Linear prediction analysis of speech based on a pole-zero representation

Bishnu S. Atal

Bell Laboratories, Murray Hill, New Jersey 07974

M. R. Schroeder

Drittes Physikalisches Institut, University of Göttingen, F. R. Germany

(Received 3 February 1978; revised 21 July 1978)

Speech analysis and synthesis by linear prediction is based on the assumption that the short-time spectral envelope of speech can be represented by a number of poles. An all-pole representation does not provide an accurate description of speech spectra, particularly for nasals and nasalized sounds. This paper presents a method for characterizing speech in terms of the parameters of a pole-zero model. In this method,an impulse response representing the composite filtering action of the glottal wave, the vocal tract, the radiation, and the speech recording system is first constructed from the speech signal.This impulse response is obtained by performing several stages of all-pole LPC analysis. The pole-zero parameters are determined from the impulse response by solving a set of simultaneous linear equations. The method, being noniterative, is very suitable for automatic analysis of speech. The method has been applied to real speech data and the results show that the speech spectra derived from the pole-zero model agree very closely with the actual spectra derived by direct Fourier analysis.

PACS numbers: 43.70.Gr

INTRODUCTION

Speech analysis by linear prediction is based on the assumption that the short-time spectral envelope of speech waveform can be represented by a number of poles (Atal and Schroeder, 1967, 1970; Atal and Hanauer, 1971; Itakura and Saito, 1968; Makhoul, 1975; Markel and Gray, 1976). The all-pole model of the speech spectrum is quite accurate for vowel and vowel-like sounds. The assumption of the all-pole representation, however, puts a restriction on the linear prediction method in performing accurate analysis of speech when there are additional zeros in the transfer function of the vocal tract. These zeros can arise if the nasal tract is coupled to the main vocal tract through the velar opening—as is the case for nasal consonants and nasalized vowels—or if the source of excitation is not at the glottis but is in the interior of the vocal tract (Flanagan 1972). In addition, the zeros can be introduced by the transmission characteristics of the environment where the recordings are made.

Recently, there has been considerable interest in extending the linear prediction analysis to include both poles and zeros of the speech transfer function. Identification of linear systems based on pole and zero models has been discussed extensively in both the statistical and the signal processing literature (Durbin, 1959; Hannan, 1969; Hsia and Landgrebe, 1967; Shanks, 1967; Steiglitz and McBride, 1965). Possible applications of some of these techniques to speech analysis are discussed by Makhoul (1974), and Steiglitz (1977). Another approach to this problem based on homomorphic deconvolution is described by Oppenheim *et al.* (1976); the reference also provides an excellent bibliography on various methods of pole-zero analysis.

We present, in this paper, a speech analysis method for estimating both poles and zeros of the speech transfer function (Atal and Schroeder, 1974, 1975; Atal, 1975). By taking account of the zeros in the transfer function, we hope to eliminate one of the limitations of the linear prediction method in analyzing nasalized sounds. The approach used in this paper is similar to one used for the all-pole analysis. It is shown that the pole-zero parameters which minimize the mean-squared prediction error are determined as solutions to a set of *nonlinear* equations. However, the pole-zero parameters can be determined by solving a set of linear equations provided an estimate of the impulse response—representing the composite filtering action of the glottal wave, the vocal tract, the radiation, and the speech recording system—can be first estimated from the speech wave. Such an impulse response is constructed by performing several stages of all-pole linear prediction analysis.

This paper consists of three sections. In the first section, we discuss the influence of the zeros in the transfer function on the spectral envelope of the speech signal. In Sec. II, we present a pole-zero model for speech synthesis and describe a procedure for determining the parameters of the pole-zero model. In Sec. III, we present results of applying this procedure to both synthetic and natural speech signals. For synthetic speech, the spectral envelope is known exactly. Thus, it is possible to compare the analyzed spectral envelope with the actual envelope used in the synthesis. For natural speech, the exact spectral envelope is not known. In this case, we compare the spectral envelope with the harmonic spectrum obtained by Fourier analysis of the speech signal.

I. INFLUENCE OF SPECTRAL ZEROS ON LPC SPECTRUM

A spectral zero, unless it is cancelled by a closely located pole, produces two separate effects in the spectrum: a dip in the spectrum in the neighborhood of the

0001-4966/78/6405-1310$00.80

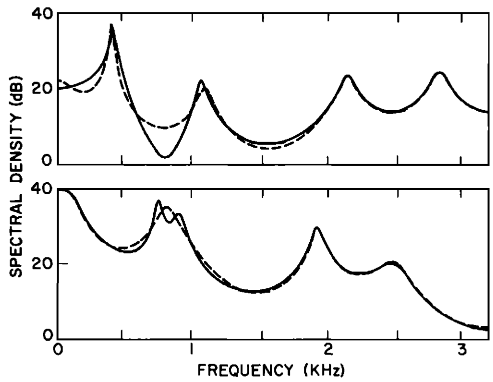

FIG. 1. Two examples showing the influence of zeros on the spectra obtained by LPC analysis using an all-pole model. The solid curve in each case shows spectrum computed directly from the transfer function of the filter. The spectrum derived from 12-pole LPC analysis is shown by the dashed curves.

frequency of the zero ("anti-resonance frequency") and an asymptotic 12-dB/octave rise in the spectrum beyond the anti-resonance frequency. In principle, the spectral properties of a zero can be approximated with arbitrary precision by additional poles in the all-pole model (Atal and Hanauer, 1971). Such an approximation, however, is likely to be inefficient except for highly damped zeros. Furthermore, it is difficult to produce a local dip in the spectrum by using additional poles without introducing ripples at other frequencies in the spectrum. In general, the 12-dB/octave rise in spectrum can be approximated satisfactorily by additional poles. However, significant errors are often introduced in the frequency region in the vicinity of the zero.

