HST.582J / 6.555J / 16.456J Biomedical Signal and Image Processing
Spring 2007

Chapter 8 - LINEAR PREDICTION

Introduction

Linear prediction is a widely-used signal processing technique in which the current value of a discrete-time signal is approximated by a finite weighted sum of its past values. It is mathematically equivalent to the techniques of *autoregressive spectral estimation,* and *maximum-entropy spectral estimation.* The central idea behind these techniques is that the signal is modeled as the response of an all-pole (autoregressive) filter to a white signal. Spectral estimation is then equivalent to estimating the coefficients of the all-pole filter. These techniques are well suited to signals whose spectra show sharp peaks such as the formants of speech.

In the case of speech signals, linear prediction takes a special significance in the context of the source-filter model of speech production. According to this model, speech is the output of a time-varying filter (representing the vocal tract resonances and radiation characteristics) excited by either a voicing source or a noise source (Fig. 1a). Under certain assumptions, linear prediction can separate the contributions of the source and the filter to the speech signal, i.e. it can *deconvolve* the source signal from the impulse response of the filter. This property contrasts with those of the short-time Fourier transform, which provides a spectral representation of speech in which effects of the source and the filter are scrambled, and which is heavily dependent on the choice of an analysis window. The deconvolution property of linear prediction is useful in biomedical applications because the source and the filter correspond to different anatomical structures, and are therefore affected by different clinical conditions. Linear prediction is also widely used in telecommunication engineering and automatic speech recognition because it represents speech in terms of a small number of parameters that contain the most important information for intelligibility.

8.1 All-pole model of speech

8.1.1 From the source-filter model to the all-pole model

In order to show how linear prediction can separate the contributions of the source and the filter to speech, we need to simplify somewhat the speech-production model of Fig. 1a by making two additional assumptions: (1) that the filter is *all-pole* (purely-recursive, or autoregressive) and (2) that the source is "white" in the sense that it has a flat spectral envelope.

The all-pole assumption makes intuitive sense because the transfer function of the vocal tract shows sharp peaks associated with the formant frequencies, which are well modeled by all-pole filters (digital resonators). In fact, acoustic theory shows that, if the source is at one end of the

vocal tract (as occurs for voiced sounds), and if the vocal tract does not have any side branches, then the transfer function of the vocal tract is all-pole. Although these conditions do not exactly hold for all speech sounds, the all-pole model is a reasonable approximation that gives useful, mathematically-simple results. The all-pole model is particularly useful for modeling formant frequencies of speech, which are very important perceptually.

The second assumption of linear prediction can be justified for voiced sounds by noting that the source signal can be considered as the output of a filter excited by a periodic train of impulses (Fig. 1b). This filter can be combined with the vocal-tract transfer function and the radiation characteristics to generate the simplified model of Fig. 1c: Voiced speech sounds are now the output of a combined filter excited by a periodic train of impulses. For voiceless sounds, no changes are necessary to the model of Fig. 1a because the noise source is nearly white over the frequency range of interest. For both voiced and voiceless sounds, the important characteristic of the model of Fig. 1c is that the spectrum of the source has a flat envelope,[1] so that any frequency dependence in the spectral envelope of speech must be due to the filter, rather than the source.

If the all-pole model of Fig. 1c holds, the speech signal $s[n]$ can be generated from the source signal $u[n]$ by means of a purely-recursive difference equation:

$$s[n] \;=\; \sum_{k=1}^{p} a_k \; s[n-k] \;+\; G \; u[n] \tag{8.1a}$$

The frequency response of the all-pole filter is

$$H(f) \;=\; \frac{G}{1 \;-\; \sum_{k=1}^{p} a_k \; e^{-j2\pi kf}} \tag{8.1b}$$

Our goal is to estimate the filter coefficients $a_k,\ 1 \leq k \leq p$ and the *gain* G from the speech signal $s[n]$, assuming (for the moment) that the *model order* p is known. If the source signal $u[n]$ were known, this would be a simple problem that could be solved, for example, by division in the frequency domain. However, in most applications, the source signal is not available, and the model parameters have to be determined from the speech signal $s[n]$ alone. While such *blind deconvolution* problems cannot be solved in general, this is possible in the case of speech because of the additional assumptions that the filter is all-pole and that the source has flat spectral characteristics.

8.1.2 Relation of linear prediction to the all-pole model

In order to understand why linear prediction provides an estimate of the all-pole model parameters, it helps to assume that the source signal $u[n]$ in (8.1a) is small. The speech signal is then approximately a weighted sum of its p past values:

$$s[n] \;\approx\; \sum_{k=1}^{p} a_k \; s[n-k] \qquad\qquad \text{if } u[n] \;\approx\; 0 \tag{8.2}$$

[1] Recall that the transform of a periodic impulse train is a periodic train of impulses in frequency, which has a flat spectral envelope.

The assumption of small $u[n]$ seems plausible for voiced sounds because periodic impulse trains are in fact zero for most times. We are thus led to introduce a *linear predictor* $\hat{s}[n]$ of the speech signal from its p past values:

$$\hat{s}[n] \;\stackrel{\triangle}{=}\; \sum_{k=1}^{p} \hat{a}_k \; s[n-k] \tag{8.3}$$

The $\hat{a}_k,\ 1 \le k \le p$ are called the *predictor coefficients.* They can be considered as estimates of the true parameters $a_k,\ 1 \le k \le p$ of the all-pole model. The *prediction error signal* $e[n]$ is the difference between the actual signal and the predictor:

$$e[n] \;\stackrel{\triangle}{=}\; s[n] \;-\; \hat{s}[n] \;=\; s[n] \;-\; \sum_{k=1}^{p} \hat{a}_k \; s[n-k] \tag{8.4}$$

The error signal is the response of an FIR filter to the speech signal $s[n]$. The frequency response of this filter is

$$\hat{A}(f) \;=\; 1 \;-\; \sum_{k=1}^{p} \hat{a}_k \; e^{-j2\pi kf} \tag{8.5}$$

If the predictor coefficients $\hat{a}_k,\ 1 \le k \le p$ were equal to the actual coefficients $a_k,\ 1 \le k \le p$ of the all-pole filter $H(f)$, one would then have:

$$e[n] \;=\; G \; u[n] \tag{8.6a}$$

and

$$H(f) = \frac{\hat{G}}{A}(f), \qquad \text{if} \;\; \hat{a}_k = a_k \;\; \text{for} \;\; 1 \le k \le p$$

For this reason, the prediction-error filter $\hat{A}(f)$ is also called the *inverse filter* for the all-pole model filter $H(f)$. The relationship between the source signal $u[n]$, the speech signal $s[n]$ and the error signal $e[n]$ is shown in Fig. 2a.

