Lecture Notes in Speech Production, Speech Coding, and Speech Recognition

Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign

February 17, 2000

Chapter 4

Linear Predictive Coding

4.1 All-Pole Model of the Speech Transfer Function

The speech spectrum $S(z)$ is produced by passing an excitation spectrum, $U(z)$, through an all-pole transfer function $H(z) = G/A(z)$:

$$S(z) = H(z)U(z) = \left(\frac{G}{A(z)}\right) U(z) \tag{4.1}$$

$$A(z) \equiv 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{4.2}$$

If we write the corresponding difference equation, we get

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{4.3}$$

Where the input signal $u(n)$ is

$$u(n) = \begin{cases} \frac{1}{N_0} \sum_r \delta(n - rN_0) & \text{Voiced Speech} \\ \text{Uncorrelated Gaussian Random Noise} & \text{Unvoiced Speech} \end{cases} \tag{4.4}$$

$A(z)$ can be written in terms of either the direct-form coefficients a_k or the roots r_i:

$$A(z) \equiv 1 - \sum_{k=1}^{p} a_k z^{-k} \equiv \prod_{i=1}^{p} (1 - r_i z^{-1}) \tag{4.5}$$

4.2 Normal Equations

Suppose we want to pick an $A(z)$ which "explains" as much as possible of $S(z)$. If we force $A(z)$ to "explain" as much as possible, then the remaining excitation signal, $e(n) = Gu(n)$ will be as "simple" as possible. We can do this by choosing $A(z)$ in order to minimize the energy of $e(n)$:

$$E = \sum_{m=-\infty}^{\infty} e^2(m) = \sum_{m=-\infty}^{\infty} G^2 u^2(m) \tag{4.6}$$

Plugging in equation 4.3, we get

$$E = \sum_{m=-\infty}^{\infty} \left(s(m) - \sum_{k=1}^{p} a_k s(m-k) \right)^2 \tag{4.7}$$

Solving the equation $\partial E/\partial a_k = 0$ yields the *generalized normal equations*:

$$\sum_{m=-\infty}^{\infty} s(m)s(m-i) = \sum_{k=1}^{p} a_k \sum_{m=-\infty}^{\infty} s(m-k)s(m-i), \quad i = 1,\ldots,p \tag{4.8}$$

The generalized normal equations contain infinite sums, which makes it uncomputable. In order to make it computable, we need to truncate it somehow.

4.2.1 Autocorrelation Method

Solving the Normal Equations

In the autocorrelation method, we window the signal:

$$s_n(m) \equiv w(m)s(n+m) \tag{4.9}$$

Then minimize the following criterion:

$$E_n = \sum_{m=-\infty}^{\infty} \left(s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right)^2 \tag{4.10}$$

Solving the equation $\partial E_n/\partial a_k = 0$ yields the autocorrelation normal equations:

$$R_n(i) = \sum_{k=1}^{p} a_k R_n(|i-k|) \tag{4.11}$$

where

$$R_n(i) \equiv \sum_{m=-\infty}^{\infty} s_n(m)s_n(|m-i|) \tag{4.12}$$

This can also be written in matrix form:

$$\bar{r}_p = \bar{\bar{R}}\bar{a} \tag{4.13}$$

$$\bar{r}_p = \begin{bmatrix} R(1) \\ R(2) \\ \ldots \\ R(p) \end{bmatrix}, \quad \bar{\bar{R}} = \begin{bmatrix} R(0) & R(1) & \ldots & R(p-1) \\ R(1) & R(0) & \ldots & R(p-2) \\ \ldots & & & \\ R(p-1) & R(p-2) & \ldots & R(0) \end{bmatrix}, \quad \bar{a} = \begin{bmatrix} a_1 \\ \ldots \\ a_p \end{bmatrix} \tag{4.14}$$

Since the matrix $\bar{\bar{R}}$ is "Toeplitz" (all elements on a given diagonal are the same), it can be inverted efficiently ($\mathcal{O}(p^2)$) using the Levinson-Durbin recursion.

- Advantage of autocorrelation method: the normal equations can be solved using a computationally efficient algorithm.
- Disadvantage of autocorrelation method: if $x(n)$ is actually a steady-state signal, autocorrelation will not find the true spectrum.

Levinson-Durbin Recursion

The normal equations for autocorrelation can be solved by inverting a Toeplitz matrix:

$$\bar{r} = \bar{\bar{R}}\bar{a} \quad \Rightarrow \quad \bar{a} = \bar{\bar{R}}^{-1}\bar{r} \tag{4.15}$$

$$\bar{r}_p = \begin{bmatrix} R(1) \\ R(2) \\ \ldots \\ R(p) \end{bmatrix}, \quad \bar{\bar{R}} = \begin{bmatrix} R(0) & R(1) & \ldots & R(p-1) \\ R(1) & R(0) & \ldots & R(p-2) \\ \ldots & & & \\ R(p-1) & R(p-2) & \ldots & R(0) \end{bmatrix}, \quad \bar{a} = \begin{bmatrix} a_1 \\ \ldots \\ a_p \end{bmatrix} \tag{4.16}$$

This inversion can be done efficiently using the following recursive algorithm, called the "Levinson-Durbin recursion:"

$$
\begin{aligned}
E^{(0)} &= R(0) && (4.17)\\
k_i &= \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E^{(i-1)}}, \quad 1 \le i \le p && (4.18)\\
a_i^{(i)} &= k_i && (4.19)\\
a_j^{(i)} &= a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \quad 1 \le j \le i-1 && (4.20)\\
E^{(i)} &= (1 - k_i^2) E^{(i-1)} && (4.21)\\
& && (4.22)
\end{aligned}
$$

Finding the Minimum Energy

The minimum energy is

$$E_{min} = \left(s(m) - \sum_{k=1}^{p} a_k s(m-k) \right)^2 = R_n(0) - \sum_{k=1}^{p} a_k R_n(k) \tag{4.23}$$

Combined Equation for Energy and Coefficients

The equations for LPC coefficients and minimum error can be written together in matrix form:

$$\gamma = \mathbf{R}_p \mathbf{a} \tag{4.24}$$

$$\gamma = \begin{bmatrix} E_{min} \\ 0 \\ \dots \\ 0 \end{bmatrix}, \quad \mathbf{R}_p = \begin{bmatrix} R(0) & R(1) & \dots & R(p) \\ R(1) & R(0) & \dots & R(p-1) \\ \dots & & & \\ R(p) & R(p-1) & \dots & R(0) \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} 1 \\ -a_1 \\ \dots \\ -a_p \end{bmatrix} \tag{4.25}$$

