Linear Prediction: A Tutorial Review

JOHN MAKHOUL, MEMBER, IEEE

Invited Paper

Abstract—This paper gives an exposition of linear prediction in the analysis of discrete signals. The signal is modeled as a linear combination of its past values and present and past values of a hypothetical input to a system whose output is the given signal. In the frequency domain, this is equivalent to modeling the signal spectrum by a pole-zero spectrum. The major part of the paper is devoted to all-pole models. The model parameters are obtained by a least squares analysis in the time domain. Two methods result, depending on whether the signal is assumed to be stationary or nonstationary. The same results are then derived in the frequency domain. The resulting spectral matching formulation allows for the modeling of selected portions of a spectrum, for arbitrary spectral shaping in the frequency domain, and for the modeling of continuous as well as discrete spectra. This also leads to a discussion of the advantages and disadvantages of the least squares error criterion. A spectral interpretation is given to the normalized minimum prediction error. Applications of the normalized error are given, including the determination of an "optimal" number of poles. The use of linear prediction in data compression is reviewed. For purposes of transmission, particular attention is given to the quantization and encoding of the reflection (or partial correlation) coefficients. Finally, a brief introduction to pole-zero modeling is given.

I. Introduction

A. Overview

THE MATHEMATICAL analysis of the behavior of general dynamic systems (be they engineering, social, or economic) has been an area of concern since the beginning of this century. The problem has been pursued with accelerated vigor since the advent of electronic digital computers over two decades ago. The analysis of the outputs of dynamic systems was for the most part the concern of "time series analysis," which was developed mainly within the fields of statistics, econometrics, and communications. Most of the work on time series analysis was actually done by statisticians. More recently, advances in the analysis of dynamic systems have been made in the field of control theory based on state-space concepts and time domain analysis.

This paper is a tutorial review of one aspect of time series analysis: linear prediction (defined here). The exposition is based on an intuitive approach, with emphasis on the clarity of ideas rather than mathematical rigor. Although the large body of related literature available on this topic often requires advanced knowledge of statistics and/or control theory concepts, this paper employs no control theory concepts *per se* and only the basic notions of statistics and random processes. For example, the very important statistical concepts of *consistency* and *efficiency* [74], [75] in the estimation of parameters will not be dealt with. It is hoped this paper will serve as a simple introduction to some of the tools used in time series analysis, as well as be a detailed analysis of those aspects of linear prediction of interest to the specialist.

Manuscript received July 21, 1974; revised November, 1974. This work was supported by the Information Processing Techniques Branch of the Advanced Research Projects Agency under Contract DAHC15-17-C-0088.

The author is with Bolt Beranek and Newman, Inc., Cambridge, Mass. 02138.

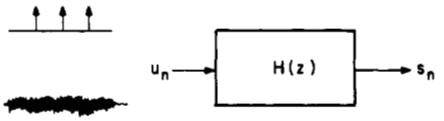

Fig. 1. Discrete speech production model.

B. Current Applications

Before we delve into signal analysis, we shall give three examples of the types of problems that are of current interest.[1] These examples will then serve to illustrate some of the concepts that are developed.

Neurophysics [15], [32], [36], [102]: The spontaneous electrical brain activity is normally measured by means of electrodes placed on the patient's scalp. The recordings, known as electroencephalograms (or EEG signals), show certain periodicities (sharp resonances) accompanied by some randomness. These signals are believed to carry information about the medical status of the brain and are used by physicians as a means of diagnosis. It is of interest to detect the presence, position, and strength of the different resonances, known as rhythms. The three most common rhythms are known as the alpha, beta, and delta rhythms. Therefore, the basic interest here is to describe the spectrum in a simple mathematical manner that would yield the characteristics of the different rhythms.

Geophysics [84]-[87], [114]: In one of the successful methods of oil exploration, a charge of dynamite is exploded in the earth, and the resulting vibrations at various points on the surface of the ground are recorded by a seismograph as seismic traces. The job of the geophysicist is to use these traces in the determination of the structure of the sedimentary rock layers. Such information is then used to decide on the presence of oil in that area. Of interest here are the direct arrival times and strengths of the deep reflections of the explosion, which are then used to determine the layered structure. If somehow one is able to remove (deconvolve) the impulse response of the structure from the seismic trace, the desired arrival times should appear as impulses of different phases and amplitudes.

Speech Communication [10], [33], [47], [50], [51], [62], [68], [89]: In EEG analysis, the spectrum of the recorded signal was of interest. In seismic analysis, the spectral properties of the seismic trace were of interest only to facilitate the deconvolution process in order to obtain the desired impulses. In the analysis of speech, both types of information are of interest.

Fig. 1 shows a rather successful model of speech production. The model consists of a filter that is excited by either a quasi-periodic train of impulses or a random noise source. The periodic source produces voiced sounds such as vowels and

[1] For applications to economic and industrial time series, see for example [17].

nasals, and the noise source produces unvoiced or fricated sounds such as the fricatives (*f*, *th*, *s*, *sh*). The parameters of the filter determine the identity (spectral characteristics) of the particular sound for each of the two types of excitation.

Given a particular speech signal, it is of interest to determine the general type of sound it is, voiced or fricated, and if voiced what the pitch period is (i.e., distance between pitch pulses). In addition, one is interested in the identity of the sound which can be obtained from the spectrum. Such derived information can then be used in an automatic speech recognition system or a speech compression system.

C. Linear Prediction

In applying time series analysis to the aforementioned applications, each continuous-time signal $s(t)$ is sampled to obtain a discrete-time[2] signal $s(nT)$, also known as a time series, where n is an integer variable and T is the sampling interval. The sampling frequency is then $f_s = 1/T$. (Henceforth, we shall abbreviate $s(nT)$ by s_n with no loss in generality.)

A major concern of time series analysis [6], [11], [12], [14], [17], [43], [45], [46], [54], [105], [112] has been the estimation of power spectra, cross-spectra, coherence functions, autocorrelation and cross-correlation functions. A more active concern at this time is that of system modeling. It is clear that if one is successful in developing a parametric model for the behavior of some signal, then that model can be used for different applications, such as prediction or forecasting, control, and data compression.

One of the most powerful models currently in use is that where a signal s_n is considered to be the output of some system with some unknown input u_n such that the following relation holds:

$$s_n = -\sum_{k=1}^{p} a_k \hat{s}_{n-k} + G \sum_{l=0}^{q} b_l u_{n-l}, \qquad b_0 = 1 \tag{1}$$

where a_k, $1 \leqslant k \leqslant p$, b_l, $1 \leqslant l \leqslant q$, and the gain G are the parameters of the hypothesized system. Equation (1) says that the "output" s_n is a linear function of past outputs and present and past inputs. That is, the signal s_n is *predictable* from *linear* combinations of past outputs and inputs. Hence the name *linear prediction*.

Equation (1) can also be specified in the frequency domain by taking the z transform on both sides of (1). If $H(z)$ is the transfer function of the system, as in Fig. 1, then we have from (1):

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

where

$$S(z) = \sum_{n=-\infty}^{\infty} s_n z^{-n} \tag{3}$$

is the z transform of s_n, and $U(z)$ is the z transform of u_n. $H(z)$ in (2) is the general *pole-zero model*. The roots of the numerator and denominator polynomials are the zeros and poles of the model, respectively.

There are two special cases of the model that are of interest:

1) all-zero model: $a_k = 0$, $1 \leqslant k \leqslant p$
2) all-pole model: $b_l = 0$, $1 \leqslant l \leqslant q$.

The all-zero model is known in the statistical literature as the *moving average* (MA) model, and the all-pole model is known as the *autoregressive* (AR) model [17]. The pole-zero model is then known as the *autoregressive moving average* (ARMA) model. In this paper we shall use the pole-zero terminology since it is more familiar to engineers.

The major part of this paper will be devoted to the all-pole model. This has been, by far, the most widely used model. Historically, the first use of an all-pole model in the analysis of time series is attributed to Yule [115] in a paper on sunspot analysis. Work on this subject, as well as on time series analysis in general, proceeded vigorously after 1933 when Kolmogorov laid a rigorous foundation for the theory of probability. Later developments by statisticians, such as Cramér and Wold, culminated in the parallel and independent work of Kolmogorov [58] and Norbert Wiener [107] on the prediction and filtering of stationary time series. For a bibliography on time series through the year 1959, see the encyclopedic work edited by Wold [113]. For a discussion of all-pole (autoregressive) models see, for example, [17], [45], [105], [112].

Much of the recent work on system modeling has been done in the area of control theory under the subjects of estimation and system identification. Recent survey papers with extensive references are those of Åström and Eykhoff [8] and Nieman *et al.* [73]. The December 1974 issue of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL is devoted to the subject of system identification. Another relevant survey paper is that of Kailath [55] on linear filtering theory. Related books are those of Lee [60], Sage and Melsa [88], and Eykhoff [30].

D. Paper Outline

Sections II–V deal exclusively with the all-pole model. In Section II, the estimation of model parameters is derived in the time domain by the method of least squares. The resulting normal equations are obtained for deterministic as well as random signals[3] (both stationary and nonstationary). Direct and iterative techniques are presented for the computation of the predictor coefficients, and the stability of the all-pole filter $H(z)$ is discussed. The response of the all-pole filter is then analyzed for two important types of input excitation: a deterministic impulse and statistical white noise.

In Section III, the all-pole modeling of a signal is derived completely in the frequency domain. The method of least squares translates into a spectral matching method where the signal spectrum is to be matched or fitted by a model spectrum. This formulation allows one to perform arbitrary spectral shaping before modeling. This viewpoint has special relevance today with the availability of hardware spectrum analyzers and fast Fourier transform techniques [21]. (We point out that all-pole modeling by linear prediction is identical to the method of maximum entropy spectral estimation [18], [96].)

Section IV gives a detailed discussion of the advantages and disadvantages of the least squares error criterion. The properties of the normalized error are reviewed. Its use is discussed

[2] See [80] for an exposition of the terminology in digital signal processing.

[3] See [12] for a description of deterministic and random signals.

in measuring the ill-conditioning of the normal equations, and in determining an optimal number of poles.

Section V discusses the use of linear prediction in data compression. Alternate representations of the linear predictor are presented and their properties under quantization are discussed. Particular emphasis is given to the quantization and encoding of the reflection (or partial correlation) coefficients.

Finally, in Section VI, a brief discussion of pole-zero modeling is given, with emphasis on methods presented earlier for the all-pole case.

II. Parameter Estimation

A. All-Pole Model

In the all-pole model, we assume that the signal s_n is given as a linear combination of past values and some input u_n:

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{4}$$

where G is a gain factor. This model is shown in Fig. 2 in the time and frequency domains. The transfer function $H(z)$ in (2) now reduces to an all-pole transfer function

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{5}$$

Given a particular signal s_n, the problem is to determine the predictor coefficients a_k and the gain G in some manner.

The derivations will be given using an intuitive least squares approach, assuming first that s_n is a deterministic signal and then that s_n is a sample from a random process. The results are identical to those obtained by the method of maximum likelihood [6], [74], [75] with the assumption that the signal is Gaussian [60], [73]. The reader is reminded of the existence of more general least squares methods such as weighted and *a priori* least squares [16], [90].

B. Method of Least Squares

Here we assume that the input u_n is totally unknown, which is the case in many applications, such as EEG analysis. Therefore, the signal s_n can be predicted only approximately from a linearly weighted summation of past samples. Let this approximation of s_n be $\tilde{s}_n$, where

$$\tilde{s}_n = -\sum_{k=1}^{p} a_k s_{n-k}. \tag{6}$$

Then the error between the actual value s_n and the predicted value $\tilde{s}_n$ is given by

$$e_n = s_n - \tilde{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k}. \tag{7}$$

e_n is also known as the *residual*. In the method of least squares the parameters a_k are obtained as a result of the minimization of the mean or total squared error with respect to each of the parameters. (Note that this problem is identical to the problem of designing the optimal one-step prediction digital Wiener filter [85].)

The analysis will be developed along two lines. First, we assume that s_n is a deterministic signal, and then we give analogous derivations assuming that s_n is a sample from a random process.

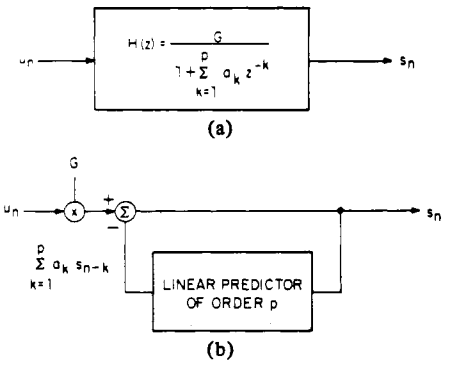

Fig. 2. (a) Discrete all-pole model in the frequency domain. (b) Discrete all-pole model in the time domain.

1) Deterministic Signal: Denote the total squared error by E, where

$$E = \sum_n e_n^2 = \sum_n \left(s_n + \sum_{k=1}^{p} a_k s_{n-k} \right)^2. \tag{8}$$

The range of the summation in (8) and the definition of s_n in that range is of importance. However, let us first minimize E without specifying the range of the summation. E is minimized by setting

$$\frac{\partial E}{\partial a_i} = 0, \qquad 1 \leqslant i \leqslant p. \tag{9}$$

From (8) and (9) we obtain the set of equations:

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = -\sum_n s_n s_{n-i}, \qquad 1 \leqslant i \leqslant p. \tag{10}$$

Equations (10) are known in least squares terminology as the *normal equations*. For any definition of the signal s_n, (10) forms a set of p equations in p unknowns which can be solved for the predictor coefficients $\{a_k, 1 \leqslant k \leqslant p\}$ which minimize E in (8).

The minimum total squared error, denoted by E_p, is obtained by expanding (8) and substituting (10). The result can be shown to be

$$E_p = \sum_n s_n^2 + \sum_{k=1}^{p} a_k \sum_n s_n s_{n-k}. \tag{11}$$

We shall now specify the range of summation over n in (8), (10), and (11). There are two cases of interest, which will lead to two distinct methods for the estimation of the parameters.

a) Autocorrelation method: Here we assume that the error in (8) is minimized over the infinite duration $-\infty < n < \infty$. Equations (10) and (11) then reduce to

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), \qquad 1 \leqslant i \leqslant p \tag{12}$$

$$E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{13}$$

where

$$R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+i} \tag{14}$$

is the autocorrelation function of the signal s_n. Note that $R(i)$ is an even function of i, i.e.,

$$R(-i) = R(i). \tag{15}$$

Since the coefficients $R(i-k)$ form what often is known as an autocorrelation matrix, we shall call this method the *autocorrelation method*. An autocorrelation matrix is a symmetric Toeplitz matrix. (A Toeplitz matrix is one where all the elements along each diagonal are equal [42].)