The linear-prediction problem consists in finding the predictor coefficients $\hat{a}_k,\ 1 \le k \le p$ that minimize the energy E_e in the error signal. The basic *assumption* of linear-prediction analysis is that such minimization yields estimates of the true $a_k,\ 1 \le k \le p$ coefficients of the all-pole model. As argued above, this assumption seems reasonable for voiced sounds because the voicing source $u[n]$ is an impulse train which is zero for most times. In fact, it can be shown that, if the all-pole model holds, and if the source signal is either stationary white noise or a unit sample, then minimization of the energy in the prediction error signal does give the correct coefficients a_k of the all-pole model. Even though these conditions are not exactly verified, this minimization criterion is still justified in that it provides results that are both useful and computationally simple.

The all-pole filter $\hat{H}(f)$ obtained by replacing the $a_k,\ 1 \le k \le p$ in (8.1b) by the optimal predictor coefficient $\hat{a}_k,\ 1 \le k \le p$ is called the *LP model filter:*

$$\hat{H}(f) \;\stackrel{\triangle}{=}\; \frac{\hat{G}}{\hat{A}(f)} \;=\; \frac{\hat{G}}{1 \;-\; \sum_{k=1}^{p} \hat{a}_k \; e^{-j2\pi kf}} \tag{8.7}$$

The significance of the model filter is that it can be used to generate synthetic speech $s\prime[n]$ by filtering a synthetic source signal $u\prime[n]$ (Fig. 2b). This is the principle for the synthesis stage in analysis-synthesis systems based on linear prediction.

Minimization of the energy in the prediction error signal can be implemented by several methods that differ slightly in their assumptions. These differences arise from the fact that, because the all-pole filter is time-varying, the energy cannot be minimized over the entire duration of the speech signal, but separate minimizations must be carried out for different short-time segments or *frames* of the signal. The exact manner in which these frames are defined and assumptions about the behavior of the signal outside of the frame yield somewhat different methods of linear prediction. Every method gives, for each frame, optimal prediction coefficients $\hat{a}_k,\ 1 \leq k \leq p$, an inverse filter $\hat{A}(f)$ and a model filter $\hat{H}(f)$. We give here one specific implementation, called the *autocorrelation method* of linear prediction, which has the advantage of always giving stable solutions. Alternative techniques are described by Makhoul (1975) and Marple (1987).

8.2 Autocorrelation method of linear prediction

8.2.1 Deterministic autocorrelation functions

The autocorrelation method of linear prediction takes its name from the *autocorrelation function.* Broadly speaking, an autocorrelation function measures the similarity between a signal $x[n]$ and a delayed version of itself $x[n-k]$. There exist different definitions of autocorrelation functions, each one best suited for a particular type of signal. The definition applicable to stationary random signals is introduced in Chapter 11. Such true autocorrelation functions play a key role in predicting the response of linear systems to random signals, and are also important for detecting unknown periodicities in noisy signals. Because we are dealing here with finite-duration speech frames rather than stationary signals, the appropriate form of autocorrelation is the *deterministic,* or "raw" autocorrelation function:

$$\tilde{R}_x[k] \ \stackrel{\triangle}{=} \ \sum_{n=-\infty}^{\infty} \ x[n] \ x[n-k] \tag{8.9}$$

Even though the sum in (8.9) goes from $-\infty$ to $+\infty$, it is in practice a sum of the finite duration of a speech frame, typically 10-40 msec. The deterministic autocorrelation function has three important properties:

1. The deterministic autocorrelation function evaluated at lag zero is the energy in the signal:

$$\tilde{R}_x[0] \ = \ \sum_{n=-\infty}^{\infty} \ x[n]^2 \ \stackrel{\triangle}{=} E_x \tag{8.10}$$

2. The autocorrelation is an even function of lag, i.e. it is symmetric with respect to the origin:

$$\tilde{R}_x[-k] \ = \ \sum_{n=-\infty}^{\infty} \ x[n] \ x[n+k] \ = \ \sum_{m=-\infty}^{\infty} \ x[m-k] \ x[m] \ = \ \tilde{R}_x[k] \tag{8.11}$$

where we made the change of variable $m \ = \ n+k$.

3. The deterministic autocorrelation function is always maximum at the origin:

$$|\tilde{R}_x[k]| \leq \tilde{R}_x[0] \;=\; E_x \tag{8.12}$$

The latter result is another form of the Cauchy-Schwarz inequality.

8.2.2 The Yule-Walker equations

The autocorrelation method of linear prediction consists in minimizing the total energy E_e in the error signal:

$$E_e \;\stackrel{\triangle}{=}\; \sum_{n=-\infty}^{\infty} e[n]^2 \;=\; \sum_{n=-\infty}^{\infty} (s[n] \;-\; \hat{s}[n])^2 \tag{8.17}$$

Again, because speech is time-varying, this assumption is meaningful only if the signal $s[n]$ in (8.17) is a short-time frame. The autocorrelation method gives best results when each frame is multiplied by a window tapered at both ends (e.g. Hamming) to avoid large prediction errors near the ends of the frame.

Replacing the predictor $\hat{s}[n]$ in (8.17) by its value from (8.11), we obtain

$$E_e \;=\; \sum_{n=-\infty}^{\infty} \left[s[n] \;-\; \sum_{k=1}^{p} \hat{a}_k \; s[n-k] \right]^2 \tag{8.18}$$

In order to minimize (8.18), we set to zero the partial derivatives of E_e with respect to the $\hat{a}_k$:

$$\frac{\partial E_e}{\partial \hat{a}_k} \;=\; -2 \sum_{n=-\infty}^{\infty} s[n-k] \left[s[n] \;-\; \sum_{l=1}^{p} \hat{a}_l \; s[n-l] \right] \;=\; 0 \qquad \text{for } 1 \leq k \leq p$$

Interchanging the order of summations over n and l, and noting that

$$\sum_{n=-\infty}^{\infty} s[n-k] \; s[n-l] \;=\; \tilde{R}_s[k-l],$$

the deterministic autocorrelation function of the windowed signal $s[n]$, we obtain:

$$\sum_{l=1}^{p} \hat{a}_l \tilde{R}_s[k-l] = \tilde{R}_s[k] \qquad \text{for } 1 \leq k \leq p \tag{8.19a}$$

