7. Linear Prediction

J. Benesty, J. Chen, Y. Huang

Linear prediction plays a fundamental role in all aspects of speech. Its use seems natural and obvious in this context since for a speech signal the value of its current sample can be well modeled as a linear combination of its past values. In this chapter, we attempt to present the most important ideas on linear prediction. We derive the principal results, widely recognized by speech experts, in a very intuitive way without sacrificing mathematical rigor.

7.1 Fundamentals

Linear prediction (LP) is a fundamental tool in many diverse areas such as adaptive filtering, system identification, economics, geophysics, spectral estimation, and speech. Recently, a nice history of LP in the context of speech coding was written by Atal [7.1]. Readers are invited to consult this reference for more information about this topic and how it has evolved.

Linear prediction is widely used in speech applications (recognition, compression, modeling, etc.) [7.2, 3]. This is due to the fact that the speech production process is well modeled with LP. Indeed, it is well recognized that a speech signal can be written in the following form [7.4, 5],

\[ x(k) = \sum_{l=1}^{L} a_l x(k-l) + Gu(k), \]  

(7.1)

where \( k \) is the time index, \( L \) represents the number of coefficients in the model (the order of the predictor), \( a_l, l = 1, \ldots, L, \) are defined as the linear prediction coefficients, \( G \) is the gain of the system, and \( u(k) \) is the excitation signal, which can be either a quasiperiodic train of impulses or a random noise source (also a combination of both signals for voiced fricatives such as ‘v’, ‘z’, and ‘zh’). The periodic source produces voiced sounds such as vowels and nasals, and the noise source produces unvoiced or fricated sounds such as the fricatives. The parameters, \( a_l \), determine the spectral characteristics of the particular sound for each of the two types of excitation and are widely used directly in many speech coding schemes and automatic speech recognition systems [7.4].

Equation (7.1) can be rewritten in the frequency domain, by using the \( z \)-transform. If \( H(z) \) is the transfer function of the system, we have:

\[ H(z) = \frac{G}{1 - \sum_{l=1}^{L} a_l z^{-l}} = \frac{G}{A(z)}, \]

(7.2)

which is an all-pole transfer function. This filter \([H(z)]\) is a good model of the human vocal tract [7.2]. Our main concern is to determine the predictor coefficients, \( a_l, l = 1, 2, \ldots, L, \) and to study the properties of the filter \( A(z) \).

The applications of LP are numerous. Before addressing the estimation of LP coefficients, we give some examples to show the importance of LP. In many aspects of speech processing (noise reduction, speech separation, speech dereverberation, speech coding, etc.), it is of great interest to compare the closeness of the spectral envelope of two speech signals (the desired and
Consider a stationary random signal \( x(k) \). The objective of the forward linear prediction is to predict the value of the sample \( x(k) \) from its past values, i.e., \( x(k-1), x(k-2), \) etc. We define the forward prediction error as [7.10, 11],

\[
e_{f,L}(k) = x(k) - \hat{x}(k) = x(k) - \sum_{l=1}^{L} a_{L,l} x(k-l),
\]

where the superscript ‘\( T \)’ denotes transposition, \( \hat{x}(k) \) is the predicted sample,

\[
a_L = [a_{L,1} \ a_{L,2} \ \cdots \ a_{L,L}]^T
\]

is the forward predictor of length \( L \), and

\[
x(k-1) = [x(k-1) \ x(k-2) \ \cdots \ x(k-L)]^T
\]

is a vector containing the \( L \) most recent samples starting with and including \( x(k-1) \).

Like the Itakura distance, this measure is not symmetric either; therefore, it is not a true metric.

The Itakura–Saito distance has many interesting properties. It has been shown that this measure is highly correlated with subjective quality judgements [7.6]. For example, a recent report on speech codec evaluation reveals that, if the Itakura–Saito measure between two speech signals is less than 0.5, the difference in their mean opinion score would be less than 1.6 [7.9]. Many other reported experiments also confirmed that when the Itakura–Saito distance between two speech signals is below 0.1, they would be perceived nearly identically by human ears. As a result, the Itakura–Saito distance, which is based on LP, is often used as an objective measure of speech quality. It is probably the most widely used measure of similarity between speech signals.

The two previous examples of the vocal-tract filter and the speech quality measure clearly show the importance of LP in speech applications.

In this chapter, we study the theory of linear prediction and derive the most important LP techniques that are often encountered in many speech applications. We assume here that all signals of interest are real, stationary, and zero mean.

### 7.2 Forward Linear Prediction

We would like to find the optimal Wiener predictor. For that, we seek to minimize the mean-square error (MSE):

\[
J_f(a_L) = E\{e_{f,L}^2(k)\},
\]

where \( E[\cdot] \) denotes mathematical expectation. Taking the gradient of \( J_f(a_L) \) with respect to \( a_L \) and equating to \( \theta_{Lx1} \) (a vector of length \( L \) containing only zeroes), we easily find the Wiener–Hopf equations:

\[
R_L a_{o,L} = r_{f,L},
\]

where the subscript ‘\( o \)’ in \( a_{o,L} \) stands for optimal,

\[
R_L = E[x(k-1)x^T(k-1)] = E\{x(k)x^T(k)\}
\]