In practice, the signal s_n is known over only a finite interval, or we are interested in the signal over only a finite interval. One popular method is to multiply the signal s_n by a *window* function w_n to obtain another signal s_n' that is zero outside some interval $0 \leqslant n \leqslant N-1$:

$$s_n' = \begin{cases} s_n w_n, & 0 \leqslant n \leqslant N-1 \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The autocorrelation function is then given by

$$R(i) = \sum_{n=0}^{N-1-i} s_n' s_{n+i}', \qquad i \geqslant 0. \tag{17}$$

The shape of the window function w_n can be of importance. The subject is discussed further in Section III.

b) Covariance method: In contrast with the autocorrelation method, here we assume that the error E in (8) is minimized over a finite interval, say, $0 \leqslant n \leqslant N-1$. Equations (10) and (11) then reduce to

$$\sum_{k=1}^{p} a_k \varphi_{ki} = -\varphi_{0i}, \qquad 1 \leqslant i \leqslant p \tag{18}$$

$$E_p = \varphi_{00} + \sum_{k=1}^{p} a_k \varphi_{0k} \tag{19}$$

where

$$\varphi_{ik} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k} \tag{20}$$

is the covariance of the signal s_n in the given interval. The coefficients φ_{ki} in (18) form a covariance matrix, and, therefore, we shall call this method the *covariance method*. From (20) it can be easily shown that the covariance matrix φ_{ik} is symmetric, i.e., $\varphi_{ik} = \varphi_{ki}$. However, unlike the autocorrelation matrix, the terms along each diagonal are not equal. This can be seen by writing from (20)

$$\varphi_{i+1,k+1} = \varphi_{ik} + s_{-i-1} s_{-k-1} - s_{N-1-i} s_{N-1-k}. \tag{21}$$

Note from (21) also that values of the signal s_n for $-p \leqslant n \leqslant N-1$ must be known: a total of $p+N$ samples. The covariance method reduces to the autocorrelation method as the interval over which n varies goes to infinity.

We point out here that the covariance method is similar to the method of Prony [49], [71] where a signal is approximated by the summation of a set of damped exponentials.

2) Random Signal: If the signal s_n is assumed to be a sample of a random process, then the error e_n in (7) is also a sample of a random process. In the least squares method, we minimize the expected value of the square of the error. Thus

$$E = \mathcal{E}(e_n^2) = \mathcal{E}\left(s_n + \sum_{k=1}^{p} a_k s_{n-k}\right)^2. \tag{22}$$

Applying (9) to (22), we obtain the normal equations:

$$\sum_{k=1}^{p} a_k \mathcal{E}(s_{n-k} s_{n-i}) = -\mathcal{E}(s_n s_{n-i}), \qquad 1 \leqslant i \leqslant p. \tag{23}$$

The minimum average error is then given by

$$E_p = \mathcal{E}(s_n^2) + \sum_{k=1}^{p} a_k \mathcal{E}(s_n s_{n-k}). \tag{24}$$

Taking the expectations in (23) and (24) depends on whether the process s_n is stationary or nonstationary.

a) Stationary case: For a stationary process s_n, we have

$$\mathcal{E}(s_{n-k} s_{n-i}) = R(i-k) \tag{25}$$

where $R(i)$ is the autocorrelation of the process. Equations (23) and (24) now reduce to equations identical to (12) and (13), respectively. The only difference is that here the autocorrelation is that of a stationary process instead of a deterministic signal. For a stationary (and ergodic) process the autocorrelation can be computed as a time average [12]. Different approximations have been suggested in the literature [54] for estimating $R(i)$ from a finite known signal s_n. One such approximation is given by (17).[4] Using this estimate in the stationary case gives the same solution for the coefficients a_k as the autocorrelation method in the deterministic case.

b) Nonstationary case: For a nonstationary process s_n, we have

$$\mathcal{E}(s_{n-k} s_{n-i}) = R(n-k, n-i) \tag{26}$$

where $R(t, t')$ is the nonstationary autocorrelation between times t and t'. $R(n-k, n-i)$ is a function of the time index n. Without loss of generality, we shall assume that we are interested in estimating the parameters a_k at time $n = 0$. Then, (23) and (24) reduce to

$$\sum_{k=1}^{p} a_k R(-k, -i) = -R(0, -i) \tag{27}$$

$$E_p' = R(0,0) + \sum_{k=1}^{p} a_k R(0,k). \tag{28}$$

In estimating the nonstationary autocorrelation coefficients from the signal s_n, we note that nonstationary processes are not ergodic, and, therefore, one cannot substitute the ensemble average by a time average. However, for a certain class of nonstationary processes known as *locally stationary processes* [12], [92], it is reasonable to estimate the autocorrelation function with respect to a point in time as a short-time average. Examples of nonstationary processes that can be considered to be locally stationary are speech and EEG signals.

[4] Usually the estimate given by (17) is divided by N, but that does not affect the solution for the predictor coefficients.

In a manner analogous to the stationary case, we estimate $R(-k, -i)$ by φ_{ik} in (20). Using this approximation for the nonstationary autocorrelation leads to a solution for the parameters a_k in (27) that is identical to that given by (18) in the covariance method in the deterministic case.

Note that for a stationary signal: $R(t, t') = R(t - t')$, and therefore, the normal equations (27) and (28) reduce to (12) and (13).

3) Gain Computation: Since in the least squares method we assumed that the input was unknown, it does not make much sense to determine a value for the gain G. However, there are certain interesting observations that can be made.

Equation (7) can be rewritten as

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + e_n. \tag{29}$$

Comparing (4) and (29) we see that the only input signal u_n that will result in the signal s_n as output is that where $Gu_n = e_n$. That is, the input signal is proportional to the error signal. For any other input u_n, the output from the filter $H(z)$ in Fig. 2 will be different from s_n. However, if we insist that whatever the input u_n, the energy in the output signal must equal that of the original signal s_n, then we can at least specify the total energy in the input signal. Since the filter $H(z)$ is fixed, it is clear from the above that the total energy in the input signal Gu_n must equal the total energy in the error signal, which is given by E_p in (13) or (19), depending on the method used.

Two types of input that are of special interest are: the deterministic impulse and stationary white noise. By examining the response of the filter $H(z)$ to each of these two inputs we shall gain further insight into the time domain properties of the all-pole model. The input gain is then determined as a by-product of an autocorrelation analysis.

a) Impulse input: Let the input to the all-pole filter $H(z)$ be an impulse or unit sample at $n = 0$, i.e. $u_n = \delta_{n0}$, where δ_{nm} is the Kronecker delta. The output of the filter $H(z)$ is then its impulse response h_n, where

$$h_n = -\sum_{k=1}^{p} a_k h_{n-k} + G\delta_{n0}. \tag{30}$$

The autocorrelation $\hat{R}(i)$ of the impulse response h_n has an interesting relationship to the autocorrelation $R(i)$ of the signal s_n. Multiply (30) by h_{n-i} and sum over all n. The result can be shown to be [10], [62]:

$$\hat{R}(i) = -\sum_{k=1}^{p} a_k \hat{R}(i-k), \qquad 1 \leqslant |i| \leqslant \infty \tag{31}$$

$$\hat{R}(0) = -\sum_{k=1}^{p} a_k \hat{R}(k) + G^2. \tag{32}$$

Given our condition that the total energy in h_n must equal that in s_n, we must have

$$\hat{R}(0) = R(0) \tag{33}$$

since the zeroth autocorrelation coefficient is equal to the total energy in the signal. From (33) and the similarity between (12) and (31) we conclude that [62]

$$\hat{R}(i) = R(i), \qquad 0 \leqslant i \leqslant p. \tag{34}$$

This says that the first $p + 1$ coefficients of the autocorrelation of the impulse response of $H(z)$ are identical to the corresponding autocorrelation coefficients of the signal. The problem of linear prediction using the autocorrelation method can be stated in a new way as follows. Find a filter of the form $H(z)$ in (5) such that the first $p + 1$ values of the autocorrelation of its impulse response are equal to the first $p + 1$ values of the signal autocorrelation, and such that (31) applies.

From (32), (34), and (13), the gain is equal to

$$G^2 = E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{35}$$

where G^2 is the total energy in the input $G\delta_{n0}$.

b) White noise input: Here the input u_n is assumed to be a sequence of uncorrelated samples (white noise) with zero mean and unit variance, i.e., $\mathcal{E}(u_n) = 0$, all n, and $\mathcal{E}(u_n u_m) = \delta_{nm}$. Denote the output of the filter by $\hat{s}_n$. For a fixed filter $H(z)$, the output $\hat{s}_n$ forms a stationary random process:

$$\hat{s}_n = -\sum_{k=1}^{p} a_k \hat{s}_{n-k} + Gu_n. \tag{36}$$

Multiply (36) by $\hat{s}_{n-i}$ and take expected values. By noting that u_n and $\hat{s}_{n-i}$ are uncorrelated for $i > 0$, the result can be shown [17] to be identical to (31) and (32), where $\hat{R}(i) = \mathcal{E}(\hat{s}_n \hat{s}_{n-i})$ is the autocorrelation of the output $\hat{s}_n$. Therefore, (31) and (32) completely specify an all-pole random process as well. Equations (31) are known in the statistical literature as the *Yule-Walker equations* [17], [98], [115].

For the random case we require that the average energy (or variance) of the output $\hat{s}_n$ be equal to the variance of the original signal s_n, or $\hat{R}(0) = R(0)$, since the zeroth autocorrelation of a zero-mean random process is the variance. By a reasoning similar to that given in the previous section, we conclude that (34) and (35) also apply for the random case.

From the preceding, we see that the relations linking the autocorrelation coefficients of the output of an all-pole filter are the same whether the input is a single impulse or white noise. This is to be expected since both types of input have identical autocorrelations and, of course, identical flat spectra. This dualism between the deterministic impulse and statistical white noise is an intriguing one. Its usefulness surfaces very elegantly in modeling the speech process, as in Fig. 1, where both unit impulses as well as white noise are actually used to synthesize speech.

C. Computation of Predictor Parameters

1) Direct Methods: In each of the two formulations of linear prediction presented in the previous section, the predictor coefficients a_k, $1 \leqslant k \leqslant p$, can be computed by solving a set of p equations with p unknowns. These equations are (12) for the autocorrelation (stationary) method and (18) for the covariance (nonstationary) method. There exist several standard methods for performing the necessary computations, e.g., the Gauss reduction or elimination method and the Crout reduction method [49]. These general methods require $p^3/3 + O(p^2)$ operations (multiplications or divisions) and p^2 storage locations. However, we note from (12) and (18) that the matrix of coefficients in each case is a covariance matrix. Covariance matrices are symmetric and in general positive semidefinite, although in practice they are usually positive definite. Therefore, (12) and (18) can be solved more ef-

ficiently by the square-root or Cholesky decomposition method [31], [39], [59], [110]. This method requires about half the computation $p^3/6 + O(p^2)$ and about half the storage $p^2/2$ of the general methods. The numerical stability properties of this method are well understood [109], [111]; the method is considered to be quite stable.

Further reduction in storage and computation time is possible in solving the autocorrelation normal equations (12) because of their special form. Equation (12) can be expanded in matrix form as

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_1 & R_0 & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{bmatrix} \tag{37}$$

Note that the $p \times p$ autocorrelation matrix is symmetric and the elements along any diagonal are identical (i.e., a Toeplitz matrix). Levinson [61] derived an elegant recursive procedure for solving this type of equation. The procedure was later reformulated by Robinson [85]. Levinson's method assumes the column vector on the right hand side of (37) to be a general column vector. By making use of the fact that this column vector comprises the same elements found in the autocorrelation matrix, another method attributed to Durbin [25] emerges which is twice as fast as Levinson's. The method requires only $2p$ storage locations and $p^2 + O(p)$ operations: a big saving from the more general methods. Durbin's recursive procedure can be specified as follows:

$$E_0 = R(0) \tag{38a}$$

$$k_i = -\left[R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)\right] \Big/ E_{i-1} \tag{38b}$$

$$a_i^{(i)} = k_i$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}, \quad 1 \leqslant j \leqslant i-1 \tag{38c}$$

$$E_i = (1 - k_i^2) E_{i-1}. \tag{38d}$$

Equations (38b)–(38d) are solved recursively for $i = 1, 2, \cdots, p$. The final solution is given by

$$a_j = a_j^{(p)}, \quad 1 \leqslant j \leqslant p. \tag{38e}$$

Note that in obtaining the solution for a predictor of order p, one actually computes the solutions for all predictors of order less than p. It has been reported [78] that this solution is numerically relatively unstable. However, most researchers have not found this to be a problem in practice.

It should be emphasized that, for many applications, the solution of the normal equations (12) or (18) does not form the major computational load. The computation of the autocorrelation or covariance coefficients require pN operations, which can dominate the computation time if $N \gg p$, as is often the case.

The solution to (37) is unaffected if all the autocorrelation coefficients are scaled by a constant. In particular, if all $R(i)$ are normalized by dividing by $R(0)$, we have what are known as the *normalized autocorrelation coefficients* $r(i)$:

$$r(i) = \frac{R(i)}{R(0)} \tag{39}$$

which have the property that $|r(i)| \leqslant 1$. This can be useful in the proper application of scaling to a fixed point solution to (37).

