Università degli Studi di Parma
Ingegneria delle Telecomunicazioni
Corso di Elaborazione di Segnali Audio e Video
RELAZIONE

Analisi e sintesi di vocali tramite LPC

eseguita da

Alessandro Colazzo
Fabio Patanisi

Indice

1 Segnale vocale - Generalità **5**

1.1 Produzione del segnale vocale 5

1.2 Classificazione e caratteristiche dei suoni 7

1.3 Spettro a breve termine 8

1.4 Suono vocalico 10

1.5 Tipi di codificatori di segnale vocale 12

1.6 Predizione lineare 12

1.6.1 15

1.7 Vocali: estrazione delle formanti 17

Capitolo 1

Segnale vocale - Generalità

1.1 Produzione del segnale vocale

La comunicazione orale fra due persone avviene per mezzo della trasmissione di un segnale vocale dal parlatore all'ascoltatore. Il segnale viene generato dal parlatore tramite il suo *sistema di produzione della voce*, costituito dall'insieme degli *organi di fonazione*, cioè da tutti gli organi che intervengono in qualche maniera nella produzione (lingua, corde vocali, polmoni etc...). Tramite il sistema di produzione, il parlatore induce una variazione nella pressione acustica dell'aria; questa variazione costituisce il segnale vocale propriamente detto. Il segnale si propaga ed è ricevuto dall'ascoltatore tramite il suo *sistema di percezione della voce*, costituito dall'insieme degli organi che intervengono in qualche maniera nell'ascolto (orecchio esterno, timpano etc...).
Il sistema di produzione è rappresentato schematicamente in figura (1.1). Per generare un segnale vocale, la massa d'aria contenuta nei polmoni viene separata dalla massa d'aria esterna, tramite l'occlusione di uno o più punti del collegamento fra le due, e compressa: si crea cosí una differenza di pressione fra le due masse. Successivamente, l'occlusione viene parzialmente aperta, permettendo all'aria contenuta nei polmoni di fluire verso l'esterno e creare una variazione (nel tempo e nello spazio) della pressione acustica.

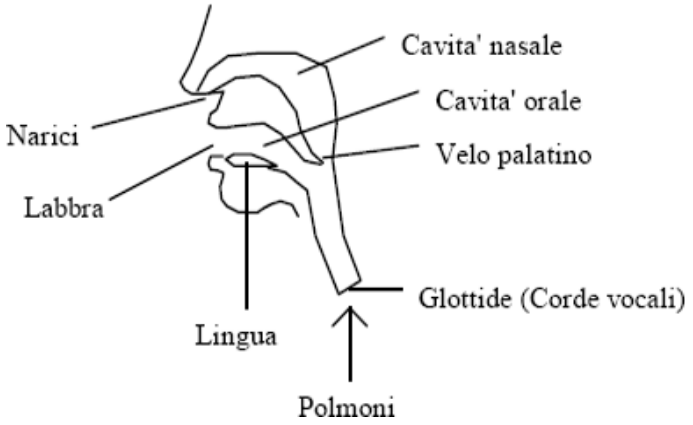

Figura 1.1: **Schema anatomico dell'apparato di produzione della voce**

Nell'occlusione è possibile identificare la *sorgente del suono*, ovvero il punto in cui inizia a variare la pressione acustica. Nelle immediate vicinanze dell'occlusione la pressione acustica sarà una funzione del tempo: questa forma d'onda è detta *segnale di sorgente*.

La sorgente può essere realizzata in primo luogo alle corde vocali (due membrane situate nella glottide) che, se tese si oppongono alla fuoriuscita dell'aria ed entrano in vibrazione. Se invece le corde vocali sono rilassate, l'aria fluisce liberamente attraverso di esse e la sorgente è realizzata occludendo il tratto in qualche altro punto.
Il segnale di sorgente si propaga verso l'esterno ed attraversa il tratto vocale, composto da due cavità, quella orale e quella nasale. Il segnale viene modificato dal passaggio per queste cavità, la cui forma e dimensioni dipendono dalla posizione dei vari organi di fonazione e sono controllate dal parlatore; in particolare, la cavità nasale può essere del tutto esculsa tramite l'innalzamento del velo palatino (e in questo caso il segnale non avrà alcuna componente nasale). Il segnale giunge infine alle labbra ed alle narici e da qui viene irradiato nello spazio.

1.2 Classificazione e caratteristiche dei suoni

Il segnale vocale è il supporto fisico del messaggio linguistico che si forma nella mente di un parlatore. Tale messaggio linguistico è costituito dalla successione di unità fonetiche (fonemi) caratteristici di una lingua data. Nella produzione di voce, il parlatore traduce ogni fonema in una serie di 'comandi articolatori', trasmessi dal cervello agli organi di fonazione, che risultano nella emissione di una successione di *suoni* diversi, opportunamente concatenati a causa dell'inerzia degli articolatori stessi. Ad ogni fonema corrisponde dunque, a livello acustico, un suono dotato di caratteristiche specifiche. Una prima classificazione dei suoni li suddivide in sonori e sordi, a seconda che la sorgente del suono sia costituita o no dalle corde vocali. Un suono sonoro è prodotto quando le corde vocali sono tese, e quindi entrano in vibrazione al passaggio dell'aria, aprendosi e chiudendosi regolarmente. Nel caso dei suoni sonori il segnale di sorgente è una sequenza di forme d'onda elementari (impulsi glottali) ed il segnale vocale ha una struttura pseudo-periodica, il cui periodo (pari alla distanza temporale fra due aperture successive delle corde vocali) è detto *periodo di pitch*. I suoni sonori sono ulteriormente classificati in base all'organo di fonazione che da il maggior contributo alla formazione del suono e/o al tipo di suono prodotto.

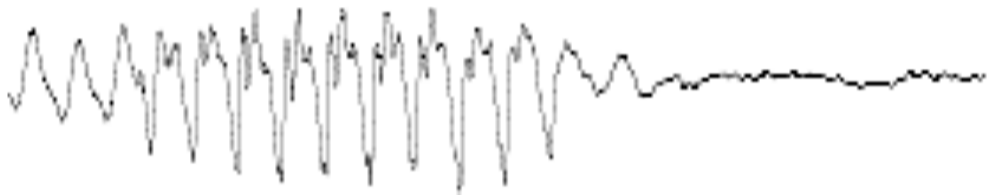

Figura 1.2: **Esempio di segnale vocale**

Un suono sordo è prodotto quando le corde vocali sono aperte, e quindi l'aria fluisce liberamente dai polmoni attraverso la glottide. Nel caso dei suoni sordi la sorgente del suono è una costrizione del tratto vocale, che provoca un moto turbolento e caotico delle molecole del flusso d'aria che la attraversa. Il segnale di sorgente è in questo caso di tipo rumoroso. Anche

i suoni sordi sono ulteriormente classificabili in base all'organo di fonazione che da il maggior contributo alla formazione del suono o in base al modo di articolazione. Le due sorgenti possono essere infine entrambe presenti, e dare luogo a suoni misti (per esempio nella 'g' di 'viaggio' le corde vocali sono tese e vibrano, ma la lingua è schiacciata sul palato e forma una costrizione che provoca la sovrapposizione al segnale pseudo-periodico di una componente di tipo rumore). In figura (1.2) è riportato un esempio di segnale vocale, in cui si distinguono una prima parte sonora ed una seconda sorda.
In definitiva, la frequenza del suono prodotto dipende quindi dalla frequenza di oscillazione delle corde vocali, la quale, a sua volta, dipende dalla loro tensione, dalla loro densità, e dalla loro lunghezza. Nei maschi adulti le corde vocali sono lunghe circa 17-25 mm, mentre nelle femmine circa 12.5-17.5 mm, il che spiega la differenza di tessitura tra maschi e femmine rispettivamente attorno a 125 Hz e 210 Hz.

1.3 Spettro a breve termine

Il segnale vocale non è un segnale stazionario: le caratteristiche del segnale variano rapidamente nel tempo e si modificano a seconda del particolare suono emesso. All'interno di uno stesso suono, il segnale può essere, con buona approssimazione, considerato stazionario per durate di circa 10-30 msec. Su finestre temporali di questa durata ha quindi senso effettuare una analisi spettrale per esaminare come l'energia si distribuisce alle varie frequenze, e si parla di 'Spettro di energia a breve termine'. E' però più comodo pensare che il tratto finestrato sia ottenuto da un segnale stazionario su tutto l'asse dei tempi, con caratteristiche statistiche uguali a quelle del tratto finestrato, e considerare lo spettro di densità di potenza di questo segnale come spettro della finestra. Lo spettro di energia a breve termine si ottiene dalla convoluzione di questo spettro con il modulo quadro della trasformata della finestra. Lo spettro di un segmento di segnale è determinato dallo spettro del segnale di sorgente e dalla conformazione del tratto

vocale. Il tratto vocale agisce sul segnale di sorgente come un filtro, la cui funzione di trasferimento (variabile a seconda della conformazione del tratto) presenta un certo numero di picchi che corrispondono, nel sistema fisico, alle frequenze di risonanza o *formanti* del sistema. Lo spettro del segnale finale è ottenuto dalla moltiplicazione dello spettro del segnale di sorgente con lo spettro (modulo quadro della funzione di trasferimento) del tratto vocale. L'andamento spettrale è nettamente diverso a seconda che il suono sia sonoro o sordo. Per suoni sonori, il segnale di sorgente è periodico (o quasi), quindi il suo spettro è a righe, ed ha un andamento genericamente passabasso (maggiore energia alle basse frequenze); lo spettro è riportato in figura 3a assieme al suo inviluppo. La distanza fra due righe è pari all'inverso del periodo di pitch ed è detta frequenza di pitch. Lo spettro del segnale finale, dopo il passaggio attraverso il tratto vocale, è mostrato in figura 3b assieme al suo inviluppo: si notino i picchi dell'inviluppo in corrispondenza delle frequenze formanti. Lo spettro può avere componenti fino ad una ventina di Khz, ma la maggior parte dell'energia è concentrata nei primi 4 Khz. Per suoni sordi lo spettro del segnale di sorgente non è a righe ma continuo, visto che tale segnale non è periodico. Di solito i suoni sordi presentano maggiore energia alle alte frequenze rispetto ai suoni sonori. Anche in questo caso il tratto vocale può esaltare o attenuare alcune frequenze e lo spettro del segnale emesso assume l'andamento mostrato in figura 3c. Lo spettro può avere componenti fino ad una ventina di Khz, ma la maggior parte dell'energia è concentrata nei primi 4 Khz.

Come è noto, ogni segnale di energia può essere ottenuto come somma (integrale) di un certo numero di sinusoidi con ampiezza e fase opportuna. Lo spettro di energia di un segnale dà informazione sulla ampiezza di queste sinusoidi, ma non sulle loro fasi. Le fasi sono descritte dallo spettro di fase, che è necessario per una ricostruzione esatta della forma d'onda. Una importante proprietà del sistema di percezione umano è che l'orecchio, nella distinzione dei diversi suoni, è poco sensibile allo spettro di fase, mentre è molto sensibile allo spettro di energia. In altre parole segnali con lo stesso spettro

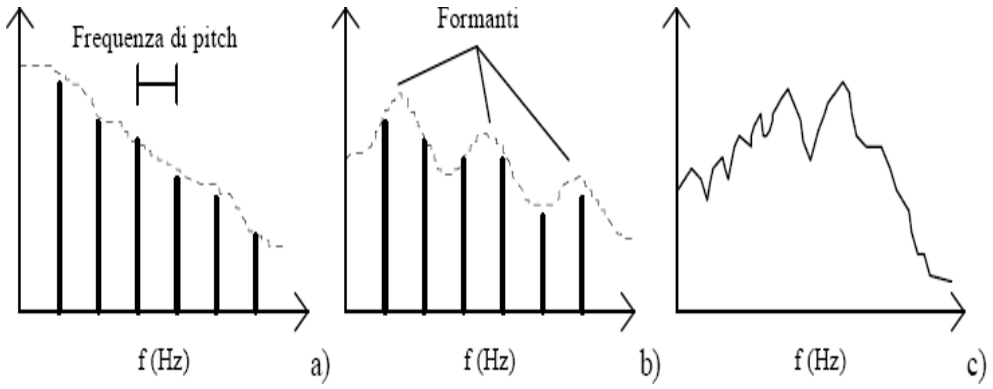

Figura 1.3: **Spettri: a) del segnale di sorgente sonoro, b) di un suono sonoro, c) di un suono sordo**

di energia vengono percepiti come appartenenti allo stesso suono, anche se lo spettro di fase è completamente diverso e quindi le forme d'onda non si assomigliano affatto. Questa proprietà, come vedremo, può essere utilmente sfruttata nella compressione del segnale vocale.

1.4 Suono vocalico

Il suono delle vocali è quello più facile da descrivere in termini fisici, perché è un suono quasi stazionario, prodotto senza che vi compaiano evidenti componenti di rumore, caratteristica tipica invece, ad esempio, dei suoni consonantici.
Cosa distingue una vocale da un'altra? I suoni vocalici hanno colori diversi, ma si tratta di una componente del timbro sonoro che non ha nulla a che fare col timbro di voce del parlante. L'utilità di questo fatto è evidente: ci permette di riconoscere una A da una I quasi indipendentemente dalla particolare qualità di voce del parlante. Ispezionando i sonogrammi (in figura (1.4) è riportato un esempio di sonogramma) è evidente che le diverse vocali abbiano suoni diversi.
Nelle analisi del linguaggio si usa, come già accennato in precedenza, descrivere più efficacemente la differenza tra le vocali in termini delle cosiddette

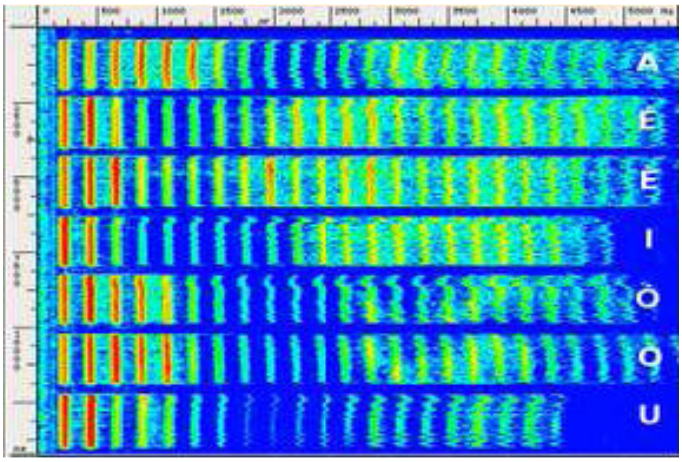

Figura 1.4: **Sonogramma**

formanti. Anziché esaminare ogni singola armonica, si preferisce suddividere lo spettro fino a 5500 Hz (5000 per la voce maschile) in bande di larghezza pari a 1000 Hz, e studiare la posizione del baricentro spettrale in ciascuna banda. Questa posizione indica come è distribuita l'energia sonora in ciascuna banda, e, insieme alla variabile tempo, è sufficiente ad identificare univocamente i diversi fonemi in un sonogramma. L'analisi in bande permette di riferirsi alle stesse grandezze, le formanti appunto, indipendentemente dall'altezza della nota pronunciata.

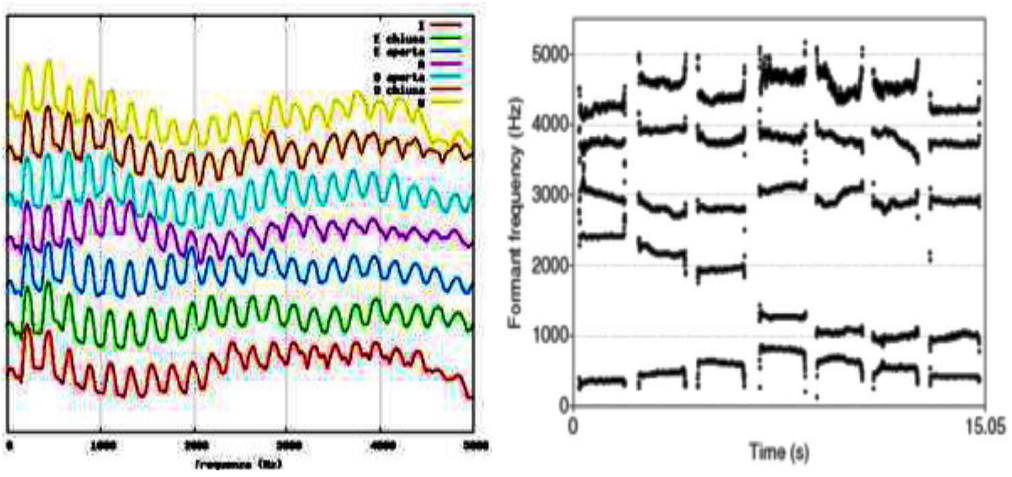

Figura 1.5: **Spettri e formanti**

Nella figura (1.5) è rappresentato un esempio di estrazione delle formanti dal sonogramma a lato. Da notare che gli assi, come vuole la convenzione in fonologia, sono scambiati: l'orizzontale indica il tempo, e il verticale la frequenza suddivisa nelle bande formanti. Nelle seguenti immagini si evidenzia l'utilità pratica di ragionare per formanti: gli spettri di ogni vocale hanno troppi dettagli, per essere confrontati uno ad uno, mentre le formanti danno un'immagine immediata relativa solo alle differenze dovute alla pronuncia. Si noti sempre che, per convenzione l'asse delle frequenze per gli spettri è orizzontale, mentre per le formanti verticale.

1.5 Tipi di codificatori di segnale vocale

I codificatori di segnale vocale si dividono in due principali categorie, waveform (di forma d'onda) e model based. Un codificatore waveform (per esempio un codificatore APC) fa in modo che la forma d'onda del segnale sintetico sia il più simile possibile a quella del segnale originale. Un codificatore model based (per esempio un Vocoder LPC) non si interessa affatto della forma d'onda; estrae invece, dal segnale da codificare, i parametri di un modello dell'apparato di produzione che vengono trasmessi e utilizzati in ricezione per pilotare il modello, e riprodurre un segnale che viene percepito dall'orecchio come simile a quello originale, anche se non esiste nessuna garanzia di somiglianza fra le due forme d'onda. I codificatori model based permettono di ottenere fattori di compressione maggiori, a scapito però di una perdita sulla qualità del segnale sintetico.

1.6 Predizione lineare

I codificatori di gran lunga più usati sono quelli di tipo LPC (Linear Prediction Coder), basati sulla predizione lineare.
L'idea della predizione lineare è di modellizzare un segnale come una combinazione lineare di: a) propri valori passati e b) valori passati e presenti di

un ipotetico ingresso ad un sistema che fornisce in uscita il segnale cercato. Tradotto matematicamente:

$$s[n] = -\sum_{k=1}^{K} a_k s[n-k] + G\sum_{l=0}^{L} b_l u[n-l] \qquad b_0 = 1 \tag{1.1}$$

dove a_k, b_l e il guadagno G sono i parametri dell'ipotetico sistema e $s[n-k]$ sono i valori passati del segnale e $u[n-l]$ sono i campioni del segnale in ingresso.
Nel dominio della frequenza, applicando la trasformata Z alla (1.1), si ottiene:

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1+\sum_{l=1}^{L} b_l z^{-l}}{1+\sum_{k=1}^{K} a_k z^{-k}} \tag{1.2}$$

In altre parole l'equazione (1.1) afferma che un segnale $s[n]$ è predicibile dal suo passato e da un qualche segnale di ingresso attraverso il sistema *H(z)*.
Esistono due casi speciali di interesse:

1. $a_k = 0$ per 1<k<K
2. $b_l = 0$ per 1<l<L

La 1 è conosciuto come *modello tutti-zeri* o modello moving average (MA). Il secondo è conosciuto come *modello tutti-poli* o modello autoregressivo (AR). Il caso generale cioè il *modello poli-zeri* è quindi anche chiamato (ARMA). Noi utilizzeremo il modello tutti-poli, cioè:

$$s[n] = -\sum_{k=1}^{K} a_k s[n-k] + Gu[n] \tag{1.3}$$

che in frequenza è:

$$H(z) = \frac{G}{1+\sum_{k=1}^{K} a_k z^{-k}} \tag{1.4}$$

cioè assumiamo che il segnale di ingresso $u[n]$ è totalmente sconosciuto e quindi il segnale $s[n]$ può essere predetto solo dalla somma dei suoi campioni passati. Indicando con $\tilde{s}[n]$ il valore predetto si ha quindi:

$$\tilde{s}[n] = -\sum_{k=1}^{K} a_k s[n-k] \tag{1.5}$$

La differenza fra il segnale e la sua predizione è detta errore o *residuo* di predizione, e rappresenta la parte di segnale che non può essere predetta dal suo passato:

$$e[n] = s[n] - \tilde{s}[n] = s[n] + \sum_{k=1}^{K} a_k s[n-k] \tag{1.6}$$

Osservando la formula precedente si nota che la sequenza $e[n]$ è ottenuta dalla sequenza $s[n]$ facendola passare in un filtro FIR, dotato di funzione di trasferimento $1+\sum_{k=1}^{K} a_k z^{-k}$. I coefficienti di predizione vengono calcolati in modo che la sequenza predetta $\tilde{s}[n]$ sia il più simile possibile, in senso quadratico, alla sequenza $s[n]$, cioè in modo che l'energia del segnale differenza sia più piccola possibile; analiticamente occorre minimizzare la quantità Err:

$$Err = E\left(e[n]^2\right) = E\left[\left(s[n] + \sum_{k=1}^{K} a_k s[n-k]\right)^2\right] \tag{1.7}$$

che rappresenta l'errore quadratico medio.
Per minimizzare Err si deriva rispetto ai coefficienti $a[k]$ e si eguagliano a zero le derivate:

$$\frac{\partial(Err)}{\partial a_k} = 0 \qquad 1 \leq k \leq K \tag{1.8}$$

si ottiene quindi:

$$-E(s[n]s[n-i]) = \sum_{k=1}^{K} a_k E\left(s[n-k]s[n-i]\right) \tag{1.9}$$

Il minimo errore quadratico è dato quindi da:

$$Err_K = E\left(s[n]^2\right) + \sum_{k=1}^{K} a_k E\left(s[n]s[n-k]\right) \tag{1.10}$$

Se $s[n]$ è un processo stazionario la sua autocorrelazione è definita come:

$$E(s[n-k]s[n-i]) = R(i-k) \tag{1.11}$$

le equazioni (1.9) e (1.10) diventano quindi:

$$-R(i) = \sum_{k=1}^{K} a_k R(i-k) \qquad 1 \leq i \leq K \tag{1.12}$$

$$Err_k = R(0) + \sum_{k=1}^{K} a_k R(k) \tag{1.13}$$

dove le (1.15) sono note come le equazioni di Yule-Walker.
Nel caso in cui $s[n]$ non è un processo stazionario si ha:

$$E\left(s[n-k]s[n-i]\right) = R(n-i, n-k) \tag{1.14}$$

Senza perdita di generalità, si può considerare l'autocorrelazione al tempo $n = 0$ e la (1.14) diventa:

$$-R(0,-i) = \sum_{k=1}^{K} a_k R(-k,-i) \qquad 1 \leq i \leq K \tag{1.15}$$

$$Err_k = R(0,0) + \sum_{k=1}^{K} a_k R(0,k) \tag{1.16}$$

E' importante notare che per alcune classi di processi non stazionari conosciuti come processi localmente stazionari, è ragionevole stimare l'autocorrelazione in un istante di tempo come una media a breve termine; si parla in questo caso di ergodicità locale. La voce è un esempio di processo non stazionario, che però può essere considerato come un processo localmente stazionario.

1.6.1

Si considerino ora due particolari tipi di ingressi $u[n]$: l'impulso deterministico e il rumore bianco stazionario:

a) **Impulso**: Supponendo che all'ingresso del filtro $H(z)$ sia presente un impulso di Dirac, cioè $u[n] = \delta[n]$, l'uscita del filtro è allora la sua risposta impulsiva $h[n]$:

$$h[n] = -\sum_{k=1}^{K} a_k h[n-k] + G\delta[n] \tag{1.17}$$

Moltiplicando la (1.17) per $h[n-i]$ e sommando su tutti gli n si ottiene:

$$-R_h(i) = \sum_{k=1}^{K} a_k R_h(i-k) \qquad i \leq |i| \leq \infty \tag{1.18}$$

e

$$R_h(0) = -\sum_{k=1}^{K} a_k R_h(k) + G^2 \qquad i = 0 \tag{1.19}$$

dove R_h è l'autocorrelazione della risposta impulsiva $h[n]$.
Se si impone la condizione che l'energia totale in $h[n]$ deve essere uguale a quella in $s[n]$, si ottiene:

$$R_h(0) = R_s(0) = R(0) \tag{1.20}$$

Dalla (1.20) e dalla similitudine tra la (1.15) e la (1.18) si può oncludere che:

$$R_h(i) = R_s(i) = R(i) \qquad 0 < i < K \tag{1.21}$$

Si scopre che i primi $K+1$ valori dell'autocorrelazione della risposta impulsiva di $H(z)$ sono identici ai corrispondenti valori di autocorrelazione del segnale. Quindi il problema può essere riformulato nella ricerca di un filtro $H(z)$ tale che i primi $p+1$ valori dell'autocorrelazione della sua risposta impulsiva sono uguali ai primi $p+1$ valori dell'autocorrelazione del segnale.
Dalla (1.10), (1.19) e (1.21) il guadagno risulta essere:

$$G^2 = Err_K = R(0) + \sum_{k=1}^{K} a_k R(k) \tag{1.22}$$

b) **Rumore bianco**: Supponendo in ingresso campioni incorrelati, a media nulla e varianza unitaria, l'uscita del filtro sarà:

$$s_w[n] = -\sum_{k=1}^{K} a_k s_w[n-k] + Gu[n] \tag{1.23}$$

moltiplicando quest'ultima per $s_w[n-i]$, prendendo i valori attesi e notando che $u[n]$ e $s[n]$ sono incorrelati per $i > 0$ si ottengono come

in precedenza le usuali equazioni di Yule-Walker che permettono di trovare i coefficienti di predizione a_k per $1 < k < K$, che rendono minima l'energia del residuo, risolvendo un sistema di K equazioni in K incognite utilizzando le autocorrelazioni. Scritte in forma matriciale e ricorrendo alle equazioni di Yule-Walker la (1.23) si traduce in:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(K-1) \\ R(1) & R(0) & R(1) & \cdots & R(K-2) \\ R(2) & R(1) & R(0) & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R(K-1) & R(K-2) & R(K-3) & \cdots & R(0) \end{bmatrix} \star \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_K \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(K) \end{bmatrix}$$

Quest'ultima oltre ad essere una matrice simmetrica è anche una matrice di Toeplitz, gli elementi su qualsiasi diagonale sono identici.

1.7 Vocali: estrazione delle formanti

L'analisi delle formanti si può affrontare con tecniche diverse. Si possono citare, a tal proposito, la tecnica di selezione del picco spettrale e l'estrazione della radice. Il metodo dell'individuazione dei picchi spettrali e le sue varianti sono state largamente utilizzate poichè dotate di bassa complessità computazionale, tuttavia, esse presentano spesso dei problemi dovuti alla fusione dei picchi spettrali, rendendo complicata la possibilità di distinguere due formanti adiacenti. Il metodo dell'estrazione delle radici, invece, cerca di trovare le posizioni delle radici risolvendo un polinomio ottenuto dai coefficienti di predizione lineare (LP), che ovviamente richiede maggiore calcolo. Tuttavia, l'accuratezza di quest'ultimo metodo può essere difficilmente alta perchè non è sempre chiaro determinare se una radice costituisce una formante o meno. La tecnica da noi utilizzata è quella dell'estrazione delle radici. Il primo passo è stato quello di calcolare i coefficienti LP