\[
= \begin{pmatrix} r(0) & r(1) & \cdots & r(L-1) \\ r(1) & r(0) & \cdots & r(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(L-1) & r(L-2) & \cdots & r(0) \end{pmatrix}
\]
is the correlation matrix, and
\[ r_{t,L} = E\{x(k-1)x(k)\} = [r(1) \ r(2) \cdots r(L)]^T \]
(7.10)
is the correlation vector. The matrix \( R_L \) has a Toeplitz structure (i.e., all the entries along the diagonals are the same); assuming that it is nonsingular, we deduce the optimal forward predictor:
\[ a_{o,L} = R_L^{-1} r_{t,L}. \]
(7.11)
Expanding \( e_{o,L}^2(k) \) in (7.7) and using (7.8) shows that the minimum mean-square error (MMSE),
\[ J_{t,\min} = J_t(a_{o,L}) = r(0) - r_{t,L}^T a_{o,L} = E_{t,L}. \]
(7.12)
This is also called the forward prediction-error power. Define the augmented correlation matrix:
\[ R_{L+1} = \begin{pmatrix} r(0) & r_{t,L} \\ r_{t,L}^T & R_L \end{pmatrix}. \]
(7.13)
equations (7.8) and (7.12) may be combined in a convenient way:
\[ R_{L+1} \begin{pmatrix} 1 \\ -a_{o,L} \end{pmatrix} = \begin{pmatrix} E_{t,L} \\ 0_{L,1} \end{pmatrix}. \]
(7.14)
We refer to (7.14) as the augmented Wiener–Hopf equations of a forward predictor of order \( L \). From (7.13) we derive that,
\[ \det(R_{L+1}) = E_{t,L} \det(R_L). \]
(7.15)
where ‘det’ stands for determinant.
Let us now write the forward prediction errors for the optimal predictors of orders \( L \) and \( L-i \):
\[ e_{t,o,L}(k) = x(k) - \sum_{l=1}^{L} a_{o,L,l} x(k-l), \]
(7.16)
\[ e_{t,o,L-i}(k) = x(k) - \sum_{l=1}^{L-i} a_{o,L-i,l} x(k-l). \]
(7.17)
From the principle of orthogonality [7.11], we know that:
\[ E(e_{t,o,L}(k)x(k-1)) = \theta_{L,1}. \]
(7.18)
For \( 1 \leq i \leq L \), we can verify by using (7.18), that:
\[ E[e_{t,o,L}(k)e_{t,o,L-i}(k-i)] = 0. \]
(7.19)
As a result,
\[ \lim_{L \to \infty} E[e_{t,o,L}(k)e_{t,o,L-i}(k-i)] = E[e_{t,o}(k)e_{t,o}(k-i)] = 0. \]
(7.20)
This indicates that the signal \( e_{t,o}(k) \) is a white noise. So the optimal forward predictor has this important property of being able to whiten a stationary random process, provided that the order of the predictor is high enough.

### 7.3 Backward Linear Prediction

The aim of the backward linear prediction is to predict the value of the sample \( x(k-L) \) from its future values, i.e., \( x(k), x(k-1), \ldots, x(k-L+1) \). We define the backward prediction error as,
\[ e_{b,L}(k) = x(k-L) - \hat{x}(k-L) = x(k-L) - \sum_{l=1}^{L} b_{L,l} x(k-l+1), \]
(7.21)
where \( \hat{x}(k-L) \) is the predicted sample,
\[ b_L = [b_{L,1} \ b_{L,2} \cdots b_{L,L}]^T \]
is the backward predictor of order \( L \), and
\[ x(k) = [x(k) \ x(k-1) \cdots x(k-L+1)]^T. \]
The minimization of the MSE,
\[ J_b(b_L) = E\{ e_{b,L}(k) \}, \]
(7.22)
leads to the Wiener–Hopf equations:
\[ R_L b_{o,L} = r_{b,L}, \]
(7.23)
where
\[ r_{b,L} = E(x(k)x(k-L)) = [r(L) \ r(L-1) \cdots r(1)]^T. \]
(7.24)
Therefore, the optimal backward predictor is:
\[ b_{o,L} = R_L^{-1} r_{b,L}. \]
(7.25)
The MMSE for backward prediction,
\[ J_{b,\min} = J_b(b_{o,L}) = r(0) - r_{b,L}^T b_{o,L} = E_{b,L}, \]
(7.26)
is also called the backward prediction-error power. Define the augmented correlation matrix:
\[ R_{L+1} = \begin{pmatrix} R_L & r_{b,L} \\ r_{b,L}^T & r(0) \end{pmatrix}. \]
(7.27)
equations (7.23) and (7.26) may be combined in a convenient way:

\[
R_{L+1} \begin{pmatrix} -b_{0,l} & 1 \\
1 & \end{pmatrix} = \begin{pmatrix} \theta_{L+1} \\
E_{b,L} \end{pmatrix} .
\]  

(7.28)

We refer to this expression as the augmented Wiener–Hopf equations of a backward predictor of order \(L\).

One important property of backward prediction is that the error signals of different orders with the optimal predictors are uncorrelated, i.e., \(E\{e_{b,0,0}(k) e_{b,0,l}(k)\} = 0, i \neq l, i, l = 0, 1, \ldots, L - 1\). To prove this, let us rewrite the error signal in vector form:

\[
e_{b,0}(k) = \mathbf{Lx}(k),
\]  

(7.29)

where

\[
e_{b,0}(k) = [e_{b,0,0}(k) \ e_{b,0,1}(k) \ \cdots \ e_{b,0,L-1}(k)]^T
\]  

and

\[
\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\
-b_{0,1}^T & 1 & 0 & \cdots & 0 \\
-b_{0,2}^T & 1 & \cdots & \cdots & \cdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-b_{0,L-1}^T & 1 & \end{pmatrix}
\]  

(7.30)

is a lower triangular matrix with 1s along its main diagonal. The covariance matrix corresponding to the vector signal \(e_{b,0}(k)\) is:

\[
E\{e_{b,0}(k)e_{b,0}^T(k)\} = \mathbf{LR}_{L}^T \mathbf{L}^T.
\]  

(7.32)

By definition, the previous matrix is symmetric. The matrix product \(\mathbf{LR}_{L}^T \mathbf{L}^T\) is a lower triangular matrix because of (7.28) and the main diagonal contains the backward prediction-error powers \(E_{b,l}\) \((0 \leq l \leq L - 1)\). Since \(\mathbf{L}\) is also a lower triangular matrix, the product between the two matrices \(\mathbf{L}^T \mathbf{R}_{L}^T\) should have the same structure and, since it has to be symmetric, the only possibility is that this resulting matrix is diagonal:

\[
E\{e_{b,0}(k)e_{b,0}^T(k)\} = \text{diag}[E_{b,0}, E_{b,1}, \ldots, E_{b,L-1}],
\]  

(7.33)

and hence the prediction errors are uncorrelated.