A byproduct of the solution in (38) is the computation of the minimum total error E_i at every step. It can easily be shown that the minimum error E_i decreases (or remains the same) as the order of the predictor increases [61]. E_i is never negative, of course, since it is a squared error. Therefore, we must have

$$0 \leqslant E_i \leqslant E_{i-1}, \quad E_0 = R(0). \tag{40}$$

If the autocorrelation coefficients are normalized as in (39), then the minimum error E_i is also divided by $R(0)$. We shall call the resulting quantity the *normalized error* V_i:

$$V_i = \frac{E_i}{R(0)} = 1 + \sum_{k=1}^{i} a_k r(k). \tag{41}$$

From (40) it is clear that

$$0 \leqslant V_i \leqslant 1, \quad i \geqslant 0. \tag{42}$$

Also, from (38d) and (41), the final normalized error V_p is

$$V_p = \prod_{i=1}^{p} (1 - k_i^2). \tag{43}$$

The intermediate quantities k_i, $1 \leqslant i \leqslant p$, are known as the *reflection coefficients*. In the statistical literature, they are known as *partial correlation coefficients* [6], [17]. k_i can be interpreted as the (negative) partial correlation between s_n and s_{n+i} holding $s_{n+1}, \cdots, s_{n+i-1}$ fixed. The use of the term "reflection coefficient" comes from transmission line theory, where k_i can be considered as the reflection coefficient at the boundary between two sections with impedances Z_i and Z_{i+1}. k_i is then given by

$$k_i = \frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i}. \tag{44}$$

The transfer function $H(z)$ can then be considered as that of a sequence of these sections with impedance ratios given from (44) by

$$\frac{Z_{i+1}}{Z_i} = \frac{1 + k_i}{1 - k_i}, \quad 1 \leqslant i \leqslant p. \tag{45}$$

The same explanation can be given for any type of situation where there is plane wave transmission with normal incidence in a medium consisting of a sequence of sections or slabs with different impedances. In the case of an acoustic tube with p sections of equal thickness, the impedance ratios reduce to the inverse ratio of the consecutive cross-sectional areas. This fact has been used recently in speech analysis [10], [52], [97]. Because of the more familiar "engineering interpretation" for k_i, we shall refer to them in this paper as reflection coefficients.

2) Iterative Methods: Beside the direct methods for the solution of simultaneous linear equations, there exist a number of iterative methods. In these methods, one begins by an initial guess for the solution. The solution is then updated by

adding a correction term that is usually based on the gradient of some error criterion. In general, iterative methods require more computation to achieve a desired degree of convergence than the direct methods. However, in some applications [100] one often has a good initial guess, which might lead to the solution in only a few iterations. This can be a big saving over direct methods if the number of equations is large. Some of the iterative methods are the gradient method, the steepest descent method, Newton's method, conjugate gradient method and the stochastic approximation method [81], [108].

Up till now we have assumed that the whole signal is given all at once. For certain real time applications it is useful to be able to perform the computations as the signal is coming in. Adaptive schemes exist which update the solution based on every new observation of the signal [106]. The update is usually proportional to the difference between the new observation and the predicted value given the present solution. Another application for adaptive procedures is in the processing of very long data records, where the solution might converge long before all the data is analyzed. It is worth noting that Kalman filtering notions [56] are very useful in obtaining adaptive solutions [60].

3) Filter Stability: After the predictor parameters are computed, the question of the stability of the resulting filter $H(z)$ arises. Filter stability is important for many applications. A causal all-pole filter is stable if all its poles lie inside the unit circle (in which case it is also a filter with minimum phase). The poles of $H(z)$ are simply the roots of the denominator polynomial $A(z)$, where

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{46}$$

and

$$H(z) = \frac{G}{A(z)}. \tag{47}$$

$A(z)$ is also known as the *inverse filter*.

If the coefficients $R(i)$ in (12) are positive definite [79] (which is assured if $R(i)$ is computed from a nonzero signal using (17) or from a positive definite spectrum[5]), the solution of the autocorrelation equation (12) gives predictor parameters which guarantee that all the roots of $A(z)$ lie inside the unit circle, i.e., a stable $H(z)$ [42], [85], [104]. This result can also be obtained from orthogonal polynomial theory. In fact, if one denotes the inverse filter at step i in iteration (38) by $A_i(z)$, then it can be shown that the polynomials $A_i(z)$ for $i = 0, 1, 2, \cdots$, form an orthogonal set over the unit circle [35], [42], [93]:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega)\, A_n(e^{j\omega})\, A_m(e^{-j\omega})\, d\omega = E_n \delta_{nm},$$

$$n, m = 0, 1, 2, \cdots \tag{48}$$

where E_n is the minimum error for an nth order predictor, and $P(\omega)$ is any positive definite spectrum whose Fourier transform results in the autocorrelation coefficients $R(i)$ that are used in (12). The recurrence relation for these polynomials is as follows:

$$A_i(z) = A_{i-1}(z) + k_i z^{-i} A_{i-1}(z^{-1}) \tag{49}$$

which is the same as the recursion in (38c).

The positive definiteness of $R(i)$ can often be lost if one uses a small word length to represent $R(i)$ in a computer. Also, roundoff errors can cause the autocorrelation matrix to become ill-conditioned. Therefore, it is often necessary to check for the stability of $H(z)$. Checking if the roots of $A(z)$ are inside the unit circle is a costly procedure that is best avoided One method is to check if all the successive errors are positive. In fact, the condition $E_i > 0$, $1 \leq i \leq p$, is a necessary and sufficient condition for the stability of $H(z)$. From (38d) and (40) it is clear that an equivalent condition for the stability of $H(z)$ is that

$$|k_i| < 1, \quad 1 \leq i \leq p. \tag{50}$$

Therefore, the recursive procedure (38) also facilitates the check for the stability of the filter $H(z)$.

The predictor parameters resulting from a solution to the covariance matrix equation (18) cannot in general be guaranteed to form a stable filter. The computed filter tends to be more stable as the number of signal samples N is increased, i.e., as the covariance matrix approaches an autocorrelation matrix. Given the computed predictor parameters, it is useful to be able to test for the stability of the filter $H(z)$. One method is to compute the reflection coefficients k_i from the predictor parameters by a backward recursion, and then check for stability using (50). The recursion is as follows:

$$k_i = a_i^{(i)}$$

$$a_j^{(i-1)} = \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2}, \quad 1 \leq j \leq i-1 \tag{51}$$

where the index i takes values $p, p-1, \cdots, 1$ in that order. Initially $a_j^{(p)} = a_j$, $1 \leq j \leq p$. It is interesting to note that this method for checking the stability of $H(z)$ is essentially the same as the Lehmer–Schur method [81] for testing whether or not the zeros of a polynomial lie inside the unit circle. An unstable filter can be made stable by reflecting the poles outside the unit circle inside [10], such that the magnitude of the system frequency response remains the same. Filter instability can often be avoided by adding a very small number to the diagonal elements in the covariance matrix.

A question always arises as to whether to use the autocorrelation method or covariance method in estimating the predictor parameters. The covariance method is quite general and can be used with no restrictions. The only problem is that of the stability of the resulting filter, which is not a severe problem generally. In the autocorrelation method, on the other hand, the filter is guaranteed to be stable, but problems of parameter accuracy can arise because of the necessity of windowing (truncating) the time signal. This is usually a problem if the signal is a portion of an impulse response. For example, if the impulse response of an all-pole filter is analyzed by the covariance method, the filter parameters can be computed accurately from only a finite number of samples of the signal. Using the autocorrelation method, one cannot obtain the exact parameter values unless the whole infinite impulse response is used in the analysis. However, in practice, very good approximations can be obtained by truncating the impulse response at a point where most of the decay of the response has already occurred.

[5] A spectrum that can be zero at most at a countable set of frequencies.

III. Spectral Estimation

In Section II, the stationary and nonstationary methods of linear prediction were derived from a time domain formulation. In this section we show that the same normal equations can be derived from a frequency domain formulation. It will become clear that linear prediction is basically a correlation type of analysis which can be approached either from the time or frequency domain. The insights gained from the frequency domain analysis will lead to new applications for linear predictive analysis. This section and the following are based mainly on references [62]-[64].

A. Frequency Domain Formulations

1) Stationary Case: The error e_n between the actual signal and the predicted signal is given by (7). Applying the z transform to (7), we obtain

$$E(z) = \left[1 + \sum_{k=1}^{p} a_k z^{-k}\right] S(z) = A(z)\, S(z) \tag{52}$$

where $A(z)$ is the inverse filter defined in (46), and $E(z)$ and $S(z)$ are the z transforms of e_n and s_n, respectively. Therefore, e_n can be viewed as the result of passing s_n through the inverse filter $A(z)$. Assuming a deterministic signal[6] s_n, and applying Parseval's theorem, the total error to be minimized is given by

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 \, d\omega \tag{53}$$

where $E(e^{j\omega})$ is obtained by evaluating $E(z)$ on the unit circle $z = e^{j\omega}$. Denoting the power spectrum of the signal s_n by $P(\omega)$, where

$$P(\omega) = |S(e^{j\omega})|^2 \tag{54}$$

we have from (52)-(54)

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) A(e^{j\omega}) A(e^{-j\omega}) \, d\omega. \tag{55}$$

Following the same procedure as in Section II, E is minimized by applying (9) to (55). The result can be shown [64] to be identical to the autocorrelation normal equations (12), but with the autocorrelation $R(i)$ obtained from the signal spectrum $P(\omega)$ by an inverse Fourier transform

$$R(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \cos(i\omega) \, d\omega. \tag{56}$$

Note that in (56) the cosine transform is adequate since $P(\omega)$ is real and even. The minimum squared error E_p can be obtained by substituting (12) and (56) in (55), which results in the same equation as in (13).

2) Nonstationary Case: Here the signal s_n and the error e_n are assumed to be nonstationary. If $R(t, t')$ is the nonstationary autocorrelation of s_n, then we define the nonstationary two-dimensional (2-D) spectrum $Q(\omega, \omega')$ of s_n by [12], [64], [67], [79]

$$Q(\omega, \omega') = \sum_{t'=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} R(t, t') \exp\left[-j(\omega t - \omega' t')\right]. \tag{57}$$

[6] A similar development assuming a random signal gives the same results.

$R(t, t')$ can then be recovered from $Q(\omega, \omega')$ by an inverse 2-D Fourier transform

$$R(t, t') = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} Q(\omega, \omega') \exp\left[j(\omega t - \omega' t')\right] d\omega \, d\omega'. \tag{58}$$

As in the time domain formulation, we are interested in minimizing the error variance for time $n = 0$, which is now given by [64]

$$E = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} Q(\omega, \omega') A(e^{j\omega}) A(e^{-j\omega'}) \, d\omega \, d\omega'. \tag{59}$$

Applying (9) to (59) results in equations identical to the nonstationary normal equations (27), where $R(t, t')$ is now defined by (58). The minimum error is then obtained by substituting (27) and (58) in (59). The answer is identical to (28).

B. Linear Predictive Spectral Matching

In this section we shall examine in what manner the signal spectrum $P(\omega)$ is approximated by the all-pole model spectrum, which we shall denote by $\hat{P}(\omega)$. From (5) and (47):

$$\hat{P}(\omega) = |H(e^{j\omega})|^2 = \frac{G^2}{|A(e^{j\omega})|^2} \tag{60}$$

$$= \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jk\omega}\right|^2}.$$

From (52) and (54) we have

$$P(\omega) = \frac{|E(e^{j\omega})|^2}{|A(e^{j\omega})|^2}. \tag{61}$$

By comparing (60) and (61) we see that if $P(\omega)$ is being modeled by $\hat{P}(\omega)$, then the error power spectrum $|E(e^{j\omega})|^2$ is being modeled by a flat spectrum equal to G^2. This means that the actual error signal e_n is being approximated by another signal that has a flat spectrum, such as a unit impulse, white noise, or any other signal with a flat spectrum. The filter $A(z)$ is sometimes known as a "whitening filter" since it attempts to produce an output signal e_n that is white, i.e., has a flat spectrum.

From (53), (60), and (61), the total error can be written as

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} \, d\omega. \tag{62}$$

Therefore, minimizing the total error E is equivalent to the minimization of the integrated ratio of the signal spectrum $P(\omega)$ to its approximation $\hat{P}(\omega)$. (This interpretation of the least squares error was proposed in a classic paper by Whittle [103]. An equivalent formulation using maximum likelihood estimation has been given by Itakura [50], [51].) Now, we can back up and restate the problem of linear prediction as follows. Given some spectrum $P(\omega)$, we wish to model it by another spectrum $\hat{P}(\omega)$ such that the integrated ratio between the two spectra as in (62) is minimized. The parameters of the model spectrum are computed from the normal equations (12), where the needed autocorrelation coefficients $R(i)$ are easily computed from $P(\omega)$ by a simple Fourier transform. The gain factor G is obtained by equating the total energy in the two

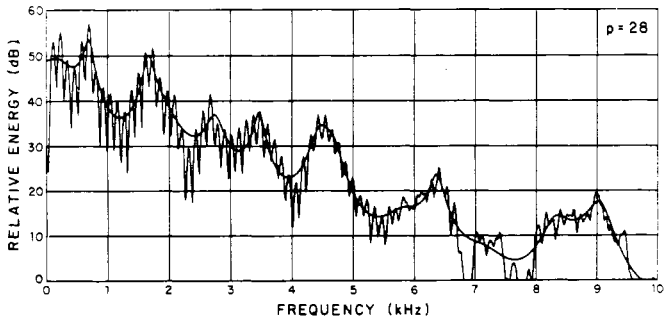

Fig. 3. A 28-pole fit to an FFT-computed signal spectrum. The signal was sampled at 20 kHz.

spectra, i.e., $\hat{R}(0) = R(0)$, where

$$\hat{R}(i) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{P}(\omega)\cos(i\omega)\,d\omega. \tag{63}$$

Note that $\hat{R}(i)$ is the autocorrelation of the impulse response of $H(z)$, which is given by (31) and (32). As then, the gain is computed from (35).

The manner in which the model spectrum $\hat{P}(\omega)$ approximates $P(\omega)$ is largely reflected in the relation between the corresponding autocorrelation functions. From (34), we have $\hat{R}(i) = R(i)$, $0 \leqslant i \leqslant p$. Since $P(\omega)$ and $\hat{P}(\omega)$ are Fourier transforms of $R(i)$ and $\hat{R}(i)$, respectively, it follows that increasing the value of the order of the model p increases the range over which $R(i)$ and $\hat{R}(i)$ are equal, resulting in a better fit of $\hat{P}(\omega)$ to $P(\omega)$. In the limit, as $p \to \infty$, $\hat{R}(i)$ becomes identical to $R(i)$ for all i, and hence the two spectra become identical:

$$\hat{P}(\omega) = P(\omega), \qquad \text{as } p \to \infty. \tag{64}$$

This statement says that we can approximate any spectrum arbitrarily closely by an all-pole model.

Another important conclusion is that since linear predictive analysis can be viewed as a process of spectrum or autocorrelation matching, one must be careful how to estimate the spectrum $P(\omega)$ or the corresponding autocorrelation that is to be modeled. Since the signal is often weighted or windowed[7] before either the autocorrelation or the spectrum is computed, it can be quite important to properly choose the type and width of the data window to be used. The choice of window depends very much on the type of signal to be analyzed. If the signal can be considered to be stationary for a long period of time (relative to the effective length of the system impulse response), then a rectangular window suffices. However, for signals that result from systems that are varying relatively quickly, the time of analysis must necessarily be limited. For example, in many transient speech sounds, the signal can be considered stationary for a duration of only one or two pitch periods. In that case a window such as Hamming or Hanning [14] is more appropriate. See [13], [14], [26], [54], [64], [99], [101] for more on the issue of windowing and spectral estimation in general.