4.2.2 Covariance Method

Solving the Normal Equations

Instead of windowing the signal, we use the original signal $s_n(m)$, but window the error:

$$E_n = \sum_{m=n}^{n+N-1} \left(s(m) - \sum_{k=1}^{p} a_k s(m-k) \right)^2 \tag{4.26}$$

Solving the equation $\partial E_n / \partial a_k = 0$ yields the autocorrelation normal equations:

$$\Phi_n(i,0) = \sum_{k=1}^{p} a_k \Phi_n(i,k) \tag{4.27}$$

where

$$\Phi_n(i,k) \equiv \sum_{m=n}^{n+N-1} s(m-i) s(m-k) \tag{4.28}$$

This can also be written as

$$\bar{\gamma} = \bar{\bar{\Phi}} \bar{a} \tag{4.29}$$

$$\bar{\gamma} = \begin{bmatrix} \Phi(1,0) \\ \Phi(2,0) \\ \dots \\ \Phi(p,0) \end{bmatrix}, \quad \bar{\bar{\Phi}} = \begin{bmatrix} \Phi(1,1) & \Phi(1,2) & \dots & \Phi(1,p) \\ \Phi(2,1) & \Phi(2,2) & \dots & \Phi(2,p) \\ \dots & & & \\ \Phi(p,1) & \Phi(p,2) & \dots & \Phi(p,p) \end{bmatrix} \tag{4.30}$$

The matrix $\bar{\bar{\Phi}}$ is symmetric, but not Toeplitz, so inverting it requires $\mathcal{O}(p^3)$ operations (using an algorithm called the Cholesky decomposition).

- Advantage: if $x(n)$ is steady-state, covariance can find the true spectrum.
- Disadvantage: solving the normal equations requires an $\mathcal{O}(p^3)$ matrix inversion operation.

Finding the Minimum Energy

The minimum error energy is:

$$E_{min} = \Phi_n(0,0) - \sum_{k=1}^{p} a_k \Phi_n(0,k) \tag{4.31}$$

Combined Equation for Energy and Coefficients

The equations for LPC coefficients and minimum error can be written together in matrix form:

$$\gamma = \mathbf{\Phi}_p \mathbf{a} \tag{4.32}$$

$$\mathbf{\Phi}_p = \begin{bmatrix} \Phi(0,0) & \Phi(0,1) & \dots & \Phi(0,p) \\ \Phi(1,0) & \Phi(1,1) & \dots & \Phi(1,p) \\ \dots & & & \\ \Phi(p,0) & \Phi(p,1) & \dots & \Phi(p,p) \end{bmatrix} \tag{4.33}$$

4.2.3 Choosing the LPC Order

The LPC order needs to be large enough to represent each formant using a complex-conjugate pole pair. There also need to be an extra 2 or 3 poles to represent spectral tilt. With everything together, we have:

$$p \approx 2 \times (\text{Number of Formants}) + (2 \text{ to } 3) \tag{4.34}$$

The number of formants in the spectrum is the Nyquist rate ($F_s/2$), divided by the average spacing between neighboring formants:

$$\text{Number of Formants} = \frac{F_s/2}{\text{average}(F_{n+1} - F_n)} \tag{4.35}$$

The spacing between neighboring formant frequencies is approximately

$$\text{average}(F_{n+1} - F_n) \approx \frac{c}{2l} \tag{4.36}$$

where $c = 35400$cm/s is the speed of sound, and l is the length of the vocal tract. The length of a male vocal tract is close to 17.7cm, so there is approximately one formant per 1000Hz:

$$p \approx \left(\frac{F_s}{1000\text{Hz}}\right) + (2 \text{ to } 3) \tag{4.37}$$

The length of a female vocal tract is close to 14.75cm, so there is approximately one formant per 1200Hz:

$$p \approx \left(\frac{F_s}{1200\text{Hz}}\right) + (2 \text{ to } 3) \tag{4.38}$$

4.2.4 Choosing the LPC Gain

The LPC excitation is defined by the formula:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{4.39}$$

The LPC error is defined by the formula:

$$e(n) \equiv s(n) - \sum_{k=1}^{p} a_k s(n-k) = Gu(n) \tag{4.40}$$

If we define

$$\sum_{k=0}^{N-1} u^2(n) \equiv 1 \tag{4.41}$$

then

$$G^2 = \frac{\sum e^2(n)}{\sum u^2(n)} = E_{min} \tag{4.42}$$

4.3 Frequency-Domain Interpretation of LPC

The LPC error is

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{4.43}$$

so the error spectrum is

$$E(z) = S(z)(1 - \sum_{k=1}^{p} a_k z^{-k}) = S(z)A(z) \tag{4.44}$$

Using Parseval's theorem, we get that

$$E_n = \sum_m e^2(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E_n(e^{j\omega})|^2 \, d\omega \tag{4.45}$$

Substituting in the form of $E(z)$, we get

$$E_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X_n(e^{j\omega})|^2 |A(e^{j\omega})|^2 \, d\omega = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{|X_n(e^{j\omega})|^2}{|H(e^{j\omega})|^2} \, d\omega \tag{4.46}$$

Since X is in the numerator inside the integral, any algorithm which minimizes E_n will automatically try to produce an $H(e^{j\omega})$ which does a good job of modeling X at frequencies where X is large. In other words, LPC models spectral peaks better than spectral valleys.

4.4 Lattice Filtering

Lattice filtering analyzes or synthesizes speech using ith-order forward prediction error $e^{(i)}(m)$ and backward prediction error $b^{(i)}(m)$:

$$e^{(i)}(m) = s(m) - \sum_{k=1}^{i} a_k s(m-k) = e^{(i-1)}(m) - k_i b^{(i-1)}(m-1) \tag{4.47}$$

$$b^{(i)}(m) = s(m-i) - \sum_{k=1}^{i} a_k s(m+k-i) = b^{(i-1)}(m-1) - k_i e^{(i-1)}(m) \tag{4.48}$$

In LPC synthesis, $s(n)$ and $b^{(p)}(m)$ are calculated recursively from $e^{(p)}(m)$, as shown in figure 4.4.