Two examples, typical of the all-pole approximation of spectrum in the presence of a single complex zero-pair at a frequency of 800 Hz, are shown in Fig. 1. The transfer function in each case consists of five pairs of complex conjugate poles and a pair of complex conjugate zeros. The solid curve represents the theoretical spectrum computed from the transfer function. The dashed curve is the spectrum obtained by a 12-pole linear prediction (LPC) analysis (Atal and Hanauer, 1971). The predictor coefficients were obtained from the samples of the autocorrelation function computed directly from the transfer function. A sampling frequency of 10 kHz was used in the above computations. Note that, in these examples, the only source of error in the LPC analysis is the presence of the zero. Two kinds of errors are evident in the all-pole spectrum: the spectrum in the vicinity of the zero is incorrect and the locations of formants close to the zero are shifted. In the second example, the zero is located in between two closely-spaced formants. In this case, these formants merge into a single formant in the LPC-derived spectrum. Similar behavior is also shown by the all-pole spectrum for natural speech. An example comparing the all-pole spectrum with the spectrum derived by Fourier analysis[1] for a nasal consonant excerpted from natural speech is shown in Fig. 2.

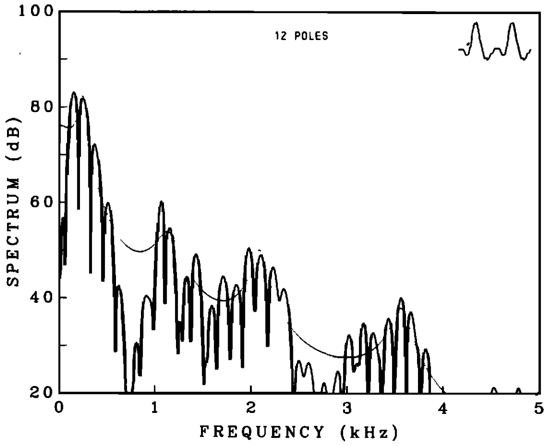

FIG. 2. An example (shown as a dotted curve) of the spectral envelope from linear prediction analysis based on an all-pole model for the nasal consonant /m/. The speech waveform used in the analysis is shown at the upper right corner of the figure. For comparison, the spectrum obtained by direct Fourier analysis of the waveform is also shown as a solid curve.

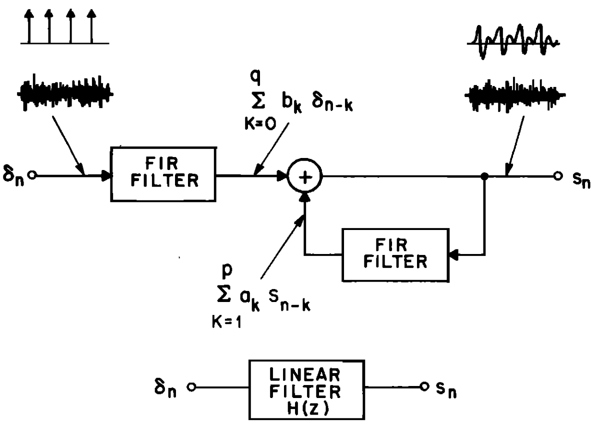

FIG. 3. A pole-zero model of speech production.

II. LINEAR PREDICTION WITH SPECTRAL ZEROS

A. Pole-zero model

In linear prediction based on the all-pole model, every speech sample is predicted as a linear combination of the previous p samples of the signal. Let us assume now that there are q zeros in the transfer function in addition to p poles. A functional model of speech production including both poles and zeros is shown in Fig. 3. For a discrete linear filter with p poles and q zeros, the nth sample of its output is expressed as

$$s_n=\sum_{k=1}^{p} a_k s_{n-k}+\sum_{k=1}^{q} b_k \delta_{n-k}+\delta_n , \qquad (1)$$

where δ_n is the nth sample of the excitation at the input of the linear filter, the coefficients a_k represent the contribution of the poles, and the coefficients b_k represent the additional contribution of the q zeros. Let us assume that the excitation δ_n has a "flat" spectrum.[2] The predicted value of the nth speech sample is then given by

$$\hat{s}_n=\sum_{k=1}^{p} a_k s_{n-k}+\sum_{k=1}^{q} b_k(s_{n-k}-\hat{s}_{n-k}). \qquad (2)$$

Note that the predicted value of a speech sample is represented as a sum of the linear combination of both the past p speech samples and past q prediction-error samples. For comparison, in the all-pole case, the predicted value is equal to the linear combination of only the past p speech samples. The mean-squared prediction error is given by

$$E = \left\langle \left[s_n - \sum_{k=1}^{p} a_k s_{n-k} - \sum_{k=1}^{q} b_k (s_{n-k} - \hat{s}_{n-k}) \right]^2 \right\rangle \tag{3}$$

where $\langle\ \rangle$ indicates averaging over a number of speech samples included in a speech segment in which the filter coefficients can be considered to be approximately constant. A speech segment 10–20 ms in duration is generally suitable for this purpose. The mean-squared prediction error is minimized by setting the various partial derivatives of E with respect to the filter coefficients equal to zero. It is to be noted that the mean-squared prediction error given in Eq. (3) is not a quadratic function of the unknowns (due to the presence of the term $\hat{s}_{n-k}$ on the right side), as is the case in the absence of zeros. The filter coefficients which minimize the mean-squared prediction error are thus obtained by solving a set of simultaneous *nonlinear* equations. Such equations must be solved by iterative methods. Iterative methods are usually quite complex. Moreover, it is not always possible to guarantee their convergence to a global minimum. Iterative methods are therefore not very convenient to use for automatic analysis of speech. Noniterative and direct methods of obtaining the filter coefficients are much more desirable. Of course, the resulting solution is not optimum any more.

One possible method of avoiding nonlinear equations would be to consider the case when the input δ_n to the filter is known. The predicted value of the nth speech sample is then given by the right side of Eq. (1). Since the predicted value now is a linear function of the filter coefficients, the resulting set of equations obtained from minimizing the mean-squared prediction error are linear. It will be shown later in this section that a precise knowledge of the input is not necessary for determining the filter parameters; it is sufficient to know the correlations between the input and the output of the linear filter. Since these correlations are not known exactly, the resulting accuracy of the filter coefficients will depend on the accuracy with which the correlations can be estimated from the speech samples. We will now discuss this suboptimum but direct method of estimating the filter coefficients.