An example of linear predictive (LP) spectral estimation is shown in Fig. 3, where the original spectrum $P(\omega)$ was obtained by computing the fast Fourier transform (FFT) of a 20-ms, Hamming windowed, 20-kHz sampled speech signal. The speech sound was the vowel [æ] as in the word "bat." The harmonics due to the periodicity of the sound are evident in the FFT spectrum. Fig. 3 also shows a 28-pole fit ($p = 28$) to the signal spectrum. In this case the autocorrelation coefficients needed to solve the normal equations (12) were computed directly from the time signal. The all-pole spectrum $\hat{P}(\omega)$ was computed from (60) by dividing G^2 by the magnitude squared of the FFT of the sequence: $1, a_1, a_2, \cdots, a_p$. Arbitrary frequency resolution in computing $\hat{P}(\omega)$ can be obtained by simply appending an appropriate number of zeros to this sequence before taking the FFT. An alternate method of computing $\hat{P}(\omega)$ is obtained by rewriting (60) as

$$\hat{P}(\omega) = \frac{G^2}{\rho(0) + 2\sum_{i=1}^{p} \rho(i)\cos(i\omega)} \tag{65}$$

where

$$\rho(i) = \sum_{k=0}^{p-i} a_k a_{k+i}, \qquad a_0 = 1, \quad 0 \leqslant i \leqslant p \tag{66}$$

is the autocorrelation of the impulse response of the inverse filter $A(z)$. From (65), $\hat{P}(\omega)$ can be computed by dividing G^2 by the real part of the FFT of the sequence: $\rho(0), 2\rho(1), 2\rho(2), \cdots, 2\rho(p)$. Note that the slope of $\hat{P}(\omega)$ is always zero at $\omega = 0$ and $\omega = \pi$.

Another property of $\hat{P}(\omega)$ is obtained by noting that the minimum error $E_p = G^2$, and, therefore, from (62) we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)}\,d\omega = 1. \tag{67}$$

(This relation is a special case of a more general result (48) relating the fact that the polynomials $A_0(z), A_1(z), \cdots, A_p(z), \cdots$, form a complete set of orthogonal polynomials with weight $P(\omega)$.) Equation (67) is true for all values of p. In particular, it is true as $p \to \infty$, in which case from (64) we see that (67) becomes an identity. Another important case where (67) becomes an identity is when $P(\omega)$ is an all-pole spectrum with p_0 poles, then $\hat{P}(\omega)$ will be identical to $P(\omega)$ for all $p \geqslant p_0$. Relation (67) will be useful in discussing the properties of the error measure in Section IV.

The transfer functions $S(z)$ and $H(z)$ corresponding to $P(\omega)$ and $\hat{P}(\omega)$ are also related. It can be shown [62] that as $p \to \infty$, $H(z)$ is given by

$$H_\infty(z) = \frac{G}{1 + \sum_{k=1}^{\infty} a_k z^{-k}} = \sum_{n=0}^{N-1} h_\infty(n) z^{-n}, \qquad p \to \infty \tag{68}$$

where $h_\infty(n)$, $0 \leqslant n \leqslant N-1$, is the minimum phase sequence corresponding to s_n, $0 \leqslant n \leqslant N-1$. Note that the minimum phase sequence is of the same length as the original signal. Fig. 4 shows a signal ($N = 256$) and its approximate minimum phase counterpart, obtained by first performing a LP analysis for $p = 250$, and then computing the sequence $h(n)$ by long division.

C. Selective Linear Prediction

The major point of the previous section was that LP analysis can be regarded as a method of spectral modeling. We had tacitly assumed that the model spectrum spans the same frequency range as the signal spectrum. We now generalize the

[7] Note that here we are discussing *data* windows which are applied directly to the signal, as opposed to *lag* windows, which statisticians have traditionally applied to the autocorrelation.

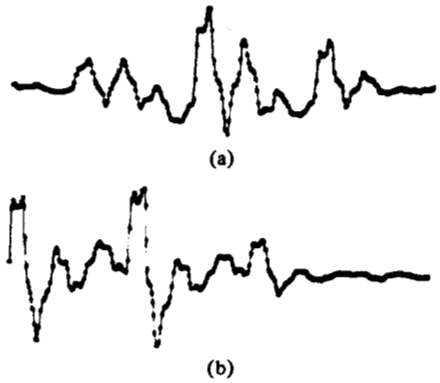

Fig. 4. (a) A 256-sample windowed speech signal. (b) The corresponding approximate minimum phase sequence obtained using a linear predictor of order $p = 250$.

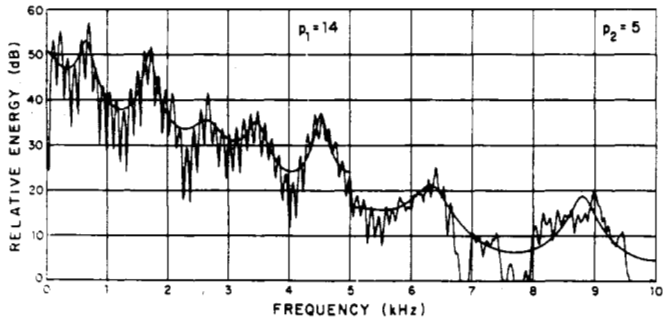

Fig. 5. Application of selective linear prediction to the same signal spectrum as in Fig. 3, with a 14-pole fit to the 0–5 kHz region and a 5-pole fit to the 5–10 kHz region.

LP spectral modeling method to the case where we wish to fit only a selected portion of a given spectrum.

Suppose we wish to model the spectrum $P(\omega)$ only in the region[8] $\omega_\alpha \leq \omega \leq \omega_\beta$ by an all-pole spectrum given by (60). Call the signal spectrum in that region $P'(\omega)$. In order to compute the parameters of the model spectrum $\hat{P}(\omega)$, we first perform a linear mapping of the given region onto the upper half of the unit circle in the z plane. This can be accomplished by the mapping $\omega' = \pi(\omega - \omega_\alpha)/(\omega_\beta - \omega_\alpha)$, so that the given region is mapped onto $0 \leq \omega' \leq \pi$. In addition, let $P'(-\omega') = P'(\omega')$ define the spectrum over the lower half of the unit circle. The model parameters are then computed from the normal equations (12), where the autocorrelation coefficients are obtained by a Fourier transform with $P'(\omega')$ replacing $P(\omega)$ and ω' replacing ω in (56).

Selective linear prediction has had applications in speech recognition and speech compression [63]. An example of its usage is shown in Fig. 5. For speech recognition applications, the 0–5 kHz region is more important than the 5–10 kHz. Even when the 5–10 kHz region is important, only a rough idea of the shape of the spectrum is sufficient. In Fig. 5, the signal spectrum is the same as in Fig. 3. The 0–5 kHz region is modeled by a 14-pole spectrum, while the 5–10 kHz region is modeled independently by only a 5-pole model.

An important point, which should be clear by now, is that since we assume the availability of the signal spectrum $P(\omega)$, any desired frequency shaping or scaling can be performed directly on the signal spectrum before linear predictive modeling is applied.

[8] The remainder of the spectrum is simply neglected.

D. Modeling Discrete Spectra

Thus far we have assumed that the spectrum $P(\omega)$ is a continuous function of frequency. More often, however, the spectrum is known at only a finite number of frequencies. For example, FFT-derived spectra and those obtained from many commercially available spectrum analyzers have values at equally spaced frequency points. On the other hand, filter bank spectra, and, for example, third-octave band spectrum analyzers have values at frequencies that are not necessarily equally spaced. In order to be able to model these discrete spectra, only one change in our analysis need be made. The error measure E in (62) is defined as a summation instead of an integral. The rest of the analysis remains the same except that the autocorrelation coefficients $R(i)$ are now computed from

$$R(i) = \frac{1}{M} \sum_{m=0}^{M-1} P(\omega_m) \cos(i\omega_m) \tag{69}$$

where M is the total number of spectral points on the unit circle. The frequencies ω_m are those for which a spectral value exists, and they need not be equally spaced. Below we demonstrate the application of LP modeling for filter bank and harmonic spectra.

Fig. 6(a) shows a typical 14-pole fit to a spectrum of the fricative [s] that was FFT computed from the time signal. Fig. 6(b) shows a similar fit to a line spectrum that is typical of filter bank spectra. What we have actually done here is to simulate a filter bank where the filters are linearly spaced up to 1.6 kHz and logarithmically spaced thereafter. Note that the all-pole spectrum for the simulated filter bank is remarkably similar to the one in the top figure, even though the number of spectral points is much smaller.

The dashed curve in Fig. 7(a) is a 14-pole spectrum. If one applied LP analysis to this spectrum, the all-pole model for $p = 14$ would be identical to the dashed spectrum. The situation is not so favorable for discrete spectra. Let us assume that the dashed spectrum corresponds to the transfer function of a 14-pole filter. If this filter is excited by a periodic train of impulses with fundamental frequency F_0, the spectrum of the output signal will be a discrete line spectrum with spectral values only at the harmonics (multiples of F_0). The line spectrum for $F_0 = 312$ Hz is shown in Fig. 7(a). Note that the dashed spectrum is an envelope of the harmonic spectrum. The result of applying a 14-pole LP analysis to the harmonic spectrum is shown as the solid curve in Fig. 7(a). The discrepancy between the two all-pole spectra is obvious. In general, the types of discrepancies that can occur between the model and original spectra include merging or splitting of pole peaks, and increasing or decreasing of pole frequencies and bandwidths. Pole movements are generally in the direction of the nearest harmonic. As the fundamental frequency decreases, these discrepancies decrease, as shown in Fig. 7(b) for $F_0 = 156$ Hz.

It is important to note in Fig. 7 that the dashed curve is the *only* possible 14-pole spectrum that coincides with the line spectrum at the harmonics.[9] It is significant that the all-pole spectrum resulting from LP modeling does not yield the spectrum we desire. The immediate reason for this is that the solution for the model parameters from (12) depends on the values of the signal autocorrelation, which for the periodic signal are

[9] In general this is true only if the period between input impulses is greater than twice the number of poles in the filter.

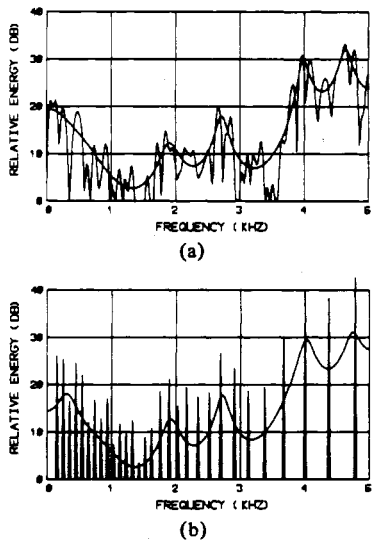

Fig. 6. Application of LP modeling to a filter bank spectrum. (a) A 14-pole fit to the original spectrum. (b) A 14-pole fit to the simulated filter bank spectrum.

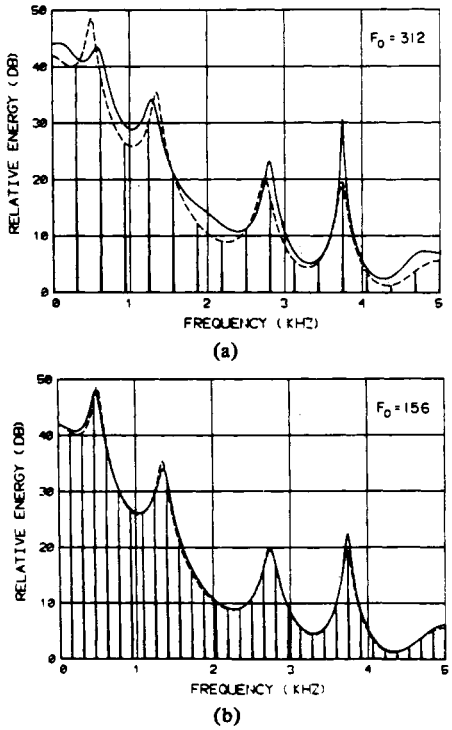

Fig. 7. Application of LP modeling to harmonic spectra. Dashed curve: 14-pole filter spectrum. Vertical lines: Corresponding harmonic spectrum for (a) and (b). (a) F_0 = 312 Hz. (b) F_0 = 156 Hz. Solid curve: 14-pole fit to the discrete harmonic spectrum. (For display purposes, the energy in the model spectrum (solid curve) was set equal to the energy in the filter spectrum (dashed curve).)

different from that for the single impulse response. However, the major underlying reason lies in the properties of the error measure used. This is the topic of the next section.

IV. Error Analysis

An important aspect of any fitting or matching procedure is the properties of the error measure that is employed, and whether those properties are commensurate with certain objectives. In this section we shall examine the properties of the error measure used in LP analysis and we shall discuss its strengths and weaknesses in order to be able to fully utilize its capabilities. The analysis will be restricted to the stationary (autocorrelation) case, although the conclusions can be extrapolated to the nonstationary (covariance) case.

The error measure used in Section II-B to determine the predictor parameters is the least squares error measure due to Gauss, who first reported on it in the early 1800's. This error measure has been used extensively since then, and is quite well understood. Its major asset is its mathematical tractability. Its main characteristic is that it puts great emphasis on large errors and little emphasis on small errors. Purely from the time domain, it is often difficult to say whether such an error measure is a desirable one or not for the problem at hand. Many would probably agree that it does not really matter which error measure one uses as long as it is a reasonable function of the magnitude of the error at each point. For the linear prediction problem, we are fortunate that the error measure can also be written in the frequency domain and can be interpreted as a goodness of fit between a given signal spectrum and a model spectrum that approximates it. The insights gained in the frequency domain should enhance our understanding of the least squares error criterion.

A. The Minimum Error

For each value of p, minimization of the error measure E in (62) leads to the minimum error E_p in (13), which is given in terms of the predictor and autocorrelation coefficients. Here we derive an expression for E_p in the frequency domain, which will help us determine some of its properties. Other properties will be discussed when we discuss the normalized minimum error.