This gives a set of p linear equations to solve for the p predictor coefficients $\hat{a}_k$, $1 \leq k \leq p$. Thus, to derive the optimum predictor coefficients, it suffices to know the deterministic autocorrelation function $\tilde{R}_s[k]$ for $0 \;\leq k \leq\; p$. By expressing the energy in (8.18) as a function of the $\tilde{R}_s[k]$, it can be shown that, if the system of equations (8.19a) is satisfied, the prediction error becomes:

$$E_e \;=\; \tilde{R}_s[0] \;-\; \sum_{k=1}^{p} \hat{a}_k \; \tilde{R}_s[k] \tag{8.19b}$$

Equations (8.19a) and (8.19b) can be combined into a single set of $p+1$ linear equations written in matrix notation:

$$\begin{bmatrix} \tilde{R}_s[0] & \tilde{R}_s[1] & \tilde{R}_s[2] & \ldots & \tilde{R}_s[p] \\ \tilde{R}_s[1] & \tilde{R}_s[0] & \tilde{R}_s[1] & \ldots & \tilde{R}_s[p-1] \\ \tilde{R}_s[2] & \tilde{R}_s[1] & \tilde{R}_s[0] & \ldots & \tilde{R}_s[p-2] \\ & & \ldots & \ldots & \\ \tilde{R}_s[p] & \tilde{R}_s[p-1] & \tilde{R}_s[p-2] & \ldots & \tilde{R}_s[0] \end{bmatrix} \begin{bmatrix} 1 \\ -\hat{a}_1 \\ -\hat{a}_2 \\ \ldots \\ -\hat{a}_p \end{bmatrix} \;=\; \begin{bmatrix} E_e \\ 0 \\ 0 \\ \ldots \\ 0 \end{bmatrix} \tag{8.20}$$

These equations are known as the (deterministic) *Yule-Walker equations.* These equations contain all information needed to solve the linear prediction problem. Note the unusual position of the unknown E_e on the right hand side of the first equation.

8.2.3 Determination of the gain

In order to obtain a complete specification of the linear-prediction model, it is necessary to estimate the gain G in addition to the prediction coefficients $\hat{a}_k,\ 1 \leq k \leq p$. For this purpose, remember from (8.15a) that, if the predictor coefficients $\hat{a}_k,\ 1 \leq k \leq p$ were exactly equal to the true filter coefficients $a_k,\ 1 \leq k \leq p$, then the error signal $e[n]$ would be equal to $G\ u[n]$. Further, the energy in the source signal $u[n]$ is always unity by definition. This convention is useful for synthesis. Therefore, if the model were exactly verified, the energy in the error signal E_e would be equal the energy in $G\ u[n]$, which is G^2. This consideration leads us to conclude that the gain estimate $\hat{G}$ is the square-root of the prediction error E_e:

$$\hat{G} \ = \ \sqrt{E_e} \tag{8.21}$$

8.2.4 Example

As an example, we will solve the Yule-Walker equations for a model of order 1. These equations are:

$$\begin{bmatrix} \tilde{R}_s[0] & \tilde{R}_s[1] \\ \tilde{R}_s[1] & \tilde{R}_s[0] \end{bmatrix} \begin{bmatrix} 1 \\ -\hat{a}_1 \end{bmatrix} \ = \ \begin{bmatrix} E_e \\ 0 \end{bmatrix}$$

The solutions are:

$$\hat{a}_1 \ = \ \frac{\tilde{R}_s[1]}{\tilde{R}_s[0]}$$

$$E_e \ = \ \frac{\tilde{R}_s[0]^2 \ - \ \tilde{R}_s[1]^2}{\tilde{R}_s[0]}$$

As always, the gain $\hat{G}$ is the square root of E_e. The model filter,

$$\hat{H}(f) \ = \ \frac{\hat{G}}{1 \ - \ \hat{a}_1\ e^{-j2\pi f}}$$

is stable so long that $|\hat{a}_1| < 1$, implying $|\tilde{R}_s[1]| < \tilde{R}_s[0]$. From (8.12), this condition is always verified. It can be shown that, for the autocorrelation method, this condition is met regardless of model order.

8.2.5 Efficient solution by the Levinson-Durbin algorithm

The $(p+1)\times(p+1)$ matrix in (8.20) is said to be *Toeplitz* because all terms along its diagonals are the same. Because of this special structure, the Yule-Walker equations can be solved recursively by the highly efficient *Levinson-Durbin* algorithm. This method provides solutions for all models of order $i < p$ before giving the solution for order p. This is useful when the model order is

not known a priori, so that different orders may be tried out before settling on a final one. In describing the Levinson-Durbin algorithm, the notation $\hat{a}_k^{(i)}$ is used to refer to the k^{th} optimal predictor coefficient for a model of order i. Similarly, $E_e^{(i)}$ refers to the energy in the prediction error signal for a model of order i. The recursion is started by setting $E_e^{(0)} = \tilde{R}_s[0]$. As shown in the Appendix, the recursive formulas for deriving solutions for model order i going from those for order $i-1$ are:

$$\hat{a}_i^{(i)} = k_i \triangleq \frac{\tilde{R}_s[i] - \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \tilde{R}_s[i-k]}{E_e^{(i-1)}} \tag{8.22a}$$

$$\hat{a}_k^{(i)} = \hat{a}_k^{(i-1)} - k_i \hat{a}_{i-k}^{(i-1)} \qquad \text{for } 1 \le k \le i-1 \tag{8.22b}$$

$$E_e^{(i)} = (1 - k_i^2) E_e^{(i-1)} \tag{8.22c}$$

It is apparent that knowing the *reflection coefficients* k_i for all orders $1 \le i \le p$ completely specifies the predictor coefficients $\hat{a}_k$, $1 \le k \le p$ for order p. Due to the special properties of the autocorrelation matrix in (8.20), it can be shown that the k_i are always between -1 and 1. Together with (8.22c), this implies that the prediction error $E_e^{(i)}$ always decreases when the model order is increased. It can also be shown that the condition $-1 < k_i < 1$ guarantees that the model filter $\hat{H}(f)$ is stable.