Furthermore,

\[
\mathbf{LR}_{L}^T \mathbf{L}^T = \text{diag}[E_{b,0}, E_{b,1}, \ldots, E_{b,L-1}],
\]  

(7.34)

taking the inverse of the previous equation,

\[
\mathbf{L}^T \mathbf{R}_{L}^{-1} \mathbf{L}^T = \text{diag}[E_{b,0}^{-1}, E_{b,1}^{-1}, \ldots, E_{b,L-1}^{-1}],
\]  

(7.35)

we finally get:

\[
\mathbf{R}_{L}^{-1} = \mathbf{L}^T \text{diag}[E_{b,0}^{-1}, E_{b,1}^{-1}, \ldots, E_{b,L-1}^{-1}] \mathbf{L}.
\]  

(7.36)

Expression (7.36) defines the Cholesky factorization of the inverse matrix \(\mathbf{R}_{L}^{-1}\) [7.10, 12].

### 7.4 Levinson–Durbin Algorithm

The Levinson–Durbin algorithm is an efficient way to solve the Wiener–Hopf equations for the forward and backward prediction coefficients. This efficient method can be derived thanks to the Toeplitz structure of the correlation matrix \(\mathbf{R}_L\). This algorithm was first invented by Levinson [7.13] and independently reformulated at a later date by Durbin [7.14, 15]. Burg gave a more-elegant presentation [7.16]. Before describing this algorithm, we first need to show some important relations between the forward and backward predictors.

We define the co-identity matrix as:

\[
\mathbf{J}_L = \begin{pmatrix} 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 0 \\
1 & 0 & \cdots & 0 \end{pmatrix}
\]
Furthermore,
\[
E_{b,L} = r(0) - r_{b,L}^T b_{o,L} \\
= r(0) - r_{b,L}^T J_L b_{o,L} \\
= r(0) - r_{i,L}^T a_{o,L} \\
= E_{i,L} = E_L.
\]  
(7.40)

Therefore, for a stationary process, the forward and backward prediction-error powers are equal and the coefficients of the optimal forward predictor are the same as those of the optimal backward predictor, but in a reverse order.

The Levinson–Durbin algorithm is based on recursions of the orders of the prediction equations. Consider the following expression,
\[
\begin{pmatrix} R_L & r_{b,L}^T \\ r_{b,L} & r(0) \end{pmatrix} \begin{pmatrix} 1 \\ -a_{o,L-1} \end{pmatrix} = \begin{pmatrix} E_{L-1} \\ 0 \end{pmatrix},
\]
(7.41)
where
\[
K_L = r(L) - a_{o,L-1}^T r_{b,L-1} \\
= r(L) - a_{o,L-1}^T J_{L-1} r_{i,L-1}.
\]
(7.42)

We define the reflection coefficient as,
\[
k_L = \frac{K_L}{E_{L-1}}.
\]
(7.43)

From backward linear prediction, we have:
\[
\begin{pmatrix} r(0) & r_{i,L}^T \\ r_{i,L} & R_L \end{pmatrix} \begin{pmatrix} 1 \\ -b_{o,L-1} \end{pmatrix} = \begin{pmatrix} K_L \\ 0 \end{pmatrix}.
\]
(7.44)

Multiplying both sides of the previous equation by \(k_L\), we get,
\[
R_{L+1} \begin{pmatrix} 0 \\ -k_L b_{o,L-1} \end{pmatrix} = \begin{pmatrix} k_L E_{L-1} \\ 0 \end{pmatrix}.
\]
(7.45)

If we now subtract (7.45) from (7.41), we obtain,
\[
R_{L+1} \begin{pmatrix} 1 \\ -k_L \end{pmatrix} b_{o,L-1} = \begin{pmatrix} E_{L-1} - k_L^2 \end{pmatrix}.
\]
(7.46)

Assuming that \(R_{L+1}\) is nonsingular and identifying (7.46) with (7.14), we can deduce the recursive equation:
\[
a_{o,L} = \begin{pmatrix} a_{o,L-1} - \kappa_l b_{o,L-1} \\ 0 \end{pmatrix} \quad \text{and} \quad E_L = E_{L-1}(1 - k_L^2).
\]
(7.47)

Iterating on the prediction-error power given in (7.48), we find that,
\[
E_L = r(0) \prod_{l=1}^{L} (1 - k_l^2),
\]
(7.50)
and since \(E_L \geq 0\), this implies that,
\[
|k_l| \leq 1, \quad \forall l \geq 1.
\]
(7.51)

Also, from (7.48) we see that we have,
\[
0 \leq E_l \leq E_{L-1}, \quad \forall l \geq 1.
\]
(7.52)
so, as the order of the predictors increases, the prediction-error power decreases.

Table 7.1 summarizes the Levinson–Durbin algorithm, whose arithmetic complexity is proportional to \(L^2\). This algorithm is much more efficient than standard methods such as the Gauss elimination technique, whose complexity is on the order of \(L^3\). The saving in number of operations to find the optimal Wiener predictor can be very important, especially when \(L\) is large. The other advantage of the Levinson–Durbin algorithm is that it gives the predictors of all orders and the algorithm can be stopped if the prediction-error power is under a threshold, which can be very useful in practice when the choice of the predictor order is not easy to get in advance. A slightly more-efficient approach, called the split Levinson algorithm, can be found in [7.17]. This algorithm requires roughly half the number of multiplications and the same number of additions as the classical Levinson–Durbin algorithm. Even more-efficient algorithms have been proposed (see, for example, [7.18]) but they are numerically unstable, which is not acceptable in most speech applications.