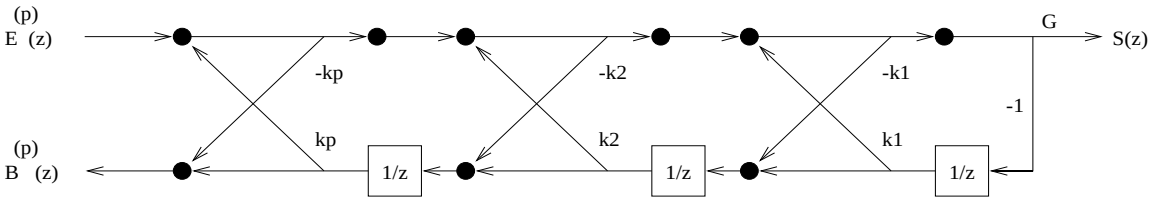

Figure 4.1: LPC synthesis using a lattice filter structure.

4.4.1 How to Calculate Reflection Coefficients

1. Reflection coefficients can be estimated directly from the speech signal by minimizing an error metric of the form

$$\tilde{E}^{(i)} = \quad \text{some function of } \; e^{(i)}(m), \quad b^{(i)}(m) \tag{4.49}$$

   Direct calculation of reflection coefficients is computationally expensive, but sometimes leads to better estimates of the transfer function (e.g. Burg's method).

2. Any LPC synthesis filter can be implemented as a lattice filter. The relationship between a_i and k_i is given by the Levinson-Durbin recursion.

4.4.2 Why Use Lattice Filter Instead of Direct-Form LPC?

Tests for stability of $1/A(z)$:

- Roots of $A(z)$: $|r_i| < 1$.
- Reflection coefficients: $|k_i| < 1$.
- Direct form coefficients a_k: No simple test.

4.4.3 Equivalence of Lattice and Concatenated-Tube Models

Reflection coefficients in the lattice filter are equivalent to reflection coefficients in a lossless tube model of the vocal tract. There are many ways to make the two structures correspond. One convenient formulation numbers the tube areas A_i backward from the lips:

$$k_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}}, \qquad A_i = \text{Area of } i\text{th tube section} \tag{4.50}$$

$$A_0 = \infty \quad \text{(Area of the space beyond the lips — lossless termination)} \tag{4.51}$$

$$A_{p+1} = A_p \quad \text{(Area of the glottis — lossy termination)} \tag{4.52}$$

The length l of each tube section is determined by the sampling period T and the speed of sound c:

$$T = \frac{2l}{c} \tag{4.53}$$

4.5 Stability of the LPC Filter

4.5.1 Stability of the Unquantized Filter $H(z)$

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} \alpha_i z^{-i}} = \prod_{i=1}^{p} \frac{1}{1 - r_i z^{-1}} \tag{4.54}$$

- H(z) is stable iff $|r_i| < 1$. This is equivalent to saying that the reflection coefficients are $|k_i| < 1$.
- Autocorrelation and lattice methods guarantee a stable filter, because they choose values of $|k_i| < 1$.
- Covariance method does not guarantee a stable filter.

4.5.2 Stability of the Quantized Filter $\hat{H}(z)$

$$\hat{H}(z) = \frac{1}{1 - \sum_{i=1}^{p} \hat{\alpha}_i z^{-i}} = \prod_{i=1}^{p} \frac{1}{1 - \hat{r}_i z^{-1}} \tag{4.55}$$

- Filter is stable iff all $|\hat{r}_i| < 1 \Leftrightarrow |\hat{k}_i| < 1$.
- If the synthesis filter $\hat{H}(z)$ is unstable, energy of the decoded speech will rapidly MAX OUT.
- Since filter state is carried forward from frame to frame, this erroneously high energy can last for a long time after the offending frame.
- This means that a single error can wipe out a whole second or more of speech, and maybe the D/A and the speakers as well.

4.5.3 Quantizing Direct-Form Coefficients Leads to Unstable Filters

A small quantization error in one of the direct-form coefficients $\hat{\alpha}_i$ can easily make $\hat{H}(z)$ unstable. For example, this filter is stable:

$$H(z) = \frac{1}{1 - 0.4z^{-1} + 0.1z^{-2} + 0.28z^{-3} + 0.49z^{-4}} \tag{4.56}$$

If α_4 is changed from 0.49 to 0.52, the filter is unstable:

$$\hat{H}(z) = \frac{1}{1 - 0.4z^{-1} + 0.1z^{-2} + 0.28z^{-3} + 0.52z^{-4}} \tag{4.57}$$

If α_1 is different, however, the same change in α_4 leaves a stable filter:

$$\hat{H}(z) = \frac{1}{1 + 0.4z^{-1} + 0.1z^{-2} + 0.28z^{-3} + 0.52z^{-4}} \tag{4.58}$$

4.5.4 THE SOLUTION

Quantize either k_i or r_i, and design the quantizer levels so that $|\hat{k}_i|$ or $|\hat{r}_i|$ is always less than 1.0.

4.6 Log Area Ratios

Goal: Quantize k_i such that $|\hat{k}_i| < 1$.

4.6.1 The Problem

Changes in k_i have a much larger effect on the synthesized speech spectrum if $|k_i| \approx 1$ than if $|k_i| << 1$, as shown in figure 4.6.1.

4.6.2 The Solution: Companded Quantization

$$k_i \rightarrow \boxed{\text{Expand}} \rightarrow g_i \rightarrow \boxed{\text{Linear PCM}} \rightarrow \hat{g}_i \rightarrow \boxed{\text{Compress}} \rightarrow \hat{k}_i \tag{4.59}$$

$$g_i = \log\left(\frac{1 - k_i}{1 + k_i}\right) \tag{4.60}$$

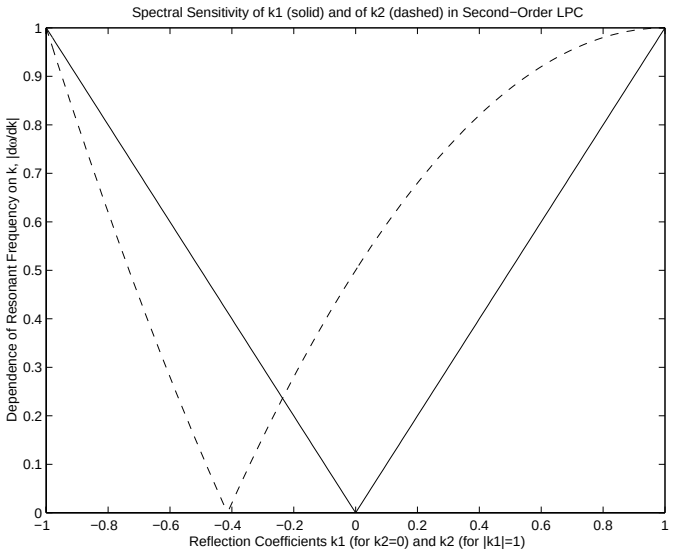

Figure 4.2: Spectral sensitivity to changes in the reflection coefficients.