Let

$$\hat{c}_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log \hat{P}(\omega)\, d\omega \tag{70}$$

be the zeroth coefficient (quefrency) of the cepstrum (inverse Fourier transform of log spectrum) [38], [77] corresponding to $\hat{P}(\omega)$. From (60), (70) reduces to

$$\hat{c}_0 = \log G^2 - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log |A(e^{j\omega})|^2\, d\omega. \tag{71}$$

$A(z)$ has all its zeros inside the unit circle. Therefore, the integral in (71) is equal to zero [64], [69], [103]. Since $G^2 = E_p$, we conclude from (71) that

$$E_p = e^{\hat{c}_0}. \tag{72}$$

From (72) and (70), E_p can be interpreted as the geometric mean of the model spectrum $\hat{P}(\omega)$. From (40) we know that E_p decreases as p increases. The minimum occurs as $p \to \infty$, and is equal to

$$E_{\min} = E_\infty = e^{c_0} \tag{73}$$

where c_0 is obtained by substituting $P(\omega)$ for $\hat{P}(\omega)$ in (70).[10] Therefore, the absolute minimum error is a function of $P(\omega)$ only, and is equal to its geometric mean, which is always positive for positive definite spectra.[11] This is a curious result, because it says that the minimum error can be nonzero even

[10] If $P(\omega)$ is a p_0-pole spectrum then $E_p = E_{\min}$ for all $p \geq p_0$.

[11] $E_{\min}$ is equal to zero only if $P(\omega)$ is zero over a noncountable set of frequencies (i.e., over a line segment). In that case, the signal is perfectly predictable and the prediction error is zero [107].

when the matching spectrum $\hat{P}(\omega)$ is identical to the matched spectrum $P(\omega)$. Therefore, although E_p is a measure of fit of the model spectrum to the signal spectrum, it is not an absolute one. The measure is always relative to $E_{\min}$. The nonzero aspect of $E_{\min}$ can be understood by realizing that, for any p, E_p is equal to that portion of the signal energy that is not predictable by a pth order predictor. For example, the impulse response of an all-pole filter is perfectly predictable *except* for the initial nonzero value. It is the energy in this initial value that shows up in E_p. (Note that in the covariance method one can choose the region of analysis to exclude the initial value, in which case the prediction error would be zero for this example.)

B. Spectral Matching Properties

The LP error measure E in (62) has two major properties:[12] a global property and a local property.

1) Global Property: Because the contributions to the total error are determined by the *ratio* of the two spectra, the matching process should perform uniformly over the whole frequency range, irrespective of the general shaping of the spectrum.

This is an important property for spectral estimation because it makes sure that the spectral match at frequencies with little energy is just as good, on the average, as the match at frequencies with high energy (see Fig. 3). If the error measure had been of the form $\int |P(\omega) - \hat{P}(\omega)|\, d\omega$, the spectral matches would have been best at high energy frequency points.

2) Local Property: This property deals with how the match is done in each small region of the spectrum.

Let the ratio of $P(\omega)$ to $\hat{P}(\omega)$ be given by

$$E(\omega) = \frac{P(\omega)}{\hat{P}(\omega)}. \tag{74}$$

Then from (67) we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} E(\omega)\, d\omega = 1, \qquad \text{for all } p. \tag{75}$$

$E(\omega)$ can be interpreted as the "instantaneous error" between $P(\omega)$ and $\hat{P}(\omega)$ at frequency ω. Equation (75) says that the arithmetic mean of $E(\omega)$ is equal to 1, which means that there are values of $E(\omega)$ greater and less than 1 such that the average is equal to 1.[13] In terms of the two spectra, this means that $P(\omega)$ will be greater than $\hat{P}(\omega)$ in some regions and less in others such that (75) applies. However, the contribution to the total error is more significant when $P(\omega)$ is greater than $\hat{P}(\omega)$ than when $P(\omega)$ is smaller, e.g., a ratio of $E(\omega) = 2$ contributes more to the total error than a ratio of 1/2. We conclude that:

> after the minimization of error, we expect a better fit of $\hat{P}(\omega)$ to $P(\omega)$ where $P(\omega)$ is greater than $\hat{P}(\omega)$, than where $P(\omega)$ is smaller (on the average).

For example, if $P(\omega)$ is the power spectrum of a quasi-periodic signal (such as in Fig. 3), then most of the energy in $P(\omega)$ will exist in the harmonics, and very little energy will reside between harmonics. The error measure in (62) insures that the approximation of $\hat{P}(\omega)$ to $P(\omega)$ is far superior at the harmonics than between the harmonics. If the signal had been generated by exciting a filter with a periodic sequence of impulses, then the system response of the filter must pass through all the harmonic peaks. Therefore, with a proper choice of the model order p, minimization of the LP error measure results in a model spectrum that is a good approximation to that system response. This leads to one characteristic of the local property:

> minimization of the error measure E results in a model spectrum $\hat{P}(\omega)$ that is a good estimate of the *spectral envelope* of the signal spectrum $P(\omega)$.

Fig. 6 shows that this statement also applies in a qualitative way when the excitation is random noise. It should be clear from the above that the importance of the local property is not as crucial when the variations of the signal spectrum from the spectral envelope are much less pronounced.

In the modeling of harmonic spectra, we showed an example in Fig. 7(a) where, although the all-pole spectrum resulting from LP modeling was a reasonably good estimate of the harmonic spectral envelope, it did not yield the unique all-pole transfer function that coincides with the line spectrum at the harmonics. This is a significant *disadvantage* of LP modeling, and is an indirect reflection of another characteristic of the local property: *the cancellation of errors*. This is evident from (75) where the instantaneous errors $E(\omega)$ are greater and less than 1 such that the average is 1. To help elucidate this point, let us define a new error measure E' that is the logarithm of E in (62):

$$E' = \log\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)}\, d\omega\right] \tag{76}$$

where the gain factor has been omitted since it is not relevant to this discussion. It is simple to show that the minimization of E' is equivalent to the minimization of E. For cases where $P(\omega)$ is smooth relative to $\hat{P}(\omega)$ and the values of $P(\omega)$ are not expected to deviate very much from $\hat{P}(\omega)$, the logarithm of the average of spectral ratios can be approximated by the average of the logarithms, i.e.,

$$E' \cong \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\frac{P(\omega)}{\hat{P}(\omega)}\, d\omega. \tag{77}$$

From (77) it is clear that the contributions to the error when $P(\omega) > \hat{P}(\omega)$ *cancel* those when $P(\omega) < \hat{P}(\omega)$.

The above discussion suggests the use of an error measure that takes the magnitude of the integrand in (77). One such error measure is

$$E'' = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\log\frac{P(\omega)}{\hat{P}(\omega)}\right]^2 d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}[\log P(\omega) - \log\hat{P}(\omega)]^2\, d\omega. \tag{78}$$

E'' is just the mean squared error between the two log spectra. It has the important property that the minimum error of zero occurs if and only if $\hat{P}(\omega)$ is identical to $P(\omega)$. Therefore, if we use (the discrete form of) E'' in modeling the harmonic spectrum in Fig. 7(a), the resulting model spectrum (for $p = 14$) will be identical to the dashed spectrum, since the minimum error of zero is achievable by that spectrum. However, while the error measure E'' solves one problem, it introduces an-

[12] Itakura [50], [51] discusses a maximum likelihood error criterion having the same properties.

[13] Except for the special case when $P(\omega)$ is all-pole, the condition $E(\omega) = 1$ for all ω is true only as $p \to \infty$.

other. Note that the contributions to the total error in (78) are equally important whether $P(\omega) > \hat{P}(\omega)$ or vice versa. This means that if the variations of $P(\omega)$ are large relative to $\hat{P}(\omega)$ (such as in Fig. 3), the resulting model spectrum will *not* be a good estimate of the spectral envelope. In addition, the minimization of E'' in (78) results in a set of nonlinear equations that must be solved iteratively, thus increasing the computational load tremendously.

Our conclusion is that the LP error measure in (62) is to be preferred in general, except for certain special cases (as in Fig. 7(a)) where an error measure such as E'' in (78) can be used, provided one is willing to carry the extra computational burden.

The global and local properties described here are properties of the error measure in (62) and do not depend on the details of $P(\omega)$ and $\hat{P}(\omega)$. These properties apply *on the average* over the whole frequency range. Depending on the detailed shapes of $P(\omega)$ and $\hat{P}(\omega)$, the resulting match can be better in one spectral region than in another. For example, if $\hat{P}(\omega)$ is an all-pole model spectrum and if the signal spectrum $P(\omega)$ contains zeros as well as poles, then one would not expect as good a match at the zeros as at the poles. This is especially true if the zeros have bandwidths of the same order as the poles or less. (Wide bandwidth zeros are usually well approximated by poles.) On the other hand, if $\hat{P}(\omega)$ is an all-zero spectrum then the preceding statement would have to be reversed.

C. The Normalized Error

The normalized error has been a very useful parameter for the determination of the optimal number of parameters to be used in the model spectrum. This subject will be discussed in the following section. Here we shall present some of the properties of the normalized error, especially as they relate to the signal and model spectra.

1) Relation to the Spectral Dynamic Range: The normalized error was defined in Section II as the ratio of the minimum error E_p to the energy in the signal $R(0)$. Keeping in mind that $R(0) = \hat{R}(0)$, and substituting for E_p from (72), we obtain

$$V_p = \frac{E_p}{R(0)} = \frac{e^{\hat{c}_0}}{\hat{R}(0)}. \tag{79}$$

Also, from (73), we have in the limit as $p \to \infty$:

$$V_{\min} = V_\infty = \frac{e^{c_0}}{R(0)}. \tag{80}$$

Therefore, the normalized error is always equal to the normalized zero quefrency of the model spectrum. From (40) and (79) it is clear that V_p is a monotonically decreasing function of p, with $V_0 = 1$ and $V_\infty = V_{\min}$ in (80). Fig. 8 shows plots of V_p as a function of p for two speech sounds (sampled at 10 kHz) whose spectra are similar to those in Figs. 3 and 6.

It is instructive to write V_p as a function of $\hat{P}(\omega)$. From (63) and (70), (79) can be rewritten as

$$V_p = \frac{\exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \hat{P}(\omega)\, d\omega\right]}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{P}(\omega)\, d\omega}. \tag{81}$$

It is clear from (81) that V_p depends completely on the shape of the model spectrum, and from (80), $V_{\min}$ is determined solely by the shape of the signal spectrum. An interesting way

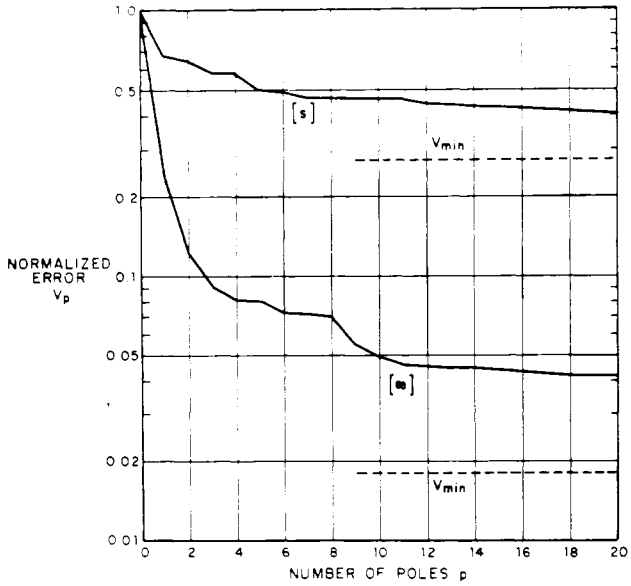

Fig. 8. Normalized error curves for the sounds [s] in the word "list" and [æ] in the word "potassium."

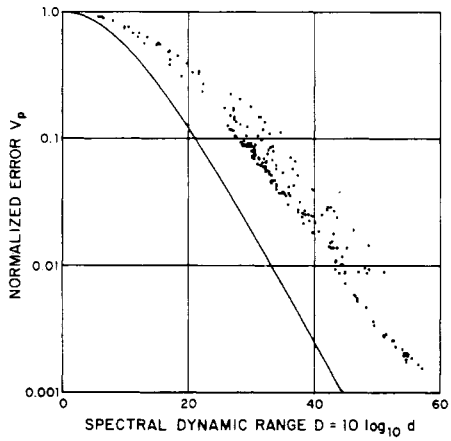

Fig. 9. Two-pole normalized error versus spectral dynamic range for 200 different two-pole models. The solid curve is V_a, the absolute lower bound on the normalized error.

to view (81) is that V_p is equal to the ratio of the geometric mean of the model spectrum to its arithmetic mean. This ratio has been used in the past as a measure of the spread of the data [22], [48]. When the spread of the data is small, the ratio is close to 1. Indeed, from (81) it is easy to see that if $\hat{P}(\omega)$ is flat, $V_p = 1$. On the other hand, if the data spread is large, then V_p becomes close to zero. Again, from (81) we see that if $\hat{P}(\omega)$ is zero for a portion of the spectrum (hence a large spread), then $V_p = 0$. (Another way of looking at V_p is in terms of the flatness of the spectrum [40].)

Another measure of data spread is the dynamic range. We define the spectral dynamic range d as the ratio of the highest to the lowest amplitude points on the spectrum:

$$d = H/L$$

where

$$H = \max_\omega \hat{P}(\omega) \qquad L = \min_\omega \hat{P}(\omega). \tag{82}$$

The relation between the normalized error and the spectral dynamic range is illustrated in Fig. 9. The dark dots in the

figure are plots of the normalized error versus the spectral dynamic range (in decibels) for 2-pole models of 200 different speech spectra. The solid curve in Fig. 9 is an absolute lower bound on the geometric-to-arithmetic mean ratio for *any* spectrum with a given dynamic range. The curve is a plot of the following relation [22], [64]

$$V_a = \gamma e^{(1-\gamma)} \tag{83}$$

where

$$\gamma = \frac{\log d}{d-1} \tag{84}$$

and V_a stands for the absolute lower bound on V_p for a given d. The overall impression from Fig. 9 is that the normalized error generally decreases as the dynamic range of the spectrum increases. This is apparent in Fig. 8 where V_p for the vowel [æ] is less than that for the fricative [s], and [æ] has a much higher spectral dynamic range than [s].