Table 7.1 Levinson–Durbin algorithm

<table>
<thead>
<tr>
<th>Initialization: (E_0 = r(0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>For (1 \leq l \leq L)</td>
</tr>
<tr>
<td>(k_l = \frac{1}{E_{l-1}}[r(l) - a_{o,l-1}^T J_l r_{i,l-1}])</td>
</tr>
<tr>
<td>(a_{o,l} = \begin{pmatrix} a_{o,l-1} - \kappa_l \end{pmatrix} \quad \text{and} \quad E_l = E_{l-1}(1 - k_l^2))</td>
</tr>
</tbody>
</table>
7.5 Lattice Predictor

In this section, we will show that the order-recursive structure of the forward and backward prediction errors has the form of a ladder, which is called a lattice predictor.

Inserting (7.47) into the forward prediction error for the optimal predictor of order \( L \),

\[
\epsilon_{\ell, 0, L}(k) = x(k) - a_{0, L}^T x(k-1),
\]

we obtain,

\[
\epsilon_{\ell, 0, L}(k) = \epsilon_{\ell, 0, L-1}(k) - \kappa_L (b_{\ell, L-1}^T 1) x(k-1).
\]

The second term (without the reflection coefficient) on the right-hand side of (7.54) is the backward prediction error for the optimal predictor of order \( L - 1 \). Therefore, (7.54) can be rewritten

\[
\epsilon_{\ell, 0, L}(k) = \epsilon_{\ell, 0, L-1}(k) - \kappa_L \epsilon_{b, 0, L-1}(k-1).
\]

If we insert (7.47) again into the backward prediction error for the optimal predictor of order \( L \),

\[
\epsilon_{b, 0, L}(k) = x(k-L) - b_{0, L}^T x(k) = x(k-L) - a_{0, L}^T J_L x(k),
\]

we get,

\[
\epsilon_{b, 0, L}(k) = \epsilon_{b, 0, L-1}(k-1) - \kappa_L \epsilon_{\ell, 0, L-1}(k-1).
\]

If we put (7.55) and (7.57) into a matrix form, we have,

\[
\begin{pmatrix}
\epsilon_{\ell, 0, L}(k) \\
\epsilon_{b, 0, L}(k)
\end{pmatrix} =
\begin{pmatrix}
1 & -\kappa_L \\
-\kappa_L & 1
\end{pmatrix}
\begin{pmatrix}
\epsilon_{\ell, 0, L-1}(k) \\
\epsilon_{b, 0, L-1}(k-1)
\end{pmatrix} = 
\prod_{l=1}^{L} \begin{pmatrix}
1 & -\kappa_l \\
-\kappa_l & 1
\end{pmatrix}
\begin{pmatrix}
x(k) \\
x(k-1)
\end{pmatrix},
\]

where we have taken for initial conditions (order 0), \( \epsilon_{\ell, 0, 0}(k) = x(k) \) and \( \epsilon_{b, 0, 0}(k-1) = x(k-1) \). Figure 7.1 depicts the \( l \)-th stage of a lattice predictor. For the whole lattice predictor, \( L \) of these stages are needed and are connected in cascade, one to each other, starting from order 0 to order \( L \).

Now let us compute the variance of \( \epsilon_{b, 0, L}(k) \) from (7.57),

\[
E\{\epsilon_{b, 0, L}(k)^2\} = E_L = E_{L-1} + \kappa_L^2 E_{L-1} - 2\kappa_L E\{\epsilon_{\ell, 0, L-1}(k)\epsilon_{b, 0, L-1}(k-1)\} = E_{L-1} - \kappa_L^2.
\]

Developing the previous expression, we obtain,

\[
k_L = \frac{E\{\epsilon_{\ell, 0, L-1}(k)\epsilon_{b, 0, L-1}(k-1)\}}{E_{L-1}} \cdot \frac{E\{\epsilon_{b, 0, L-1}(k-1)\epsilon_{b, 0, L-1}(k-1)\}}{E\{\epsilon_{b, 0, L-1}(k-1)\epsilon_{b, 0, L-1}(k-1)\}}.
\]

We see from (7.60) that the reflection coefficients are also the normalized cross-correlation coefficients between the forward and backward prediction errors, which is why they are also often called partial correlation (PARCOR) coefficients [7.3, 19, 20]. These coefficients are linked to the zeroes of the forward prediction-error FIR filter of order \( L \), whose transfer function is

\[
A_{0, L}(z) = 1 - \sum_{l=1}^{L} a_{0, L} z^{-l} = \prod_{l=1}^{L} (1 - z_{0, l} z^{-1}),
\]

where \( z_{0, l} \) are the roots of \( A_{0, L}(z) \). Since \( \kappa_L = a_{0, L} \), we have,

\[
k_L = (-1)^{L+1} \prod_{l=1}^{L} z_{0, l}.
\]

The filter \( A_{0, L}(z) \) can be shown to be minimum phase, i.e., \( |z_{0, l}| \leq 1 \). As a result [because of the relation (7.39)], the filter \( B_{0, L}(z) \) corresponding to the backward predictor is maximum phase. We will now show this very important property that the forward predictor is minimum phase. As far as we know, this simple and elegant proof was first shown by M. Mohan Sondhi but was never published. A similar proof can be found in [7.21] and [7.22].

To avoid cumbersome notation, redefine the coefficients \( w_l = -a_{0, L} \), with \( w_0 = 1 \), so that the polynomial...
becomes,
\[ A_{o,L}(z) = \sum_{l=0}^{L} w_l z^{-l} . \]  (7.63)

Also, define the vector,
\[ w = \begin{bmatrix} w_0 & w_1 & \cdots & w_L \end{bmatrix}^T . \]

We know that,
\[ R_{L+1}w = \begin{bmatrix} E_L \\ \theta_{L \times 1} \end{bmatrix} . \]  (7.64)

If \( \lambda \) is a root of the polynomial, it follows that,
\[ A_{o,L}(z) = (1 - \lambda z^{-1}) \sum_{l=0}^{L-1} g_l z^{-l}, \text{ with } g_0 = 1 . \]  (7.65)