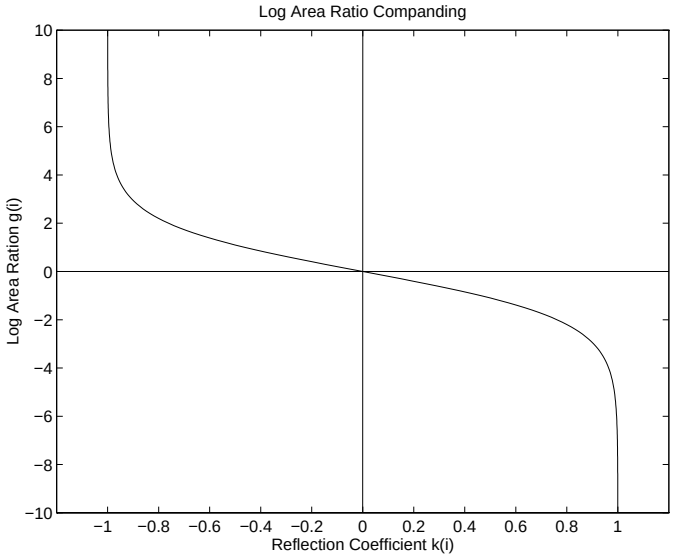

Figure 4.3: LAR companding.

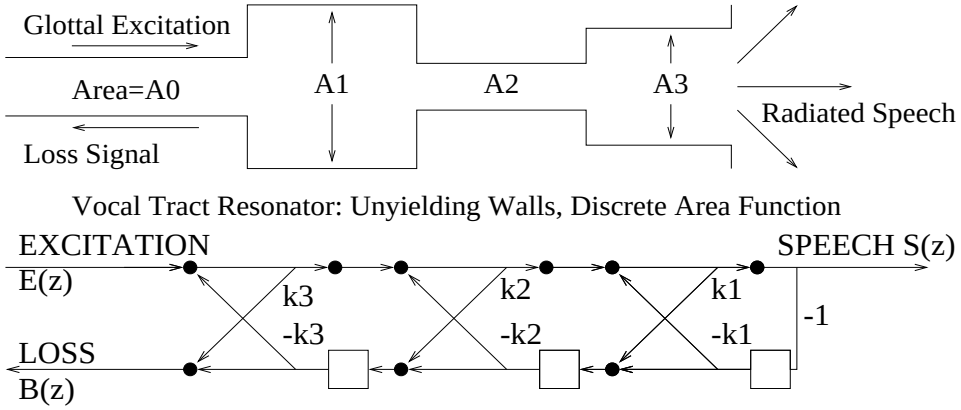

Figure 4.4: Acoustic resonator and lattice model with a matched impedance termination at the glottis.

4.6.3 Interpretation: Log Area Ratios

Remember that the reflection coefficients k_i can be used to define a stylized vocal tract model, with reflection coefficients of $r_L = 1$ at the lips, $r_g = 0$ at the glottis, and $r_i = -k_{p-i}$ $(i = 0, \ldots, p-1)$, as shown in figure 4.6.3.

If the cross-sectional areas are A_i, A_{i+1}, then

$$r_i = -k_{p-i} = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \tag{4.61}$$

and

$$\frac{A_{i+1}}{A_i} = \frac{1 - k_{p-i}}{1 + k_{p-i}} \tag{4.62}$$

So the "expanded" reflection coefficient is really a "log area ratio:"

$$g_i = \log\left(\frac{1 - k_i}{1 + k_i}\right) = \log\left(\frac{A_{p-i+1}}{A_{p-i}}\right) \tag{4.63}$$

4.7 Line Spectral Frequencies

Goal: Quantize r_i such that $|r_i| < 1$.

4.7.1 The Impedance Interpretation of the LSFs

There is no regular relationship between the real and imaginary parts of r_i (the frequencies and bandwidths of the roots), so they are difficult to quantize.

Solution

Instead, we find two adjunct polynomials $Q(z)$ and $P(z)$ whose roots e^{jq_n} and e^{jp_n} are on the unit circle if and only if the roots of $A(z)$ are inside the unit circle ($|r_i| < 1$). If the real numbers p_n and q_n have nice properties, then they might be easier to quantize than the complex numbers r_i.

Concise Definition

The reflection coefficients k_i can be used to define a stylized vocal tract model, with reflection coefficients of $r_L = 1$ at the lips, $r_g = 0$ at the glottis, and $r_i = -k_{p-i}$.

The LSFs are the zero and pole frequencies of the impedance of this vocal tract model as seen from the glottis, which is:

$$Z_T(z) = \frac{Q(z)}{P(z)} = \frac{(1-e^{j0}z^{-1})}{(1-e^{j\pi}z^{-1})}\prod_{n=1}^{p/2}\frac{(1-e^{jq_n}z^{-1})(1-e^{-jq_n}z^{-1})}{(1-e^{jp_n}z^{-1})(1-e^{-jp_n}z^{-1})} \tag{4.64}$$

Since the termination at the lips is lossless, and every tube section is lossless, the impedance $Z_T(z)$ must also be lossless. This means that the poles and zeros e^{jp_n} and e^{jq_n} are on the unit circle.

Nice Characteristics of LSFs

1. The pole frequencies p_n ($Z_T(e^{jp_n}) = \infty$) approximately equal the formant frequencies.
2. The zeros and poles alternate:
$$0 < p_1 < q_1 < p_2 < q_2 < \dots \tag{4.65}$$
$$\Rightarrow \text{Range}(q_n, p_n) \text{ is limited.} \tag{4.66}$$
$$\Rightarrow q_n, p_n \text{ easy to quantize.} \tag{4.67}$$
3. q_n, p_n are correlated with each other, so intra-frame prediction and block quantization are possible.
4. q_n, p_n change slowly from frame to frame, so inter-frame prediction is also possible.