8.2.6 Choice of model order

We have assumed so far that the model order p is known *a priori.* This is rarely the case in practice, and the choice of model order is often a crucial question in linear prediction. An empirical method for choosing p (which is easily implemented by the Levinson-Durbin algorithm) is to track the energy in the error signal $E_e^{(p)}$ as a function of p, and stop increasing the order when the energy reaches a plateau. For a signal of length N, this method can be formalized using *Akaike's information criterion:*

$$AIC(p) \triangleq \ln \frac{E_e^{(p)}}{E_e^{(0)}} + \frac{2p}{N} \tag{8.23}$$

The value of p which minimizes (8.23) is chosen as the model order. The first term in the AIC represents how well the model fits the data, and decreases monotonically as the model order is increased. The second term $2p/N$ is a penalty factor for increasing the model order. This penalty is introduced because, given a finite signal $s[n]$ of length N, it is always possible to exactly fit its energy spectrum $|S(f)|^2$ by an all-pole model if p approaches N. Thus, minimizing the AIC represents a compromise between getting a better fit to the data and using no more parameters than can be justified on the basis of the available data.

For speech signals, an alternative method for determining the model order is to use knowledge about speech-production mechanisms. The resonant frequencies of a typical vocal tract are expected to be separated by intervals of approximately 1 kHz for a male voice. Thus, if the signal is sampled at 10 kHz, there should be about five resonances within the 5-kHz range of frequency analysis. One needs two predictor coefficients (complex conjugate poles) to model each of these resonances, plus approximately 4 coefficients to model the voicing source spectrum and the radiation characteristics. In fact, linear prediction of speech with a model of order

14 does give useful results for voiced speech sampled at 10 kHz. For example, Figure 3 shows results of a 14-th order linear prediction analysis by the autocorrelation method for a 20-ms windowed segment of a vowel produced by a male speaker. Consistent with the assumptions of the all-pole model of speech production, the prediction error signal shows two peaks, one for each pitch period, and its spectrum has a flat envelope. For voiceless sounds, not all resonances of the vocal tract are excited by the source, so that linear predictions of orders 8 to 10 usually suffice for speech sampled at 10 kHz. Lower model orders are also suitable if the sampling rate is less than 10 kHz because the number of resonances in the modeled frequency band decreases.

8.3 Frequency-domain interpretation of linear prediction

Further insight into linear prediction can be gained by frequency-domain analysis. In fact, in many applications, linear prediction is used to obtain a smooth spectral representation of speech rather than to explicitly separate the source from the filter. This usage is consistent with the more general use of autoregressive models in spectral estimation (Sec. 8.5).

Parseval's theorem gives an expression for the energy E_e in the error signal $e[n]$ in terms of its Fourier transform $E(f)$:

$$E_e \;=\; \sum_{-\infty}^{\infty} e[n]^2 \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} |E(f)|^2 \, df \tag{8.24}$$

Because $e[n]$ is generated by passing the signal $s[n]$ through the inverse filter $\hat{A}(f)$, i.e. $e[n] = \hat{a}[n] \;*\; s[n]$, one has:

$$E(f) \;=\; \hat{A}(f)\; S(f) \tag{8.25}$$

where $S(f)$ is the Fourier transform of $s[n]$ (which is guaranteed to exist because $s[n]$ is a finite-duration speech frame). Therefore, the prediction error becomes:

$$E_e \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{A}(f)|^2 \; |S(f)|^2 \, df \tag{8.26}$$

If we recall from (8.16) that the frequency response of the model filter is

$$\hat{H}(f) \;=\; \frac{\hat{G}}{\hat{A}(f)},$$

we obtain

$$E_e \;=\; \hat{G}^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{|S(f)|^2}{|\hat{H}(f)|^2} \, df \tag{8.27}$$

Thus, minimizing the energy in the error signal is equivalent to minimizing the integral over frequency of the ratio of the energy spectrum of the speech segment to the energy spectrum of the model filter. It can be shown that, when the model order p approaches the signal length N, the energy spectrum $|\hat{H}(f)|^2$ for the model filter approaches the energy spectrum $|S(f)|^2$ of the original signal*. [2] In practice, the closeness of the approximation of the signal spectrum by the all-pole model can be controlled by varying the model order p (Fig. 4). For low model orders,

[2]*Note however that the predicted signal $\hat{s}[n]$ does not necessarily approach the actual signal $s[n]$ because the phase responses can differ.

the model spectrum is too broad to adequately represent all the spectral peaks corresponding to formant frequencies. In contrast, if the order is too high, the model spectrum begins to match the fine structure of the source spectrum, such as individual harmonics of the fundamental frequency. Thus, there exists an optimal model order for which the model spectrum approximates the true filter spectrum with little contamination from the source spectrum. As argued above, this optimal order is approximately 14 for voiced speech sampled at 10 kHz.

It should also be noted that, because the quantity that is being minimized in (8.27) is a *ratio* of the actual spectrum to the model spectrum, the approximation provided by the all-pole model is better when the signal spectrum shows a peak than when it shows a valley, i.e. linear prediction provides a good match to the *spectral envelope.* This property, which is apparent in Figs. 3 and 4, is desirable in speech analysis because spectral peaks are perceptually more important than spectral valleys. This perceptual unimportance of spectral valleys is due to the masking properties of the ear.

8.4 Applications of linear prediction

8.4.1 Spectral analysis

One of the most common application of linear prediction is generating smooth spectral representations of short-time segments of speech. To the extent that the all-pole model of speech production is valid, these smooth spectra represent the frequency response of the filter in the speech production model, and are therefore suitable for estimating the formant frequencies, for example by determining local maxima in the linear-prediction spectrum. Tracking of formant frequencies throughout an utterance can be achieved by measuring linear-prediction spectra for successive frames. The spectral-analysis applications of linear prediction are not limited to speech, but are also advantageous for any signal whose spectrum shows sharp resonances.

8.4.2 Analysis-synthesis systems

Linear-prediction vocoders (Fig. (5) are based on the idea that the speech signal can be replaced by a model signal without affecting intelligibility. This model signal $s\prime[n]$ is the response of the all-pole model filter $\hat{H}(f)$ to a synthetic source signal $u\prime[n]$ which is either a periodic train of impulses for voiced speech or white noise for voiceless speech (Fig. 2b). Thus, the model signal is entirely specified by a small number of parameters: the predictor coefficients $\hat{a}_k, 1 \leq k \leq p$, the gain $\hat{G}$, and, for voiced sounds, the frequency of the impulse train (which is the fundamental frequency). For a 20-ms segment of speech sampled at 10 kHz, this represents about 16 parameters, an order of magnitude less than the number of samples (200) in the raw signal. Because signal synthesis requires knowledge of the fundamental frequency of voice, the analysis stage of a linear-prediction vocoder must estimate the fundamental frequency (pitch) in addition to the linear-prediction parameters. This can be done for example by detecting local maxima in the short-time autocorrelation function of the signal. In practice, it is more convenient to estimate the pitch of the prediction error signal $e[n]$ than that of the speech signal $s[n]$ because the error signal usually shows one sharp peak for each pitch period (Fig. 5). Indeed,

the linear-prediction error signal is often used for pitch estimation even in applications that do not involve signal synthesis from linear-prediction parameters.