(Note that since \( \lambda \) can be complex, the coefficients \( g_l \) are, in general, complex.) Thus the vector \( w \) can be written
\[ w = g - \lambda \hat{g} , \]  (7.66)

where
\[ g = [1 \ g_1 \ g_2 \ \cdots \ g_{L-1} \ 0]^T = [g^T \ 0]^T , \]
\[ \hat{g} = [0 \ 1 \ g_1 \ g_2 \ \cdots \ g_{L-1}]^T = [0 \ g^T]^T . \]

Substituting (7.66) in (7.64), we obtain,
\[ R_{L+1}g = \lambda R_{L+1} \hat{g} + \begin{bmatrix} E_L \\ \theta_{L \times 1} \end{bmatrix} . \]  (7.67)

Now, premultiplying by \( \hat{g}^H \) (where the superscript \( H \) denotes conjugate transpose) gives,
\[ \hat{g}^H R_{L+1}g = \lambda \hat{g}^H R_{L+1} \hat{g} . \]  (7.68)

Thus,
\[ \left| \hat{g}^H R_{L+1} \hat{g} \right|^2 = |\lambda|^2 \left( \hat{g}^H R_{L+1} \hat{g} \right)^2 . \]  (7.69)

Using the Schwartz inequality,
\[ \left| \hat{g}^H R_{L+1} \hat{g} \right|^2 \leq (\hat{g}^H R_{L+1} \hat{g}) (\hat{g}^H R_{L+1} \hat{g}) . \]  (7.70)

However,
\[ \hat{g}^H R_{L+1} \hat{g} = \begin{bmatrix} 0 & g^H \end{bmatrix} \begin{bmatrix} r(0) \\ r_{1:L} \end{bmatrix} \begin{bmatrix} 0 \\ g' \end{bmatrix} = g^H R_L g' . \]  (7.71)

Similarly,
\[ \hat{g}^H R_{L+1}g = \begin{bmatrix} g^H \end{bmatrix} \begin{bmatrix} R_L \\ r_{b:L} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = g^H R_L g' . \]  (7.72)

Therefore, \( \hat{g}^H R_{L+1} \hat{g} = g^H R_{L+1}g \), and the Schwartz inequality becomes,
\[ \left| \hat{g}^H R_{L+1} \hat{g} \right|^2 \leq (\hat{g}^H R_{L+1} \hat{g})^2 . \]  (7.73)

From (7.69) we see that \( |\lambda|^2 \leq 1 \). This completes the proof.

This property allows one easily to ensure that the all-pole system in (7.2) is stable (when the correlation matrix is positive definite) by simply imposing the constraint that the PARCOR coefficients are less than 1 in magnitude. As a result, in speech communication, transmitting PARCOR coefficients is more advantageous than directly transmitting linear prediction coefficients.

### 7.6 Spectral Representation

It is important to understand the link between the spectrum of a speech signal and its prediction coefficients. Let us again take the speech model given in Sect. 7.1,
\[ x(k) = \sum_{l=1}^{L} a_l x(k-l) + Gu(k) , \]  (7.74)

where we now assume that \( u(k) \) is a white random signal with variance \( \sigma_u^2 = 1 \). Since \( x(k) \) is the output of the filter \( H(z) \) (see Sect. 7.1), whose input is \( u(k) \), its spectrum is [7.11],
\[ S_x(\omega) = |H(e^{j\omega})|^2 S_u(\omega) , \]  (7.75)

where \( \omega \) is the angular frequency, \( H(e^{j\omega}) \) is the frequency response of the filter \( H(z) \), and \( S_u(\omega) \) is the spectrum of \( u \). We have \( S_u(\omega) = 1 \) (\( u \) is white). Using (7.2), we deduce the spectrum of \( x \),
\[ S_x(\omega) = \frac{G^2}{|A(e^{j\omega})|^2} \left| 1 - \sum_{l=1}^{L} a_l e^{-j\omega l} \right|^2 . \]  (7.76)

Therefore, the spectrum of a speech signal can be modeled by the frequency response of an all-pole filter,
where the elements are the prediction coefficients [7.23–25].

Consider the prediction error signal,
\[ e_L(k) = x(k) - a_T^T x(k - 1). \]

(7.77)

Taking the \( z \)-transform of (7.77) and setting \( z = e^{j\omega} \), we obtain:
\[ S_{x,L}(\omega) = \frac{|E_L(e^{j\omega})|^2}{|A_{x,L}(e^{j\omega})|^2}, \]

(7.78)

where \( |E_L(e^{j\omega})|^2 \) is the spectrum of \( e_L(k) \). From Sect. 7.2, we know that, for a large order \( L \), linear prediction tends to whiten the signal, so the power spectrum \( E_L(e^{j\omega})^2 \) of the error signal, \( e_L(k) \), will tend to be flat. Hence,
\[ \lim_{L \to \infty} |E_L(e^{j\omega})|^2 = G^2. \]

(7.79)

As a result,
\[ \lim_{L \to \infty} S_{x,L}(\omega) = \frac{G^2}{|1 - \sum_{l=1}^{\infty} a_l e^{-j\omega l}|^2}. \]

(7.80)

This confirms that (7.76) can be a very good approximation of the spectrum of a speech signal, as long as the order of the predictor is large enough.

## 7.7 Linear Interpolation

Linear interpolation can be seen as a straightforward generalization of forward and backward linear predictions. Indeed, in linear interpolation, we try to predict the value of the sample \( x(k - i) \) from its past and future values [7.26, 27]. We define the interpolation error as
\[ e_i(k) = x(k - i) - \hat{x}(k - i) \]
\[ = x(k - i) - \sum_{l=0,\not= l}^{L} c_l x(k - l) \]
\[ = c_i^T x_{L+1}(k), \quad i = 0, 1, \ldots, L, \]

(7.81)

where \( \hat{x}(k - i) \) is the interpolated sample,
\[ c_i = [-c_{i,0}, -c_{i,1}, \ldots, c_{i,i}, \ldots, -c_{i,L}]^T \]

is a vector of length \( L + 1 \) containing the interpolation coefficients, with \( c_{i,i} = 1 \), and
\[ x_{L+1}(k) = [x(k), x(k - 1), \ldots, x(k - L)]^T. \]

The special cases \( i = 0 \) and \( i = L \) are the forward and backward prediction errors, respectively.