4.7.2 Derivation of the LSF Polynomials

Interpretation

Lattice filter is like a vocal tract resonator with no reflections at the glottis, as shown in figure 4.6.3. In the figure, the "glottal excitation signal" is the forward LPC prediction error $E^{(p)}(z)$:

$$E^{(p)}(z) = \left(1 - \sum_{i=1}^{p}\alpha_i z^{-i}\right)X(z) \tag{4.68}$$

and the "glottal loss signal" is the backward LPC prediction error $z^{-(p+1)}B^{(p)}(z)$, which can be obtained by time-reversing the LPC analysis filter (as proven in R&S 8.3.3):

$$z^{-(p+1)}B^{(p)}(z) = z^{-(p+1)}\left(1 - \sum_{i=1}^{p}\alpha_i z^{i}\right)X(z) \tag{4.69}$$

Notice that, since $r_L = 1$, the only loss in this vocal tract model is at the glottis. The impedance of the rest of the vocal tract, as seen from the glottis, is pure imaginary:

$$\Re\{Z_T(e^{j\omega})\} = 0 \tag{4.70}$$

That means the zeros and poles of $Z_T(e^{j\omega})$ are on the unit circle:

$$Z_T(e^{jq_n}) = 0 \tag{4.71}$$

$$Y_T(e^{jp_n}) = \frac{1}{Z_T(e^{jp_n})} = 0 \tag{4.72}$$

...where q_n and p_n are real numbers.

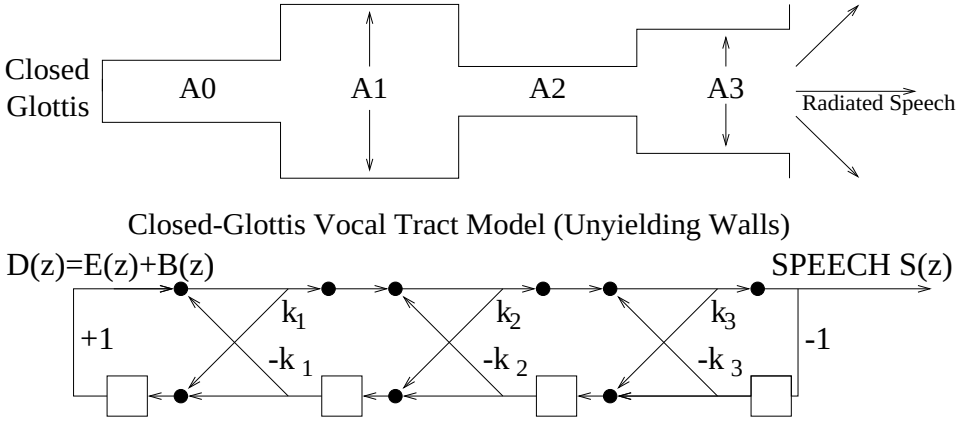

Figure 4.5: Acoustic resonator and lattice filter model with a zero-admittance termination at the glottis.

Finding the Poles

$$Y_T(e^{jp_n}) = 0 \Rightarrow Y_g(e^{jp_n}) + Y_T(e^{jp_n}) = 0 \tag{4.73}$$

where $Y_g(e^{j\omega})$ is the admittance of a rigid wall,

$$Y_g(e^{j\omega}) = 0 \quad \forall \ \omega \tag{4.74}$$

This is the same as saying that p_n are the resonant frequencies of the system shown in figure 4.7.2.

Find the resonances of the augmented lattice filter:

$$D(z) = E^{(p)}(z) + z^{-(p+1)} B^{(p)}(z) \tag{4.75}$$

$$= \left(1 - \sum_{i=1}^{p} \alpha_i z^{-i}\right) X(z) + z^{-(p+1)} \left(1 - \sum_{i=1}^{p} \alpha_i z^{i}\right) X(z) \tag{4.76}$$

$$X(z) = \frac{D(z)}{A(z) + z^{-(p+1)} A(z^{-1})} \tag{4.77}$$

So the resonances e^{jp_n} are the roots of the polynomial

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{4.78}$$

Finding the Zeros

$$Z_T(e^{jq_n}) = 0 \Rightarrow Z_g(e^{jq_n}) + Z_T(e^{jq_n}) = 0 \tag{4.79}$$

where $Z_g(e^{j\omega})$ is the impedance of open space,

$$Z_g(e^{j\omega}) = 0 \quad \forall \ \omega \tag{4.80}$$

This is the same as saying that q_n are the resonant frequencies of the system shown in figure 4.7.2.

Find the resonances of the augmented lattice filter:

$$D(z) = E^{(p)}(z) - z^{-(p+1)} B^{(p)}(z) \tag{4.81}$$

$$= \left(1 - \sum_{i=1}^{p} \alpha_i z^{-i}\right) X(z) - z^{-(p+1)} \left(1 - \sum_{i=1}^{p} \alpha_i z^{i}\right) X(z) \tag{4.82}$$

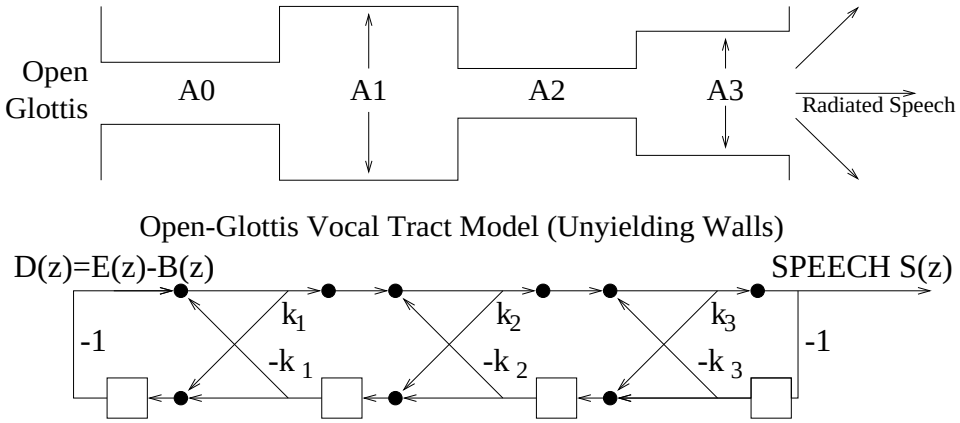

Figure 4.6: Acoustic resonator and lattice filter model with a zero-impedance termination at the glottis.