Speech produced by standard linear-prediction vocoders as described above is intelligible, but shows noticeable degradation compared to natural speech. High-quality linear prediction synthesis is possible by using more realistic source signals than simple periodic pulse trains. In *multipulse LPC vocoders* the source signal can contain multiple pulses in each pitch period. The locations and amplitudes of these pulses are adjusted to find the best match between the synthetic speech and the original speech. In *code-excited LPC vocoders,* the source signal is chosen for each frame among a finite codebook of source waveforms. Optimization techniques are used both in selecting the codebook and in selecting the source waveform from the codebook for each frame. Both techniques provide synthetic speech essentially undistinguishable from natural speech at the price of a moderate increase in bit rate over standard linear prediction.

8.5 Autoregressive spectral estimation

In Chapter 13, we introduce general techniques for estimating the power spectra of stationary random signals. An important limitation of these methods is a trade-off between bias and statistical stability. Estimates of the autocorrelation function need to be windowed in order to obtain stable spectral estimates, On the other hand, short autocorrelation windows lead to biases in spectral estimates, and in particular to underestimation of spectral peaks. As a result of this trade-off, very long data records are needed to reliably estimate the power spectra of signals with sharp spectral features.

Autoregressive spectral estimation is an alternative technique in which an all-pole model is fit to the data sample. Because this technique is model-based, its limitations differ from those the traditional techniques of Chapter 11. Autoregressive spectral estimation is particularly advantageous for signals that have sharp spectral peaks because it can reliably estimate these peaks based on relatively short data records. Mathematically, autoregressive spectral estimation (a.k.a. maximum-entropy spectral estimation) is equivalent to linear prediction, but the emphasis is more on getting a reliable spectral estimate than on deconvolution of a source signal from a filter. Historically, autoregressive spectral estimation was developed first by statisticians (in the early 1900's). Its relationship to linear prediction only became widely recognized in the 1970's.

8.5.1 The Yule-Walker equations

In autoregressive spectral estimation, a stationary random signal $x[n]$ is modelled as the output of an all-pole (autoregressive) filter excited by white noise $w[n]$:

$$x[n] \;=\; \sum_{k=1}^{p} a_k \; x[n-k] \;+\; w[n] \tag{8.28}$$

The white noise is further assumed to be zero-mean, with variance σ_w^2. The frequency response of the all-pole filter is given by:

$$H(f) \;=\; \frac{1}{1 \;-\; \sum_{k=1}^{p} \; a_k \; e^{-j2\pi fk}} \tag{8.29}$$

Therefore, according to the autoregressive model, the power spectrum of $x[n]$ is given by:

$$S_x(f) \;=\; |H(f)|^2 \; S_w(f) \;=\; \frac{\sigma_w^2}{\left|\; 1 \;-\; \sum_{k=1}^{p} \; a_k \; e^{-j2\pi fk} \;\right|^2} \tag{8.30}$$

To estimate the power spectrum $S_x(f)$, it suffices to determine the filter coefficients $a_k,\ 1 \leq k \leq p$, and the variance σ_w^2. This can be done by writing the autocorrelation function of $x[n]$:

$$R_x[k] =< x[n] \; x[n-k] >=< \left(\sum_{l=1}^{p} \; a_l \; x[n-l] \;+\; w[n] \right) \; x[n-k] > \tag{8.31}$$

Applying the linearity and time-invariance properties of time averages, this becomes:

$$R_x[k] \;=\; \sum_{l=1}^{p} \; a_l \; R_x[k-l] \;+\; R_{xw}[k] \qquad\qquad \text{for } 0 \leq k \leq p \tag{8.32}$$

Any sample of the white noise $w[n]$ is uncorrelated with its past values, and therefore with past values of $x[n]$, which are themselves weighted sums of past values of $w[n]$. Therefore,

$$R_{xw}[k] = \begin{bmatrix} \sigma_w^2 & \text{if } k = 0 \\ 0 & \text{if } k > 0 \end{bmatrix} \tag{8.33}$$

Replacing $R_{xw}[k]$ by its value in (8.32) for $0 \leq k \leq p$ gives a set of $p+1$ linear equations in the $p+1$ unknowns $a_k,\ 1 \leq k \leq p$, and σ_w^2:

$$\begin{bmatrix} R_x[0] & R_x[1] & R_x[2] & ... & R_x[p] \\ R_x[1] & R_x[0] & R_x[1] & ... & R_x[p-1] \\ R_x[2] & R_x[1] & R_x[0] & ... & R_x[p-2] \\ & & ... & ... & \\ R_x[p] & R_x[p-1] & R_x[p-2] & ... & R_x[0] \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ -a_2 \\ ... \\ -a_p \end{bmatrix} \;=\; \begin{bmatrix} \sigma_w^2 \\ 0 \\ 0 \\ ... \\ 0 \end{bmatrix} \tag{8.34}$$

This system of linear equations is called the (stochastic) *Yule-Walker equations.* They are equivalent to the deterministic form of the Yule-Walker equations (8.20) if the deterministic autocorrelation function $\tilde{R}_s[k]$ is substituted for the true autocorrelation function $R_x[k]$, and the prediction error E_e for the variance of the white noise σ_w^2. Therefore, they can also be solved using the computationally-efficient Levinson-Durbin algorithm.

As argued above, the principal advantage of autoregressive spectral estimation over the conventional spectral estimation techniques described above is that it can provide good frequency resolution with short data records. One disadvantage of this technique is that it is relatively computationally intensive. A further disadvantage is that results can be hard to interpret, partly because the estimate depends appreciably on the exact manner in which the autocorrelation function is estimated, partly because the model order p is usually not known a priori, and partly because the autoregressive model does not apply equally well to all signals.