To find the optimal Wiener interpolator, we need to minimize the cost function,
\[ J_i(c_i) = E\{e_i^2(k)\} \]
\[ = c_i^T R_{L+1} c_i, \]

(7.82)

subject to the constraint
\[ c_i^T v_i = c_{i,i} = 1, \]

(7.83)

where
\[ v_i = [0 \ 0 \ 0 \ 1 \ 0 \ \ldots \ 0]^T \]

is a vector of length \( L + 1 \) with its \( i \)-th component equal to one and all others equal to zero. By using a Lagrange multiplier, it is easy to see that the solution to this optimization problem is
\[ R_{L+1} c_{0,i} = E_i v_i, \]

(7.84)

where
\[ E_i = c_{0,i}^T R_{L+1} c_{0,i} \]
\[ = \frac{1}{v_i^T R_{L+1}^{-1} v_i} \]

(7.85)

is the interpolation-error power.

From (7.84) we find,
\[ c_{0,i} = \frac{1}{E_i} \]
\[ R_{L+1}^{-1} v_i, \]

(7.86)

hence the \( i \)-th column of \( R_{L+1}^{-1} \) is \( E_{0,i} / E_i \). We can now see that \( R_{L+1}^{-1} \) can be factorized as follows [7.28]:
\[ R_{L+1}^{-1} = \begin{pmatrix}
1 & -c_{0,1,0} & \ldots & -c_{0,L,0} \\
-c_{0,0,1} & 1 & \ldots & -c_{0,L,1} \\
\vdots & \vdots & \ddots & \vdots \\
-c_{0,0,L} & -c_{0,1,L} & \ldots & 1
\end{pmatrix}
\]
\[ \begin{pmatrix}
1/E_0 & 0 & \ldots & 0 \\
0 & 1/E_1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1/E_{L-1}
\end{pmatrix}
\]
\[ = C_o^T D_e^{-1}. \]

(7.87)
Furthermore, since $\mathbf{R}_{L+1}^{-1}$ is a symmetric matrix, (7.87) can be written as,

$$
\mathbf{R}_{L+1}^{-1} = \begin{pmatrix}
1/E_0 & 0 & \cdots & 0 \\
0 & 1/E_1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1/E_{L-1}
\end{pmatrix} 
$$

$$
= \begin{pmatrix}
1 & -c_{0.1} & \cdots & -c_{0,L} \\
-c_{0.1} & 1 & \cdots & -c_{0.1} \\
\vdots & \vdots & \ddots & \vdots \\
-c_{0.L} & -c_{0.1} & \cdots & 1
\end{pmatrix}
= \mathbf{D}^{-1} \mathbf{C}_{0}.
$$

Therefore, we deduce that,

$$
\frac{c_{0,i,l}}{E_i} = \frac{c_{0,i,l}}{E_i}, \quad i, l = 0, 1, \ldots, L.
$$

(7.89)

The first and last columns of $\mathbf{R}_{L+1}^{-1}$ contain, respectively, the normalized forward and backward predictors and all the columns between contain the normalized interpolators.

We are now going to show how the condition number of the correlation matrix depends on the interpolators. The condition number of the matrix $\mathbf{R}_{L+1}$ is defined as [7.29]:

$$
\chi(\mathbf{R}_{L+1}) = \|\mathbf{R}_{L+1}\| \|\mathbf{R}_{L+1}^{-1}\|,
$$

(7.90)

where $\| \cdot \|$ can be any matrix norm. Note that $\chi(\mathbf{R})$ depends on the underlying norm. Let us compute $\chi(\mathbf{R}_{L+1})$ using the Frobenius norm:

$$
\|\mathbf{R}_{L+1}\|_F = \left(\text{tr}(\mathbf{R}_{L+1}^2)\right)^{1/2}
= \left(\text{tr}(\mathbf{R}_{L+1}^2)\right)^{1/2}
$$

(7.91)

and

$$
\|\mathbf{R}_{L+1}^{-1}\|_F = \left(\text{tr}(\mathbf{R}_{L+1}^{-2})\right)^{1/2}.
$$

(7.92)

From (7.86), we have,

$$
c_{0,i,l} \frac{c_{0,i,l}}{E_i} = \mathbf{v}_i^T \mathbf{R}_{L+1}^{-2} \mathbf{v}_i,
$$

(7.93)

which implies that,

$$
\sum_{i=0}^{L} c_{0,i}^T c_{0,i} E_i = \sum_{i=0}^{L} \mathbf{v}_i^T \mathbf{R}_{L+1}^{-2} \mathbf{v}_i
= \text{tr}(\mathbf{R}_{L+1}^{-2})
= \text{tr}(\mathbf{R}_{L+1}^{-2}).
$$

(7.94)

Also, we can easily check that,

$$
\text{tr}(\mathbf{R}_{L+1}^2) = (L + 1)r^2(0) + 2 \sum_{i=1}^{L} (L + 1 - i)r^2(i).
$$

(7.95)

Therefore, the square of the condition number of the correlation matrix associated with the Frobenius norm is

$$
\chi^2(\mathbf{R}_{L+1}) = \left( (L + 1)r^2(0) + 2 \sum_{i=1}^{L} (L + 1 - i)r^2(i) \right)
\times \sum_{i=0}^{L} \frac{c_{0,i}^T c_{0,i}}{E_i}.
$$

(7.96)

Some other interesting relations between the forward predictors and the condition number can be found in [7.30].

To conclude this section, we would like to let readers know that several algorithms exist to compute the optimal predictors efficiently, see for example, [7.31] and [7.32]. All these algorithms are based on Levinson–Durbin recursions.

### 7.8 Line Spectrum Pair Representation

Line spectrum pair (LSP) representation, first introduced by Itakura [7.33], is a more-robust way to represent the coefficients of linear predictive models. The LSP polynomials have some very interesting properties shown in [7.34].

A polynomial $P(z)$ of order $L$ is said to be symmetric if

$$
P(z) = z^{-L} P(z^{-1})
$$

(7.97)

and a polynomial $Q(z)$ is antisymmetric if

$$
Q(z) = -z^{-L} Q(z^{-1}).
$$

(7.98)

Let

$$
A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_L z^{-L}
$$

(7.99)

be the optimal polynomial predictor of order $L$. It is well known that in speech compression the coefficients of this polynomial are inappropriate for quantization because of their relatively large dynamic range and also because, as stated earlier, quantization can change a stable LPC filter into an unstable one [7.35]. From (7.99), we can construct two artificial $(L + 1)$-th-order (symmetric and antisymmetric) polynomials by setting the $(L + 1)$-th
7.9 Multichannel Linear Prediction

Multichannel linear prediction can be very useful in stereo or multichannel speech compression. In an increasing number of speech or audio applications, we have at least two channels available, which are often highly correlated with each other. Therefore, it makes sense to take this interchannel correlation into account in order to obtain more-efficient compression schemes. Multichannel linear prediction is the best way to do this.

Let

\[
\chi(k) = \left[ x_1(k) \ x_2(k) \ \cdots \ x_M(k) \right]^T
\]

be a real, zero-mean, stationary \(M\)-channel time series. We define the multichannel forward prediction error vector as,

\[
e^+(k) = x(n) - \frac{1}{2} [x(n-1) + x(n)]^T A^+ + \frac{1}{2} [x(n-1) - x(n)]^T A^-
\]

It is shown in [7.37] and [7.38] that the LSP polynomials, whose trivial zeroes have been removed, are equivalent to the two optimal Wiener predictors \(a^+_o\) and \(a^-_o\). This is easy to see if we rewrite, \(e^+(k)\) for example as

\[
e^+(k) = \left[ 1 - \frac{a^+T}{2} \right] x_{L+1}(n) + \frac{a^+T}{2} x_{L+1}(n) = g^T I_1 x_{L+1}(n),
\]

where

\[
I_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}
\]

and

\[
g = \begin{pmatrix} 1 \\ -\frac{a^+}{2} \end{pmatrix}.
\]

By minimizing the MSE, \(E(e^{+2}(k))\), with respect to \(g\), with the constraint \(g^T e_1 = 1\), one can find the most important results. For readers who are interested in more details on the properties of LSP polynomials, we recommend the paper by Bäckström and Magi [7.39].
To derive the optimal Wiener forward predictors, we need to minimize the MSE,
\[ J_I(A_L) = E\{e_I^T(k)e_I(k)\}. \tag{7.109} \]
We find the multichannel Wiener–Hopf equations:
\[ R_L A_{o,L} = R_L(1/L), \tag{7.110} \]
where
\[ R_L = E\{x(k-1)x^T(k-1)\} = E\{x(k)x^T(k)\} \tag{7.111} \]
\[ = \begin{pmatrix} R(0) & R(1) & \cdots & R(L-1) \\ R^T(1) & R(0) & \cdots & R(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ R^T(L-1) & R^T(L-2) & \cdots & R(0) \end{pmatrix} \tag{7.112} \]
is the block-Toeplitz covariance matrix of size \(ML \times ML\),
\[ R(I) = E\{\hat{x}(k-1)\hat{x}^T(k-I)\}, \quad I = 0, 1, \ldots, L-1, I \]
\[ R(-I) = E\{\hat{x}(k-I)\hat{x}^T(k)\} = R^T(I), \]
and
\[ R_L(1/L) = [R(1)\ R(2)\ \cdots \ R(L)]^T \tag{7.113} \]
is the intercorrelation matrix of size \(ML \times M\).
Using the augmented block-Toeplitz covariance matrix of size \((ML+M) \times (ML+M)\):
\[ R_{L+1} = \begin{pmatrix} R_L & R_L^T(1/L) \\ R_L(1/L) & R_L \end{pmatrix}, \tag{7.114} \]
we deduce the augmented multichannel Wiener–Hopf equations:
\[ R_{L+1} \begin{pmatrix} I_{M \times M} \\ -A_{o,L} \end{pmatrix} = E_{f,L}, \tag{7.115} \]
where \(I_{M \times M}\) is the identity matrix of size \(M \times M\) and
\[ E_{f,L} = E\{e_{I,o,L}(k)e_{I,o,L}^T(k)\} = R(0) - R_{L}^T(1/L)A_{o,L} \tag{7.116} \]
is the forward error covariance matrix of size \(M \times M\), with
\[ e_{I,o,L}(k) = \chi(k) - A_{o,L}^T x(k-1). \tag{7.117} \]
We will proceed with the same philosophy to derive important equations for the multichannel backward prediction. We define the multichannel backward prediction error vector as
\[ e_{b,L}(k) = \chi(k-L) - \hat{\chi}(k-L) \]
\[ = \chi(k-L) - \sum_{l=1}^{L} B_{L,l}\chi(k-l+1) \tag{7.118} \]
where
\[ B_L = [B_{L,1} \ B_{L,2} \ \cdots \ B_{L,L}]^T \tag{7.119} \]
is the backward predictor matrix of size \(ML \times M\) with each one of the square submatrices \(B_{L,l}\) being of size \(M \times M\).
The minimization of the MSE,
\[ J_b(B_L) = E\{e_{b,L}^T(k)e_{b,L}(k)\} \tag{7.120} \]
leads to the multichannel Wiener–Hopf equations for the backward prediction:
\[ R_L B_{o,L} = R_b(1/L), \tag{7.121} \]
where
\[ R_b(1/L) = E\{x(k)\hat{x}^T(k-L)\} \tag{7.122} \]
\[ = [R^T(L) \ R^T(L-1) \ \cdots \ R^T(1)]^T. \tag{7.123} \]
By using the augmented block-Toeplitz covariance matrix:
\[ R_{L+1} = \begin{pmatrix} R_{L} & R_b(1/L) \\ R_b(1/L) & R(0) \end{pmatrix}, \tag{7.124} \]
we find the augmented multichannel Wiener–Hopf equations:
\[ R_{L+1} \begin{pmatrix} -B_{o,L} \\ I_{M \times M} \end{pmatrix} = E_{b,L}, \tag{7.125} \]
where
\[ E_{b,L} = E\{e_{b,o,L}(k)e_{b,o,L}^T(k)\} = R(0) - R_{b}^T(1/L)B_{o,L} \tag{7.126} \]
is the backward error covariance matrix of size \(M \times M\), with
\[ e_{b,o,L}(k) = \chi(k-L) - B_{o,L}^T x(k). \tag{7.127} \]
To solve the multichannel Wiener–Hopf equations efficiently, we need to derive some important relations [7.40]. Consider the following system,

\[
\begin{pmatrix}
R_L & R_b(1/L) \\
R_b^T(1/L) & R(0)
\end{pmatrix}
\begin{pmatrix}
I_{M \times M} \\
-A_{b,L-1}
\end{pmatrix}
= \begin{pmatrix}
\theta_{M \times M} \\
K_{f,L}
\end{pmatrix},
\]

where

\[
K_{f,L} = R^T(L) - R_b^T(1/L - 1)A_{b,L-1}.
\]

Consider the other system,

\[
\begin{pmatrix}
R(0) & R_b^T(1/L) \\
R_b(1/L) & R_L
\end{pmatrix}
\begin{pmatrix}
\theta_{M \times M} \\
-M_{b,L-1}
\end{pmatrix}
= \begin{pmatrix}
K_{b,L} \\
E_{b,L-1}
\end{pmatrix},
\]

where

\[
K_{b,L} = R(L) - R_b^T(1/L - 1)B_{b,L-1}.
\]

If we post-multiply both sides of (7.127) by \(E_{b,L-1}^{-1}K_{f,L}\), we get:

\[
R_{L+1} = -B_{o,L-1}E_{b,L-1}^{-1}K_{f,L}
\]

Subtracting (7.129) from (7.125) and identifying the resulting system with the augmented multichannel Wiener–Hopf equations for forward prediction [eq. (7.114)], we deduce the two recursions:

\[
\begin{align*}
R_{L+1} & = \begin{pmatrix}
I_{M \times M} \\
-A_{b,L-1}
\end{pmatrix}
E_{f,L-1}^{-1}K_{b,L} \\
= \begin{pmatrix}
K_{b,L} \\
-M_{b,L-1}
\end{pmatrix}
E_{f,L-1}^{-1}K_{b,L} \\
& = \begin{pmatrix}
K_{b,L} \\
-M_{b,L-1}
\end{pmatrix}
E_{f,L-1}^{-1}K_{b,L}.
\end{align*}
\]

Relations (7.130), (7.131), (7.133), and (7.134) were independently discovered by Whittle [7.41] and Wiggins and Robinson [7.42]. Another important relation needs to be found. Indeed, using (7.116) and (7.124), we can easily verify,

\[
\begin{align*}
E \{ e_{f,o,L-1}(k)e_{b,o,L-1}(k-1) \} &= K_{b,L}, \\
E \{ e_{b,o,L-1}(k-1)e_{f,o,L-1}(k) \} &= K_{f,L}.
\end{align*}
\]

which implies that,

\[
K_{b,L} = K_{f,L}^T.
\]

Table 7.2 summarizes the Levinson–Wiggins–Robinson algorithm [7.42–44], which is a generalization of the Levinson–Durbin algorithm to the multichannel case.

**Table 7.2** Levinson–Wiggins–Robinson algorithm

<table>
<thead>
<tr>
<th>Initialization: (E_{o,0} = E_{b,0} = R(0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>For (1 \leq l \leq L)</td>
</tr>
<tr>
<td>(K_{b,l} = R(l) - R_b^T(l/L - 1)B_{b,l-1})</td>
</tr>
</tbody>
</table>
| \(A_{b,l} = \begin{pmatrix}
A_{b,l-1} \\
-M_{b,L-1}
\end{pmatrix}
E_{b,l-1}^{-1}K_{b,l}^T\) |
| \(B_{b,l} = \begin{pmatrix}
A_{b,l-1}^{-1} \\
-M_{b,L-1}
\end{pmatrix}
E_{b,l-1}^{-1}K_{b,l}^T\) |
| \(E_{b,l} = E_{b,l-1} - K_{b,l}E_{b,l-1}^{-1}K_{b,l}\) |
| \(E_{b,l} = E_{b,l-1} - K_{b,l}E_{b,l-1}^{-1}K_{b,l}\) |
7.10 Conclusions

In this chapter, we have tried to present the most important results in linear prediction for speech. We have explained the principle of forward linear prediction and have shown that the optimal prediction error signal tends to be a white signal. We have extended the principle of forward linear prediction to backward linear prediction and derived the Cholesky factorization of the inverse correlation matrix. We have developed the classical Levinson–Durbin algorithm, which is a very efficient way to solve the Wiener–Hopf equations for the forward and backward prediction coefficients. We have explained the idea behind the lattice predictor. We have shown how the spectrum of a speech signal can easily be estimated thanks to the prediction coefficients. We have given some notions of linear interpolation and have demonstrated how the condition number of the correlation matrix is related to the optimal interpolators. We have also presented some notions of line spectrum pair polynomials. Finally, in the last section, we have generalized some of these ideas to the multichannel case.

References

7.10 M.G. Bellanger: Adaptive Digital Filters and Signal analysis (Marcel Dekker, New York 1987)
Springer Handbook of Speech Processing
Benesty, J.; Sondhi, M.M.; Huang, Y. (Eds.)
2008, XXXVI, 1176 p. 456 illus. With DVD., Hardcover
ISBN: 978-3-540-49125-5