B. Minimization of prediction error

The mean-squared prediction error for known filter input is given by

$$E = \left\langle \left(s_n - \sum_{k=1}^{p} a_k s_{n-k} - \sum_{k=0}^{q} b_k \delta_{n-k} \right)^2 \right\rangle , \tag{4}$$

where b_0 is 1. On setting the partial derivatives of the error E in Eq. (4) with respect to the unknown parameters a_k, $k=1,2,\ldots,p$, and b_k, $k=1,2,\ldots,q$, to zero, one obtains the following set of equations:

$$\sum_{k=1}^{p} a_k \langle s_{n-k} s_{n-r} \rangle = \langle s_n s_{n-r} \rangle - \sum_{k=1}^{q} b_k \langle \delta_{n-k} s_{n-r} \rangle, \quad 1 \leq r \leq p, \tag{5}$$

$$\sum_{k=1}^{q} b_k \langle \delta_{n-k} \delta_{n-r} \rangle = \langle s_n \delta_{n-r} \rangle - \sum_{k=1}^{p} a_k \langle s_{n-k} \delta_{n-r} \rangle, \quad 1 \leq r \leq q, \tag{6}$$

These equations include three different kinds of correlations: the correlations between different samples of the speech signal, the correlations between different samples of the input, and the correlations between the input and the speech samples. The terms consisting of correlations between speech samples are of course the same as in the all-pole case. Let us now consider the terms in Eqs. (5) and (6) involving correlations between the input and the speech samples. The term $\langle \delta_{n-k} s_{n-r} \rangle$ represents the correlation between the input and output of the filter. The correlation between the input and the output of a linear filter is proportional to the impulse response of the linear filter provided the input has a flat spectrum (Schwarz and Friedland, 1965). Let h_n be the nth sample of the impulse response of the linear filter of Fig. 3. Then

$$\langle s_n \delta_{n-r} \rangle = \sum_{k=1}^{\infty} h_k \langle \delta_{n-k} \delta_{n-r} \rangle, \quad 1 \leq r \leq q. \tag{7}$$

Furthermore, the term $\langle \delta_{n-k} \delta_{n-r} \rangle$ is zero, if $k \neq r$, and equals ϵ_0, if $k = r$, where ϵ_0 is the minimum value of the mean-squared prediction error. Equation (7) is then reduced to

$$\langle s_n \delta_{n-r} \rangle = \epsilon_0 h_r, \quad 1 \leq r \leq q. \tag{8}$$

Similarly,

$$\langle \delta_{n-k} s_{n-r} \rangle = \epsilon_0 h_{k-r}, \quad 1 \leq r \leq p, \quad 1 \leq k \leq q, \tag{9}$$

and

$$\langle s_{n-k} \delta_{n-r} \rangle = \epsilon_0 h_{r-k}, \quad 1 \leq r \leq q, \quad 1 \leq k \leq p. \tag{10}$$

On substituting for these correlations, Eq. (6) simplifies to

$$b_r = h_r - \sum_{k=1}^{p} a_k h_{r-k}, \quad 1 \leq r \leq q. \tag{11}$$

On substituting for b_k, $k=1,2,\ldots,q$, from Eq. (11) and changing the subscript k to i on the right side, Eq. (5) is rewritten as

$$\sum_{k=1}^{p} a_k \langle s_{n-k} s_{n-r} \rangle = \langle s_n s_{n-r} \rangle - \epsilon_0 \sum_{i=1}^{q} h_{i-r} \left(h_i - \sum_{k=1}^{p} a_k h_{i-k} \right), \quad 1 \leq r \leq p. \tag{12}$$

The terms $\langle s_{n-k} s_{n-r} \rangle$ and $\langle s_n s_{n-r} \rangle$ in Eq. (12) can be replaced by the appropriate correlations between the different samples of the impulse response of the linear filter.

$$\langle s_{n-k} s_{n-r} \rangle = \epsilon_0 \sum_{i=1}^{\infty} h_{i-k} h_{i-r}, \quad 1 \leq k \leq p, \quad 1 \leq r \leq p, \tag{13}$$

and

$$\langle s_n s_{n-r} \rangle = \epsilon_0 \sum_{i=1}^{\infty} h_i h_{i-r}, \quad 1 \leq r \leq p. \tag{14}$$

The coefficients a_k and b_k are therefore obtained by solving the equations

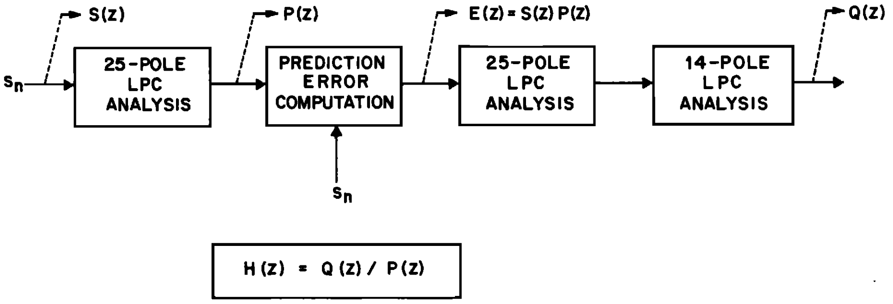

FIG. 4. Block diagram showing the method used for computing from the speech waveform an effective impulse response representing the composite filtering action of the vocal tract, glottal wave, radiation, and the speech recording system.

$$\sum_{k=1}^{p} a_k \sum_{i=q+1}^{\infty} h_{i-k} h_{i-r} = \sum_{i=q+1}^{\infty} h_i h_{i-r}, \quad 1 \le r \le p, \tag{15}$$

and

$$b_r = h_r - \sum_{k=1}^{r} a_k h_{r-k}, \quad 1 \le r \le q, \tag{16}$$

respectively. The solution given by Eq. (15) for the pole parameters a_k is the same as suggested by Shanks (1967). The parameters b_k representing zeros are obtained somewhat differently in Shank's method by solving another least-square minimization problem. It is interesting to note that the solution given in Eq. (15) is also obtained by the application of the covariance method (Atal and Hanauer, 1971) to the impulse response. These equations can be solved for the unknowns a_k and b_k provided the impulse response of the filter can be determined first. We will now discuss a procedure for estimating this impulse response directly from the speech signal.