2) A Measure of Ill-Conditioning: In solving the autocorrelation normal equations (12), the condition of the autocorrelation matrix is an important consideration in deciding the accuracy of the computation needed. An ill-conditioned matrix can cause numerical problems in the solution. An accepted measure of ill-conditioning in a matrix is given by the ratio

$$d' = \lambda_{\max}/\lambda_{\min} \tag{85}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of the matrix [27], [81]. Grenander and Szegö [41], [42] have shown that all the eigenvalues of an autocorrelation matrix lie in the range $\lambda_i \in [H, L]$, $1 \leqslant i \leqslant p$, where H and L are defined in (82). In addition, as the order of the matrix p increases, the eigenvalues become approximately equal to $\hat{P}(\omega)$ evaluated at equally spaced points with separation $2\pi/(p+1)$. Therefore, the ratio d' given in (85) can be well approximated by the dynamic range of $\hat{P}(\omega)$:

$$d' \cong d. \tag{86}$$

Therefore, the spectral dynamic range is a good measure of the ill-conditioning of the autocorrelation matrix. The larger the dynamic range, the greater is the chance that the matrix is ill-conditioned.

But in the previous section we noted that an increase in d usually results in a decrease in the normalized error V_p. Therefore, V_p can also be used as a measure of ill-conditioning: the ill-conditioning is greater with decreased V_p. The problem becomes more and more serious as $V_p \to 0$, i.e., as the signal becomes highly predictable.

If ill-conditioning occurs sporadically, then one way of patching the problem is to increase the values along the principal diagonal of the matrix by a small fraction of a percent. However, if the problem is a regular one, then it is a good idea if one can reduce the dynamic range of the signal spectrum. For example, if the spectrum has a general slope, then a single-zero filter of the form $1 + az^{-1}$ applied to the signal can be very effective. The new signal is given by

$$s'_n = s_n + a s_{n-1}. \tag{87}$$

An optimal value for a is obtained by solving for the filter $A(z)$ that "whitens" (flattens) s'_n. This is, of course, given by the first order predictor, where

$$a = -\frac{R(1)}{R(0)}. \tag{88}$$

$R(1)$ and $R(0)$ are autocorrelation coefficients of the signal s_n. The filtered signal s'_n is then guaranteed to have a smaller spectral dynamic range. The above process is usually referred to as *preemphasis*.

One conclusion from the above concerns the design of the low-pass filter that one uses before sampling the signal to reduce aliasing. In order to ensure against aliasing, it is usually recommended that the cutoff frequency of the filter be lower than half the sampling frequency. However, if the cutoff frequency is appreciably lower than half the sampling frequency, then the spectral dynamic range of the signal spectrum increases, especially if the filter has a sharp cutoff and the stop band is very low relative to the pass band. This increases problems of ill-conditioning. Therefore, if one uses a lowpass filter with a sharp cutoff, the cutoff frequency should be set as close to half the sampling frequency as possible.

D. Optimal Number of Poles

One of the important decisions that usually has to be made in fitting of all-pole models is the determination of an "optimal" number of poles. It is a nontrivial exercise to define the word "optimal" here, for as we have seen, the fit of the model "improves" as the number of poles p increases. The problem is where to stop. Clearly we would like the minimum value of p that is adequate for the problem at hand, both to reduce our computation and to minimize the possibility of ill-conditioning (which increases with p since V_p decreases).

If the signal spectrum is an all-pole spectrum with p_0 poles, then we know that $V_p = V_{p_0}$, $p \geqslant p_0$, and $k_p = 0$, $p > p_0$, i.e., the error curve remains flat for $p > p_0$. Therefore, if we expect the signal spectrum to be an all-pole spectrum, a simple test to obtain the optimal p is to check when the error curve becomes flat. But, if the signal is the output of a p_0-pole filter with white noise excitation, then the suggested test will not work, because the estimates of the poles are based on a finite number of data points and the error curve will not be flat for $p > p_0$. In practice, however, the error curve will be almost flat for $p > p_0$. This suggests the use of the following threshold test

$$1 - \frac{V_{p+1}}{V_p} < \delta. \tag{89}$$

This test must succeed for several consecutive values before one is sure that the error curve has actually flattened out.

The use of the ratio V_{p+1}/V_p has been an accepted method in the statistical literature [6], [17], [112] for the determination of the optimal p. The test is based on hypothesis testing procedures using maximum likelihood ratios. A critical review of hypothesis testing procedures has been given recently by Akaike [5]. Akaike's main point is that model fitting is a problem where multiple decision procedures are required rather than hypothesis testing. The fitting problem should be stated as an estimation problem with an associated measure of fit. Akaike suggests the use of an information theoretic criterion that is an estimate of the mean log-likelihood [3], [5].[14] This is given by

$$I(p) = -2 \log (\text{maximum likelihood}) + 2p. \tag{90}$$

The value of p for which $I(p)$ is minimum is taken to be the optimal value. In our problem of all-pole modeling, if we assume that the signal has a Gaussian probability distribution,

[14] An earlier criterion used by Akaike is what he called the "final prediction error" [1], [2], [37].

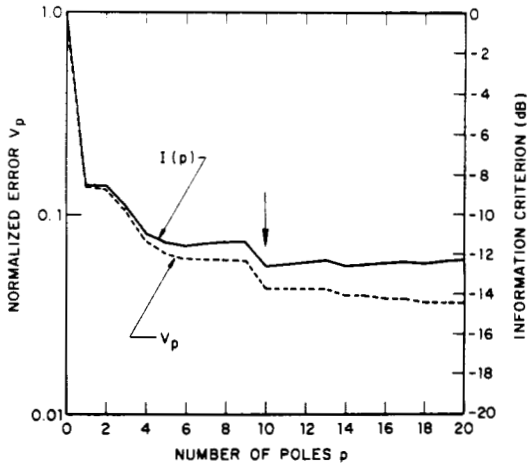

Fig. 10. A plot of Akaike's information criterion versus the order of the predictor p. Here, $I(p) = 10 \log_{10} V_p + (8.686p/0.4N)$, ($N = 200$, Hamming windowed). The "optimal" value of p occurs at the minimum of $I(p)$, shown by the arrow at $p = 10$.

then (90) reduces to (neglecting additive constants and dividing by N) [6], [17], [51]

$$I(p) = \log V_p + \frac{2p}{N_e} \tag{91}$$

where N_e is the "effective" number of data points in the signal. The word "effective" is used to indicate that one must compensate for possible windowing. The effective width of a window can be taken as the ratio of the energy under the window relative to that of a rectangular window. For example, for a Hamming window, $N_e = 0.4N$.

Note that the first term in (91) decreases as a function of p, and the second term increases. Therefore, a minimum can occur. In practice, there are usually several local minima, then the value of p corresponding to the absolute minimum of $I(p)$ is taken as the optimal value. Usually $I(p)$ is computed up to the maximum value of interest,[15] and the minimum of $I(p)$ is found in that region.

Fig. 10 shows an example of the application of Akaike's criterion. The dotted curve is the usual error curve and the solid curve is a plot of $I(p)$ in (91) multiplied by $10 \log_{10} e$ to obtain the results in decibels. In Fig. 10, the optimal predictor order is $p = 10$. Note that $I(p)$ for $p > 10$ slopes upward, but very gently. This indicates that the actual absolute minimum is quite sensitive to the linear term in (91). In practice, the criterion in (91) should not be regarded as an absolute, because it is based on several assumptions which might not apply for the signal of interest. For example, the assumptions of uncorrelated noise excitation and Gaussian distributions might not hold. Therefore, the experimenter should feel free to adjust the criterion to suit one's application. One simple way of "tuning" the criterion is to multiply N_e by an appropriate factor.

V. Data Compression by Linear Prediction

The methods outlined in Section II for the modeling of the behavior of a signal can be very useful in data compression. The process of signal or system modeling is essentially one of redundancy removal, which is the essence of data compression.

[15] Akaike informed me that he usually recommends $p_{max} < 3N^{1/2}$ as the maximum value of p that should be used if one is interested in a reliable estimate.

The idea of attempting to predict the value of a signal from previous sample values has been labeled in communications as "predictive coding" [28]. Adaptive linear prediction has been used extensively in speech and video transmission [7], [9], [23], [44], [50], [57], [83]. For the purposes of transmission one must quantize and transmit the predictor parameters or some transformation thereof. It has been known for some time that the quantization of the predictor parameters themselves is quite inefficient since a large number of bits is required to retain the desired fidelity in the reconstructed signal at the receiver [72]. Below, several equivalent representations of the predictor are presented and their quantization properties are discussed. We shall continue to assume that $H(z)$ is always stable, and hence minimum phase. $A(z)$ is, of course, then also minimum phase.

A. Alternate Representations of Linear Predictor

The following is a list of possible sets of parameters that characterize uniquely the all-pole filter $H(z)$ or its inverse $A(z)$.

1) (a) Impulse response of the inverse filter $A(z)$, i.e., predictor parameters a_k, $1 \leq k \leq p$.

(b) Impulse response of the all-pole model h_n, $0 \leq n \leq p$, which is defined in (30). Note that the first $p + 1$ coefficients uniquely specify the filter.

2) (a) Autocorrelation coefficients of a_k, $\rho(i)$, $0 \leq i \leq p$, as defined in (66).

(b) Autocorrelation coefficients of h_n, $\hat{R}(i)$, $0 \leq i \leq p$, as defined in (31) and (32).

3) Spectral coefficients of $A(z)$, Γ_i, $0 \leq i \leq p$ (or equivalently spectral coefficients of $H(z)$, G^2/Γ_i):

$$\Gamma_i = \rho(0) + 2 \sum_{j=1}^{p} \rho(j) \cos \frac{2\pi ij}{2p+1}, \qquad 0 \leq i \leq p \tag{92}$$

where $\rho(i)$ are as defined in (66). In other words, $\{\Gamma_i\}$ is obtained from $\{\rho(i)\}$ by a discrete Fourier transform.

4) Cepstral coefficients of $A(z)$, c_n, $1 \leq n \leq p$ (or equivalently cepstral coefficients of $H(z)$, $-c_n$):

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log A(e^{j\omega}) e^{jn\omega} \, d\omega. \tag{93}$$

Since $A(z)$ is minimum phase, (93) reduces to [38], [77]

$$c_n = a_n - \sum_{m=1}^{n-1} \frac{m}{n} c_m a_{n-m}, \qquad 1 \leq n \leq p. \tag{94}$$

Equation (94) is an iterative method for the computation of the cepstral coefficients directly from the predictor coefficients.

5) Poles of $H(z)$ or zeros of $A(z)$, z_k, $1 \leq k \leq p$, where $\{z_k\}$ are either real or form complex conjugate pairs. Conversion of the roots to the s plane can be achieved by setting each root $z_k = e^{s_k T}$, where $s_k = \sigma_k + j\omega_k$ is the corresponding pole in the s plane, and T is the sampling period. If the root $z_k = z_{kr} + jz_{ki}$, then

$$\omega_k = \frac{1}{T} \arctan \frac{z_{ki}}{z_{kr}}$$

$$\sigma_k = \frac{1}{2T} \log (z_{kr}^2 + z_{ki}^2) \tag{95}$$

where z_{kr} and z_{ki} are the real and imaginary parts of z_k, respectively.

6) Reflection coefficients k_i, $1 \leqslant i \leqslant p$, which are obtained as a byproduct of the solution of the autocorrelation normal equations, as in (38), or from the backward recursion (51).

Some of the preceding sets of parameters have $p + 1$ coefficients while others have only p coefficients. However, for the latter sets the gain G needs to be specified as well, thus keeping the total number of parameters as $p + 1$ for all the cases.

For purposes of data transmission, one is usually interested in recovering the predictor coefficients from the parameters that are chosen for transmission. The required transformations are clear for most of the above parameters, except perhaps for the parameters $\rho(i)$ and Γ_i. Through an inverse DFT, the spectral coefficients Γ_i can be converted to autocorrelation coefficients $\rho(i)$. One method of recovering the predictor parameters from $\{\rho(i)\}$ is as follows. Apply a DFT to the sequence $\{\rho(i)\}$ after appending it with an appropriate number of zeros to achieve sufficient resolution in the resulting spectrum of $A(z)$. Divide G^2 by this spectrum to obtain the spectrum of the filter $H(z)$. Inverse Fourier transformation of the spectrum of $H(z)$ yields the autocorrelation coefficients $\hat{R}(i)$. The first $p + 1$ coefficients $\hat{R}(i)$, $0 \leqslant i \leqslant p$, are then used to compute the predictor coefficients via the normal equations (12) with $R(i) = \hat{R}(i)$.

B. Quantization Properties

Although the sets of parameters given above provide equivalent information about the linear predictor, their properties under quantization are different. For the purpose of quantization, two desirable properties for a parameter set to have are: 1) filter stability upon quantization and 2) a natural ordering of the parameters. Property 1) means that the poles of $H(z)$ continue to be inside the unit circle even after parameter quantization. By 2), we mean that the parameters exhibit an inherent ordering, e.g., the predictor coefficients are ordered as $a_1, a_2, \cdots, a_p$. If a_1 and a_2 are interchanged then $H(z)$ is no longer the same in general, thus illustrating the existence of an ordering. The poles of $H(z)$, on the other hand, are not naturally ordered since interchanging the values of any two poles does not change the filter. When an ordering is present, a statistical study on the distribution of individual parameters can be used to develop better encoding schemes. Only the poles and the reflection coefficients insure stability upon quantization, while all the sets of parameters except the poles possess a natural ordering. Thus only the reflection coefficients possess both of these properties.

In an experimental study [63] of the quantization properties of the different parameters, it was found that the impulse responses $\{a_k\}$ and $\{h_n\}$ and the autocorrelations $\{\rho(i)\}$ and $\{\hat{R}(i)\}$ are highly susceptible to causing instability of the filter upon quantization. Therefore, these sets of parameters can be used only under minimal quantization, in which case the transmission rate would be excessive.

In the experimental investigation of the spectral and cepstral parameters, it was found that the quantization properties of these parameters are generally superior to those of the impulse responses and autocorrelation coefficients. The spectral parameters often yield results comparable to those obtained by quantizing the reflection coefficients. However, for the cases when the spectrum consists of one or more very sharp peaks (narrow bandwidths), the effects of quantizing the spectral coefficients often cause certain regions in the reconstructed spectrum (as described in the previous section) to become negative, which leads to instability of the computed filter. Quantization of the cepstral coefficients can also lead to instabilities. It should be noted here that the quantization properties of these parameters give better results if the spectral dynamic range of the signal is limited by some form of preprocessing.

Filter stability is preserved under quantization of the poles. But poles are expensive to compute, and they do not possess a natural ordering.

The conclusion is that, of the sets of parameters given in the preceding, the reflection coefficients are the best set to use as transmission parameters. In addition to ease of computation, stability under quantization, and natural ordering, the values of the reflection coefficients k_i, $i < p$, do not change as p is increased, unlike any of the other parameters. In the following, we discuss the optimal quantization of the reflection coefficients.

