di tubi verrà a risuonare a diverse frequenze. Le principali formano le formanti, ma anche quelle più piccole contribuiscono al suono che percepiamo.
- Nella forma più comune, il modello considera 22 tubi interconnessi ed è capace di modellare i più importanti dettagli del tratto vocale umano.
- Un modello esteso a 44 tubi viene usato per arrivare sino ai 20 kHz di banda.
- Mediante la scansione di volontari con MRI (magnetic resonance imaging) mentre pronunciavano diversi fonemi è stato possibile collegare questo modello alla geometria reale del tratto vocale.

Il modello a tubi

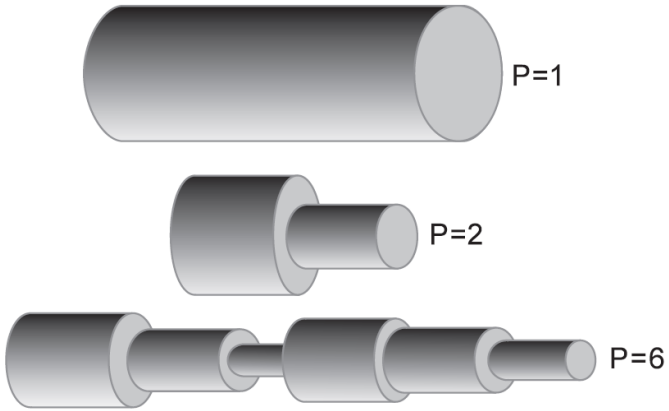


Figure 6.12 First-, second- and sixth-order tube models. Imagine sound entering from one end (the 'glottal' end), resonating through the tubes and exiting at the opposite ('lip' or 'mouth') end.

Il modello a tubi

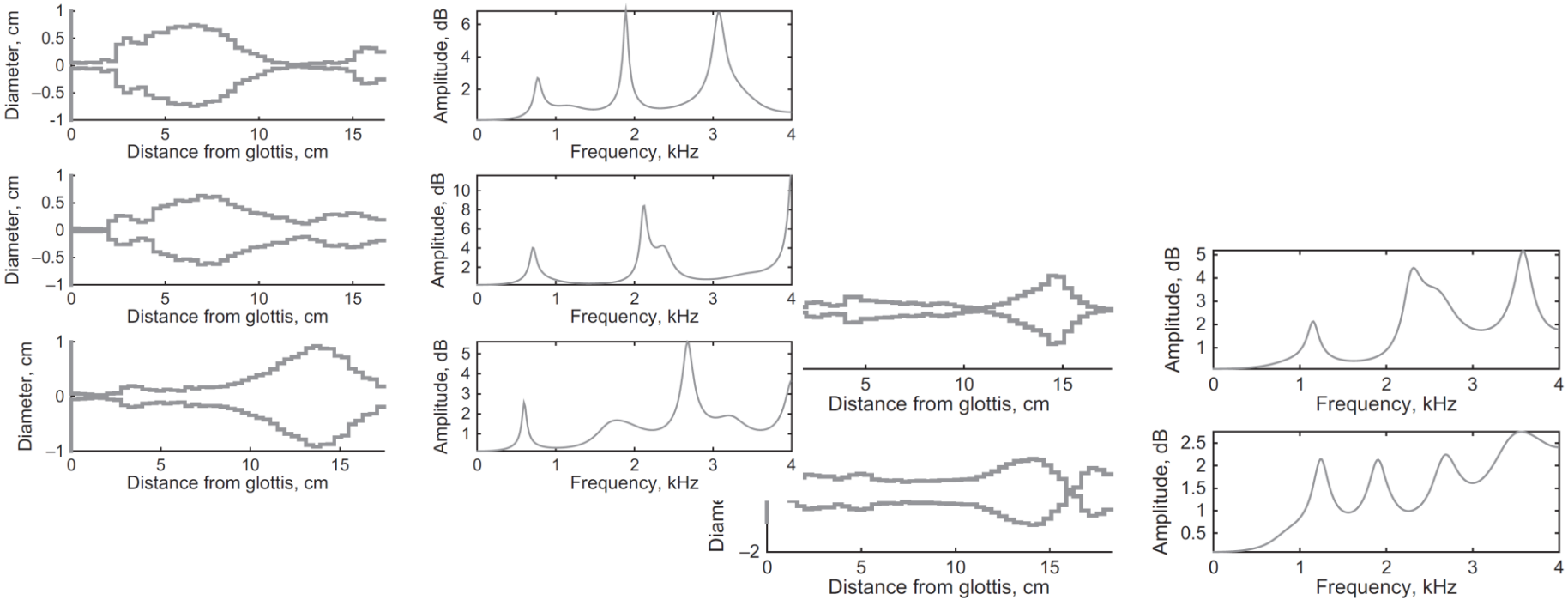


Figure 6.13 Tube cross-sections on the left, and frequency responses on the right for vowels /i/, /ɪ/, /ʌ/, /o/ and consonant /l/ (all in IPA notation).

Line spectral pairs (LSP)

- Le linee spettrali accoppiate LSP consistono in una trasformazione matematica diretta dei parametri LPC, usata in molti sistemi di compressione moderni.
- Le LSP sono divenute popolari per le loro eccellenti caratteristiche di quantizzazione e l'efficienza della rappresentazione.
- Sono anche chiamate «line spectral frequencies».
- Le LSP descrivono le due condizioni limite che possono nascere nel modello a tubi interconnessi del tratto vocale, con il modello che a livello di glottide o è completamente aperto o completamente chiuso.
- Le due condizioni danno origine a due insiemi di frequenze di risonanza, con il numero di risonanze in ciascun insieme dato dal numero di tubi interconnessi.
- Le frequenze di risonanza dei due insiemi si alternano determinano le linee spettrali pari e dispari. Le risonanze reali cadono in mezzo alle coppie di LSP.

Line spectral pairs (LSP)

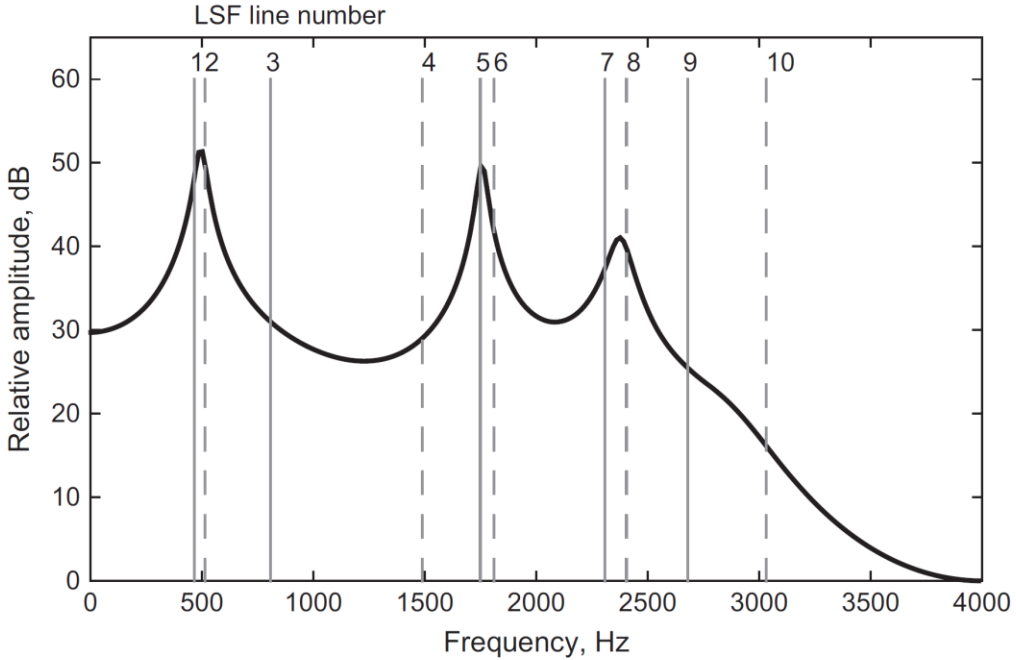


Figure 6.14 Plot of a sample LPC spectrum with the corresponding LSP positions overlaid. Odd lines are drawn solid and even lines are drawn dashed.

Line spectral pairs (LSP)

- Sono ottenute dal filtro LPC che rappresenta le risonanze del tratto vocale:

$$A_p(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_P z^{-P}$$

- Definiremo due filtri *P(z)* e *Q(z)* le cui radici determineranno le LSP. *P(z)* sarà un polinomio antisimmetrico (glottide completamente chiusa), *Q(z)* invece sarà simmetrico (glottide completamente aperta).

- *P(z)* e *Q(z)* hanno ordine *P+1* e sono tali che $A_p(z) = \frac{P(z) + Q(z)}{2}$

- Si ottengono da

$$P(z) = A_p(z) - z^{-(P+1)} A_p(z^{-1}),$$
$$Q(z) = A_p(z) + z^{-(P+1)} A_p(z^{-1}).$$

Line spectral pairs (LSP)

- Se il filtro LPC è stabile (ovvero se *A(z)* ha radici entro il circolo di raggio unitario) si dimostra che le radici di *P(z)* e *Q(z)* cadono tutte sul circolo di raggio unitario e si alternano su tale circolo.
- Qualunque set di radici che si alternano rappresenterà un filtro stabile.
- Le radici possono essere ottenute mediante diversi metodi, ad es., Newton-Raphson, metodi a griglia che cercano l'inversione del segno, etc.
- Trovate le radici, le frequenze LSP sono ottenute risolvendo:

$$\omega_k = \tan^{-1}\left(\frac{\mathrm{Re}\{\theta_k\}}{\mathrm{Im}\{\theta_k\}}\right)$$

- Da queste possiamo ricostruire *P(z), Q(z) (*e quindi *A(z))* con

$$P(z) = (1 - z^{-1}) \prod_{k=2,4,\ldots,P} (1 - 2z^{-1} \cos\ \omega_k + z^{-2}),$$

$$Q(z) = (1 + z^{-1}) \prod_{k=1,3,\ldots,P-1} (1 - 2z^{-1} \cos\ \omega_k + z^{-2}),$$

Line spectral pairs (LSP)

Table 6.1 SEGSNR resulting from different degrees of uniform quantisation of LPC and LSP parameters.

Bits/parameter	LPC	LSP
4	–	−6.26
5	−535	−2.14
6	−303	1.24
7	−6.04	8.28
8	−10.8	15.9
10	19.7	20.5
12	22.2	22.2
16	22.4	22.4

Line spectral pairs (LSP)

- Le LSP sono rappresentabili sia nel dominio della frequenza che del coseno.
- Ciascuna linea può essere quantizzata indipendentemente su una scala uniforme o non, anche adattata dinamicamente, oppure le linee possono essere raggruppate insieme e quantizzate vettorialmente.
- Sia la quantizzazione scalare che vettoriale può essere applicata ai valori LSP diretti o quelli differenziali, dove la differenza può essere tra la posizione attuale e quella nel frame precedente, o tra posizione attuale e quella media.
- A seconda dell'importanza delle LSP possiamo dedicare diversi bit per la loro quantizzazione, spendendo di più ad esempio per le formanti F1 e F2.

Line spectral pairs (LSP)

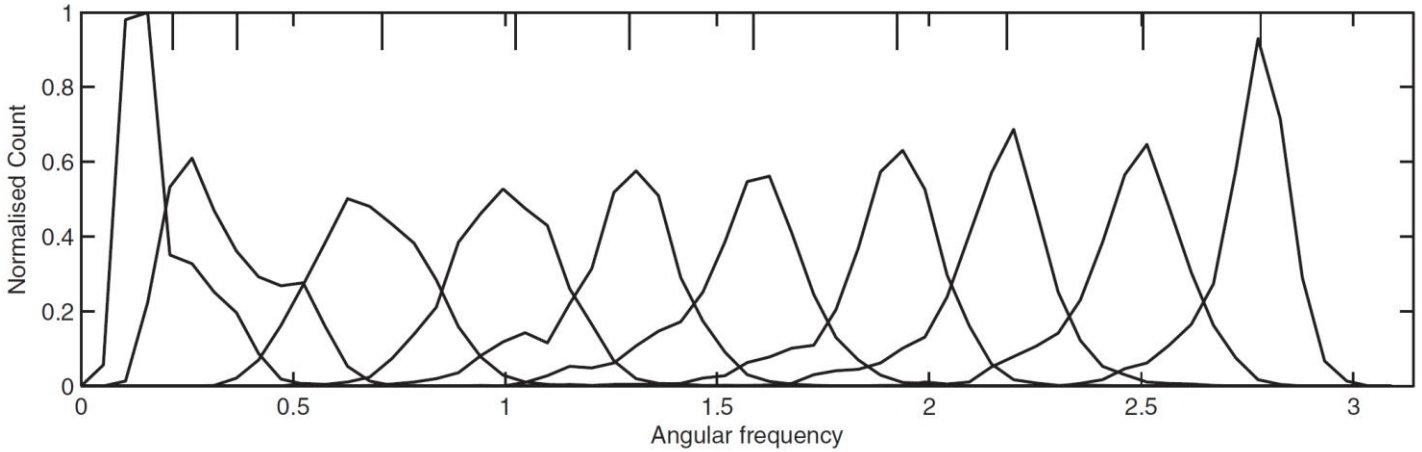


Figure 6.15 A histogram of the relative frequency of each LSP for tenth-order analysis of speech, showing tick marks at the top of the plot corresponding to the mean angular frequency of the ordered lines.

Line spectral pairs (LSP)

Table 6.2 Average frequency, standard deviation and median frequency for ten line frequencies.

No.	Average (Hz)	σ (Hz)	Median (Hz)
1	385	117	364
2	600	184	727
3	896	241	1091
4	1269	272	1455
5	1618	299	1818
6	1962	306	2182
7	2370	284	2545
8	2732	268	2909
9	3120	240	3272
10	3492	156	3636

Line spectral pairs (LSP)

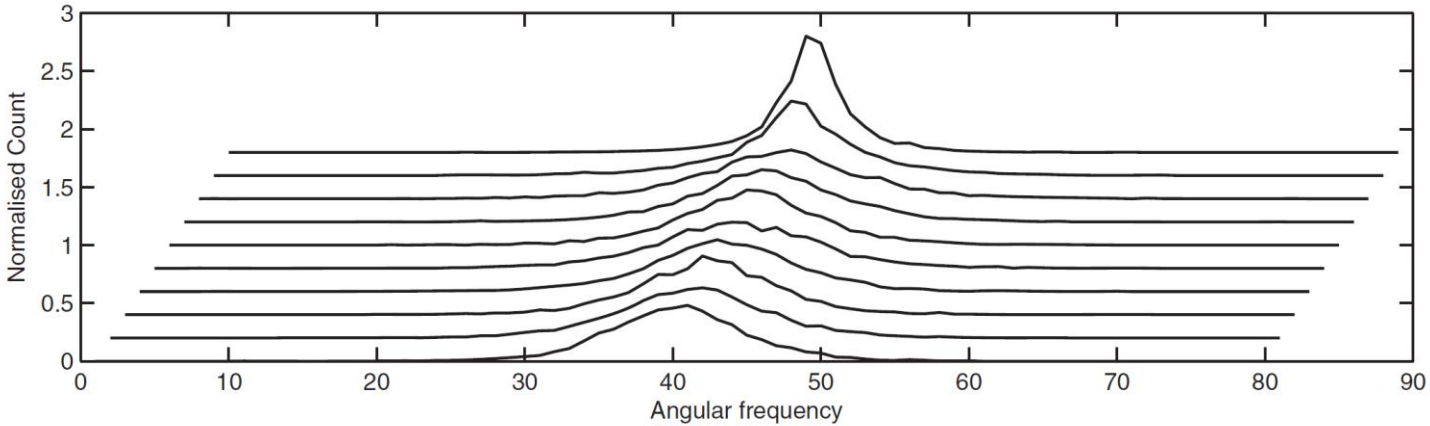


Figure 6.16 A histogram of the relative frequency difference between LSP line locations for consecutive analysis frames, plotted using 80 analysis bins.

Line spectral pairs (LSP)

- Nella quantizzazione vettoriale VQ, un vettore generato da un set di parametri viene confrontato con ciascun elemento di una tabella di vettori (codebook) di uguale lunghezza.
- Viene generalmente calcolata la distanza euclidea.
- L'indice del vettore più vicino al nostro viene scelto e trasmesso al ricevitore.

- La *Split VQ* divide le LSP in sotto-vettori che vengono confrontati con sotto-tabelle, i cui indici saranno inviati al ricevitore.

Modelli per il pitch

- Il *source-filter model* che modella direttamente il processo di generazione della voce è uno dei modelli più utilizzati per la parametrizzazione della voce.
- Il modello utilizza il modello LPC/LSP per il tratto vocale, un rumore aleatorio o similare per l'eccitazione dei polmoni, e un *pitch filter* per ricreare l'effetto della glottide.
- Misure del sistema di produzione del pitch mediante sensori a microonde e raggi X, hanno mostrato che l'azione della glottide non genera toni sinusoidali, ma una sequenza di impulsi.

Modelli per il pitch

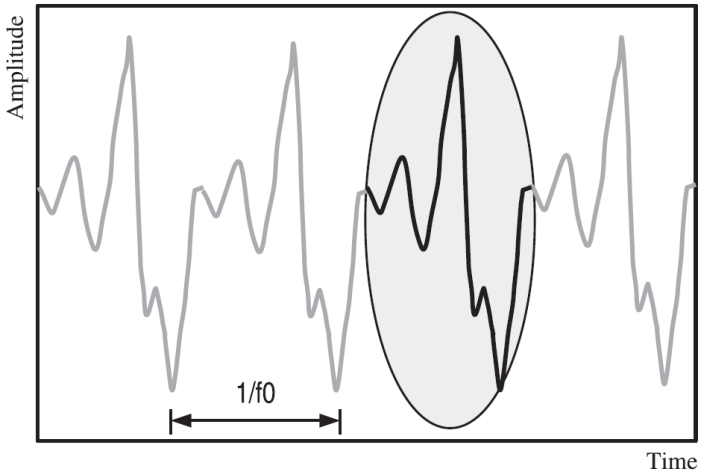


Figure 6.18 A pitch waveform showing several individual pitch pulses. One is circled, and the spacing between pulses, shown, determines the perceived pitch frequency.

Modelli per il pitch

- Ci sono diverse informazioni importanti sul pitch, che sono trattate in modo diverso dai sistemi di compressione voce:
  - La reale forma del pitch,
  - Le ampiezze relative e la posizione dei picchi positivi e negativi del pitch,
  - L'ampiezza maggiore del pitch,
  - La spaziatura degli impulsi.
- I sistemi di compressione più evoluti e di maggiore complessità considerano tutti questi aspetti.
- Alcuni sistemi tengono conto solo degli ultimi tre elementi.
- I codificatori CELP tengono conto solo degli ultimi due elementi e i sistemi RPE solo dell'ultimo.

Regular pulse excitation (RPE)

- L'RPE è una codifica parametrica del pitch della voce.
- E' stato implementato nello standard ETSI 06.10, codifica voce full rate GSM.
- Il codec FR GSM codifica frame di 160 campioni voce a 13 bit e FC = 8 kHz in 260 bit compressi.
- Dato un frame voce, viene fatta pre-enfasi e analisi LPC ottenendo 8 coefficienti che vengono trasformati in LAR.
- Il residuo di predizione viene spezzato in 4 subframes, ciascuna separatamente analizzata per trovare i parametri del pitch.
- La subframe attuale più le tre precedenti ricostruite formano un frame completo su cui viene fatta la *long term prediction (LTP).*
- Tolto il contributo a lungo termine, rimane un insieme di impulsi simili al pitch (se presente) che vengono estratti, codificati ADPCM e trasmessi.
- Se non c'è il pitch, il residuo viene rappresentato come un rumore random.

Regular pulse excitation (RPE)

- Per codificare gli impulsi: il residuo di 40 campioni viene decimato di un fattore 3, ottenendo 3 sequenze da 14, 13, 13 campioni interlacciate. Da queste si ottengono 4 sequenze da 13 campioni interlacciate.
- Viene scelta e codificata ADPCM la sequenza di massima potenza.
- L'intero processo prende il nome RPE-LTP.

Regular pulse excitation (RPE)

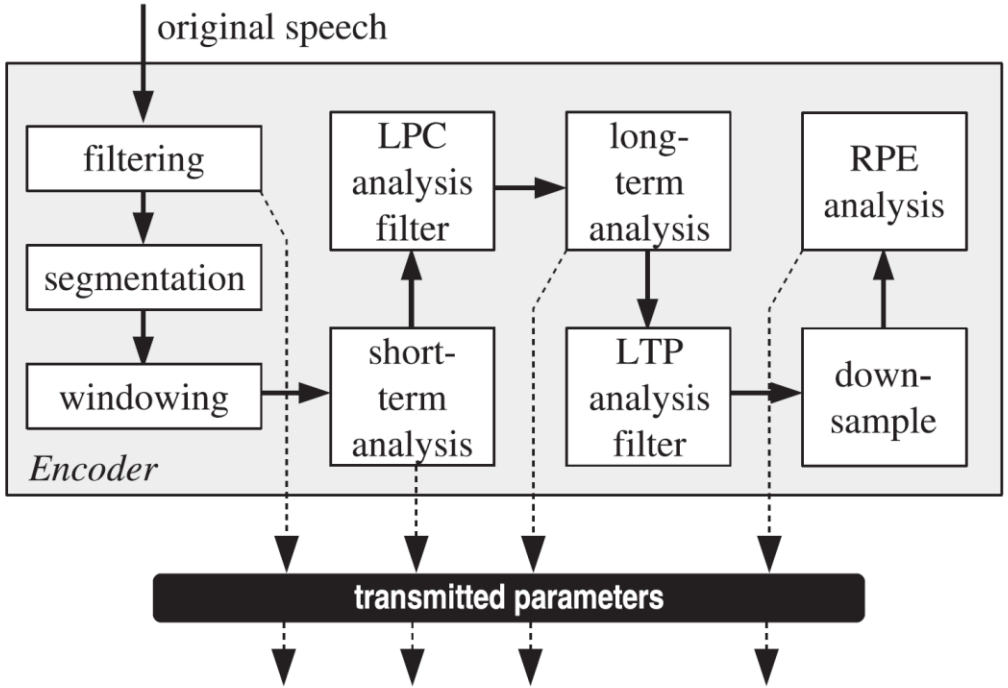


Figure 6.19 GSM RPE-LTP encoder block diagram, showing transmitted parameters.

Regular pulse excitation (RPE)

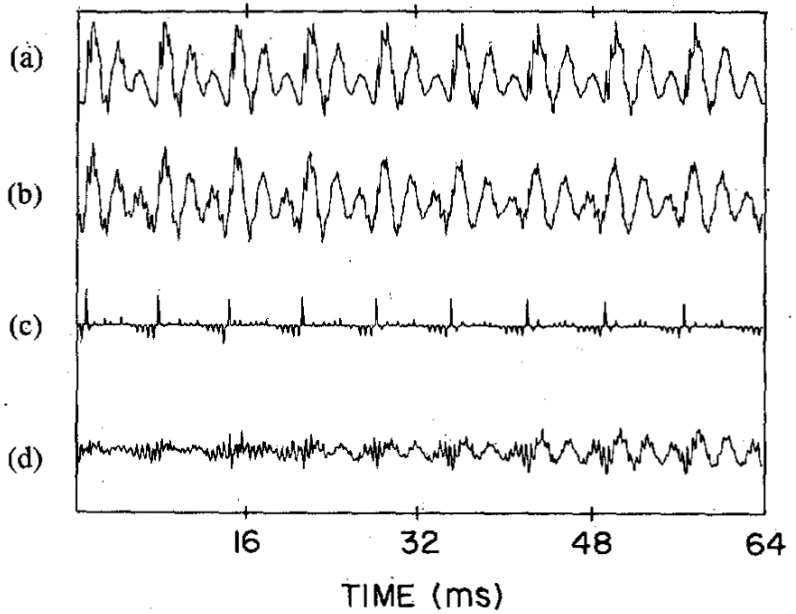


Fig. 5. (a) Speech signal $s(n)$, (b) reconstructed speech signal $\hat{s}(n)$, (c) excitation signal $v(n)$, and (d) difference signal $s(n) - \hat{s}(n)$ in the RPE coding procedure.

Da Kroon et al.

Long Term Prediction

- Dato un vettore di campioni audio *c(n)* (l'eccitazione dei polmoni) possiamo aggiungere una componente pitch secondo il seguente modello

$$x(n) = c(n) + \beta x(n - M).$$

- Dove β scala l'ampiezza e M corrisponde al periodo del pitch.
- In rappresentazioni più complesse:

$$x(n) = c(n) + \beta_1 x(n - M - 1) + \beta_2 x(n - M) + \beta_3 x(n - M + 1).$$

- Il filtro LTP è IIR e calcola il segnale sulla base dei valori passati. *M* può variare da meno di una subframe a più di una subframe (in GSM FR da 40 a 120).
- L'equazione vista consente la sintesi, vediamo come fare l'analisi.

Pitch extraction

- Detto *e(n)* il residuo LPC e *e'(n)* il segnale predetto LTP vogliamo trovare M e β tali da minimizzare il MSE:

$$E(M,\beta) = \sum_{n=0}^{N-1} \{e(n) - e'(n)\}^2,$$

$$E(M,\beta) = \sum_{n=0}^{N-1} \{e(n) - \beta e(n-M)\}^2,$$

- Derivo rispetto a β e impongo l'annullamento della derivata:

$$\frac{\delta E}{\delta \beta} = \sum_{n=0}^{N-1} \{2\beta e^2(n-M) - 2e(n)e(n-M)\} = 0,$$

$$\beta_{\text{optimum}} = \frac{\sum_{n=0}^{N-1} e(n)e(n-M)}{\sum_{n=0}^{N-1} e^2(n-M)}.$$

Pitch extraction

- Sostituendo:

$$E_{\text{optimum}}(M) = \sum_{n=0}^{N-1} e^2(n) - E'_{\text{optimum}}(M).$$

- Dobbiamo massimizzare:

$$E'_{\text{optimum}}(M) = \frac{\left[\sum_{n=0}^{N-1} e(n)e(n-M)\right]^2}{\sum_{n=0}^{N-1} e^2(n-M)}.$$

- Trovato M che lo massimizza, avremo

$$\beta = \frac{\sum_{n=0}^{N-1} e(n)e(n-M)}{\sum_{n=0}^{N-1} e^2(n-M)}.$$

Pitch extraction

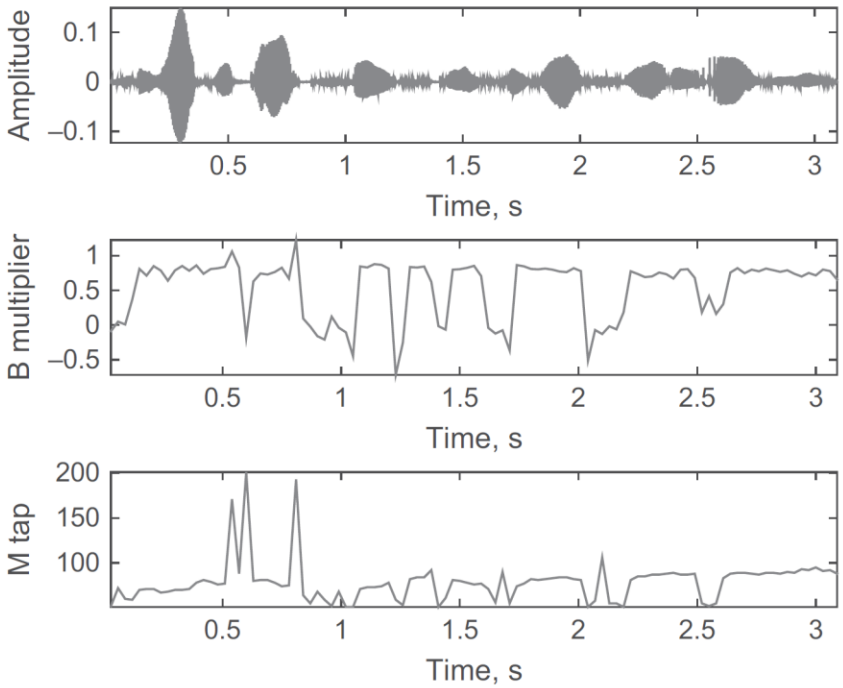


Figure 6.20 An example speech waveform (top) analysed to determine the one-tap pitch multiplier (middle) and pitch lap/lag (bottom).

Analysis by synthesis

- Un frame di campioni voce viene analizzato per estrarne i parametri.
- I parametri quantizzati vengono usati per creare il frame di campioni ricostruiti che viene confrontato con quello originale per valutare *la bontà del risultato*.
- Qualche parte del processo di estrazione dei parametri viene variata per creare un set leggermente diverso, confrontato di nuovo con l'originale.
- Il processo viene ripetuto anche molte centinaia di volte.
- Il miglior set di parametri verrà selezionato e trasmesso al ricevitore.

- Per valutare la «bontà del risultato» non è sufficiente una distanza euclidea. La maggior parte dei metodi di analisi per sintesi applica qualche metodo percettivo: una pesatura percettiva dell'errore o una misura della distanza spettrale tra i segnali.

Codifica CELP base

- La codifica CELP è molto comune ai nostri giorni e può fornire una eccellente qualità della voce.
- Usa il «source filter model» della voce parametrizzato con un guadagno, un modello del tratto vocale (LPC/LSP), il pitch (modello LPT), l'eccitazione del polmoni.
- CELP sta per *code excited linear prediction* e descrive una grande varietà di algoritmi strutturati in modo simile.

Codifica CELP base

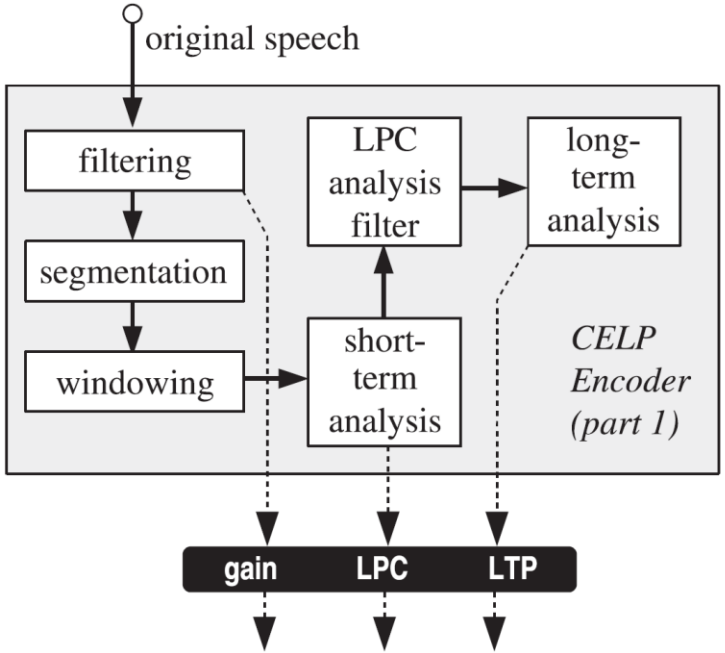


Figure 6.21 A block diagram of part of a CELP encoder, showing original speech being decomposed into gain, LPC and LTP parameters.

Codifica CELP base

- La differenza principale tra CELP e RPE è nel trattamento del segnale di eccitazione dei polmoni.
- Viene usata una grande tabella, *codebook,* di vettori di possibili eccitazioni disponibile sia al codificatore che decodificatore e viene applicato un processo iterativo per identificare quali vettori rappresentino al meglio l'eccitazione dei polmoni.

- Per un codebook tipico, si calcolano fino a 1024 errori.
- L'indice del vettore che risulta nel più piccolo errore viene usato per rappresentare il segnale e viene trasmesso dal codificatore al decodificatore.

Codifica CELP base

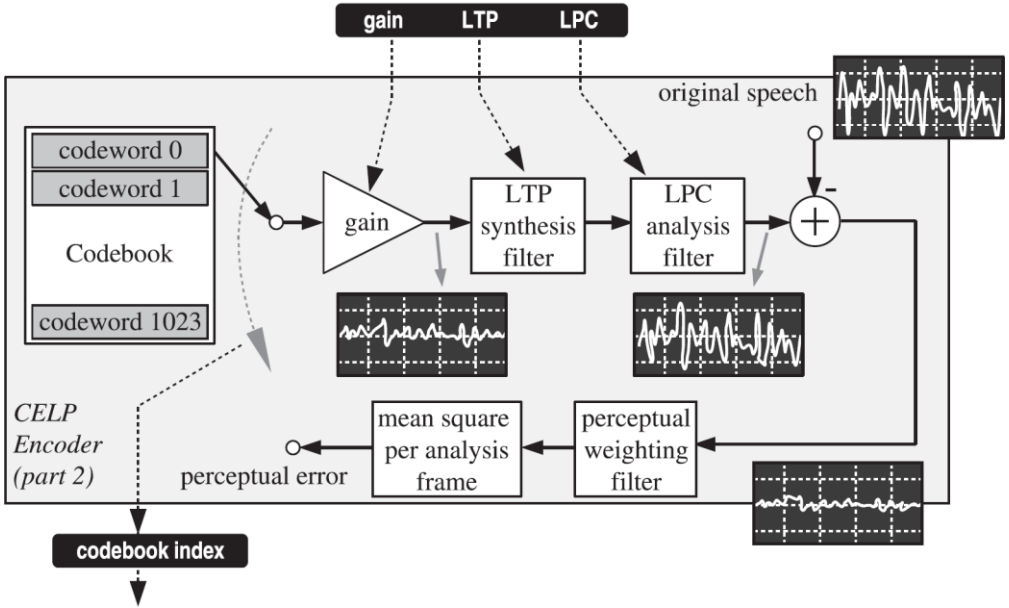


Figure 6.22 A block diagram of the remainder of the CELP encoder. Gain, LPC and LTP parameters were obtained in the first part shown in Figure 6.21, whilst the section now shown is devoted to determining the optimum codebook index that best matches the analysed speech.

Codifica CELP base

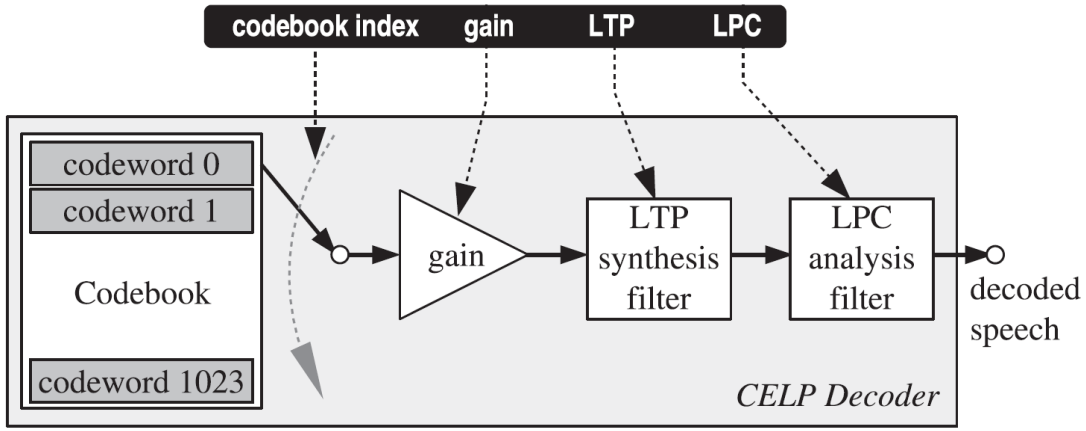


Figure 6.23 A block diagram of the remainder of the CELP decoder utilising codebook index, gain, LPC and LTP parameters to recreate a frame of speech.

ACELP – Algebraic CELP

- La complessità computazionale di un codificatore CELP è molto grande.
- Mentre molti metodi CELP cercano di ridurre il numero di iterazioni (vedasi Split CELP), l'ACELP cerca di ridurre le operazioni richieste da ciascun loop.
- Nella sintesi, dobbiamo amplificare il segnale, eseguire la sintesi LTP e LPC. Ciascuna operazione è lineare e tempo-invariante e quindi può essere eseguita nell'ordine che desideriamo.
- E' meglio fare il prodotto per il guadagno alla fine, e almeno negli ACELP conviene fare il filtraggio LPC per primo.
- Negli ACELP le codeword sono costruite in modo da essere sparse (con tanti 0) e con ciascun elemento che è +1, 0, o -1.
- Le codeword inviate al filtro LPC risultano in un numero limitato di somme e sottrazioni,
- Gli ACELP hanno l'80% dei campioni delle codeword posti a 0.

Split codebook

- Cercano di ridurre lo spazio di ricerca. Di norma questo è l'intero codebook. Riducendo in numero di candidati testati, ridurremmo la complessità.
- Ci sono due metodi possibili.
- Nel primo, il codebook viene ordinato per caratteristiche di approssimazione note. Mediante i processi predittivi possiamo determinare le caratteristiche spettrali o nel tempo del vettore desiderato. Solo quelle codeword che soddisfano queste caratteristiche vengono testate.
- Nel secondo vengono usati due o più codebook con caratteristiche ortogonali. Si cerca dapprima nel codebook 1 con il codebook 2 fissato a un valore arbitrario. Trovata la miglior codeword del codebook 1, si ricerca con quel valore il codebook 2.

Forward-Backward CELP

- La codifica voce con CELP introduce dei ritardi dal momento in cui riceviamo un campione, al momento in cui questo arriva all'ascoltatore.
- Un primo ritardo viene inserito dalla necessità di raccogliere i campioni audio in un buffer prima di poterli processare. Avremo un ritardo per le elaborazioni. Anche il buffer d'uscita potrà introdurre un ritardo simile a quello di ingresso.
- Con i primi CELP erano comuni ritardi di 200-300ms.
- La maggior parte delle persone non nota latenze di 100-150 ms, ma oltre questo limite le latenze sono un disturbo alla conversazione.
- Nella struttura forward-backward, la ricerca viene effettuata in parallelo alla bufferizzazione, usando nella parte di sintesi i parametri del guadagno, LTP e LPC trovati nella precedente frame.

Forward-Backward CELP

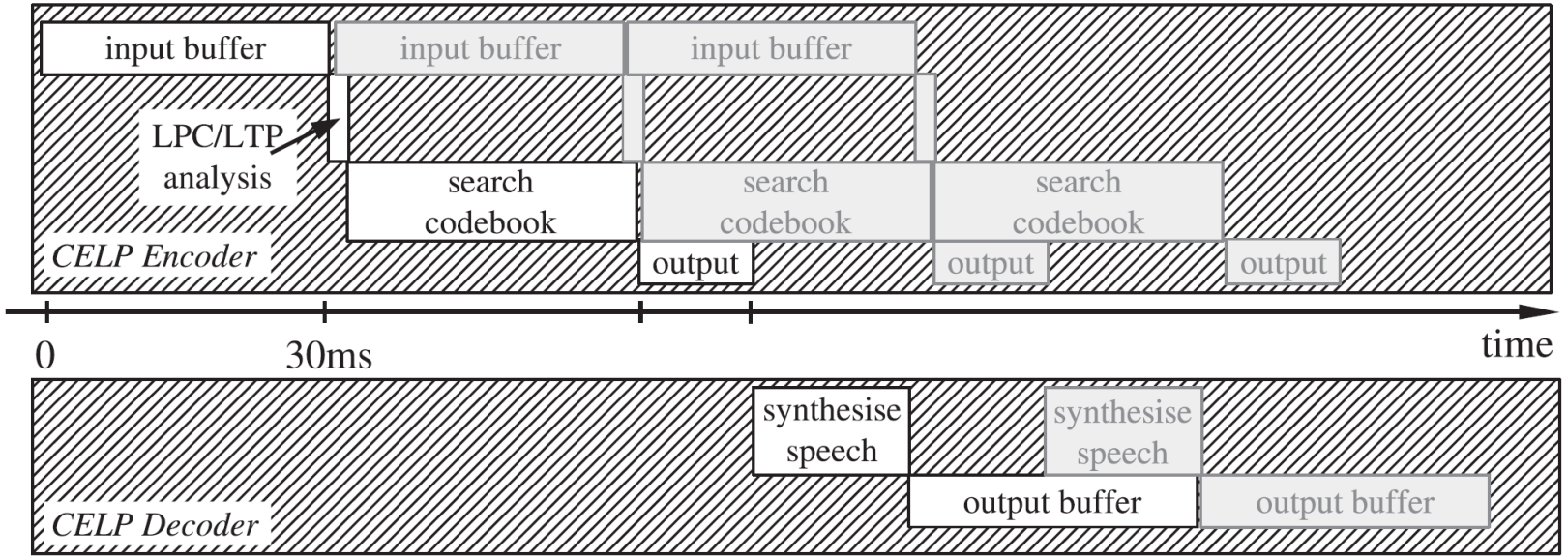


Figure 6.24 Timing diagram for basic CELP encoder and decoder processing, illustrating processing latency.

Forward-Backward CELP

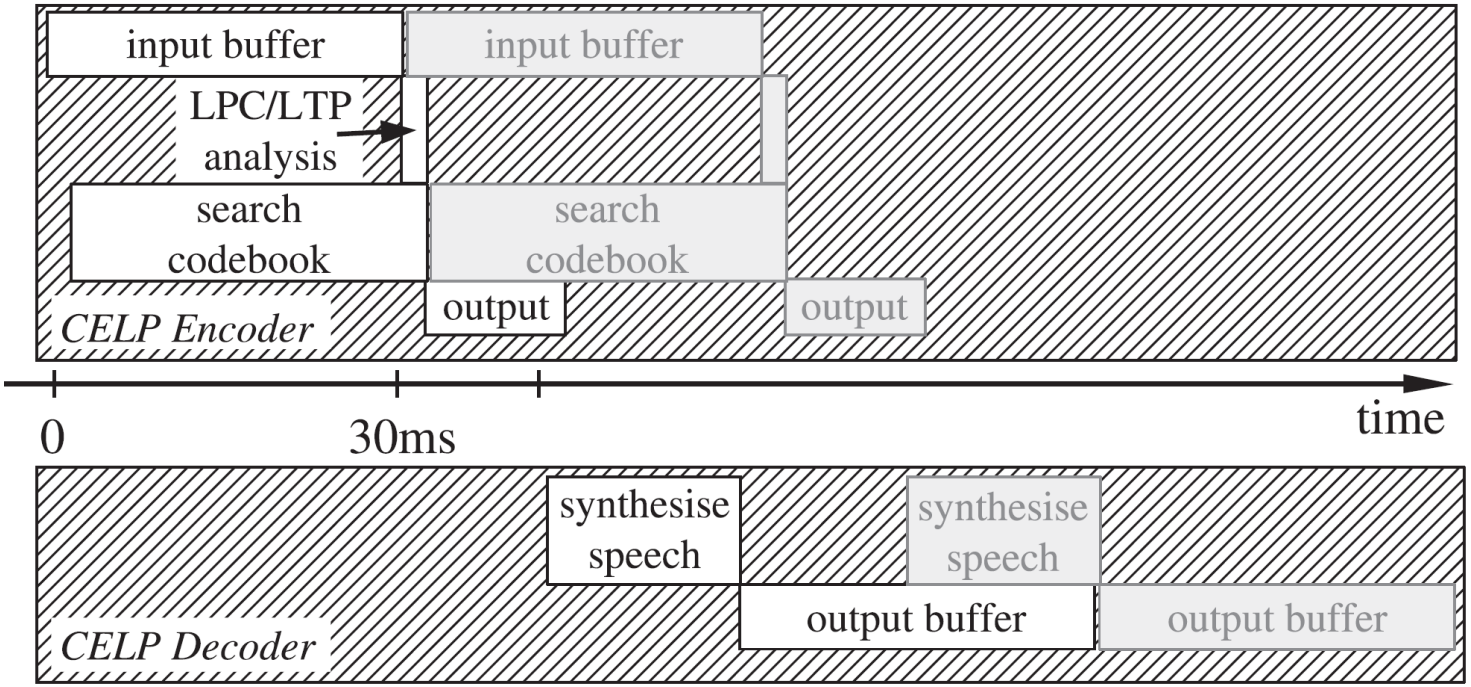


Figure 6.25 Timing diagram for forward–backward CELP encoder and decoder processing.

Perceptual weighting

- Il perceptual error weighting viene usato in molti codificatori voce CELP.
- Normalmente si usano i parametri di predizione lineare LPC per rafforzare le risonanze e quindi per incrementare la potenza delle formanti.
- Siccome la regione delle formanti è la più rilevante per la percezione, il processo di pesatura dà maggiore enfasi alla loro ricostruzione.
- Se il filtro LPC è *H(z)* viene fatto un filtraggio con

$$W(z) = \frac{1 - H(z/\zeta_1)}{1 - H(z/\zeta_2)}. \qquad \text{dove} \qquad H(z/\zeta) = \sum_{k=1}^{P} \zeta^k a_k z^{-k}.$$

- Il filtraggio può essere usato per rafforzare l'importanza dell'errore nella zona delle formanti o anche per migliorare l'intellegibilità della voce.

$$\zeta_1 = 0.95 \text{ and } \zeta_2 = 1.0$$

Vedere:

- Ian Vince McLoughlin, “Speech and Audio Processing”- Cambridge Univesity Press (2016)
  - Cap. 6