speech that includes a larger class of sounds than those we have considered.

The second question is the effect of line errors on synthesizer output when the delta modulator is used as a transmitter. This is a problem that calls for discussion even for the case when the control signals are transmitted by conventional PCM. However, with DM, the control signal sequences that could be processed in our simulation had to be limited to 200 samples; and this precluded, in our opinion, any meaningful study of the line error problem.

Acknowledgment

The author wishes to thank Dr. R. W. Schafer for supplying the control signals and the digital low-pass filter, Dr. L. R. Rabiner for synthesizing the delta-modulator outputs, and Mrs. K. Shipley for programming assistance.

References

[1] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, part 2, pp. 634-648, Feb. 1970.
[2] A. E. Rosenberg, R. W. Schafer, and L. R. Rabiner, "Effect of smoothing and quantizing parameters of formant-coded voiced speech," *J. Acoust. Soc. Amer.*, vol. 50, part 2, pp. 1532-1538, Dec. 1972.
[3] F. de Jager, "Delta modulation, a method of PCM transmission using a 1-unit code," *Phillips Res. Rep.*, no. 60, pp. 442-466, Dec. 1952.
[4] J. E. Abate, "Linear and adaptive delta modulation," *Proc. IEEE (Special Issue on Redundancy Reduction)*, vol. 55, pp. 298-308, Mar. 1967.
[5] N. S. Jayant, "Adaptive delta modulation with a one-bit memory," *Bell Syst. Tech. J.*, pp. 321-342, Mar. 1970.
[6] —, "Characteristics of a delta modulator," *Proc. IEEE* (Lett.), vol. 59, pp. 428-429, Mar. 1971.
[7] L. R. Rabiner, B. Gold, and C. A. McGonegal, "An approach to the approximation problem for nonrecursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 83-106, June 1970.
[8] L. R. Rabiner, R. W. Schafer, and J. L. Flanagan, "Computer synthesis of speech by concatenation of formant-coded words," *Bell Syst. Tech. J.*, pp. 1541-1558, May-June 1971.

Spectral Analysis of Speech by Linear Prediction

JOHN MAKHOUL

Abstract—The autocorrelation method of linear prediction is formulated in the time, autocorrelation, and spectral domains. The analysis is shown to be that of approximating the short-time signal power spectrum by an all-pole spectrum. The method is compared with other methods of spectral analysis such as analysis-by-synthesis and cepstral smoothing. It is shown that this method can be regarded as another method of analysis-by-synthesis where a number of poles is specified, with the advantages of noniterative computation and an error measure which leads to a better spectral envelope fit for an all-pole spectrum. Compared to spectral analysis by cepstral smoothing in conjunction with the chirp z transform (CZT), this method is expected to give a better spectral envelope fit (for an all-pole spectrum) and to be less sensitive to the effects of high pitch on the spectrum.

The normalized minimum error is defined and its possible usefulness as a voicing detector is discussed.

Manuscript received April 30, 1972. This work was supported by the Information Processing Techniques Branch of the Advanced Research Projects Agency.

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

I. Introduction

Although predictive coding has been used in communication for some time now, it was not applied to speech analysis until recently. During the past two years several formulations for the linear prediction of speech have been suggested, of which two least squares formulations have become prominent. We shall call these the autocorrelation and covariance methods of linear prediction. The autocorrelation method is represented by the works of Markel [1] with his digital inverse filtering formulation, and of Itakura and Saito [2] with their maximum-likelihood method. The covariance method is represented by the work of Atal and Hanauer [3]; this method is similar to what is known as Prony's method [4]. A detailed comparison between the autocorrelation and covariance methods is given in Makhoul and Wolf [5]. Although both methods can be derived in the time domain, they can also be derived from a spectral analysis-by-synthesis formulation. The autocorrelation method is derived through the analysis of a stationary short-time spectrum, while the covariance method is derived through the analysis of a nonstationary two-dimensional short-time spectrum [5]. In this paper we shall restrict our attention to the spectral analysis of speech by the autocorrelation method of linear prediction. The method is compared with traditional methods of spectral analysis such as Fourier transformation, cepstral smoothing, and analysis-by-synthesis.

II. Linear Prediction

The various methods of linear prediction have in common the assumption that a speech sample $s(nT)$, where T is the sampling interval, can be predicted approximately from a linearly weighted summation of a number of immediately preceding samples. Let this approximation of $s(nT)$ be $s'(nT)$, given by

$$s'_n = \sum_{k=1}^{p} a_k \, s_{n-k} \tag{1}$$

where $s(nT)$ and $s'(nT)$ are represented by s_n and s'_n, respectively, a_k, $1 \leqslant k \leqslant p$ is a set of real constants known as the predictor coefficients that are to be computed, and p is the order of the predictor. Let the error between the actual value and the predicted value be given by e_n, where

$$e_n = s_n - s'_n = s_n - \sum_{k=1}^{p} a_k \, s_{n-k} . \tag{2}$$

The problem is to find the predictor coefficients a_k that minimize the error e_n in some sense over the desired range of samples. The minimization of the total-squared error leads to a mathematically attractive solution. Denote the total-squared error by E, where

$$E = \sum_n e_n^2 = \sum_n (s_n - s'_n)^2 . \tag{3}$$

It is in the range over which the summation in (3) applies and the definition of the signal s_n in that range that the autocorrelation and covariance methods are different. Here we shall present only the autocorrelation method as derived by Markel [1]. Other formulations of the autocorrelation method can be found elsewhere [5].

We assume that the portion of the signal to be analyzed is multiplied by a finite (not necessarily rectangular) window of width N. We then have

$$s_n = \begin{cases} \text{some sampled signal}, & 0 \leqslant n \leqslant N-1, \\ 0, & n < 0 \text{ and } n \geqslant N. \end{cases} \tag{4}$$

Note that the windowed signal s_n in (4) is defined for all time, $-\infty < n < \infty$. Substituting (1) and (4) in (3) and setting $L = N - 1 + p$, we obtain

$$E = \sum_{n=-\infty}^{\infty} (s_n - s'_n)^2$$

$$= s_0^2 + \sum_{n=1}^{L} \left(s_n - \sum_{k=1}^{p} a_k \, s_{n-k} \right)^2 . \tag{5}$$

The condition for the minimization of the total-squared error is obtained by setting the partial derivative of E with respect to each a_k, $1 \leqslant k \leqslant p$, to zero. The result can be shown to be

$$\sum_{k=1}^{p} a_k \, R_{|i-k|} = R_i, \qquad 1 \leqslant i \leqslant p \tag{6}$$

where

$$R_i = \sum_{n=0}^{N-1-|i|} s_n \, s_{n+|i|} \tag{7}$$

is the autocorrelation function of the signal s_n. Therefore, in order to minimize the total-squared error E, compute the autocorrelation coefficients R_k, $0 \leqslant k \leqslant p$, using (7), and then solve (6) for the predictor coefficients a_k, $1 \leqslant k \leqslant p$. Henceforth, the coefficients a_k will be used to indicate those predictor coefficients that minimize the total-squared error, and are a solution to (6). Equation (6) is a matrix equation of a special form and there exist several recursive solutions for it [1], [5], [6].

The minimum total-squared error E_p can now be computed by making use of (6) in (5). The answer can be shown to be equal to

$$E_p = R_0 - \sum_{k=1}^{p} a_k \, R_k . \tag{8}$$

Of interest is the normalized minimum total-squared error which is discussed in Section V.

III. Spectral Approximation

Corresponding to the time-domain approximation of the sampled signal, as presented above, there is a frequency-domain approximation. In the rest of this paper we shall explore the nature of the spectral approximation.

Multiplying both sides of (2) by z^{-n} and summing over all n, we obtain

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

$$= S(z) \, H(z) \tag{9}$$

where $E(z)$ and $S(z)$ are the z transforms of e_n and s_n, respectively, and

$$H(z) = 1 - \sum_{k=1}^{p} a_k \, z^{-k} \tag{10}$$

where the predictor coefficients a_k are obtained from (6). Therefore, the minimum-error sequence e_n can be interpreted as the output of a filter $H(z)$ whose input is s_n. $H(z)$ is usually known as the inverse filter. (Another way to view the minimization problem in

Section II is to solve for the filter $H(z)$ which minimizes the energy $\Sigma_n e_n^2$ in the output error signal, for a given value of p.)

From (9) we have

$$S(z) = \frac{E(z)}{H(z)}. \tag{11}$$

Equation (11) is exact. Now comes the approximation: let us assume that the transfer function $S(z)$ of the signal is to be modeled by an all-pole filter $\hat{S}(z)$ of the form

$$\hat{S}(z) = \frac{A}{H(z)} = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{12}$$

where A is a gain factor to be computed. From (11), this means that $E(z)$ is approximated by another function $\hat{E}(z) = A$. In the time domain this is equivalent to the approximation of the error signal by an impulse:

$$\hat{e}_n = \begin{cases} A, & n = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The value of A is determined by the application of energy conservation between $\hat{e}_n$ and the minimum-error sequence e_n, i.e., $\Sigma_n \hat{e}_n^2 = \Sigma_n e_n^2$. But $\Sigma_n e_n^2$ is simply the minimum total-squared error. Therefore, from (13) and (8) we obtain

$$A^2 = E_p = R_0 - \sum_{k=1}^{p} a_k R_k, \tag{14}$$

and A^2 is equal to the minimum total-squared error.

Corresponding to the transfer functions $S(z)$ and $\hat{S}(z)$ we have the power spectra $P(\omega)$ and $\hat{P}(\omega)$ which are obtained by setting $z = e^{j\omega T}$ and taking the magnitude squared of the respective transfer functions.

Thus,

$$P(\omega) = |S(\omega)|^2 = \left| \sum_{n=0}^{N-1} s_n e^{-jn\omega T} \right|^2, \tag{15}$$

and

$$\hat{P}(\omega) = |\hat{S}(\omega)|^2 = \frac{A^2}{\left| 1 - \sum_{k=1}^{p} a_k e^{-jk\omega T} \right|^2}, \tag{16}$$

where A^2 is given by (14) and ω is the angular frequency. $P(\omega)$ as well as the denominator of $\hat{P}(\omega)$ can be computed for discrete values of frequency via the fast Fourier transform (FFT). The signal power spectrum $P(\omega)$ is to be approximated by an all-pole power spectrum $\hat{P}(\omega)$. The question then is: in what sense does $\hat{P}(\omega)$ approximate the signal spectrum $P(\omega)$ for speech? In the following sections we shall give interpretations in terms of: 1) the autocorrelation coefficients; 2) error minimization, the spectral envelope, and analysis-by-synthesis; and 3) the pole-zero configuration of the spectrum.

IV. Autocorrelation Analysis

Here we shall investigate the relation between the autocorrelation coefficients R_i of the speech signal and those corresponding to the approximate spectrum $\hat{P}(\omega)$ which will be denoted by $\hat{R}_i$. The coefficients R_i can be computed from (7) for all i. (Note that R_i is equal to zero for $|i| \geq N$.) $\hat{R}_i$ will be computed below as the autocorrelation function of the impulse response $\hat{s}_n$ of $\hat{S}(z)$. By rearranging (12) and taking the inverse z transform we have

$$\hat{s}_n = \begin{cases} 0, & n < 0, \\ A, & n = 0, \\ \sum_{k=1}^{p} a_k \hat{s}_{n-k}, & n > 0. \end{cases} \tag{17}$$

By definition, the autocorrelation function $\hat{R}_i$ is given by

$$\hat{R}_i = \sum_{n=-\infty}^{\infty} \hat{s}_n \hat{s}_{n+|i|} = \sum_{n=0}^{\infty} \hat{s}_n \hat{s}_{n+|i|}, \quad \text{for all } i. \tag{18}$$

From (17) and (18), it can be shown that

$$\hat{R}_i = \sum_{k=1}^{p} a_k \hat{R}_{|i-k|}, \quad 1 \leq |i| < \infty, \tag{19}$$

and

$$\hat{R}_0 = A^2 + \sum_{k=1}^{p} a_k \hat{R}_k. \tag{20}$$

Note that (19) and (6) have exactly the same form except that the range of the subscript i in (6) is finite. Therefore, both autocorrelation functions R_i and $\hat{R}_i$ obey (6). From the properties of matrix equation (6) we conclude that R_i and $\hat{R}_i$ are related by

$$\hat{R}_i = c R_i, \quad 0 \leq i \leq p \tag{21}$$

where c is a constant to be determined.

In order to conserve energy between the signal spectrum $P(\omega)$ and the approximate spectrum $\hat{P}(\omega)$, we must have

$$\hat{R}_0 = R_0 \quad \text{(energy conservation)} \tag{22}$$

since the zeroth autocorrelation coefficient is equal to the total energy.

Equation (21) applies for $i = 0$, therefore, from (22) we must have $c = 1$, and (21) reduces to

$$\hat{R}_i = R_i, \quad 0 \leq i \leq p. \tag{23}$$

This says that the first p coefficients (other than $\hat{R}_0$) of the autocorrelation function corresponding to the approximate spectrum, as computed from $\hat{S}(z)$, are identical to the first p coefficients of the autocorrelation function of the actual signal. The rest of the coefficients $\hat{R}_i$ are determined by (19). The problem of linear prediction using the autocorrelation method can be stated in a new way as follows. Find a power spectrum such that first p values of the corresponding autocorrelation function are equal to the first p values of the signal autocorrelation function, and such that (19) applies.

From (23) and (20) we conclude that $A^2 = R_0 - \sum_{k=1}^{p} a_k R_k$, which is identical to the result obtained in a different manner in (14).

The spectra $P(\omega)$ and $\hat{P}(\omega)$ are the Fourier transforms of R_i and $\hat{R}_i$, respectively. Therefore, increasing 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 power spectra become identical:

$$\hat{P}(\omega) = P(\omega), \qquad \text{as } p \to \infty. \tag{24}$$

In light of (24), it is natural to ask how the transfer functions $\hat{S}(z)$ and $S(z)$ are related as $p \to \infty$. It might seem that the two transfer functions will become equal. However, this is not true in general. As $p \to \infty$, $\hat{S}(z) = S(z)$ if and only if the signal s_n is minimum phase, i.e., $S(z)$ has no zeros or poles outside the unit circle. We know that the speech signal is generally nonminimum phase. We also know that $\hat{S}(z)$, in the autocorrelation method, is always minimum phase: all poles are inside the unit circle and there are no zeros. Furthermore, there is a unique minimum phase sequence whose spectrum is identical to $P(\omega)$. From (24) and the properties of $\hat{S}(z)$ we know that $\hat{S}(z)$, as $p \to \infty$, is the transfer function of that minimum-phase sequence:

$$\hat{S}(z) = \frac{A}{1 - \sum_{k=1}^{\infty} a_k z^{-k}}$$

$$= \sum_{n=0}^{N-1} b_n z^{-n}, \qquad p \to \infty \tag{25}$$

where $b(nT)$ is the minimum-phase sequence corresponding to $s(nT)$. Refer to [5] for methods to compute the sequence $b(nT)$.

V. Error Minimization and Analysis-by-Synthesis

Making use of Parseval's theorem, the total-squared error E defined in (3) and (5) can be rewritten as

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} |E(\omega)|^2 \, d\omega \tag{26}$$

where $E(\omega)$ is obtained by substituting $z = e^{j\omega T}$ in $E(z)$. Substituting for $E(z)$ from (9) and for $H(z)$ from (12), then replacing z by $e^{j\omega T}$ and using (15) and (16), (26) reduces to

$$E = \frac{A^2 T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{\hat{P}(\omega)} \, d\omega. \tag{27}$$

Therefore, minimizing the total-squared error E is equivalent to the minimization of the integrated ratio of the signal power spectrum $P(\omega)$ to its approximation $\hat{P}(\omega)$. Another way to look at this is that if one is interested in approximating a power spectrum $P(\omega)$ by an all-pole spectrum $\hat{P}(\omega)$ then (27) is an error measure that can be used in optimizing the approximation. This error can be minimized analytically [5] resulting in equations identical to (6) which can be solved for a_k, the parameters of the sought-for approximate spectrum $\hat{P}(\omega)$. The question, then, is what are the properties of the error measure in (27), and are these properties commensurate with requirements for the spectral analysis of speech? This is discussed below.

For speech, we usually desire that the spectrum $\hat{P}(\omega)$ approximate the *envelope* of the signal power spectrum $P(\omega)$. One important consideration in estimating the spectral envelope is the determination of an optimum value for p, the number of poles in the all-pole approximate spectrum $\hat{P}(\omega)$. This subject is discussed later in this section when we analyze the properties of the normalized minimum error. However, assuming that somehow we know this optimal value of p, there remains the question of whether the error measure in (27) will result in a good estimate of the spectral envelope. We note from (27) that spectral values of $P(\omega)$ that are greater than the corresponding values in $\hat{P}(\omega)$ will contribute to the total error in a significant manner, while spectral values of $P(\omega)$ that are much smaller than the corresponding values in $\hat{P}(\omega)$ will not affect the total error significantly. This means 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. For example, if $P(\omega)$ is the power spectrum of a quasi-periodic signal (such as a sonorant), 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 (27) insures that the approximation of $\hat{P}(\omega)$ to $P(\omega)$ is far superior at the harmonics where the energy is greater, than between the harmonics where there is very little energy. Since $\hat{P}(\omega)$ is expected to be a smooth spectrum (this is insured by choosing an appropriate value for p), we conclude that minimization of the error measure in (27) results in an approximate spectrum $\hat{P}(\omega)$ that is a good estimate of the spectral envelope of the signal power spectrum $P(\omega)$. It should be clear from the above that the importance of the goodness of the error measure is much more crucial for voiced sounds than for unvoiced sounds where the variations of the

signal spectrum from the spectral envelope are much less pronounced.

Another important property of this estimation procedure is that, because the contributions to the total error are determined by the *ratio* of the two spectra, the matching process should perform uniformly over the frequency band of interest, irrespective of the general shaping of the envelope of the speech spectrum. This property is reminiscent of the analysis-by-synthesis method of spectral reduction developed at the Massachusetts Institute of Technology [7] and more recently at Bell Laboratories [8], where the total error to be minimized is given (in our notation) by

$$\int_{\omega} \left[\log \frac{P(\omega)}{\hat{P}(\omega)}\right]^2 d\omega.$$

Indeed, spectral estimation by linear prediction of the speech signal can be regarded as yet another method of analysis-by-synthesis, but instead of specifying formants and antiformants we only specify a number of poles p. In addition, (27) gives a superior measure of error if a spectral envelope is desired. But then, analysis-by-synthesis methods have generally used already smoothed spectra, in which case it is probably of little consequence which error measure is used. The elegance of the linear prediction method is that it performs the smoothing (for a well-chosen p) as well as the analysis-by-synthesis type of computation all at once by simply solving the special matrix equation given in (6), without resorting to iterative techniques. The price that one has to pay is that the approximate spectrum $\hat{P}(\omega)$ can have only poles. Traditional analysis-by-synthesis methods were used mainly for formant tracking, and spectral fitting was just the means of refining the estimates of the formants or antiformants. So, how do the poles of the all-pole filter $\hat{S}(z)$ relate to the formants and antiformants of the signal? This is discussed in Section VI.

Normalized Minimum Error

Another interesting aspect of error minimization is the properties of the normalized minimum total-squared error V_p, which is defined as the ratio of the energy in the minimum-error sequence e_n to the energy in the speech signal s_n. From (8) and (14) we conclude that

$$V_p = \frac{E_p}{R_0} = \frac{A^2}{R_0}, \tag{28}$$

or

$$V_p = 1 - \sum_{k=1}^{p} a_k r_k, \tag{29}$$

where

$$r_k = \frac{R_k}{R_0} \tag{30}$$

are the normalized autocorrelation coefficients; they have the property that $|r_k| \leq 1$, for all k. Fig. 1 shows the normalized error curves as a function of p for the unvoiced fricative [s] in the word "list" and the vowel [æ] in the word "potassium." The speech signal was bandpassed at 4.5 kHz and sampled at 10 kHz. V_p is a monotonically decreasing function of p such that

$$0 < V_p \leq 1. \tag{31}$$

In particular, for $p = 0$, $V_0 = 1$, and as $p \to \infty$, V_p approaches a minimum value $V_{\min} = V_\infty$. The normalized error curve, including $V_{\min}$, is a function of the power spectrum $P(\omega)$ only. This can be seen from the fact that V_p in (29) is a function only of the autocorrelation coefficients. [Remember that the predictor coefficients a_k are computed from the autocorrelation coefficients in (6).] In fact, it can be shown [5] that

$$V_p = \frac{e^{\hat{c}_0}}{\hat{R}_0} = \frac{e^{\hat{c}_0}}{R_0} \tag{32}$$

where

$$\hat{c}_0 = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \log \hat{P}(\omega)\, d\omega \tag{33}$$

is the zeroth coefficient (quefrency) of the cepstrum which we define as the inverse Fourier transform of the logarithm of the power spectrum.

From (24) we know that as $p \to \infty$, $\hat{P}(\omega)$ becomes equal to $P(\omega)$. Substituting $P(\omega)$ for $\hat{P}(\omega)$ in (33) and the result in (32), we obtain an expression for the normalized error $V_{\min} = V_\infty$:

$$V_{\min} = \frac{e^{c_0}}{R_0} \tag{34}$$

where c_0 is the zeroth coefficient of the signal cepstrum, and R_0 is the energy in the signal.

Another method [5] for deriving (34) is to compute the minimum-phase sequence b_n in (25) corresponding to the signal s_n, and then

$$V_{\min} = \frac{b_0^2}{R_0}$$

and $b_0^2 = e^{c_0}$.

It is instructive to write (34) as a function of $P(\omega)$:

$$V_{\min} = \frac{\exp\left[\dfrac{T}{2\pi} \displaystyle\int_{-\pi/T}^{\pi/T} \log P(\omega)\, d\omega\right]}{\dfrac{T}{2\pi} \displaystyle\int_{-\pi/T}^{\pi/T} P(\omega)\, d\omega}. \tag{35}$$

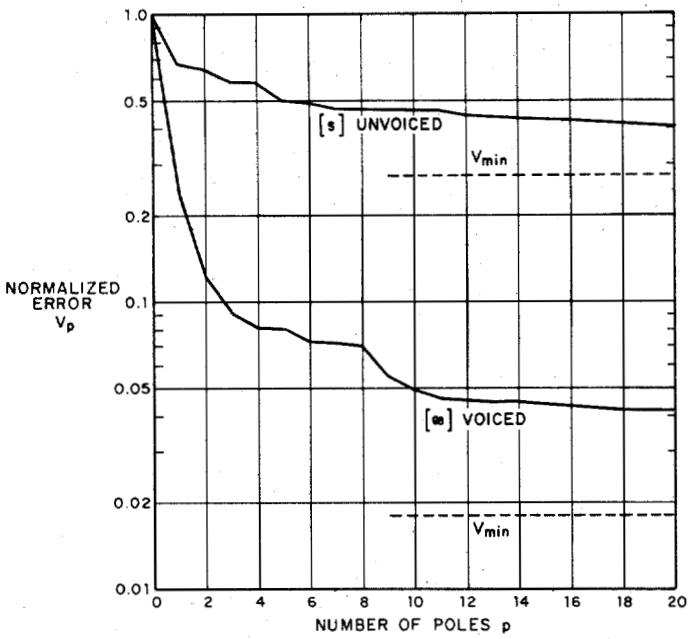

Fig. 1. Normalized error curves for [s] in the word "list" and [æ] in the word "potassium."

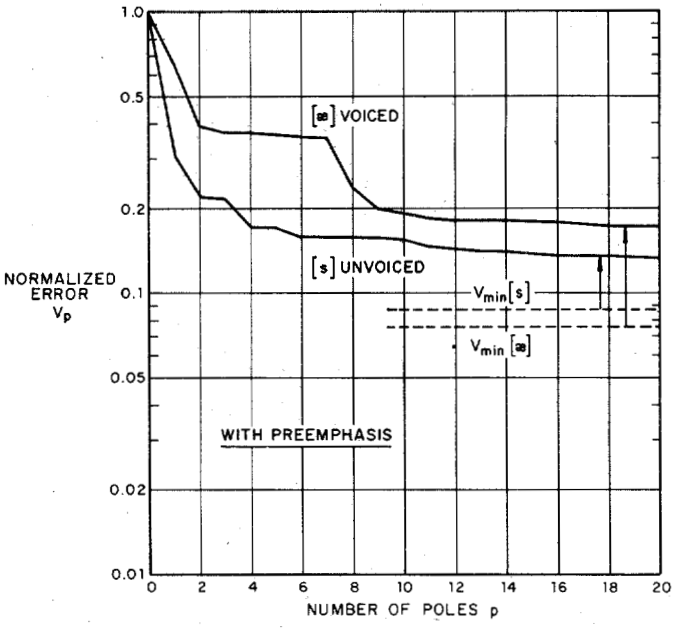

Fig. 2. Normalized error curves for the same two sounds as in Fig. 1, except that the speech signals were preemphasized by simple differencing.

It is clear from (35) that V_{min} depends completely on the shape of the signal spectrum. Similarly, from (32), V_p depends completely on the shape of the approximate spectrum. This fact is very important in interpreting the behavior of the normalized error curve for the spectra of different sounds. For example, in Fig. 1 the error curve for the unvoiced fricative [s] is much higher than that for the vowel [æ]. On the whole, unvoiced sounds have a high error curve while voiced sounds have a much lower error curve. This property of voiced versus unvoiced sounds has been observed before [1], [3], and V_p has been suggested as a possible parameter for the detection of voicing. However, with our result showing that the error curves are dependent only on the shape of the spectrum, it is clear that what makes this apparent dichotomy between voiced and unvoiced sounds has nothing to do with the fact of voicing itself, but rather with the shape of the spectra corresponding to these sounds.

By examining the behavior of (35) for V_{min} one gains insight into how the error curves change for different shapes of the spectrum. For example, it is easy to show that if the spectrum is perfectly flat, then $V_{min} = 1$, and the error curve is the highest possible. On the other hand, if all the energy is concentrated in certain regions of the spectrum and the rest of the spectrum contains zero energy, then $V_{min} = 0$, and the error curve is the lowest possible. Speech sounds lie somewhere between these two extremes. In general, voiced sounds (especially sonorants) have most of the energy concentrated in one region at low frequencies, resulting in low error curves. Unvoiced sounds, on the other hand, have the energy more spread out across the spectrum, resulting in higher error curves. However, this property cannot be relied upon all the time.

Another way to look at V_{min} is to note that (35) is the ratio of the geometric mean to the arithmetic mean of the spectrum, where we have extended the notions of the geometric and arithmetic means to the continuous case. This becomes clear if one assumes that the spectrum is discrete, in which case (35) does reduce to the ratio of a geometric mean to an arithmetic mean. It is well known that this ratio is equal to one if all the data are equal, and the value decreases as the spread of the data increases. A larger spread is equivalent to heavy energy concentrations in certain regions and a simultaneous lack of energy in the other regions of the spectrum.

It must be remembered that distortions to the speech signal that affect the shape of the spectrum can have marked effects on the error curves. For example, if the speech signal is preemphasized by simple differencing, the effect on the spectrum is a rise of approximately 6 dB/octave, and the corresponding effect on the error curves is often drastic. Fig. 2 shows the error curves for the same two sounds that are in Fig. 1, except that preemphasis was used. The error curve for the unvoiced fricative [s] became lower while that for the vowel [æ] became much higher. Another example where the normalized error cannot be used effectively as a voicing detector is in telephone speech, where much of the low frequency energy has been filtered out and the spectral dynamic range is sharply reduced. (An alternate method that

employs the first normalized autocorrelation coefficient r_1 for the detection of voicing is described elsewhere [5].)

We shall now relate the normalized error to the problem of choosing a value of p such that $\hat{P}(\omega)$ approximates the envelope of $P(\omega)$ optimally, in the sense that the formant structure of the vocal tract is just evident with a minimum of extraneous detail. Each of the error curves in Figs. 1 and 2 starts at 1 for $p = 0$ and monotonically decreases to its own $V_{\min}$ as $p \to \infty$. Also, each of the curves exhibits what might be called the "knee" of the curve. This is a value of p after which the curve slopes very slowly towards its asymptote. For example, in Fig. 1, this occurs around $p = 7$ and $p = 11$ for the upper and lower error curves, respectively. A physical explanation for this "knee" in the error curve is that around that value of p the approximate spectrum is the "optimal" approximation to the envelope of the signal spectrum. (This is true only for spectra that are well approximated by poles.) A lower value of p results in a grosser approximation to the spectral envelope while a larger value of p will superimpose detailed spectral information on top of the spectral envelope. In general, increasing p from its "optimal" value has a less drastic effect than decreasing it. Therefore, for many applications, it is usually sufficient to set p to a fixed value that is the upper limit necessary to describe the envelope of the changing speech spectrum. For speech signals band limited to 5 kHz and sampled at 10 kHz, a value of p between 10–14 kHz is chosen depending on the application.

VI. Formant Analysis

The poles of $\hat{S}(z)$ can be computed by setting the denominator in (12) to zero and solving the resultant polynomial equation in z for its roots. Conversion 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 pole in the s plane. Some or none of the roots may be real and the rest are complex conjugate pairs.

If a speech spectrum can be approximated by poles only, then the formants can be obtained from the poles of $\hat{S}(z)$ by noting that: 1) a formant consists of a pair of complex conjugate poles; 2) a formant normally has a high ratio between its frequency and bandwidth. Complex conjugate poles with very wide bandwidths can be regarded as contributing to general spectral shaping only; 3) the frequency range of a particular formant is usually known; 4) peak picking can be performed on the approximate spectrum as a double check on the formant values; and 5) continuity of formant values from one spectral frame to another can always be invoked, keeping in mind that very fast formant transitions do exist in speech.

Formant tracking by simple peak picking of the approximate spectrum $\hat{P}(\omega)$ seems to be an effective method in general [1]. However, problems can arise when two formants come close to each other. In these cases the approximate spectrum can be computed inside the unit circle in order to enhance the formant peaks. This is done by multiplying the predictor coefficients a_k by a rising exponential before computing the spectrum, then

$$\hat{P}(\sigma, \omega) = \frac{A^2}{\left| 1 - \sum_{k=1}^{p} (a_k e^{-k\sigma T}) e^{-jk\omega T} \right|^2}. \quad (40)$$

Values of $\sigma \cong -2\pi \times 75$ seem to do quite well.

Comparison with the Cepstral Smoothing Method

Schafer and Rabiner [9] have developed a system for formant analysis by a peak-picking algorithm applied to a cepstrally smoothed spectrum (i.e., a low-pass filtered log spectrum), and in cases where formants were believed to be very close to each other, they applied the chirp z transform (CZT) to the cepstrum in order to enhance the formant peaks and separate the formants. It is of interest to compare that method to linear prediction.

First, it should be pointed out that if the CZT is applied to the cepstrum corresponding to the approximate spectrum $\hat{P}(\omega)$, as obtained through linear prediction, then the enhanced peaks in the resulting spectrum should correspond to the formant frequencies which could be obtained more accurately by solving for the poles of $\hat{S}(z)$. Therefore, unlike the cepstral smoothing method where the CZT is useful in obtaining extra information about formant locations, applying the CZT in linear prediction adds no information.

Another point of comparison is that both types of spectra are smoothed versions of the original spectrum. One method does it by actually low-pass filtering the log spectrum, and the other by reducing the number of poles of an all-pole approximate spectrum. The two types of smoothing are not equivalent, however, because in linear prediction the spectral fitting is based on an all-pole model of speech which, for nonnasal sonorants, corresponds to the usual model of the vocal tract transfer function. For those sounds, we would expect linear prediction to give a better spectral fit. Fig. 3(a) shows a spectrum of a Hamming weighted 25 ms of the vowel [a] obtained from 10-kHz sampled telephone speech, and superimposed on it is the smoothed spectrum obtained by linear prediction with $p = 14$. Fig. 3(b) shows the corresponding cepstrally smoothed spectrum. (The cepstrum has unity weighting up to 1.5 ms and cosine weighting up to 3.0 ms.) Note that a simple peak picking algorithm in Fig. 3(b) would re-

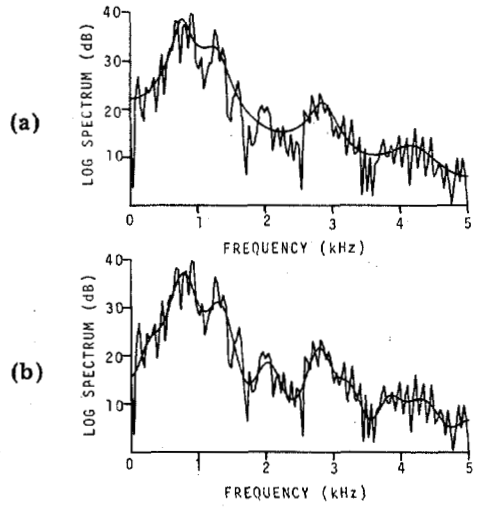

Fig. 3. Spectral smoothing of a spectrum of the vowel [a] obtained from 10-kHz sampled telephone speech, (a) by linear prediction with $p = 14$, (b) by cepstral smoothing.

sult in a false third formant at 2 kHz. Because of the spectral properties of the vowel [a], the third forment of the vocal tract is more likely at 2.8 kHz as shown in Fig. 3(a).

High-pitched speech normally gives rise to problems in formant tracking due to the fact that for voiced sounds the spectral harmonics are widely separated. It should be emphasized that this results in a basic loss of information about the formant structure, a loss that cannot be recovered even by pitch-synchronous analysis, unless new information is added. The method of linear prediction should perform quite well (with nonnasal sonorants) because of the fact that we assume an all-pole model, which amounts to additional useful information. In cepstral smoothing the cutoff point of the low-pass filter is placed below the pitch peak, which for high-pitched speech can mean a further loss of information about the formant structure. In linear prediction, the formant locations are less affected by the pitch because the harmonics are forced to fit the all-pole model. This is a well-known property of analysis-by-synthesis methods.

Although for nonnasal sonorants linear prediction is expected to give more accurate formant values than the cepstral smoothing method, the same is not necessarily true for other sounds such as nasals and fricatives, whose spectra are known to have antiformants (zeros) as well as formants. The difficulty is that the all-pole linear prediction model attempts to approximate the effects of the formants and the antiformants at the same time. A consequence is that the positions and bandwidths of the extracted formants will often be very different from their "actual" values, depending on the position of each formant with respect to the antiformants. Formants that are far from the nearest antiformants are well approximated, while those that are close to an antiformant are often poorly approximated.

VII. Preprocessing Considerations

Very often for voiced speech, the effect of the glottal source appears as a very low-frequency peak in the approximate spectrum. A peak-picking formant tracker might mistaken the peak for a low-frequency first formant. One method of eliminating this peak is to preemphasize the speech signal. Not only does preemphasis reduce these low-frequency effects, it also enhances high-frequency formants, which is often enough to separate the peaks of closely spaced formants. A side effect of preemphasis is the possible shifting of formant locations in the approximate spectrum, especially the first formant [5]. Also, we have already mentioned in Section V that the normalized minimum error becomes useless as a voicing detector if preemphasis is employed.

In computing the short-time spectrum of a portion of a speech signal it is usually necessary to multiply the signal by a nonrectangular window (e.g., Hamming or Hanning windows) if pitch-asynchronous analysis is desired. This is done to reduce the effects of including a nonintegral number of pitch periods. Since in the autocorrelation method of linear prediction the approximate spectrum is a fit to the envelope of the short-time signal spectrum, it follows that windowing is in general necessary to insure spectrally accurate results. A detailed analysis of windowing can be found in [5].

VIII. Conclusions

Spectral analysis by linear prediction of the sampled speech signal approximates the signal spectrum by an all-pole spectrum. The nature of this ap-

proximation was examined in the autocorrelation domain as well as in the spectral domain. We pointed out that this type of spectral analysis can be regarded as another method of analysis-by-synthesis with three major differences: 1) instead of specifying a number of formants or antiformants, we specify a number of poles; though it is not always clear how these poles relate to the possible existence of antiformants in the spectrum; 2) for an all-pole spectrum, the error measure leads to a better spectral envelope fit; and 3) the method for computing the poles is straightforward and noniterative.

We have also argued that compared to spectral analysis by cepstral smoothing, this method is expected to give a better spectral envelope fit to an all-pole signal spectrum and to be less sensitive to the effects of high pitch on that spectrum.

The normalized minimum error was defined and related to the zeroth coefficient of the cepstrum. We have shown that the normalized error is closely related to the shape of the spectrum. By examining the properties of the normalized error it was clear that it could be used as a voicing detector in many cases, but care must be taken if the signal had been preprocessed in any way.

In preprocessing the signal, the advantages and disadvantages of preemphasis were discussed. Preemphasis reduces low frequency effects and enhances high frequency formants, which is usually desirable for formant tracking. However, a shift in the formant frequencies might occur. Also, the usefulness of the normalized error as a voicing detector is sharply reduced.

In the autocorrelation method of linear prediction it is necessary to window the speech signal. The choice of the window depends on the signal being analyzed. For pitch-asynchronous analysis, a smooth window such as the Hamming window is sufficient.

References

[1] 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; also, Speech Commun. Res. Lab., Santa Barbara, Calif., SCRL Monograph 7, 1971.

[2] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. (Japan)*, vol. 53-A, no. 1, pp. 36-43, 1970.

[3] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637-655, Aug. 1971.

[4] F. B. Hildebrand, *Introduction to Numerical Analysis*. New York: McGraw-Hill, 1956, p. 378.

[5] J. Makhoul and J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman, Cambridge, Mass., Rep. 2304, Aug. 1972.

[6] E. A. Robinson, *Statistical Communication and Detection*. New York: Hafner, 1967, pp. 274-279.

[7] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1725-1736, Dec. 1961.

[8] J. P. Olive, "Automatic formant tracking by a Newton-Raphson technique," *J. Acoust. Soc. Amer.*, vol. 50, pp. 661-670, Aug. 1971.

[9] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634-648, Feb. 1970.