C. Optimal Quantization [53], [65]

Optimal quantization of the reflection coefficients depends on the fidelity criterion chosen. For many applications, it is important that the log spectrum of the all-pole model be preserved. In this case, it is reasonable to study the sensitivity of the log spectrum with respect to changes in the reflection coefficients. In a recent study [65], a spectral sensitivity curve was plotted versus each of the reflection coefficients k_i for many different all-pole models obtained by analyzing a large number of speech samples. The results of the study show that each sensitivity curve versus k_i has the same general shape, irrespective of the index i. Each sensitivity curve is U-shaped; it is even-symmetric about $k_i = 0$, with large values when $|k_i| \to 1$, and small values when $|k_i|$ is close to zero. These properties indicate that linear quantization of the reflection coefficients is not desirable, especially if some of them take values very close to 1, which happens when the spectrum contains sharp resonances. Nonlinear quantization of k_i is equivalent to a linear quantization of another parameter, say g_i, which is related to k_i by a nonlinear transformation. The requirement that the spectral sensitivity of the new parameters be flat resulted in the following optimal transformation [65]:

$$g_i = \log \frac{1 + k_i}{1 - k_i}, \qquad \text{all } i. \tag{96}$$

It is interesting to note from (45) that g_i is simply the logarithm of the hypothetical impedance ratios corresponding to k_i.

The optimality of the preceding transformation was based on a specific spectral fidelity criterion. Other transformations would result if other quantization fidelity criteria were adopted.[16]

The transmission rate can be reduced further without affecting the fidelity by proper encoding of each parameter. Variable word length encoding [34] (such as Huffman) can be used for this purpose if the statistical distributions of each of the parameters is known. These distributions can be obtained very simply from a representative sample of signals.

VI. Pole-Zero Modeling

Given the spectrum of some arbitrary signal, it is generally not possible to determine for certain the identity of the system that generated the signal in terms of a set of poles and

[16] Using a log likelihood criterion, Itakura informed me that a transformation proportional to arcsin (k_i) is optimal.

zeros. The problem is inherently nondeterministic, for a zero can be approximated arbitrarily closely by a large number of poles and vice versa. Indeed, we have seen in this paper that the all-pole model spectrum can approximate the signal spectrum arbitrarily closely by simply increasing the number of poles. However, if there are a number of influential zeros in the signal spectrum, the number of model poles can become very large. For data compression applications, this is an undesirable situation. Also, there are applications where the identification of the zeros is important. Therefore, it is useful to be able to model a spectrum in terms of poles and zeros.

Much effort is currently being expended on the problem of pole-zero modeling [8], [17], [19], [20], [29], [30], [88], [91]. Most of these methods are purely in the time domain. However, there seems to be a growing interest in frequency domain methods [4], [46], [94], partly due to the speed offered by the FFT. Add to this the market availability of spectrum analyzers and special hardware FFT processors. Of course, the time and frequency domain approaches should give similar results since LP analysis is actually performed in the autocorrelation domain.

The beauty of all-pole modeling is that it is relatively simple, straightforward, well understood, inexpensive, and "always" works. Unfortunately, none of these properties apply to pole-zero modeling. The main difficulty is that the pole-zero problem is nonlinear. We show this below for the stationary case. Then, we sketch out representative schemes for iterative and noniterative estimation of the pole and zero parameters. No exhaustive analysis is attempted; the reader is referred to the aforementioned references.

A. Normal Equations

The transfer function of the pole-zero model is given by $H(z)$ in (2). The corresponding model spectrum is given by

$$\hat{P}(\omega) = |H(e^{j\omega})|^2 = G^2 \frac{|B(e^{j\omega})|^2}{|A(e^{j\omega})|^2} = G^2 \frac{N(\omega)}{D(\omega)} \tag{97}$$

where $B(z)$ and $A(z)$ are the numerator and denominator polynomials in $H(z)$, and the all-zero spectra $N(\omega)$ and $D(\omega)$ form the numerator and denominator of $\hat{P}(\omega)$ and are given by

$$N(\omega) = \left| 1 + \sum_{l=1}^{q} b_l e^{-jl\omega} \right|^2 \tag{98}$$

and

$$D(\omega) = \left| 1 + \sum_{k=1}^{p} a_k e^{-jk\omega} \right|^2. \tag{99}$$

The matching error between the signal spectrum $P(\omega)$ and $\hat{P}(\omega)$ is given by (62), and from (97) is equal to

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \frac{D(\omega)}{N(\omega)} d\omega. \tag{100}$$

E can be interpreted as the residual energy obtained by passing the signal through the filter $A(z)/B(z)$. The problem is to determine $\{a_k\}$ and $\{b_l\}$ such that E in (100) is minimized.

In the sequel we shall make use of the following two relations:

$$\frac{\partial N(\omega)}{\partial b_i} = 2 \sum_{l=0}^{q} b_l \cos(i-l)\omega, \quad b_0 = 1 \tag{101}$$

$$\frac{\partial D(\omega)}{\partial a_i} = 2 \sum_{k=0}^{p} a_k \cos(i-k)\omega, \quad a_0 = 1. \tag{102}$$

In addition, we shall use the notation $R_{\alpha\beta}(i)$ to represent the autocorrelation defined by

$$R_{\alpha\beta}(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \frac{[D(\omega)]^{\beta}}{[N(\omega)]^{\alpha}} \cos(i\omega)\, d\omega. \tag{103}$$

Thus, $R_{00}(i)$ is simply the Fourier transform of $P(\omega)$.[17] Taking $\partial E/\partial a_i$ in (100) one obtains

$$\frac{\partial E}{\partial a_i} = 2 \sum_{k=0}^{p} a_k R_{10}(i-k), \quad 1 \leqslant i \leqslant p. \tag{104}$$

Similarly, one can show that

$$\frac{\partial E}{\partial b_i} = -2 \sum_{l=0}^{q} b_l R_{21}(i-l), \quad 1 \leqslant i \leqslant q. \tag{105}$$

In order to minimize E, we set $\partial E/\partial a_i = 0$, $1 \leqslant i \leqslant p$, and $\partial E/\partial b_i = 0$, $1 \leqslant i \leqslant q$, simultaneously. These, then, comprise the normal equations.

From (103), it is clear that $R_{10}(i-k)$ is not a function of a_k. Therefore, setting (104) to zero results in a set of linear equations, identical in form to the autocorrelation normal equations (12). However, $R_{21}(i-l)$ in (105) is a function of b_l, as can be deduced from (103) with $\alpha = 2$ and $\beta = 1$. Therefore, setting (105) to zero results in a set of nonlinear equations in b_l. If one wishes to solve for $\{a_k\}$ and $\{b_l\}$ simultaneously, then one solves a set of $p+q$ *nonlinear* equations.

Note that if the signal is assumed to be nonstationary, the above analysis can be modified accordingly in a manner similar to that in Section III-A for the all-pole case. The resulting equations will be very similar in form to the preceding equations, with nonstationary autocorrelations replacing the stationary autocorrelations.

B. Iterative Solutions

Since the minimization of E in (100) leads to a set of nonlinear equations, the problem of minimizing E must then be solved iteratively. There are many methods in the literature for finding the extrema of a function [30], [108], many of which are applicable in our case. In particular, gradient methods are appropriate here since it is possible to evaluate the error gradient, as in (104) and (105). One such method was used by Tretter and Steiglitz [94] in pole-zero modeling. Other schemes, such as the Newton–Raphson method, require the evaluation of the Hessian (i.e., second derivative). This can be very cumbersome in many problems, but is straightforward in our case. This is illustrated below by giving a Newton–Raphson solution.

Let $x' = [a_1 a_2 \cdots a_p b_1 b_2 \cdots b_q]$ be the transpose of a column vector x whose elements are the coefficients a_k and b_l. If $x(m)$ is the solution at iteration m, then $x(m+1)$ is given by

$$x(m+1) = x(m) - J^{-1} \left. \frac{\partial E}{\partial x} \right|_{x=x(m)} \tag{106}$$

where J is the $(p+q) \times (p+q)$ symmetric Hessian matrix

[17] For a discrete spectrum $P(\omega_n)$ the integrals in (100) and (103) are replaced by summations.

given by $J = \partial^2 E/\partial x \partial x'$. Setting $a' = [a_1 a_2 \cdots a_p]$ and $b' = [b_1 b_2 \cdots b_q]$, (106) can be partitioned, with $x' = [a' b']$, as

$$\begin{bmatrix} a(m+1) \\ b(m+1) \end{bmatrix} = \begin{bmatrix} a(m) \\ b(m) \end{bmatrix} - \left[\begin{array}{c|c} \dfrac{\partial^2 E}{\partial a \partial a'} & \dfrac{\partial^2 E}{\partial a \partial b'} \\ \hline \dfrac{\partial^2 E}{\partial b \partial a'} & \dfrac{\partial^2 E}{\partial b \partial b'} \end{array}\right]^{-1}_{\substack{a=a(m) \\ b=b(m)}} \begin{bmatrix} \dfrac{\partial E}{\partial a} \\ \dfrac{\partial E}{\partial b} \end{bmatrix}_{\substack{a=a(m) \\ b=b(m)}} . \tag{107}$$

The elements of the first-order partial derivatives in (107) are given by (104) and (105). The elements of the second partial derivatives can be shown to be equal to

$$\frac{\partial^2 E}{\partial a_i \partial a_j} = 2R_{10}(i-j) \tag{108}$$

$$\frac{\partial^2 E}{\partial a_i \partial b_j} = -2 \sum_{k=0}^{p} \sum_{l=0}^{q} a_k b_l \cdot [R_{20}(j+i-l-k) + R_{20}(j-i-l+k)] \tag{109}$$

$$\frac{\partial^2 E}{\partial b_i \partial b_j} = -2R_{21}(i-j) + 4 \sum_{k=0}^{p} \sum_{l=0}^{q} b_k b_l \cdot [R_{31}(j+i-l-k) + R_{31}(j-i-l+k)]. \tag{110}$$

Given the estimates $a(m)$ and $b(m)$, one can compute $N(\omega)$ and $D(\omega)$ from (98) and (99) using FFT's, and then use (103) to compute the autocorrelations R_{10}, R_{20}, R_{21}, and R_{31}, which can then be used in (107)–(110) to compute the new estimates $a(m+1)$ and $b(m+1)$. The iterations are halted when the error gradient goes below some prespecified threshold. The minimum error is then computed from (100).

The Newton–Raphson method works very well if the initial estimate is close to the optimum. In that case, the Hessian J is positive definite and the convergence is quadratic [30]. In the next section we discuss noniterative methods which can be used to give these good initial estimates.

C. Noniterative Solutions

One property that is common to noniterative methods is that a good estimate of the number of poles and zeros seems to be necessary for a reasonable solution. Indeed, in that case, there is not much need to go to expensive iterative methods. However, in general, such information is unavailable and one is interested in obtaining the best estimate for a given p and q. Then, noniterative methods can be used profitably to give good initial estimates that are necessary in iterative methods.

1) Pole Estimation: Assume that the signal s_n had been generated by exciting the pole-zero filter $H(z)$ in (2) by either an impulse or white noise. Then it is simple to show that the signal autocorrelation obeys the autocorrelation equation (31) for $i > q$. Therefore, the coefficients a_k can be estimated by solving (31) with $q+1 \leqslant i \leqslant q+p$.

The effect of the poles can now be removed by applying the inverse filter $A(z)$ to the signal. In the spectral domain this can be done by computing a new spectrum $P_1(\omega) = P(\omega) D(\omega)$. The problem now reduces to the estimation of the zeros in $P_1(\omega)$.

2) Zero Estimation: A promising noniterative method for pole-zero estimation is that of *cepstral prediction* [76], [95]. The basic idea is that the poles of nc_n, where c_n is the complex cepstrum, comprise the poles and zeros of the signal [77]. Therefore, for zero estimation, the problem reduces to finding the poles of nc_n which can be computed by the method just described above, where c_n here is the cepstrum corresponding to $P_1(\omega)$.

Another method for zero estimation is that of *inverse LP modeling* [63]. The idea is quite simple: Invert the spectrum $P_1(\omega)$ and apply a q-pole LP analysis. The resulting predictor coefficients are then good estimates of b_l. This method gives good results if $P_1(\omega)$ is smooth relative to the model spectrum. Problems arise if the variations of the signal spectrum about the model spectrum are large. The reason is that LP modeling attempts to make a good fit to the spectral envelope, and the envelope of the inverted spectrum is usually different from the inverse of the desired spectral envelope. One solution is to smooth the spectrum $P_1(\omega)$ before inversion. Spectral smoothing is usually performed by applying a low-pass filter to the spectrum (autocorrelation smoothing) or to the log spectrum (cepstral smoothing). Another method is all-pole smoothing. Indeed, all-pole modeling can be thought of as just another method of smoothing the spectrum, where the degree of smoothing is controlled by the order of the predictor p, which is usually chosen to be much larger than the number of zeros in the model q. We point out that zero estimation by inverse LP modeling with all-pole smoothing is similar to the method of Durbin [24], [25] in the time domain.

VII. Conclusion

Linear prediction is an autocorrelation-domain analysis. Therefore, it can be approached from either the time or frequency domain. The least squares error criterion in the time domain translates into a spectral matching criterion in the frequency domain. This viewpoint was helpful in exploring the advantages and disadvantages of the least squares error criterion.

The major portion of this paper was devoted to all-pole modeling. This type of modeling is simple, inexpensive and effective; hence its wide applicability and acceptance. In contrast, pole-zero modeling is not simple, generally expensive, and is not yet well understood. Future research should be directed at acquiring a better understanding of the problems in pole-zero modeling and developing appropriate methodologies to deal with these problems.

Acknowledgment

The author wishes to thank the following friends and colleagues who have read and commented on earlier versions of this paper: H. Akaike, R. Barakat, C. Cook, T. Fortmann, F. Itakura, G. Kopec, A. Oppenheim, L. Rabiner, R. Viswanathan, and V. Zue. He is especially grateful to R. Viswanathan for the fruitful discussions they have had throughout the writing of the paper and for his help in supplying many of the references. He also wishes to thank B. Aighes, C. Williams, and R. Schwartz for their assistance in the preparation of the manuscript.

References

[1] H. Akaike, "Power spectrum estimation through autoregressive model fitting," *Ann. Inst. Statist. Math.*, vol. 21, pp. 407–419, 1969.

[2] ——, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, vol. 22, pp. 203–217, 1970.

[3] ——, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Information Theory* (Supplement to Problems of Control and Information Theory), 1972.

[4] ——, "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika*, vol. 60, no. 2, pp. 255–265, 1973.
[5] ——, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, Dec. 1974.
[6] T. W. Anderson, *The Statistical Analysis of Time Series.* New York: Wiley, 1971.
[7] C. A. Andrews, J. M. Davies, and G. R. Schwartz, "Adaptive data compression," *Proc. IEEE*, vol. 55, pp. 267–277, Mar. 1967.
[8] K. J. Åström and P. Eykhoff, "System identification—A survey," *Automatica*, vol. 7, pp. 123–162, 1971.
[9] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, no. 6, pp. 1973–1986, Oct. 1970.
[10] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 637–655, 1971.
[11] M. S. Bartlett, *An Introduction to Stochastic Processes with Special Reference to Methods and Applications.* Cambridge, England: Cambridge Univ. Press, 1956.
[12] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures.* New York: Wiley, 1971.
[13] C. Bingham, M. D. Godfrey, and J. W. Tukey, "Modern techniques of power spectrum estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 56–66, June 1967.
[14] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra.* New York: Dover, 1958.
[15] T. Bohlin, "Comparison of two methods of modeling stationary EEG signals," *IBM J. Res. Dev.*, pp. 194–205, May 1973.
[16] S. F. Boll, "*A priori* digital speech analysis," Computer Science Div., Univ. Utah, Salt Lake City, UTEC-CSC-73-123, 1973.
[17] G. E. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control.* San Francisco, Calif.: Holden-Day, 1970.
[18] J. P. Burg, "The relationship between maximum entropy spectra and maximum likelihood spectra," *Geophysics*, vol. 37, no. 2, pp. 375–376, Apr. 1972.
[19] C. S. Burrus and T. W. Parks, "Time domain design of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 137–141, June 1970.
[20] J. C. Chow, "On estimating the orders of an autoregressive moving-average process with uncertain observations," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 707–709, Oct. 1972.
[21] W. T. Cochran *et al.*, "What is the fast Fourier transform?," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 45–55, June 1967.
[22] H. Cox, "Linear versus logarithmic averaging," *J. Acoust. Soc. Amer.*, vol. 39, no. 4, pp. 688–690, 1966.
[23] L. D. Davisson, "The theoretical analysis of data compression systems," *Proc. IEEE*, vol. 56, pp. 176–186, Feb. 1968.
[24] J. Durbin, "Efficient estimation of parameters in moving-average models," *Biometrika*, vol. 46, Parts 1 and 2, pp. 306–316, 1959.
[25] ——, "The fitting of time-series models," *Rev. Inst. Int. Statist.*, vol. 28, no. 3, pp. 233–243, 1960.
[26] A. Eberhard, "An optimal discrete window for the calculation of power spectra," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 37–43, Feb. 1973.
[27] M. P. Ekstrom, "A spectral characterization of the ill-conditioning in numerical deconvolution," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 344–348, Aug. 1973.
[28] P. Elias, "Predictive coding, Parts I and II," *IRE Trans. Inform. Theory*, vol. IT-1, p. 16, Mar. 1955.
[29] A. G. Evans and R. Fischl, "Optimal least squares time-domain synthesis of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 61–65, Feb. 1973.
[30] P. Eykhoff, *System Identification: Parameter and State Estimation.* New York: Wiley, 1974.
[31] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra.* San Francisco, Calif.: Freeman, 1963.
[32] P. B. C. Fenwick, P. Michie, J. Dollimore, and G. W. Fenton, "Mathematical simulation of the electroencephalogram using an autoregressive series," *Bio-Med. Comput.*, vol. 2, pp. 281–307, 1971.
[33] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd Edition. New York: Springer-Verlag, 1972.
[34] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.
[35] L. Y. Geronimus, *Orthogonal Polynomials.* New York: Consultants Bureau, 1961.
[36] W. Gersch, "Spectral analysis of EEG's by autoregressive decomposition of time series," *Math. Biosci.*, vol. 7, pp. 205–222, 1970.
[37] W. Gersch and D. R. Sharpe, "Estimation of power spectra with finite-order autoregressive models," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 367–369, Aug. 1973.
[38] B. Gold and C. M. Rader, *Digital Processing of Signals.* New York: McGraw-Hill, 1969.
[39] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numer. Math.*, vol. 14, pp. 403–420, 1970.
[40] A. H. Gray and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 207–216, June 1974.
[41] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, Nov. 1972.
[42] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications.* Berkeley, Calif.: Univ. California Press, 1958.
[43] U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series.* New York: Wiley, 1957.
[44] A. Habibi and G. S. Robinson, "A survey of digital picture coding," *Computer*, pp. 22–34, May 1974.
[45] E. J. Hannan, *Time Series Analysis.* London, England: Methuen, 1960.
[46] ——, *Multiple Time Series.* New York: Wiley, 1970.
[47] J. R. Haskew, J. M. Kelly, R. M. Kelly, Jr., and T. H. McKinney, "Results of a study of the linear prediction vocoder," *IEEE Trans. Commun.*, vol. COM-21, pp. 1008–1014, Sept. 1973.
[48] R. L. Hershey, "Analysis of the difference between log mean and mean log averaging," *J. Acoust. Soc. Amer.*, vol. 51, pp. 1194–1197, 1972.
[49] F. B. Hildebrand, *Introduction to Numerical Values.* New York: McGraw-Hill, 1956.
[50] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Rep. 6th Int. Congr. Acoustics*, Y. Kohasi, Ed., pp. C17–C20, Paper C-5-5, Aug. 1968.
[51] ——, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970.
[52] ——, "Digital filtering techniques for speech analysis and synthesis," in *Conf. Rec., 7th Int. Congr. Acoustics*, Paper 25 C 1, 1971.
[53] ——, "On the optimum quantization of feature parameters in the Parcor speech synthesizer," in *IEEE Conf. Rec., 1972 Conf. Speech Communication and Processing*, pp. 434–437, Apr. 1972.
[54] G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications.* San Francisco, Calif.: Holden-Day, 1968.
[55] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 146–181, Mar. 1974.
[56] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, J. Basic Eng.*, Series D82, pp. 35–45, 1960.
[57] H. Kobayashi and L. R. Bahl, "Image data compression by predictive coding I: Prediction algorithms," *IBM J. Res. Dev.*, pp. 164–171, Mar. 1974.
[58] A. Kolmogorov, "Interpolation und Extrapolation von stationären zufälligen Folgen," *Bull. Acad. Sci., U.S.S.R., Ser. Math.*, vol. 5, pp. 3–14, 1941.
[59] K. S. Kunz, *Numerical Analysis.* New York: McGraw-Hill, 1957.
[60] R. C. K. Lee, *Optimal Estimation, Identification, and Control*, Research Monograph no. 28. Cambridge, Mass.: M.I.T. Press, 1964.
[61] N. Levinson, "The Wiener RMS (root mean square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, no. 4, pp. 261–278, 1947. Also Appendix B, in N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* Cambridge, Mass.: M.I.T. Press, 1949.
[62] J. Makhoul, "Spectral analysis of speech by linear prediction," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 140–148, June 1973.
[63] ——, "Selective linear prediction and analysis-by-synthesis in speech analysis," Bolt Beranek and Newman Inc., Cambridge, Mass., Rep. 2578, Apr. 1974.
[64] J. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman Inc., Cambridge, Mass., NTIS AD-749066, Rep. 2304, Aug. 1972.
[65] J. Makhoul and R. Viswanathan, "Quantization properties of transmission parameters in linear predictive systems," Bolt Beranek and Newman Inc., Cambridge, Mass., Rep. 2800, Apr. 1974.
[66] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica*, vol. 11, nos. 3 and 4, pp. 173–220, 1943.
[67] W. D. Mark, "Spectral analysis of the convolution and filtering of non-stationary stochastic processes," *J. Sound Vib.*, no. 11, pp. 19–63, 1970.
[68] J. D. Markel, "Digital inverse filtering—A new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.

[69] J. D. Markel and A. H. Gray, "On autocorrelation equations as applied to speech analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 69–79, Apr. 1973.
[70] E. Matsui *et al.*, "An adaptive method for speech analysis based on Kalman filtering theory," *Bull Electrotech. Lab*, vol. 36, no. 3, pp. 42–51, 1972 (in Japanese).
[71] R. N. McDonough and W. H. Huggins, "Best least-squares representation of signals by exponentials," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 408–412, Aug. 1968.
[72] S. K. Mitra and R. J. Sherwood, "Estimation of pole-zero displacements of a digital filter due to coefficient quantization," *IEEE Trans. Circuits Syst.*, vol. CAS-21, pp. 116–124, Jan. 1974.
[73] R. E. Nieman, D. G. Fisher, and D. E. Seborg, "A review of process identification and parameter estimation techniques," *Int. J. Control*, vol. 13, no. 2, pp. 209–264, 1971.
[74] R. H. Norden, "A survey of maximum likelihood estimation," *Int. Statist. Rev.*, vol. 40, no. 3, pp. 329–354, 1972.
[75] ——, "A survey of maximum likelihood estimation, Part 2," *Int. Statist. Rev.*, vol. 41, no. 1, pp. 39–58, 1973.
[76] A. Oppenheim and J. M. Tribolet, "Pole-zero modeling using cepstral prediction," Res. Lab. Electronics, M.I.T., Cambridge, Mass., QPR 111, pp. 157–159, 1973.
[77] A. V. Oppenheim, R. W. Schafer and T. G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, Aug. 1968.
[78] M. Pagano, "An algorithm for fitting autoregressive schemes," *J. Royal Statist. Soc., Series C (Applied Statistics)*, vol. 21, no. 3, pp. 274–281, 1972.
[79] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 1965.
[80] L. R. Rabiner *et al.*, "Terminology in digital signal processing," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 322–337, Dec. 1972.
[81] A. Ralston, *A First Course in Numerical Analysis.* New York: McGraw-Hill, 1965.
[82] A. Ralston and H. Wilf, *Mathematical Methods for Digital Computers*, vol. II. New York: Wiley, 1967.
[83] M. P. Ristenbatt, "Alternatives in digital communications," *Proc. IEEE*, vol. 61, pp. 703–721, June 1973.
[84] E. A. Robinson, *Multichannel Time Series Analysis with Digital Computer Programs.* San Francisco, Calif.: Holden-Day, 1967.
[85] ——, *Statistical Communication and Detection.* New York: Hafner, 1967.
[86] ——, "Predictive decomposition of time series with application to seismic exploration," *Geophysics*, vol. 32, no. 3, pp. 418–484, June 1967.
[87] E. A. Robinson and S. Treitel, *The Robinson-Treitel Reader*, 3rd ed. Tulsa, Okla.: Seismograph Service Corp., 1973.
[88] A. P. Sage and J. L. Melsa, *System Identification.* New York: Academic Press, 1971.
[89] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," this issue, pp. 662–677.
[90] F. C. Schweppe, *Uncertain Dynamic Systems.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.
[91] J. L. Shanks, "Recursion filters for digital processing," *Geophysics*, vol. XXXII, no. 1, pp. 33–51, 1967.
[92] R. A. Silverman, "Locally stationary random processes," *IRE Trans. Inform. Theory*, vol. IT-3, pp. 182–187, Sept. 1957.
[93] G. Szegö, *Orthogonal Polynomials.* New York: Amer. Math. Soc. Colloquium Publ., vol. XXIII, N.Y., 1959.
[94] S. A. Tretter and K. Steiglitz, "Power-spectrum identification in terms of rational models," *IEEE Trans. Automat. Control*, vol. AC-12, pp. 185–188, Apr. 1967.
[95] J. M. Tribolet, "Identification of linear discrete systems with applications to speech processing," Master's thesis, Dep. Elec. Eng., M.I.T., Cambridge, Mass., Jan. 1974.
[96] A. Van den Bos, "Alternative interpretation of maximum entropy spectral analysis," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 493–494, July 1971.
[97] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–427, Oct. 1973.
[98] G. Walker, "On periodicity in series of related terms," *Proc. Royal Soc.*, vol. 131-A, p. 518, 1931.
[99] R. J. Wang, "Optimum window length for the measurement of time-varying power spectra," *J. Acoust. Soc. Amer.*, vol. 52, no. 1 (part 1), pp. 33–38, 1971.
[100] R. J. Wang and S. Treitel, "The determination of digital Wiener filters by means of gradient methods," *Geophysics*, vol. 38, no. 2, pp. 310–326, Apr. 1973.
[101] P. D. Welch, "A direct digital method of power spectrum estimation," *IBM J. Res. Dev.*, vol. 5, pp. 141–156, Apr. 1961.
[102] A. Wennberg and L. H. Zetterberg, "Application of a computer-based model for EEG analysis," *Electroencephalogr. Clin. Neurophys.*, vol. 31, no. 5, pp. 457–468, 1971.
[103] P. Whittle, "Some recent contributions to the theory of stationary processes," Appendix 2, in H. Wold, *A Study in the Analysis of Stationary Time Series.* Stockholm, Sweden: Almqvist and Wiksell, 1954.
[104] ——, "On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix," *Biometrika*, vol. 50, nos. 1 and 2, pp. 129–134, 1963.
[105] ——, *Prediction and Regulation by Linear Least Square Methods.* London, England: English Universities Press, 1963.
[106] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, "Adaptive antenna systems," *Proc. IEEE*, vol. 55, pp. 2143–2159, Dec. 1967.
[107] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series With Engineering Applications.* Cambridge, Mass.: M.I.T. Press, 1949.
[108] D. J. Wilde, *Optimum Seeking Methods.* Englewood Cliffs, N.J.: Prentice-Hall, 1964.
[109] J. H. Wilkinson, "Error analysis of direct methods of matrix inversion," *J. Ass. Comput. Mach.*, vol. 8, no. 3, pp. 281–330, 1961.
[110] ——, "The solution of ill-conditioned linear equations," in Ralston and Wilf, *Mathematical Methods for Digital Computers*, pp. 65–93, 1967.
[111] J. H. Wilkinson and C. Reinsch, *Linear Algebra*, vol. II. New York: Springer-Verlag, 1971.
[112] H. Wold, *A Study in the Analysis of Stationary Time Series.* Stockholm, Sweden: Almqvist and Wiksell, 1954.
[113] H. O. A. Wold, *Bibliography on Time Series and Stochastic Processes.* Cambridge, Mass.: M.I.T. Press, 1965.
[114] L. C. Wood and S. Treitel, "Seismic signal processing," this issue, pp. 649–661.
[115] G. U. Yule, "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers," *Phil. Trans. Roy. Soc.*, vol. 226-A, pp. 267–298, 1927.