11

8.5.2 Estimation of the autocorrelation function

In practice, the true autocorrelation function $R_x[k]$ is rarely known, so that it has has to be estimated from a finite data sample. When the $R_x[k]$ are replaced by appropriate estimates in the Yule-Walker equations (8.34), one obtains estimates for the filter coefficients $\hat{a}_k,\ 1 \leq k \leq p$ and the variance $\hat{\sigma}_w^2$. Substituting these estimates for their expected values in (8.30) gives the autoregressive spectral estimate $\hat{S}_x(f)$.

Several alternative methods can be used for estimating the autocorrelation function for lags $0 \leq k \leq p$ from a data record of length N. The most straightforward method is to use the unbiased estimate $\hat{R}_x[k]$ introduced in Chapter 13:

$$\hat{R}_x[k] \ = \ \frac{1}{N-|k|} \sum_{n=|k|}^{N-1} x[n]\ x[n-|k|] \tag{8.35}$$

Use of unbiased autocorrelation estimates can lead to high-resolution spectral estimates because no assumptions are made about the data outside of the available range. On the other hand, for small data records, this estimate will occasionally yield an unstable filter transfer function, and therefore an undefined spectral estimate $\hat{S}_x(f)$. It can be shown that use of the *biased* autocorrelation estimate

$$\hat{R}_x^b[k] \ = \ \frac{N-|k|}{N}\ \hat{R}_x[k] \tag{8.36}$$

guarantees the convergence of the autoregressive spectral estimate, at the price of some loss in frequency resolution. The book by Marple (1987) describes variants of autoregressive spectral estimation that provide both stability of the model filter and fine frequency resolution. The book also describes efficient techniques for solving the Yule-Walker equations.

8.6 Summary

Linear prediction is a form of spectral analysis which fits an all-pole (autoregressive) model to the data. This technique is well suited to speech signals because they are well approximated by the response of an all-pole filter to either a white-noise source or a periodic impulse train. Unlike spectral-analysis techniques based on the DFT, linear prediction can, in principle, separate the contributions of the source and the filter to the speech signal.

The best-fitting all-pole model is found by minimizing the energy in the error signa, the difference between the current signal sample and a linear predictor based on a small number of past signal samples. Minimization of this energy yields a set of linear equations for the predictor coefficients. In the frequency domain, this is equivalent to minimization of the integral over frequency of the ratio of the signal spectrum to the spectrum of the all-pole model filter. Thus, linear-prediction spectra provide a good fit to the envelope of the signal spectrum. Practically speaking, the key property of linear prediction is that it provides a representation of speech signals in terms of a small number of parameters that preserve essential information for intelligibility.

8.7 Further reading

Rabiner and Schafer, Chapter 8, Sections 1, 2, 6.
Makhoul, Proc. IEEE 63, 561-580 (1975).
Marple, Digital spectral analysis with applications. Prentice-Hall (1987), Chapters 6-8.
Quatieri, Chapter 5; Chapter 12, Section 6.

8.8 Appendix: Proof of the Levinson-Durbin algorithm

Matrix notation helps in deriving the Levinson-Durbin algorithm given in Equations (8.22a-c). We define $\hat{\mathbf{a}}^{(\mathbf{i})}$ as the column vector of coefficients on the left-hand side of (8.20) for a model of order i

$$\hat{\mathbf{a}}^{(i)} \triangleq [1 \quad -\hat{a}_1^{(i)} \quad -\hat{a}_2^{(i)} \quad \ldots \quad -\hat{a}_i^{(i)}]^T, \tag{8.A.1}$$

$\hat{\mathbf{a}}_{\mathbf{r}}^{(\mathbf{i})}$ as this same vector in reverse order

$$\hat{\mathbf{a}}_{\mathbf{r}}^{(\mathbf{i})} \triangleq [-\hat{a}_i^{(i)} \quad -\hat{a}_{i-1}^{(i)} \quad \ldots \quad -\hat{a}_1^{(i)} \quad 1\,]^T, \tag{8.A.2}$$

and $\mathbf{R}^{(\mathbf{i})}$ as the $(i+1) \times (i+1)$ Toeplitz correlation matrix on the left-hand side of (8.20). It can be readily verified that

$$\mathbf{R}^{(\mathbf{i})} \begin{bmatrix} \hat{\mathbf{a}}^{(\mathbf{i-1})} \\ 0 \end{bmatrix} = \begin{bmatrix} E_e^{(i-1)} \\ \ldots \\ \gamma_{i-1} \end{bmatrix} \tag{8.A.3a}$$

$$\mathbf{R}^{(\mathbf{i})} \begin{bmatrix} 0 \\ \hat{\mathbf{a}}_{\mathbf{r}}^{(\mathbf{i-1})} \end{bmatrix} = \begin{bmatrix} \gamma_{i-1} \\ \ldots \\ E_e^{(i-1)} \end{bmatrix} \tag{8.A.3b}$$

where γ_{i-1} is equal to

$$\gamma_{i-1} \triangleq \begin{bmatrix} \tilde{R}_s[i] & \tilde{R}_s[i-1] & \ldots & \tilde{R}_s[2] & \tilde{R}_s[1] \end{bmatrix} \hat{\mathbf{a}}^{(\mathbf{i-1})} = \tilde{R}_s[i] - \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \tilde{R}_s[i-k]$$

Since both vectors on the right hand side of (8.A.3) have only two non-zero elements, we can form a linear combination for which the last element equal to 0. Specifically, if we define

$$k_i \triangleq \frac{\gamma_{i-1}}{E_e^{(i-1)}},$$

we find that

$$\begin{bmatrix} E_e^{(i-1)} \\ \ldots \\ \gamma_{i-1} \end{bmatrix} - k_i \begin{bmatrix} \gamma_{i-1} \\ \ldots \\ E_i^{(i-1)} \end{bmatrix} = \begin{bmatrix} (1-k_i^2)E_e^{(i-1)} \\ \ldots \\ 0 \end{bmatrix}.$$

Therefore, if we have the solution $\hat{\mathbf{a}}^{(\mathbf{i-1})}$ and $E_e^{(i-1)}$ for model order $i-1$, we can compute the solution for model order i as

$$k_i = \frac{1}{E_e^{(i-1)}} \left(\tilde{R}_s[i] - \sum_{k=1}^{i-1} \hat{a}_k^{(i-1)} \tilde{R}_s[i-k] \right) \tag{8.A.4a}$$

$$\hat{\mathbf{a}}^{(\mathbf{i})} = \begin{bmatrix} \hat{\mathbf{a}}^{(\mathbf{i-1})} \\ 0 \end{bmatrix} - k_i \begin{bmatrix} 0 \\ \hat{\mathbf{a}}_{\mathbf{r}}^{(\mathbf{i-1})} \end{bmatrix} \tag{8.A.4b}$$

$$E_e^{(i)} = (1 - k_i^2) E_e^{(i-1)}. \tag{8.A.4c}$$

These equations are the same as (8.22a-c) written in matrix notation.