$$X(z) = \frac{D(z)}{A(z) - z^{-(p+1)}A(z^{-1})} \tag{4.83}$$

So the resonances e^{jq_n} are the roots of the polynomial

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{4.84}$$

4.7.3 The Augmented-Tube Interpretation of Line Spectral Frequencies

Setting $A_{p+1} = 0$ in the concatenated tube model yields:

$$\frac{S(z)}{U(z)} = \frac{S(z)}{E^{(p)}(z) + z^{-(p+1)}B^{(p)}(z)} = \frac{1}{P(z)}, \qquad P(z) \equiv A(z) + z^{-(p+1)}A(z^{-1}) \tag{4.85}$$

Setting $A_{p+1} = \infty$ yields:

$$\frac{S(z)}{U(z)} = \frac{S(z)}{E^{(p)}(z) - z^{-(p+1)}B^{(p)}(z)} = \frac{1}{Q(z)}, \qquad Q(z) \equiv A(z) - z^{-(p+1)}A(z^{-1}) \tag{4.86}$$

Because of symmetry, the roots of both $P(z)$ and $Q(z)$ are on the unit circle. If p is even:

$$P(z) = (1 + z^{-1}) \prod_{n=1}^{p/2} (1 - e^{jp_n} z^{-1})(1 - e^{-jp_n} z^{-1}), \qquad p_n \text{ real} \tag{4.87}$$

$$Q(z) = (1 - z^{-1}) \prod_{n=1}^{p/2} (1 - e^{jq_n} z^{-1})(1 - e^{-jq_n} z^{-1}), \qquad q_n \text{ real} \tag{4.88}$$

The frequencies p_n and q_n, for $1 \leq n \leq p/2$, are called the line spectral frequencies (LSFs). The LSFs have the following useful characteristics:

- LSFs are real, so they are easier to quantize than the LPC roots r_i, which are complex.
- If and only if $1/A(z)$ is stable, the LSFs satisfy: $0 < p_1 < q_1 < p_2 < q_2 < \ldots < \pi$
- The LSFs tend to track the LPC root frequencies $\arg(r_i)$, but...
- The LSFs vary more slowly and smoothly than the LPC roots r_i.
- Efficient algorithms exist for calculating the LSFs.

4.8 LPC Distance Measures

Suppose we want to calculate the distance between two all-pole spectra,

$$S_1(\omega) = |X_1(\omega)|^2 \approx \left|\frac{G_1}{A_1(\omega)}\right|^2, \quad S_2(\omega) = |X_2(\omega)|^2 \approx \left|\frac{G_2}{A_2(\omega)}\right|^2 \tag{4.89}$$

We can:

- Calculate the spectral L_2 norm by integrating $\log|G_1/A_1(\omega)|^2 - \log|G_2/A_2(\omega)|^2$ over ω.
- Find the LPC cepstrum, and calculate a weighted cepstral distance.
- Calculate an LPC likelihood distortion.

4.8.1 Itakura-Saito Distortion

Suppose that $S_1(\omega)$ is a random spectrum, produced by filtering a unit-energy noise process $U(\omega)$ through an unknown all-pole filter $A_1(\omega)$:

$$S_1(\omega) = \left|\frac{G_1}{A_1(\omega)}\right|^2 |U(\omega)|^2 \tag{4.90}$$

$$E[S_1(\omega)] = \left|\frac{G_1}{A_1(\omega)}\right|^2 \tag{4.91}$$

Suppose we don't know A_1, but we have a spectrum A_2 which might or might not be a good approximation to A_1. One question worth asking is, what is the probability that the signal $x_1(n)$ was generated using filter G_2/A_2? This probability is related to a distance called the Itakura-Saito measure of the distortion between spectra G_1/A_1 and G_2/A_2, named after the engineers who first derived it:

$$d_{IS}\left(\frac{G_1^2}{|A_1|^2}, \frac{G_2^2}{|A_2|^2}\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X_1(\omega)|^2 \frac{|A_2(\omega)|^2}{G_2^2} d\omega - \log\frac{G_1^2}{G_2^2} - 1 \tag{4.92}$$

$$\sim -\log\left(p_{\mathbf{x}_1}\left([x_1(0)\ldots x_1(L-1)] \mid G_2, A_2(\omega)\right)\right) \tag{4.93}$$

The first term in the I-S distortion is the residual energy of the random signal $x_1(n)$, filtered through the inverse filter $A_2(z)/G_2$:

$$x_1(n) \rightarrow \boxed{A_2(z)/G_2} \rightarrow e_{12}(n) \tag{4.94}$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |E_{12}(\omega)|^2 d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X_1(\omega)|^2 \frac{|A_2(\omega)|^2}{G_2^2} d\omega \tag{4.95}$$

$$= \frac{1}{G_2^2}\sum_n (x_1(n) - \sum_{k=1}^{p} a_{k,2} x_1(n-k))^2 \tag{4.96}$$

$$= \frac{1}{G_2^2}\left(R_1(0) - 2\sum_{k=1}^{p} a_{k,2} R_1(k) + \sum_{i=1}^{p}\sum_{k=1}^{p} a_{i,2} a_{k,2} R_1(|i-k|)\right) \tag{4.97}$$

$$= \frac{\mathbf{a}_2' \mathbf{R}_{p,1} \mathbf{a}_2}{G_2^2} \tag{4.98}$$

where $\mathbf{a}_2$ is the LPC coefficient vector representing the polynomial $A_2(z)$, and $\mathbf{R}_{p,1}$ is the autocorrelation matrix built out of samples of signal $x_1(n)$. Using this notation, the I-S distortion can be written in the easy-to-compute form:

$$\boxed{d_{IS}\left(\frac{G_1^2}{|A_1|^2},\frac{G_2^2}{|A_2|^2}\right)=\frac{\mathbf{a}_2'\mathbf{R}_{p,1}\mathbf{a}_2}{G_2^2}-\log\frac{G_1^2}{G_2^2}-1} \quad (4.100)$$

(4.100)

Characteristics:

- Asymmetric!

$$d_{IS}\left(\frac{G_1^2}{|A_1|^2},\frac{G_2^2}{|A_2|^2}\right)\neq d_{IS}\left(\frac{G_2^2}{|A_2|^2},\frac{G_1^2}{|A_1|^2}\right) \quad (4.101)$$

- Minimum value of zero is obtained when $A_1=A_2$, $G_1=G_2$.

$$d_{IS}\left(\frac{G_1^2}{|A_1|^2},\frac{G_1^2}{|A_1|^2}\right)=\frac{E_{min}}{G_1^2}-\log\frac{G_1^2}{G_1^2}-1=0 \quad (4.102)$$

4.8.2 Likelihood-Ratio Distortion, Itakura Distortion

Many times, we don't care about differences in the spectral energy of S_1 and S_2. The *likelihood-ratio* distortion measure is obtained by normalizing both S_1 and S_2 to unit energy, and then calculating an Itakura-Saito distortion:

$$d_{LR}\left(\frac{1}{|A_1|^2},\frac{1}{|A_2|^2}\right)=d_{IS}\left(\frac{1}{|A_2|^2},\frac{1}{|A_1|^2}\right)=\frac{1}{2\pi}\int_{-\pi}^{\pi}|A_2(\omega)|^2\frac{S_1(\omega)}{G_1^2}d\omega-1 \quad = \quad \frac{\mathbf{a}_2'\mathbf{R}_{p,1}\mathbf{a}_2}{G_1^2}-1 \quad (4.103)$$

(4.104)

A similar distortion measure called the Itakura distortion is often used instead of the likelihood-ratio distortion:

$$d_I\left(\frac{1}{|A_1|^2},\frac{1}{|A_2|^2}\right)\equiv\log\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}|A_2(\omega)|^2\frac{S_1(\omega)}{G_1^2}d\omega\right) \quad = \quad \log\left(\frac{\mathbf{a}_2'\mathbf{R}_{p,1}\mathbf{a}_2}{G_1^2}\right) \quad (4.105)$$

LPC Parameter Sets Containing Equivalent Information

Direct-Form Coefficients (a_k)

$$R_n(i) = \sum_{k=1}^{p} a_k R_n(|i-k|) \tag{4.106}$$

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{4.107}$$

LPC Roots (r_i)

$$r_i = \text{roots}(A(z)) \tag{4.108}$$

$$A(z) = \prod_{i=1}^{p} (1 - r_i z^{-1}) \tag{4.109}$$

PARCOR Coefficients (k_i)

$$\begin{aligned} k_i &= a_i^{(i)} \\ a_j^{(i-1)} &= \frac{a_j^{(i)} + k_i a_{i-j}^{(i)}}{1 - k_i^2}, \quad 1 \le j \le i-1 \end{aligned}$$

$$a_j^{(i)} = \begin{cases} k_i & j = i \\ a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} & 1 \le j \le i-1 \end{cases} \tag{4.110}$$

Log Area Ratios (g_i)

$$g_i = \log\left(\frac{1-k_i}{1+k_i}\right) \tag{4.111}$$

$$k_i = \frac{1 - e^{g_i}}{1 + e^{g_i}} \tag{4.112}$$

Line Spectral Frequencies (p_n, q_n)

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{4.113}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{4.114}$$

$$p_n = \arg(\text{roots}(P(z))) \quad \text{s.t.} \quad 0 < p_n < \pi$$

$$q_n = \arg(\text{roots}(Q(z))) \quad \text{s.t.} \quad 0 < q_n < \pi$$

$$P(z) = (1 + z^{-1}) \prod_{n=1}^{p/2} (1 - e^{jp_n} z^{-1})(1 - e^{-jp_n} z^{-1}) \tag{4.115}$$

$$Q(z) = (1 - z^{-1}) \prod_{n=1}^{p/2} (1 - e^{jq_n} z^{-1})(1 - e^{-jq_n} z^{-1}) \tag{4.116}$$

LPC Cepstrum (c_m)

$$\begin{aligned} c_0 &= \log G^2 \\ c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \le m \le p \\ c_m &= \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p \end{aligned}$$

$$\begin{aligned} G &= e^{c_0/2} \\ a_m &= c_m - \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \le m \le p \end{aligned}$$

4.9 Exercises

1. **Likelihood Ratio Distance**

   Suppose that the autocorrelation coefficients of signal $x_1(n)$ are $R(0) = 500$ and $R(1) = 400$.

   (a) Find the parameters G_1 and $A_1(z)$ of a first-order autocorrelation LPC model of $x_1(n)$.

   (b) Suppose that the spectrum of another signal, $x_2(n)$, can be modeled using the following transfer function:

   $$\frac{G_2}{A_2(z)} = \frac{10}{1 - 0.9z^{-1}} \tag{4.117}$$

   What is the likelihood ratio distance $d_{LR}\left(\frac{1}{|A_1|^2}, \frac{1}{|A_2|^2}\right)$?

2. **Vocal Tract Area Function** Remember that the PARCOR coefficients k_i can be viewed as reflection coefficients in a concatenated-tube model of the vocal tract, with tube sections numbered backward from the lips. Calculate the area A_i and length l of each tube section; you may assume that the area of the lip opening is $A_1 = 5\text{cm}^2$. Plot the area of the vocal tract $A(x)$ as a function of position x, where x is measured in centimeters. Turn in code and/or equations showing how you calculated l and $A(x)$.

3. **LPC Analysis and Synthesis**

   Write a function which reads in a frame of speech, calculates LPC coefficients, filters the input by $A(z)$ to find an LPC residual, and then filters the residual by $H(z) = 1/A(z)$ to synthesize a speech waveform. Verify that the output is identical to the non-overlapping part of the input.

   (a) It is important to carry the state of all digital filters forward from frame to frame. The state of digital filters in matlab can be accessed using the **Zf** and **Zi** arguments to "filter."

   What happens if the state of the LPC synthesis filter is carried forward from frame to frame, but not the state of the analysis filter? What happens if the state of the analysis filter is carried forward, but not the state of the synthesis filter? What happens if neither filter state is carried forward? (Hint: compare the LPC residual with and without filter state carry-over.)

   (b) Quantize the LPC coefficients using LAR quantization. Aim for an average of 4-5 bits per LPC coefficient. Verify that when you use the same quantized filter $\hat{A}(z)$ for both analysis and synthesis, LPC quantization does not introduce any errors into the reconstructed signal.

   How many bits per second are you using to quantize the LPC coefficients?

4. **Formant Tracking**

   Write a matlab function of the form **A = lpcana(X, P, N)** which performs LPC analysis of order **P** on the waveform **X** using frames of length **N** samples. The matrix **A** should contain one row for each frame; each row should contain the LPC filter coefficients for one frame.

   Write a matlab function **[F,BW] = formants(A,FS)** which finds the roots of the LPC polynomials stored in **A**, and calculates up to $p/2$ analog formant bandwidths BW_i and frequencies F_i per frame, such that

   $$r_i = e^{-\frac{\pi B W_i + j 2\pi F_i}{F_s}} \tag{4.118}$$

   where r_i is one of the roots of $A(z)$, calculated using the matlab **roots** function.

   Plot the formant frequencies as a function of time, and compare your plot to a spectrogram of the utterance. Which formants are tracked during voiced segments? What happens when there are less than $p/2$ trackable formants? What happens to the LPC-based formant estimates during unvoiced speech segments?

5. **Line Spectral Frequencies**

   Write a matlab function **[PF, QF] = tf2lsf(A)** which computes ordered line spectral frequencies corresponding to the LPC coefficients in **A**. The LSF matrices **PF** and **QF** should each contain as many rows as **A**, and $p/2$ columns.

   (a) Plot the analog P-frequencies, $p_n F_s/2\pi$, and compare to a spectrogram of the utterance. Do the P-frequencies track the formants during voiced segments? What happens if there are 5 P-frequencies, but only 4 trackable formants? What do the P-frequencies do during unvoiced segments? Do you think this behavior is likely to make the P-frequencies more or less quantizable than the LPC-based formant frequencies? Why?

   Plot the Q-frequencies $q_n F_s/2\pi$. How do the Q-frequencies relate to the P-frequencies? Is there ever a time when a formant frequency is tracked more closely by q_n than by p_n?

   How rapidly do the line spectral frequencies change as a function of time? What is the range of each individual line spectral frequency? What is the range of the difference between neighboring line spectral frequencies? Can you think of an efficient way of quantizing the line spectral frequencies?

   (b) Quantize the LPC coefficients using LSF quantization, then convert the quantized LSFs back into direct-form LPC coefficients, and synthesize speech. How does the speech sound? How can you guarantee that the quantized LPC synthesis filter will be stable?

6. **Pitch Tracking**

   Write a matlab function **[N0, B] = pitch(X, N, N0MIN, N0MAX)**. For each frame, set **B** and **N0** according to the following formulas:

$$B = \max r_x(m), \qquad N_{0,min} \le m \le N_{0,max} \tag{4.119}$$

$$N_0 = \arg\max r_x(m), \qquad N_{0,min} \le m \le N_{0,max} \tag{4.120}$$

$$\tag{4.121}$$

   (a) Try different values of $N_{0,min}$ and $N_{0,max}$. For each value you test, plot **B** and **N0** as a function of time, and compare them to the spectrogram. What values give the best pitch tracking? Is **B** always larger for voiced segments than unvoiced segments? What is the threshold value of **B** which best divides voiced and unvoiced segments?

   (b) Try pitch tracking using the autocorrelation of the LPC residual, rather than the signal autocorrelation. Do you find any improvement? Why or why not?

7. **Lattice Filter and Concatenated-Tube Model**

   In this problem, you will explore the relationship between the PARCOR lattice filter structure of R&S Fig. 8.3 and the lossless tube models of R&S Fig. 3.40.

   Recall that the lattice filter in Fig. 8.3 iteratively removes the redundancy from the speech signal $S(z)$ in order to calculate the forward prediction error, $E^{(i)}(z)$, and the backward prediction error, $B^{(i)}(z)$:

$$E^{(i)}(z) = E^{(i-1)}(z) - k_i z^{-1} B^{(i-1)}(z) \tag{4.122}$$

$$B^{(i)}(z) = z^{-1} B^{(i-1)}(z) - k_i E^{(i-1)}(z) \tag{4.123}$$

$$E^{(0)}(z) = B^{(0)}(z) = S(z) \tag{4.124}$$

   (a) Suppose you are given as inputs the forward error of order (i) and the backward error of order $(i-1)$, $E^{(i)}(z)$ and $B^{(i-1)}(z)$, and you are asked to synthesize $E^{(i-1)}(z)$ and $B^{(i)}(z)$. By re-arranging the equations above, devise a filter structure to accomplish this. Draw the filter structure.

   (b) Suppose you are given only the forward error of order (2), $E^{(2)}(z)$, and asked to synthesize the speech signal $S(z)$ and the backward error $B^{(2)}(z)$. Devise a lattice filter structure to accomplish this.

(c) Fig. 2 shows a section of a digital concatenated tube model, similar to the model shown in R&S Fig. 3.40, but with a delay on the backward arc rather than the forward arc. Suppose that the left-hand nodes are the (i)th order LPC prediction errors, $A(z) = E^{(i)}(z)$ and $B(z) = B^{(i)}(z)$. Show that the right-hand nodes are $C(z) = \gamma E^{(i-1)}(z)$ and $D(z) = \gamma B^{(i-1)}(z)$, for some constant γ. What is the value of γ?

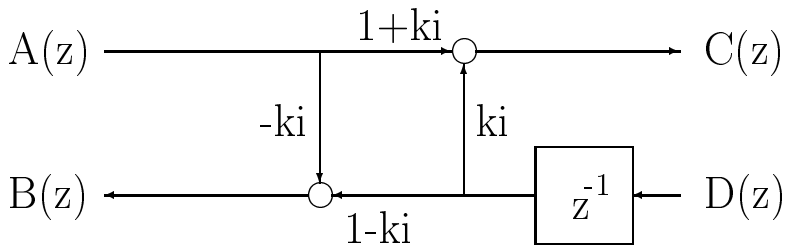

Figure 2: Section of a digital concatenated tube model.

(d) Find values of the reflection coefficients r_G, r_1, r_2, and r_L such that the transfer function of the concatenated tube model in Fig. 3.40c of your text is

$$\frac{U_l(z)}{U_g(z)} = z^{-3/2} G \frac{S(z)}{E^{(2)}(z)} \tag{4.125}$$

where $S(z)$ and $E^{(2)}(z)$ are as defined in part (b) of this problem, and G is a constant.

(e) In the concatenated tube model of part (d), what is the acoustic impedance at the lips? at the glottis? (you may express one or both impedances in terms of the cross-sectional tube areas, if necessary).

(f) The concatenated tube elements in your model of part (d) are lossless elements, and yet the impulse response of the system decays gradually over time, implying that at least one element in the system must be lossy. Where are the losses occurring? In a real vocal tract, where else would losses occur?