C. Determination of the impulse response

The impulse response is determined in two steps by successive approximation of the impulse response as a convolution of an all-pole response with an all-zero response. This procedure is illustrated in detail in Fig. 4. First, a 25-pole LPC analysis is performed on the speech signal (sampled at 10 kHz) to obtain the best all-pole approximation of its spectral envelope. The number of poles in the all-pole analysis is purposely chosen to be large to minimize errors in the resulting pole locations due to the presence of zeros. The analysis interval for minimizing the prediction error is two pitch periods. A modified form of the covariance method is used for the all-pole LPC analysis (Atal, 1977). In this method, a set of partial autocorrelation coefficients are first determined from the Cholesky decomposition of the covariance matrix. The partial correlations are then transformed to a set of predictor coefficients under the assumption of an all-pole model. This procedure always provides predictor coefficients which lead to a stable transfer function (that is, all of the poles of the transfer function are inside the unit circle). The analysis is performed on a finite length of the signal but no additional windows are used on the data. Let $P(z)$ be a polynomial in z whose kth coefficient equals the kth predictor coefficient. The prediction error is obtained by inverse filtering of the speech samples by a filter whose transfer function is $P(z)$. If the speech spectrum consisted of only poles, the short-time spectral envelope of the prediction error after the all-pole analysis would be nearly white. However, in the presence of zeros, the spectrum of the prediction error contains zeros skipped by the all-pole analysis. These zeros are determined using a spectral inversion technique originally suggested by Durbin (1959). The underlying idea in this technique is that an all-zero spectrum can be approximated to any degree of accuracy by sufficiently large number of poles. Therefore, as a first step, an all-pole LPC analysis is performed on the prediction error signal using a large number of poles (a suitable choice is 25). The spectrum of the resulting predictor coefficients (including the first term which is unity) is approximately the inverse of the spectral envelope of the error signal. To obtain a set of coefficients whose spectrum is equal to the spectral envelope of the prediction error, another all-pole LPC analysis using 14 poles is performed on the predictor coefficients. East stage of all-pole LPC analysis can be viewed here as a spectral inversion process. Two stages of LPC analysis thus produce a polynomial $Q(z)$ whose zeros approximate the zeros of the prediction error spectrum.

The impulse response $H(z)$ of the linear filter is obtained by dividing the polynomial $Q(z)$ by $P(z)$, that is,

$$H(z) = Q(z)/P(z) \ . \tag{17}$$

Once the impulse response h_n is known, Eq. (15) is solved to determine the filter parameters $a_1 \ldots, a_p$, representing the poles of the filter. The parameters $b_1, \ldots, b_q$, representing the zeros of the transfer function, are then determined by Eq. (16).

III. RESULTS

A. Synthetic speech

The method has been tested on both synthetic and natural speech signals. First, we discuss results obtained from synthetic speech. Speech samples were synthesized at a sampling frequency of 16 kHz from a digital filter with a transfer function consisting of 18 poles and two zeros. One set of frequencies and bandwidths of poles and zeros used in the synthesis are listed in Table I. The glottal pulse used for exciting the filter was Rosenberg's polynomial pulse (Rosenberg, 1970) with a duty cycle of 0.56. The waveshape of the pulse is given by

$$g(t) = \begin{cases} 2(t/t_p)^3 - 3(t/t_p)^2 & 0 \le t < t_p, \\ 1 - [(t-t_p)/(t_m - t_p)]^2 & t_p \le t \le t_m, \\ 0 & t_m < t \le t_0. \end{cases} \tag{18}$$

TABLE I. Formant frequencies and bandwidths for synthetic speech. All-pole analysis was carried out with 12 poles while pole-zero analysis was done with 12 poles and six zeros. Sampling frequency = 8 kHz. Fundamental frequency = 123 Hz.

	Synthesis data		All-pole model		Pole-zero model	
	Frequency	Bandwidth	Frequency	Bandwidth	Frequency	Bandwidth
Poles	285	37	239	19	258	48
	1200	100	1293	230	1220	109
	2084	82	2096	39	2084	74
	2495	192	2497	94	2489	156
	3624	111	3333	309	3449	342
	4575	164				
	5415	222				
	6301	266				
	7319	300				
Zero	800	50			806	46

where $t_p = 0.4t_0$, $t_m = 0.56t_0$, and t_0 is the pitch period. The pitch frequency was 123 Hz and increased from one period to the next at a rate of 15 Hz/s. The speech waveform was bandlimited to 4.0 kHz and down-sampled to 8000 samples/s prior to LPC analysis. The low-pass filter used for band-limiting the speech waveform was essentially flat to 3.0 kHz, had an attenuation of 6 dB at 3.5 kHz and an attenuation of more than 40 dB at and above 4.0 kHz.

LPC analysis based on all-pole and pole-zero models was performed on segments of speech waveform approximately two pitch periods in duration. The all-pole analysis was done using the modified covariance method (Atal, 1977). The result of an all-pole analysis with 12 poles is presented in Fig. 5. The dotted curve shows the power spectrum obtained from the predictor coefficients. The power spectrum at a frequency f is given by

$$G(f) = \epsilon_0 \Big/ \left| 1 - \sum_{k=1}^{p} a_k e^{-2\pi j f k T} \right|^2, \tag{19}$$

where ϵ_0 is the minimum mean-squared prediction error, T is the sampling interval $= 0.125$ ms, and p is the number of predictor coefficients $= 12$ (in this example). The solid curve on the figure is the power spectrum obtained by Fourier analysis of the speech waveform.[1] The analyzed speech waveform is shown on the upper right corner of the figure. The frequencies and bandwidths of the various formants as determined by the all-pole analysis are also listed in Table I. These frequencies and bandwidths were determined (Atal and Hanauer, 1971) by solving for the roots of the predictor polynomial (the inverse of the all-pole transfer function) obtained from the LPC analysis. The errors in the all-pole spectrum are obvious: the zero at 800 Hz is skipped and there is a shift in the locations of the formants in the neighborhood of the zero. The shift of the first formant from 285 to 239 Hz and narrowing of its bandwidth from 37 to 19 Hz are caused by pulling of the formant towards the second harmonic of the glottal excitation located at 246 Hz (Atal and Schroeder, 1974). The shift of the second formant and widening of its bandwidth are, however, caused by the interaction of this formant with the zero at 800 Hz. The bandwidths of the third and fourth formants are also incorrect—probably due to the zero at 800 Hz. The result of the LPC analysis based on a pole-zero model with 12 poles ($p = 12$) and 6 zeros ($q = 6$) is shown in Fig. 6. The power spectrum for a pole-zero model is determined

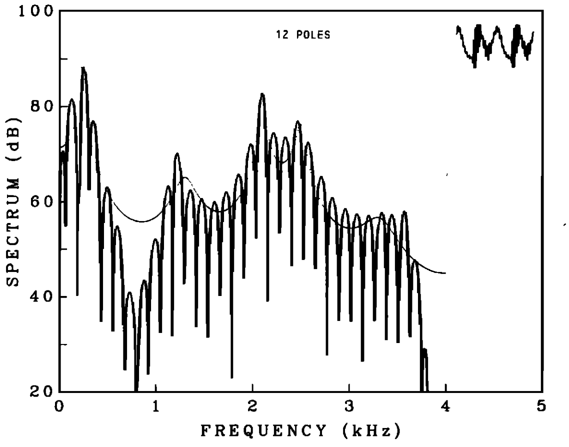

FIG. 5. 12-pole LPC spectral envelope (shown as a dotted curve) and the Fourier spectrum (shown as a solid curve) for a segment of synthetic speech. The frequencies and bandwidths of poles and zeros used to generate synthetic speech are listed in Table I.

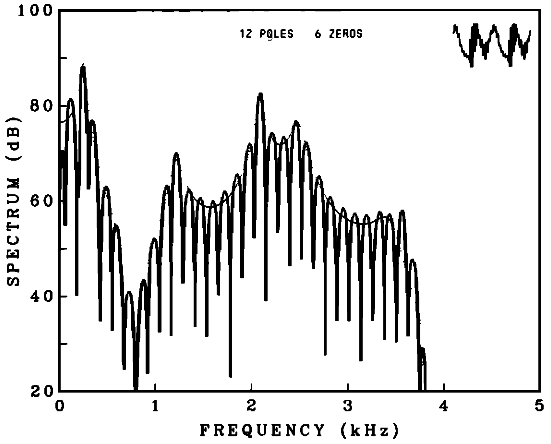

FIG. 6. Spectral envelope based on a pole-zero model with 12 poles and six zeros and the Fourier spectrum for the same segment of synthetic speech as used in Fig. 5.

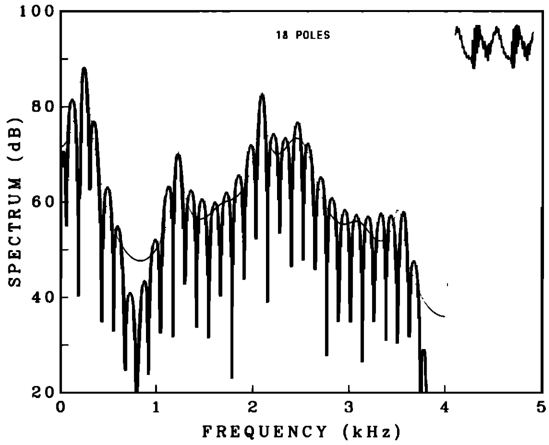

FIG. 7. Spectral envelope from an all-pole LPC analysis using 18 poles and the Fourier spectrum for the same segment of synthetic speech as used in Figs. 5 and 6.

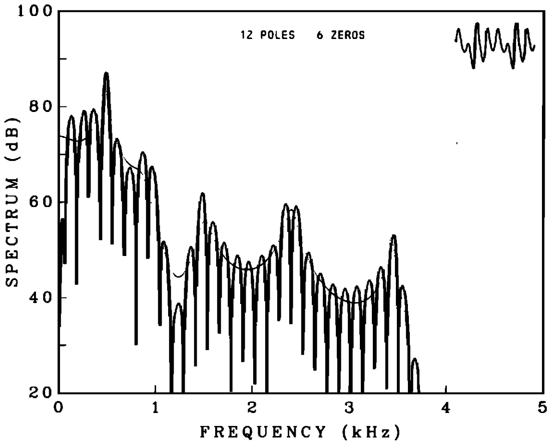

FIG. 9. Spectral envelope derived from a pole-zero model with 12 poles and six zeros and the Fourier spectrum for the same segment of synthetic speech as used in Fig. 8.

from the filter coefficients $a_1, \ldots, a_p$; $b_1, \ldots b_q$ by the equation

$$G(f) = \epsilon_0 \left| 1 + \sum_{k=1}^{q} b_k e^{-2\pi j f k T} \right|^2 \Big/ \left| 1 - \sum_{k=1}^{p} a_k e^{-2\pi j f k T} \right|^2 . \quad (20)$$

The formant frequencies and bandwidths based on the pole-zero model are listed in the last column of Table I. The frequencies of both poles and zeros (with the exception of the first formant) have been identified fairly accurately by the pole-zero model. The error in the first formant frequency is caused by the periodic nature of the excitation and is thus not corrected by the pole-zero analysis. A spectrum for the same speech segment using 18 poles is also shown in Fig. 7. The spectral fit is considerably improved in comparison to the 12-pole spectrum shown in Fig. 5. However, even with a large number of poles the all-pole spectrum is wrong in the vicinity of the zero.

Another set of results with synthetic speech data is shown in Fig. 8 for the all-pole model and in Fig. 9 for the pole-zero model. The frequencies and bandwidths of the formants and zeros as used in the synthesis, as well as those determined by the two analyses, are listed in Table II. As shown in Fig. 8, the second and third formants are almost entirely missed by the all-pole analysis. The large errors in the estimated bandwidths of the second formant in both the all-pole and pole-zero models are caused by the closeness of the formant frequency to the middle of the seventh and eight harmonics of the glottal excitation (located at 861 and 984 Hz, respectively) providing partial cancellation of the formant. The bandwidth errors for the fourth formant are also caused by a similar problem. The errors in the frequency and bandwidth of the third formant present in the all-pole case have been eliminated by the pole-zero model.

In the above examples, 12 poles and six zeros were used in the analysis. Somewhat superior results were obtained by increasing the number of zeros to 12. An important question arises—how many zeros should be used in the pole-zero analysis? It is worth noting here that zeros are also introduced in the spectra by the low-pass filter used prior to sampling in the analog-to-digital conversion, the radiation load at the lips, and the glottal excitation. Thus, a large number of zeros could

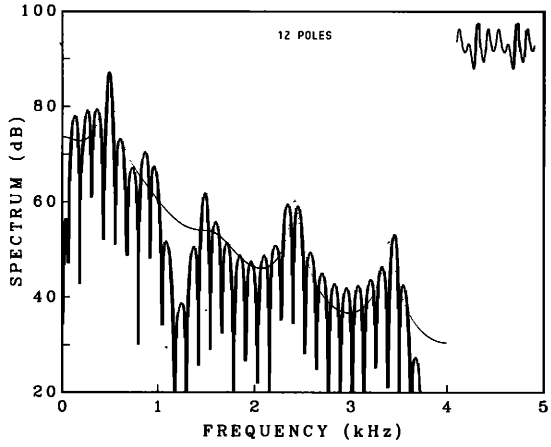

FIG. 8. 12-pole LPC spectral envelope (dotted curve) and the Fourier spectrum (solid curve) for another segment of synthetic speech. The frequencies and bandwidths of poles and zeros used to generate synthetic speech are listed in Table II.

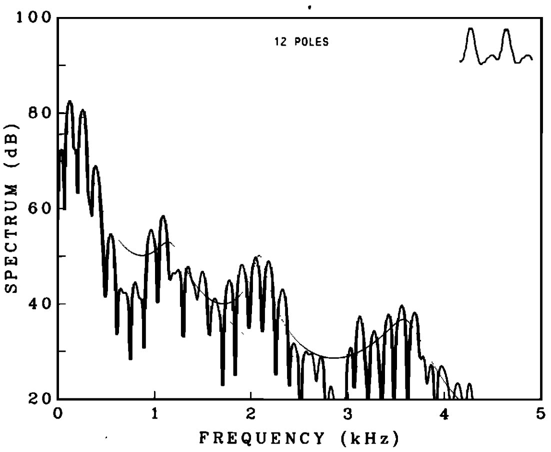

FIG. 10. Spectral envelope from LPC analysis using 12 poles and the Fourier spectrum for the nasal consonant /m/ in the word *coming* spoken as an isolated word.

TABLE II. Formant frequencies and bandwidths for synthetic speech. All-pole analysis was carried out with 12 poles while pole-zero analysis assumed 12 poles and six zeros. Sampling frequency = 8 kHz. Fundamental frequency = 123 Hz.

	Synthesis data		All-pole model		Pole-zero model	
	Frequency	Bandwidth	Frequency	Bandwidth	Frequency	Bandwidth
Poles	466	42	483	57	475	69
	937	57	863	697	843	217
	1500	100	1607	494	1511	122
	2398	39	2421	75	2402	117
	3479	69	3399	77	3418	74
	4430	86				
	5438	105				
	6423	158				
	7374	371				
Zero	1200	50			1270	410

possibly be present in the speech spectra. Our experience with real speech data suggests that 12 zeros ($q = 12$) are capable of providing sufficient accuracy in the representation of the spectral envelope for most nasal consonants and nasalized vowels. In many cases, even six zeros are sufficient.

B. Natural speech

A number of nasal consonants and nasalized vowels were excerpted from natural speech for the analysis. The speech signal was low-pass filtered to 4 kHz, sampled at 10 kHz, and quantized at 12 bits per sample. In all of the results presented in this section, 12 poles were used for the all-pole model while 12 poles and 12 zeros were used for the pole-zero model.

The first two examples are for the nasal consonants /m/ and /ŋ/ in the word *coming* spoken as an isolated word. The power spectrum based on a 12-pole all-pole model and the Fourier spectrum for /m/ are shown in Fig. 10 by dotted and solid lines, respectively. The Fourier spectrum shows a number of zeros. Two of them—at frequencies of approximately 750 and 2900 Hz—have been introduced by the nasalization. The zero at 750 Hz produces a shift in the second formant frequency in the all-pole spectrum. The pole-zero spectrum with 12 poles and 12 zeros is compared with the Fourier spectrum in Fig. 11. The zero at 750 Hz as well as the second formant appear at the correct place in the pole-zero spectrum. The spectrum with 12 poles and six zeros was found to be almost identical to the one shown in Fig. 11 with 12 poles and 12 zeros. The spectra for the final consonant /ŋ/ in *coming* based on all-pole and pole-zero models are shown in Figs. 12 and 13, respectively. The Fourier spectrum for this consonant shows a fairly large number of poles and zeros. In this case, therefore, 14 poles were used for the all-pole analysis and 14 poles with 14 zeros were used for the pole-zero analysis. Although the all-pole spectrum is allowed to have 14 poles, it seems to skip over most of the formants between 1 and 3 kHz—probably due to strong interaction between poles and zeros. The pole-zero model recovers the pole-zero structure quite well. However, there are several discrepancies. The formant at approximately 1100 Hz appears with a considerably wider bandwidth in the pole-zero spectrum—most probably due to the zero at 800 Hz. Similarly, the dip of the zero at 2200 Hz is not approximated accurately.

The third example is for the nasal consonant /n/ in the word *learn* spoken as a part of the utterance "May

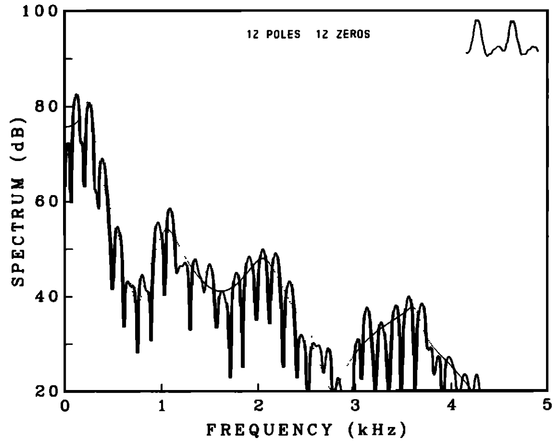

FIG. 11. Spectral envelope based on a pole-zero model with 12 poles and 12 zeros and the Fourier spectrum for the same segment of speech as used in Fig. 10.

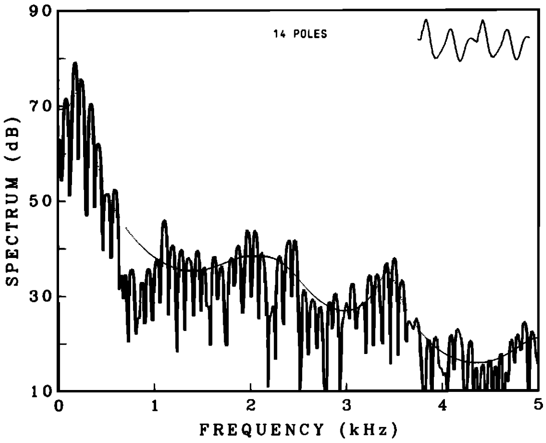

FIG. 12. Spectral envelope from LPC analysis using 14 poles and the Fourier spectrum for the nasal consonant /ŋ/ in the word *coming* spoken in isolation.

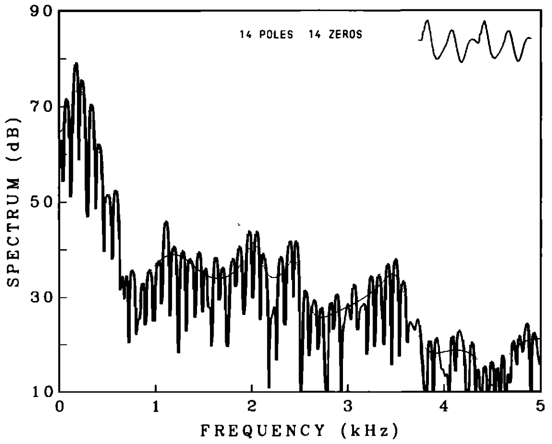

FIG. 13. Spectral envelope based on a pole-zero model with 14 poles and 14 zeros and the Fourier spectrum for the same segment of speech as used in Fig. 12.

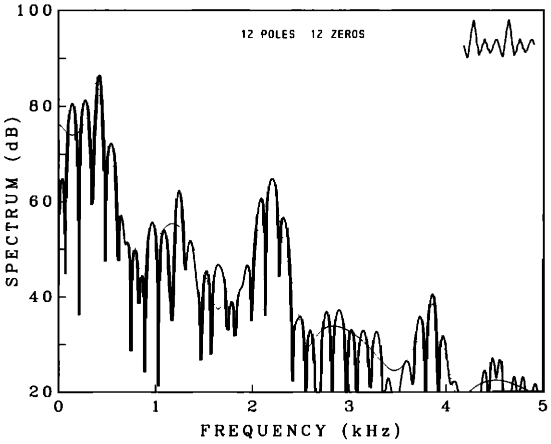

FIG. 15. Spectral envelope based on a pole-zero model with 12 poles and 12 zeros and the Fourier spectrum for the same segment of speech as used in Fig. 14.

we all learn a yellow lion roar." The all-pole spectrum drawn by dotted curve in Fig. 14 again shows the interaction between the zero at 850 Hz and the formant at approximately 1100 Hz. The pole-zero spectrum for this example is shown in Fig. 15. The spectral fit is quite good except in the frequency range 1 to 2 kHz. There appears to be more poles and zeros in this frequency range than provided by the pole-zero solution. The results presented here for the three nasal consonants are typical of the results we get with natural speech data.

How does the pole-zero analysis perform for non-nasalised sounds? In general, the pole-zero analysis does not show any significant improvement over the all-pole analysis for non-nasalised sounds except at low frequencies, where it often provides a better approximation of the zero due to radiation at the lips, or at high frequencies where it follows more accurately the fall-off response of the low-pass filter used in the process of sampling the speech signal. An example, typical of the results obtained for non-nasalized sounds, is shown in Fig. 16 for the all-pole analysis and in Fig. 17 for the pole-zero analysis.

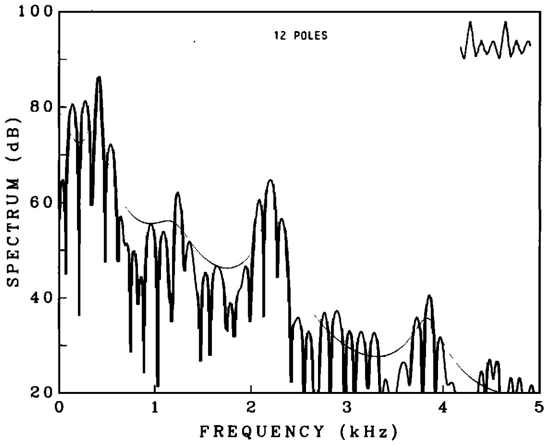

FIG. 14. Spectral envelope from LPC analysis using 12 poles and the Fourier spectrum for the nasal consonant /n/ in the word *learn* spoken in the utterance, " May we all learn a yellow lion roar."

C. Perceptual improvements

The perceptual effect of neglecting zeros on the quality of synthetic speech is not well understood. Mermelstein (1972) has reported that listeners prefer synthetic speech which includes zeros for nasals. Our experience has been that an utterance which includes many nasals often sounds "muffled" when synthesized from an all-pole model with 12 or 14 poles. Furthermore, in isolated words containing the final /n/, such as in *construction* or in *discussion*, the synthetic speech from the all-pole model sounds distinctly buzzy towards the end of the utterance. We were able to trace this buzzy quality to the extra energy present in the synthetic signal at the frequency of the predominant zero. In informal listening, these distortions were found to be absent or substantially reduced in the speech synthesized from the pole-zero model. However, in many cases, an increase in the number of poles in the all-pole model to 20 provided a partial correction of these distortions. These results, although not very conclusive, seem to suggest that the differences between a pole-zero model and an high-order all-pole model may not be that large, as perceived in the quality of the synthetic speech from the two models.

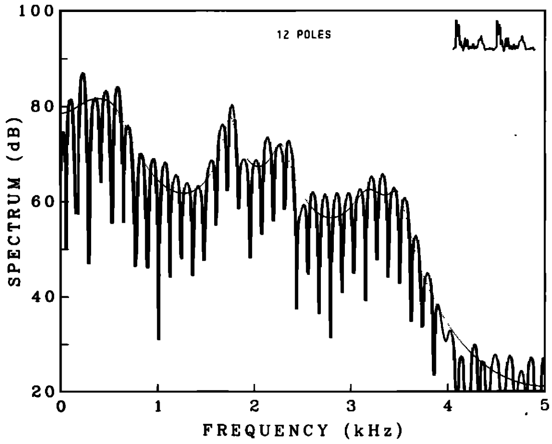

FIG. 16. Spectral envelope from LPC analysis using 12 poles and the Fourier spectrum for the vowel /ɛ/.

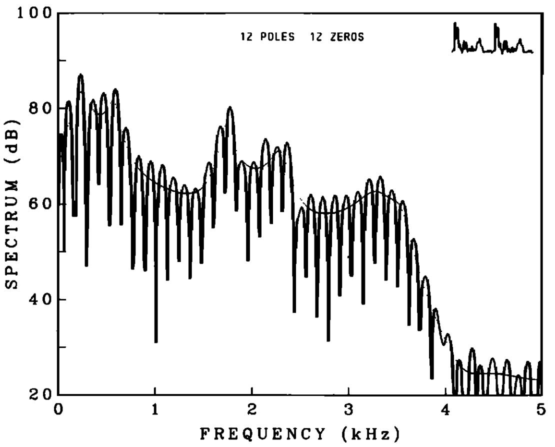

FIG. 17. Spectral envelope based on a pole-zero model with 12 poles and 12 zeros and the Fourier spectrum for the same segment of speech as used in Fig.,16.

IV. SUMMARY

We have described a method for linear prediction analysis of speech based on a pole-zero model. In this method, an impulse response representing the composite filtering action of the glottal wave, the vocal tract, the radiation, and the speech recording system is first constructed from the speech signal. This impulse response is obtained by performing several stages of all-pole LPC analysis. The pole-zero parameters are determined from the impulse response by solving a set of linear simultaneous equations.

The method was tested both on natural and synthetic speech data. The spectral envelope of the speech signal obtained from the pole-zero model was found to provide a good fit to the spectrum obtained by direct Fourier analysis of the speech waveform. Informal listening tests show that many distortions introduced by the all-pole model in synthetic speech are reduced significantly when speech is synthesized by a linear filter which includes both poles and zeros. The perceptual differences between synthetic speech signals from pole-zero models and high-order all-pole models are however small.

[1] It is a common practice in computing the Fourier spectrum to use a Hamming or some other similar window on the signal to avoid the Gibb's phenomena caused by finite truncation of the signal. Voiced speech spectra contains formant regions with high concentration of energy. For such spectra, the Gibb's phenomena can cause a significant spillover of formant energy in the inter-formant regions. However, use of Hamming window introduces distortions in the speech spectrum. We did not use such a windowing procedure in the spectral analysis. The Gibb's phenomena was avoided by using an inverse filter with zeros at the formants as a whitening filter on the speech signal prior to Fourier analysis. The coefficients of the inverse filter were determined using a 25th-order LPC analysis. The spectrum of the speech signal was obtained by dividing the Fourier spectrum of the whitened signal with the Fourier spectrum of the inverse filter. To avoid division by zero, the zeros of the inverse filter were constrained to have a minimum bandwidth of 50 Hz. For computing the logarithmic spectrum expressed in dB, the division is transformed to forming a difference between the two spectra expressed in dB.

[2] Strictly speaking, the spectrum of the excitation could be considered flat only for unvoiced speech. Due to the periodic nature of voiced speech, only the spectral envelope of the excitation, but not its fine structure, can be assumed to be flat. The assumption of flat spectrum of voiced speech introduces an additional source of error in the LPC analysis. In this paper, we are considering problems in the LPC analysis due to the presence of spectral zeros. The problems introduced by the periodicity of voiced speech will therefore be ignored.

Atal, B. S., and Schroeder, M. R. **(1967).** "Predictive coding of speech signals," Proc. 1967 Conf. on Comm. Process., pp. 360–361.

Atal, B. S., and Schroeder, M. R. **(1970).** "Adaptive predictive coding of speech signals," Bell Sys. Tech. J. **49,** 1973–1986.

Atal, B. S., and Hanauer, S. L. **(1971).** "Speech analysis and synthesis by linear prediction," J. Acoust. Soc. Am. **50,** 637–655.

Atal, B. S., and Schroeder, M. R. **(1974).** "Recent advances in predictive coding—applications to speech synthesis," in *Proceedings of the Speech Communications Seminar*, Stockholm, Vol. 1, pp. 27–31.

Atal, B. S. **(1975).** "Linear prediction of speech—recent advances with applications to speech analysis," in *Speech Recognition*, edited by D. R. Reddy (Academic, New York), 221–230.

Atal, B. S., and Schroeder, M. R. **(1975).** "Linear prediction analysis of speech based on a pole-zero model," J. Acoust. Soc. Am. **58,** S96(A) (1975).

Atal, B. S. **(1977).** "On determining partial correlation coefficients by the covariance method of linear prediction," J. Acoust. Am. **62,** S64(A).

Durbin, J. **(1959).** "Efficient estimation of parameters in moving-average models," Biometrika, **46,** parts 1 and 2, pp. 306–316.

Flanagan, J. L. **(1972)**. *Speech Analysis Synthesis and Perception* (Springer–Verlag, New York), p. 281.

Hannan, E. J. **(1969).** "The estimation of mixed moving average autoregressive systems," Biometrika, **56,** 579–593.

Hsia, T. S., and Landgrebe, D. A. **(1967).** "On a Method of Estimating power Spectra," IEEE Trans. Instrum. Meas. **IM-16,** 255–257.

Itakura, F., and Saito, S. **(1968).** "Analysis synthesis telephony based on the maximum likelihood method," *Reports of the 6th International Congress on Acoustics, Tokyo*, II, Paper C-5-5.

Makhoul, J. I. **(1974).** "Selective linear prediction and analysis-synthesis in speech analysis," Technical Report No. 2578, Bolt Beranek and Newman, Inc.

Makhoul, J. I. **(1975).** "Linear prediction: A tutorial review," Proc. IEEE **63,** 561–580.

Markel, J. D., and Gray, A. H., Jr. **(1976).** *Linear Prediction of Speech* (Springer–Verlag, New York).

Mermelstein, P. **(1972).** "Speech synthesis with the aid of a recursive filter approximating the transfer function of the nasalized vocal tract," *Conference Record* 1972 *Conference on Speech Communication and Processing*, 24–26 April, 1972, paper D7.

Oppenheim, A. V., Tribolet, J. M., and Kopec, G. E. **(1976).** "Speech analysis by homomorphic prediction," IEEE Trans. Acous. Speech Signal Process. ASSP-**24,** 327–332.

Rosenberg, A. E. **(1970).** "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **47,** No. 2, part 2, 583–590.

Schwarz, R. J., and Friedland, B. **(1965).** *Linear Systems* (McGraw–Hill, New York), pp. 320–325.

Shanks, J. L. **(1967).** "Recursion filters for digital processing," Geophysics **32,** 33–51.

Steiglitz, K., and McBridge, L. E. **(1965).** "A technique for the identification of linear systems," IEEE Trans. Automat. Contr. **AC-10,** 461–464.

Steiglitz, K. **(1977).** "On the simultaneous estimation of poles and zeros in speech analysis," IEEE Trans. Acoust. Speech, and Signal Proc. ASSP-**25,** 229–234.