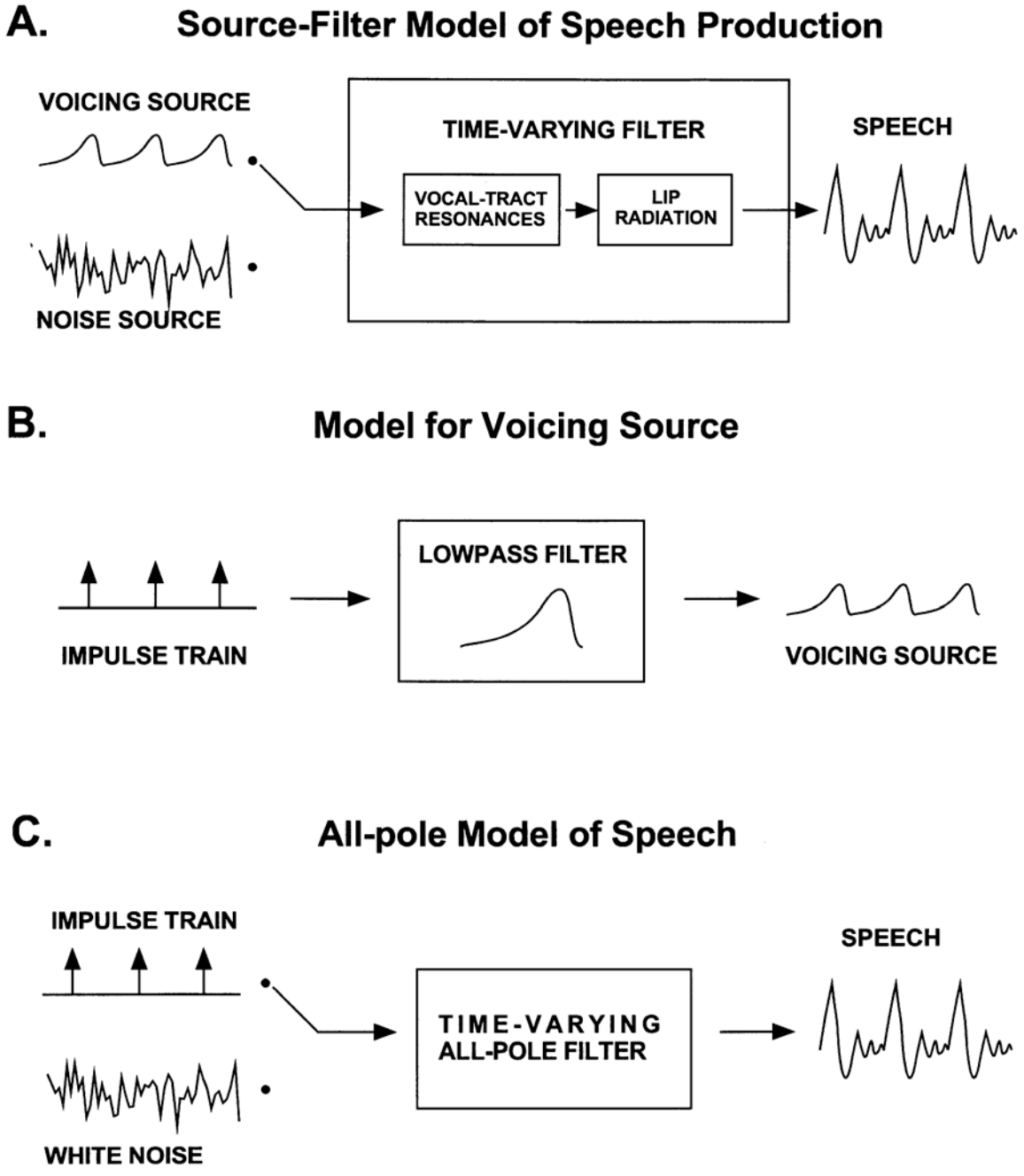

Figure 8.1:

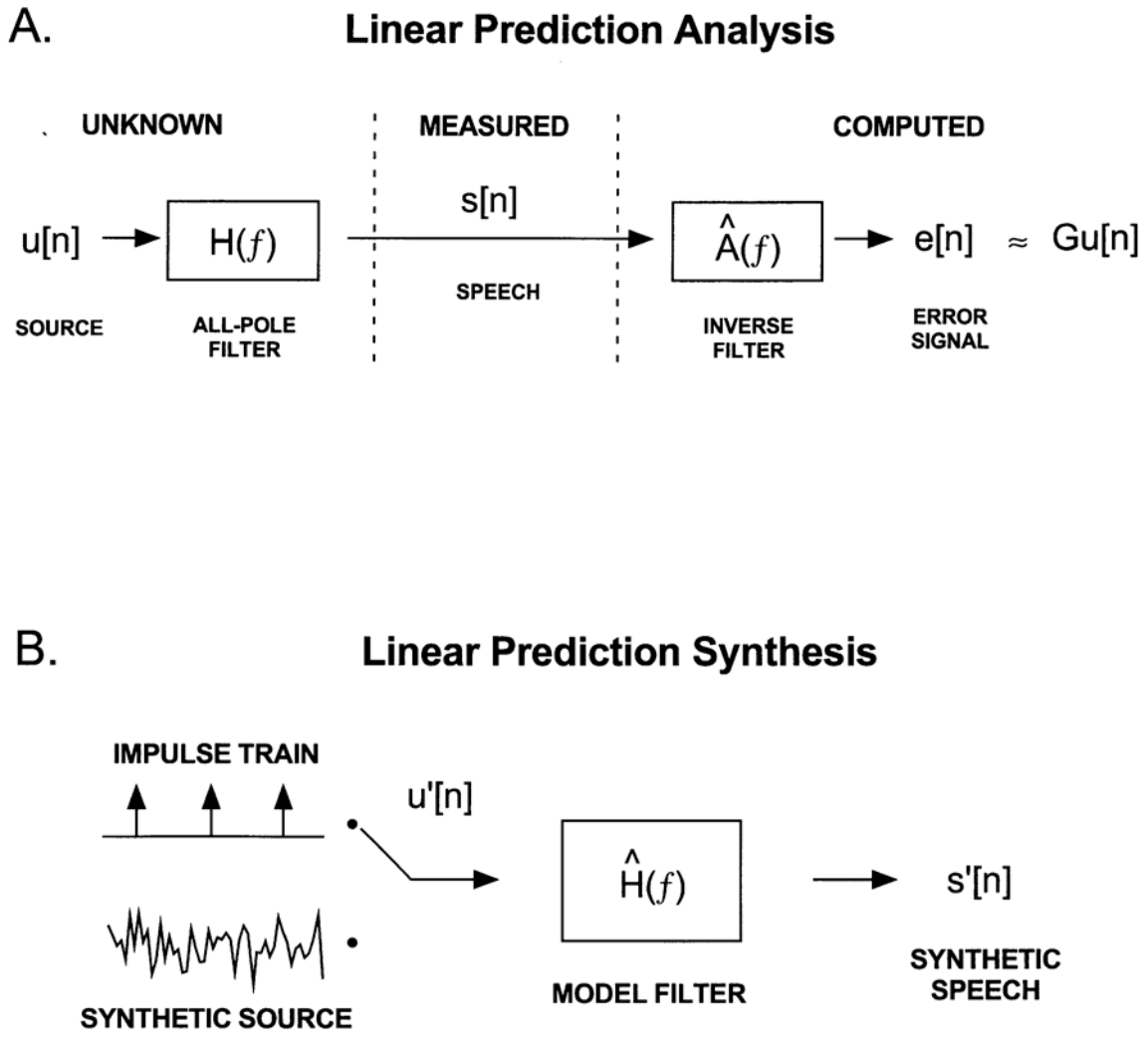

Figure 8.2:

Figure removed due to copyright restrictions.
Examples for speech samples, Fig. 8.7 and 8.9 in Rabiner and Schafer.

Figure 8.3: (from Rabiner and Schafer) Signals and spectra involved in the autocorrelation method of lineaer prediction. a) Windowed speech signal. b) Prediction error signal. c) DFT spectrum and linear-prediction spectrum of the windowed signal. d) Spectrum of the prediction error signal.

Figures removed due to copyright restrictions.
(a) Fig. 8.18 in Rabiner and Shafer.
(b) Fig. 8.17 in Rabiner and Shafer.

Figure 8.4: (from Rabiner and Schafer) a) Linear-prediction spectra of a vowel sampled at $6kHz>$ for several values of the prediction order p. b) 28-th order linear-prediction spectrum and DFT spectrum of a speech sound sampled at $20kHz$.

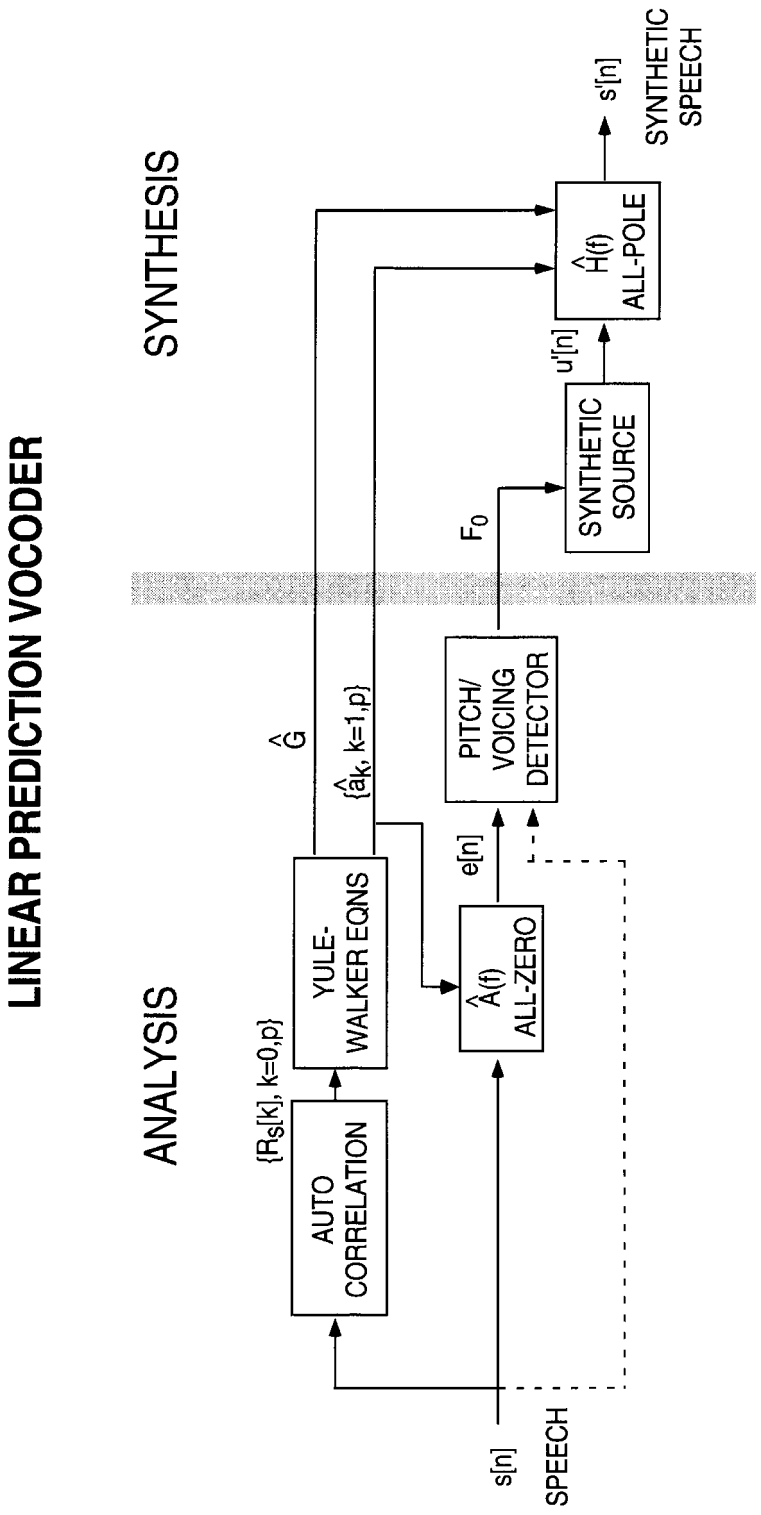

Figure 8.5: