Speech Analysis by Homomorphic Prediction

GARY E. KOPEC, STUDENT MEMBER, IEEE, ALAN V. OPPENHEIM, SENIOR MEMBER, IEEE, AND JOSÉ M. TRIBOLET, STUDENT MEMBER, IEEE

Abstract—**Linear prediction is a generally accepted method for obtaining all-pole speech representations. However, in many situations (e.g., nasalization studies) spectral zeros are important and a more general modeling procedure is required. Unfortunately, the need for pitch synchronization has limited the success of available techniques. This paper explores a novel approach to pole-zero analysis, called *homomorphic prediction*, which seems to avoid the synchronization problem. A minimum-phase estimate of the vocal-tract impulse response is obtained by homomorphic filtering of the speech waveform. Such a signal, by definition, has a known time registration. Linear prediction is applied to this waveform to identify its poles. The LPC "residual" (error signal) is computed by inverse filtering. This signal contains the information about the zeros. Its *z* transform is then approximated by a polynomial either through a weighted least squares procedure (homomorphic prediction, using Shanks' method of finding zeros), or by spectral inversion followed by a second pass of LPC (homomorphic prediction involving "inverse LPC"). Results of a preliminary evaluation on real and synthetic speech are presented.**

I. Introduction

IN MANY AREAS of speech processing, the use of a *parametric representation* of the speech signal is of central importance [1]. Most such representations are based on a model of the speech process in which a slowly time-varying linear system (representing the vocal tract) is excited by a quasi-periodic pulse train (for voiced segments) or random noise (for unvoiced segments). In one class of parametric models the vocal-tract transfer function is considered to be rational over short time intervals so that it is characterized by a finite number of poles and zeros, or equivalently by the coefficients of its numerator and denominator. The set of algorithms referred to collectively as *linear prediction* represents a highly efficient means for determining the coefficients of the denominator polynomial, and thus has been extremely successful in the all-pole modeling of speech. Furthermore, because of the spectral matching properties of linear prediction [30], it is possible to apply it directly to a speech segment consisting of several periods of voiced speech. Linear prediction tends to model the *envelope* of the spectrum without being sensitive to the fine structure caused by the periodicity [30] of the time signal. Techniques which have been proposed for simultaneous determination of both poles and zeros have been less successful for a variety of reasons. Many of them are computationally complicated and inefficient. Since speech-processing facilities tend to be minicomputer-based and oriented towards on-line interactive computation, there are definite limits on the available memory, arithmetic speed, and numerical precision.

A second consideration is that few modeling techniques which include zeros can be applied directly to a segment containing several periods of speech, since it is difficult to distinguish between zeros introduced by the vocal tract and those resulting from the excitation. Techniques for pole-zero modeling thus require some type of deconvolution of the speech waveform. One procedure commonly proposed for the deconvolution is *pitch synchronous analysis* in which individual pitch periods are extracted and analyzed. Such techniques involve specifying the location of each pitch pulse in a voiced segment. Unless this is done accurately, the resulting representation will be unsatisfactory.

In this paper we consider an approach to pole-zero modeling of speech which avoids the problem of pitch synchronization. The basic procedure is to perform pole-zero identification from a minimum-phase estimate of the vocal-tract impulse response rather than from the speech waveform itself. The impulse response of the vocal tract by definition is its response to a single pitch pulse located at the time origin. Therefore, in modeling this signal there is no problem with synchronization. The vocal-tract response is estimated by *homomorphic filtering* of the speech waveform [13]-[15], [31], [32].

Once the vocal-tract impulse response is obtained, any of a variety of methods of pole-zero analysis may be used. However, the success of linear prediction in all-pole speech modeling suggests using a suitable extension of that technique. In this spirit, two generalizations of LPC will be considered, the method of Shanks [7], and one similar to a technique which has been called "inverse LPC" [6], [8], [9].

Analysis methods which result from combining homomorphic filtering with linear prediction (or one of its generalizations) have been collectively termed *homomorphic prediction* [10]. This paper is thus a preliminary exploration of two methods of homomorphic prediction, corresponding to the two extensions to LPC mentioned above, in speech analysis. Both of the techniques were tested on several natural and synthetic speech signals and some typical results are presented. These illustrate the steps involved in applying homomorphic prediction and demonstrate some of its characteristics.

In the next sections we briefly review the techniques of homomorphic deconvolution, pole estimation by linear prediction, and zero estimation by Shanks' method and by inverse linear prediction analysis. We then discuss, within the context of several examples, pole-zero modeling by homomorphic prediction.

II. Homomorphic Deconvolution

Homomorphic filtering is a technique which has been used for separating sequences combined through multiplication or convolution [13]. Its use in speech analysis to estimate the

Manuscript received April 9, 1976; revised August 23, 1976. This work was supported in part by the Advanced Research Projects Agency monitored by ONR under Contract N00014-75-C-0951-NR 049-308, in part by the Fannie and John Hertz Foundation, and in part by the National Science Foundation under Grant ENG71-02319-A02.

The authors are with the Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

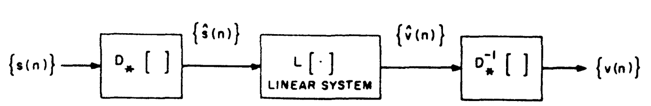

Fig. 1. Canonic form homomorphic deconvolution processor (after Oppenheim, Schafer and Stockham [13]).

vocal-tract impulse response from the speech waveform has been investigated in detail [14], [15], [31], [32].

In the usual speech-production model a segment of voiced speech, $\{s(n)\}$, is the convolution of the vocal-tract impulse response, $\{v(n)\}$, with a quasi-periodic train of pitch pulses, $\{p(n)\}$ so that

$$\{s(n)\} = \{p(n)\} * \{v(n)\}. \tag{1}$$

The homomorphic estimation of $\{v(n)\}$ is a three-step process as shown in Fig. 1. The sequence $\{\hat{s}(n)\}$, the *complex cepstrum* of $\{s(n)\}$, is defined by the equation

$$\hat{S}(e^{j\omega}) = D_*[S(e^{j\omega})] = \log\,[S(e^{j\omega})], \tag{2}$$

where $S(e^{j\omega})$ is the Fourier transform of $\{s(n)\}$, and the complex logarithm is appropriately defined [19]. The fundamental property of the complex cepstrum is that if (1) holds, then

$$\hat{s}(n) = \hat{p}(n) + \hat{v}(n), \tag{3}$$

so that the cepstra of convolved sequences are related by ordinary addition [18]. Since $\{\hat{p}(n)\}$ and $\{\hat{v}(n)\}$ occupy disjoint time intervals [31], the vocal-tract impulse response may be recovered by choosing as $L[\cdot]$ a time domain window which retains the low-time portion of the cepstrum and attenuates the long-time portion. If, in place of (2), the system $D_*[\cdot]$ is defined by

$$\hat{S}(e^{j\omega}) = D_*[S(e^{j\omega})] = \log|S(e^{j\omega})|, \tag{4}$$

then, by choosing as the linear system $L[\cdot]$ a time-domain window which is zero for negative time and retains only the low-time portion of the cepstrum for positive time, the output of the system of Fig. 1 will be a minimum-phase approximation to the vocal-tract impulse response.

III. Pole-Zero Estimation

The basic approach to pole-zero estimation by homomorphic prediction is to first obtain a minimum-phase estimate of the vocal-tract impulse response by homomorphic deconvolution, and then estimate the poles and zeros of this impulse response in two stages. First, linear prediction is used to identify the poles. Then the zeros are located, making use of the information available about the poles. This strategy is attractive because it directly exploits the fact that linear prediction has been very successful for all-pole speech modeling.

In the remainder of this section we will discuss two algorithms for pole-zero modeling based on the above ideas. When coupled with minimum-phase homomorphic deconvolution, they define two methods of speech analysis by homomorphic prediction, as shown in Fig. 2.

Pole Identification by Linear Prediction

Since linear prediction has been widely investigated and used in speech analysis we review it here only briefly to define our terminology and notation. Let $V(z)$ denote the vocal-tract transfer function and suppose that it is given *exactly* by

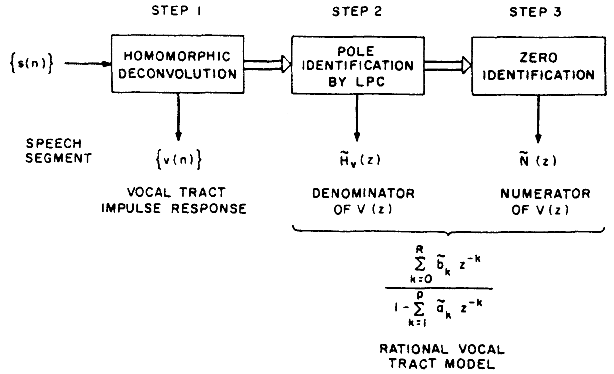
STEP 1
STEP 2
STEP 3
{s(n)}
HOMOMORPHIC DECONVOLUTION
POLE IDENTIFICATION BY LPC
ZERO IDENTIFICATION
SPEECH SEGMENT
{v(n)}
$\tilde{H}_v(z)$
$\tilde{N}(z)$
VOCAL TRACT IMPULSE RESPONSE
DENOMINATOR OF V(z)
NUMERATOR OF V(z)
$\dfrac{\sum_{k=0}^{R} \tilde{b}_k z^{-k}}{1-\sum_{k=1}^{P} \tilde{a}_k z^{-k}}$
RATIONAL VOCAL TRACT MODEL

Fig. 2. Overall strategy of pole-zero modeling by homomorphic prediction.

$$V(z) \equiv \frac{N(z)}{D(z)} = \frac{\sum_{k=0}^{R} b_k z^{-k}}{1 - \sum_{k=1}^{P} a_k z^{-k}}. \tag{5}$$

Then

$$v(n) = \sum_{k=1}^{P} a_k v(n-k) \qquad \text{for } n > R, \tag{6}$$

so that $v(n)$ is "perfectly predictable" from the preceding P values. By solving P linear equations of the form of (6), the coefficients $\{a_k\}$ can be computed from $2P$ points of $\{v(n)\}$. Note that since (6) does not involve the numerator coefficients, $\{b_k\}$, identification of the poles of $V(z)$ is independent of the presence of zeros.

The *covariance* form of linear prediction [16], [22] is an extension of this technique to the situation where the recorded signal, $\{v(n)\}$, only approximates the impulse response of a rational system. In that case,

$$v(n) = \sum_{k=1}^{P} a_k v(n-k) + e(n) \qquad n > R$$

where $e(n)$ is the *single-point prediction error,*

$$e(n) \equiv v(n) - \sum_{k=1}^{P} a_k v(n-k). \tag{7}$$

The coefficients, $\{a_k\}$, are estimated by minimizing the *total squared prediction error*

$$E \equiv \sum_{n=Q+1}^{\infty} |e(n)|^2, \tag{8}$$

where Q represents the length of an initial segment of $\{v(n)\}$ which, because of the zeros of $V(z)$, is not predictable.

The resulting estimate of the denominator of $V(z)$,

$$\tilde{H}_v(z) \equiv 1 - \sum_{k=1}^{P} \tilde{a}_k z^{-k} \tag{9}$$

is the *predictor polynomial.* From (7)

$$\tilde{E}_v(z) = V(z)\tilde{H}_v(z), \tag{10}$$

so that the "residual" *LPC error signal,* $\{\tilde{e}_v(n)\}$ can be obtained by passing $\{v(n)\}$ through the finite length impulse response filter whose transfer function is $\tilde{H}_v(z)$.

In addition to the covariance method of linear prediction reviewed above, there is an alternate formulation, known as the autocorrelation method [16], [25], which is also widely used. For all-pole modeling the two techniques are distinguished primarily by computational considerations; both methods give roughly the same result. However, if zeros are being explicitly estimated there may be a theoretical reason for preferring the covariance formulation. As shown in [16], with autocorrelation LPC as the predictor order increases the frequency response of the all-pole model asymptotically approaches the actual response $|V(e^{j\omega})|$. For this reason, the autocorrelation method performs all-pole *modeling* of $V(z)$.

In contrast, in a two-stage algorithm it is desirable that the first step be one of pole *identification.* The difference is that pole identification involves approximating only the denominator of $V(z)$ rather than the overall transfer function. The covariance method is the appropriate one in that case.

Zero Identification by Shanks' Method

At least two algorithms are available for identifying spectral zeros once the poles have been located. Shanks [7] proposed a method in which the numerator of $V(z)$ is estimated by a least squares criterion. Suppose that $V(z)$ is given by (5). Then

$$v(n) = \sum_{k=0}^{R} b_k f(n-k) \tag{11}$$

where $\{f(n)\}$ is the sequence with all-pole z transform

$$F(z) \equiv \frac{1}{D(z)}.$$

In Shanks' method, $D(z)$ is estimated by covariance LPC, and then $N(z)$ is found by minimizing

$$E \equiv \sum_{n=0}^{\infty} \left| v(n) - \sum_{k=0}^{R} b_k \tilde{f}(n-k) \right|^2 \tag{12}$$

where

$$\tilde{F}(z) \equiv \frac{1}{\tilde{H}_v(z)}.$$

This leads to the set of linear equations

$$\sum_{k=0}^{R} \tilde{b}_k \phi_{\tilde{f}\tilde{f}}(k,r) = \phi_{v\tilde{f}}(0,r), \qquad r = 0 \cdots R \tag{13}$$

in which

$$\phi_{x_1 x_2}(t,u) \equiv \sum_{n=0}^{\infty} x_1(n-t)x_2(n-u) \tag{14}$$

is the *cross-correlation* between $\{x_1(n)\}$ and $\{x_2(n)\}$.

An interpretation of Shanks' method as a least squares polynomial approximation is established by rewriting (12) in the frequency domain. Applying Parseval's theorem and the definition of $\tilde{F}(z)$,

$$E = \int_{-\pi}^{\pi} \left| V(z)\tilde{H}_v(z) - \sum_{k=0}^{R} b_k z^{-k} \right|^2 |\tilde{F}(z)|^2 \frac{d\omega}{2\pi}$$

where $z \equiv e^{j\omega}$. Recalling (10) we find that

$$E = \int_{-\pi}^{\pi} |\tilde{E}_v(z) - N_s(z)|^2 \, |\tilde{F}(z)|^2 \frac{d\omega}{2\pi}, \tag{15}$$

where

$$N_s(z) \equiv \sum_{k=0}^{R} b_k z^{-k}.$$

Thus Shanks' method estimates the numerator of $V(z)$ by fitting a polynomial, $N_s(z)$, to the z transform of the LPC error signal, $\{\tilde{e}_v(n)\}$.

Zero Identification by Inverse LPC

A second way of estimating $N(z)$ was proposed by Makhoul [6] and is similar to a technique described by Durbin [8], [9]. This approach involves inverting the spectrum of $\{v(n)\}$, and then using linear prediction to estimate its zeros. If $V(z)$ is given by (5) then

$$V^{-1}(z) \equiv \frac{1}{V(z)} = \frac{D(z)}{N(z)}$$

so that the poles of $V^{-1}(z)$ are the zeros of $V(z)$ and vice versa. When LPC is applied to $\{v^{-1}(n)\}$, the resulting predictor polynomial is an estimate of $N(z)$.

A variation of this method considered here is to invert the spectrum of the LPC error signal, $\{\tilde{e}_v(n)\}$, rather than that of $\{v(n)\}$. If $\tilde{H}_v(z) \simeq D(z)$ then

$$\tilde{E}_v(z) \equiv V(z)\tilde{H}_v(z) \simeq N(z)$$

and

$$\tilde{E}_v^{-1}(z) \simeq \frac{1}{N(z)}.$$

This is equivalent to removing the zeros of $V^{-1}(z)$ prior to LPC analysis. One reason for doing this is to facilitate comparing inverse LPC to Shanks' method. If linear prediction is applied to $\{\tilde{e}_v^{-1}(n)\}$, then both methods may be interpreted as modeling the LPC error signal as a finite length sequence.

IV. Experimental Results

In this section seven examples of pole-zero analysis by homomorphic prediction will be presented. The examples have been selected specifically for the way in which they display certain features. In particular, the speech segments chosen exhibit spectral antiresonances which are considerably more distinct than those commonly found in natural utterances. Thus, the results presented here do not represent a careful evaluation of homomorphic prediction, but only serve to illustrate the technique.

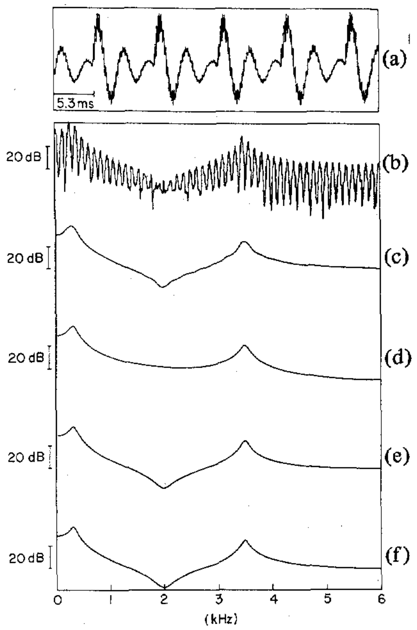

Fig. 3. Example 1: 4 pole/2 zero synthetic signal-4 pole/2 zero model. (a) Segment of impulse train response. (b) Log spectrum of (a). (c) Log spectrum after homomorphic deconvolution of (a). (d) Result of 4 pole LPC analysis of (c). (e) Log spectrum of 4 pole/2 zero model with the zeros estimated by Shanks' method. (f) 4 pole/2 zero model with zeros estimated by inverse LPC.

The first three examples involve totally synthetic digital data. Since the actual filter structures are known, the accuracy of the pole and zero estimates can be judged quantitatively. The remaining examples use real speech signals. These were extracted from sentences recited by a male speaker. The waveforms were sampled at a rate of 9.4 kHz and quantized to 12 bits. The sampling rate was selected to ensure that the frequency response of the available low-pass analog prefilter was reasonably flat everywhere within the Nyquist band. At higher sampling frequencies, the filter characteristic rolled off significantly below one-half the sampling rate and introduced apparent spectral zeros at the Nyquist frequency. Some aliasing was present with 9.4 kHz sampling, but this was felt to be less harmful than strong extraneous antiresonances.

The natural speech was digitally preemphasized using a single positive real axis zero with a bandwidth of 236* Hz.[1] The number of poles and zeros used to represent any particular speech segment was established by trial and error based on visual comparison of the actual and model spectral envelopes. No attempts were made to systematize this process or minimize the total number of parameters required for a given degree of spectral fidelity.

In Examples 1 to 3 (synthetic signals) a sampling rate of 12 kHz is assumed. This is done so that pole and zero locations may be specified in terms of frequency (hertz) and bandwidth.

Example 1–4 Pole/2 Zero Synthetic Signal-4 Pole/2 Zero Model: Fig. 3 and Table I summarize the analysis of a portion of the impulse train response of a 4 pole/2 zero digital filter. The first column of Table I lists the actual pole/zero structure of the network used to generate the synthetic signal shown in Fig. 3(a). The "pitch period" is 100 samples, corresponding to a fundamental frequency of 120 Hz. Fig. 3(c) shows the estimate of the log spectral envelope obtained by homomorphic filtering from the first 50 points of the minimum phase cepstrum of $\{s(n)\}$.

The second column of Table I gives the pole estimates obtained through fourth-order linear prediction of $\{v(n)\}$. Fig. 3(d) shows the corresponding spectrum log $|\tilde{H}_v^{-1}(e^{j\omega})|$. When Shanks' method is used to locate the zeros of $V(z)$ the resulting estimates are those given in column three of Table I. Combined with the LPC pole estimates, they result in the log spectral envelope in Fig. 3(e).

Homomorphic prediction using inverse LPC gave the zeros listed in the fourth column of Table I and produced the overall model spectrum shown in Fig. 3(f).

One interesting feature of this example is that the final 4 pole/2 zero model spectra [(Fig. 3(e) and 3(f)] exhibit resonances which are noticeably sharper than those of the estimated vocal-tract response Fig. 3(c). This phenomenon has been frequently observed both with real and synthetic signals.

Example 2–4 Pole/3 Zero Synthetic Signal-4 Pole/2 Zero Model: The number of poles and zeros used in Example 1 to approximate $V(z)$ was the same as the actual number of parameters in the filter which generated $\{s(n)\}$. In this example, homomorphic prediction is applied to a signal which is characterized by more zeros than the number included in the model. As a result, not all features of the actual signal spectrum can be represented. This situation is roughly analogous to that encountered when analyzing real speech, which fits no rational model exactly. Although the present example is much too simple to be truly prototypical of actual speech, the performances of the two methods of homomorphic prediction are similar to their behavior with natural signals.

Table II describes the test signal which was obtained by filtering $\{s(n)\}$ of Example 1 and lists the poles and zeros estimated when the previous 4 pole/2 zero analyses were repeated. The various waveforms associated with the analysis are given in Fig. 4.

By comparing Fig. 4(a) with Fig. 3(c), it may be seen that the extra zero depresses the spectrum of $\{v(n)\}$ by about 12 dB at 6000 Hz and thus represents a very wide but shallow antiresonance compared to the one at 2 kHz.

Linear prediction analysis of $\{v(n)\}$ results in pole estimates nearly identical to those obtained in Example 1. When Shanks' method is used in homomorphic prediction to locate a single zero-pair, an antiresonance is detected at 2.4 kHz. On the other hand, the inverse LPC technique identifies the zero at 2.0 kHz, although its bandwidth is significantly overestimated. In general it has been found that homomorphic prediction using inverse LPC provides somewhat better estimates of antiresonance frequency than homomorphic prediction

[1] Bandwidth is computed by the formula $BW = (F_s/2\pi) \ln (z_R^2 + z_I^2)$ where F_s is the sampling rate and z_R and z_I are the real and imaginary parts, respectively, of a pole or zero within the unit circle. When $z_I = 0$ (e.g., dc zero) one-half this value is given. To minimize confusion, such bandwidths are indicated with an asterisk. When $z_R^2 + z_I^2 > 1$, the computed bandwidth is negative.

TABLE I
SUMMARY OF EXAMPLE 1—ANALYSIS OF 4 POLE/2 ZERO SYNTHETIC SIGNAL

actual poles/zeros		est. poles	est. zeros	est. zeros
pole/zero	freq/bw(Hz)	(LPC, P=4)	(Shanks',R=2)	(inverse LPC,R=2)
P	292/79	291/118	-	-
P	3500/100	3498/128	-	-
Z	2000/200	-	2004/242	1998/240

Note: Analysis window length = 21 ms (256 points), cepstral cutoff for $\{\hat{v}_{mp}(n)\}$ = 4.2 ms (50 points).

TABLE II
SUMMARY OF EXAMPLE 2—4 POLE/3 ZERO SYNTHETIC SIGNAL

actual poles/zeros		est. poles	est. zeros	est. zeros
pole/zero	freq/bw(Hz)	(LPC, P=4)	(Shanks',R=2)	(inverse LPC,R=2)
P	292/79	291/117	-	-
P	3500/100	3499/131	-	-
Z	2000/200	-	2439/505	2040/343
Z	6000/1000*	-	-	-

Note: Analysis as in example 1.

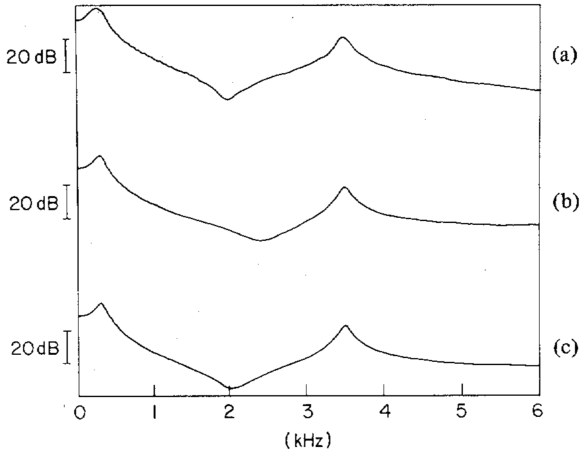

Fig. 4. Example 2: 4 pole/3 zero synthetic signal–4 pole/2 zero model. (a) Log spectral envelope obtained by homomorphic deconvolution. (b) 4 pole/2 zero model with zeros estimated by Shanks' method. (c) 4 pole/2 zero model with zeros estimated by inverse LPC.

in which the zeros are found by Shanks' method (for real speech spectra this was determined visually). Inverse LPC tends to ignore wide but shallow spectral depressions in favor of sharp antiresonances if there are insufficient parameters to represent both. This is not at all surprising since the corresponding phenomenon for LPC pole identification is well known [16].

Example 3–Artifical Voiced Speech: The sentence "May we all learn a yellow lion roar" was synthesized by rule from its phonetic transcription using a terminal-analog synthesizer [26]. In this system, the vocal tract is represented by a cascade of 6 all-pole digital resonators. Laryngeal excitation is produced by shaping an impulse train with a filter consisting of a pair of real-axis poles which impose an approximately 12 dB/octave falloff above 300 Hz. The radiation load at the mouth is simulated by first-differencing the vocal-tract output. Since the synthesizer is an all-pole system (except for the radiation zero at dc) the output waveform was filtered to introduce a fixed antiresonance of frequency 2000 Hz and bandwidth 200 Hz. Although the actual pole and zero locations are again known *a priori* in this example, the synthesizer parameters are now time varying.

The test signal, $\{s(n)\}$, is a 30 ms segment of |e| from "May" which was preemphasized to remove both of the glottal source poles. Thus it is characterized by 12 poles and 3 zeros. Over the interval of $\{s(n)\}$, the fundamental frequency is constant at 108 Hz, and the first five formants vary with time within the ranges indicated in Table V. The sixth formant is fixed, but its actual location is unknown.

Columns 2 through 4 of Table III and Fig. 5 summarize the 12 pole/3 zero analysis of $\{s(n)\}$. From Fig. 5(c) and (d) it appears that homomorphic prediction using inverse LPC provides a visually superior representation of the antiresonance structure of $V(z)$ than does homomorphic prediction involving Shanks' method.

The last column of Table III lists the estimates obtained

TABLE V
SUMMARY OF EXAMPLE 5—ANALYSIS OF NATURAL VOWEL WITH ARTIFICIAL ZERO

Original Segment			Original Segment plus z: 1994/157		
est. poles (LPC, P=11) freq/bw (Hz)	est. zeros (Shanks', R=2)	est. zeros (inv. LPC, R=2)	est. poles (LPC, P=11)	est. zeros (Shanks', R=4)	est. zeros (inv. LPC, R=4)
769/222	-	-	772/214	-	-
1300/152	-	-	1285/194	-	-
2570/166	-	-	2558/176	-	-
3389/183	-	-	3382/136	-	-
4050/899	-	-	4146/582	-	-
0/102*	-	-	0/84*	-	-
-	0/178*	0/3*	-	0/102*	0/3*
-	4704/568*	4704/132*	-	4704/293*	4704/112*
-	-	-	-	2068/211	1978/206

Note: Analysis interval = 30 ms (285 points), cepstral cutoff = 5.3 ms (50 points).

TABLE III
SUMMARY OF EXAMPLE 3—ARTIFICIAL VOICED SPEECH

actual p/z			est. poles	est. zeros	est. zeros	est. zeros
p/z	freq	bw	(LPC, P=12) freq/bw (Hz)	(Shanks', R=3)	(inv. LPC, R=3)	(Shanks', R=5)
P	413-465	72-101	468/45	-	-	-
P	1619-1727	139-189	1745/178	-	-	-
P	2359-2403	184-193	2411/169	-	-	-
P	3252-3386	223-230	3374/162	-	-	-
P	4024-4091	351-417	4092/273	-	-	-
P	?	?	5420/458	-	-	-
Z	2000	200	-	1837/1216	2014/163	2077/251
Z	0	0*	-	0/1019*	0/14*	0/74*
Z	-	-	-	-	-	5237/4762

Note: Analysis interval = 30 ms (360 points), cepstral cutoff time = 4.2 ms (50 points).

from homomorphic prediction using Shanks' method when the number of zeros in the model is increased to five.

Example 4–Natural Vowel: In this example and the next, homomorphic prediction is used to model a segment of natural speech to which a known antiresonance was added by filtering. Two sets of results are presented, for analyses performed before and after the synthetic zeros were introduced. These form Examples 4 and 5, respectively. The analysis of the original speech segment is included to illustrate pole-zero modeling in the absence of obvious antiresonances.

Fig. 6 and Table IV summarize the 11 pole/2 zero analysis of a 30 ms segment of |a| from the word "five." Note that a relatively narrow-band zero was found at dc by both methods of homomorphic prediction. Since the acoustic radiation characteristics of the mouth may be modeled roughly by a dc zero on the unit circle, it is tempting to associate the observed zero with that mechanism. However, the audio equipment used to record the speech signals contained several capacitively coupled amplifiers. Thus, it is more likely that the apparent dc zeros are due to the recording electronics.

Example 5–Natural Vowel Plus Artificial Zero: Table V and Fig. 7 were obtained after a fixed antiresonance of frequency 1994 Hz and bandwidth 157 Hz was added to the speech waveform of Example 4. The results of the previous analysis are repeated in the left half of Table V for comparison.

Example 6–Natural Nasalized Vowel ("moon"): Table VI and Fig. 8 summarize the 10 pole/6 zero modeling of a segment of |u| from the word "moon." In Fig. 8(a) there appears

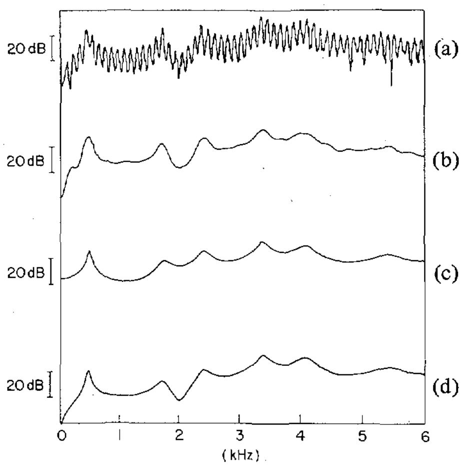

Fig. 5. Artificial voiced speech. (a) Log spectrum of preemphasized |e| from "may". (b) Log spectral envelope obtained by homomorphic filtering. (c) 12 pole/3 zero model with zeros estimated by Shanks' method. (d) 12 pole/3 zero model with zeros estimated by inverse LPC.

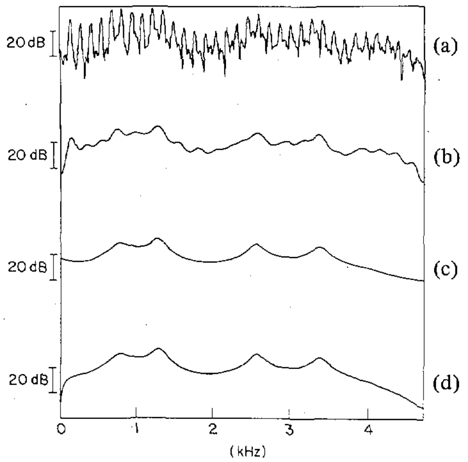

Fig. 6. Example 4: natural vowel. (a) Log spectrum of preemphasized |a| from "five". (b) Log spectral envelope obtained by homomorphic filtering. (c) 11 pole/2 zero model with zeros estimated by Shanks' method. (d) 11 pole/2 zero model with zeros estimated by inverse LPC.

TABLE IV
SUMMARY OF EXAMPLE 4—ANALYSIS OF NATURAL VOWEL

est. poles (LPC, P=11) freq/bw (Hz)	est. zeros (Shanks', R=2)	est. zeros (inv. LPC, R=2)
769/222	-	-
1300/152	-	-
2570/166	-	-
3389/183	-	-
4050/899	-	-
0/102*	-	-
-	0/178*	0/3*
-	4704/568*	4704/132*
-	-	-

Note: Analysis interval = 30 ms (285 points), cepstral cutoff = 5.3 ms (50 points).

to be a zero near 2.6 kHz. This lies within the range of the second antiresonance for a nasalized vowel [27]. Normally a first antiresonance occurs near 600 Hz [27], but there is no evidence of it in Fig. 8(a). Both methods of homomorphic prediction detected a significant zero near 2700 Hz (see Table VI), while neither technique found an antiresonance below 1 kHz.

As this example illustrates, it is possible to obtain reasonable representations of speech spectra involving one significant antiresonance with 15 or 16 parameters. Note that several of the estimated zeros do not correspond to identifiable antiresonances of the original signal. Nevertheless, they are necessary for a good spectral match since they are part of a representation of the LPC error signal. For example, it is clear that the "extraneous" zero found by homomorphic prediction using inverse LPC at 3762 Hz improves the representation above 3 kHz. The LPC error spectrum is the deviation of the estimated spectral envelope, $V(e^{j\omega})$, from the all-pole spectrum $\tilde{F}(e^{j\omega})$. In addition to the actual vocal system antiresonances, it contains features associated with pole identification errors, equipment artifacts (e.g., dc zero), the low-time cepstral window (which determines the smoothness of the estimated spectrum), and the fact that the vocal tract is not a linear time-invariant rational filter. Thus it is generally necessary to include several more zeros then just those apparent in the original speech spectrum.

Example 7–Natural Nasal Consonant (|m|): Table VII and Fig. 9 show the results of the 12 pole/10 zero analysis of a 54 ms segment of intervocalic |m| from the sentence "Say mo<u>m</u>o again." As seen in Fig. 9(a), this signal exhibits two

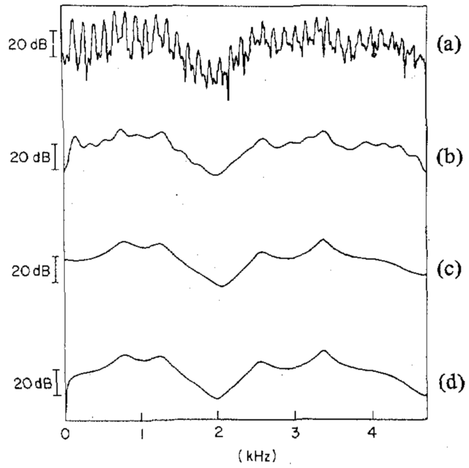

Fig. 7. Example 5: natural vowel plus artificial zero. (a) Log spectrum of preemphasized |a| from "five" with added zero. (b) Log spectral envelope obtained by homomorphic filtering. (c) 11 pole/4 zero model with zeros estimated by Shanks' method. (d) 11 pole/4 zero model with zeros estimated by inverse LPC.

TABLE VI
SUMMARY OF EXAMPLE 6—NATURAL NASALIZED VOWEL ("MOON")

est. poles	est. zeros	est. zeros
(LPC, P=10)	(Shanks', R=6)	(inv. LPC, R=6)
freq/bw (Hz)		
257/126	-	-
985/126	-	-
2391/214	-	-
3074/137	-	-
3959/1295	-	-
-	0/113*	0/27*
-	1691/603	-
-	2841/247	2673/156
-	-	3762/521
-	4704/-371*	-
-	-	0/796*

Note: Analysis interval = 30 ms (285 points), cepstral cutoff = 4.3 ms (40 points).

strong antiresonances, at about 650 Hz and 3.2 kHz. Homomorphic prediction using Shanks' method of locating zeros gave estimates of 695 Hz and 3220 Hz, while the inverse LPC technique found zeros at 596 Hz and 3180 Hz. These frequencies are comparable to those given by Fujimura [28], who reported that the first antiresonance for |m| usually falls between 700 Hz and 1200 Hz, while the second zero occurs near 3000 Hz.

V. CONCLUSIONS

Based on the preceding examples, homomorphic prediction, using either Shanks' method or inverse LPC to locate spectral zeros, appears to be useful for the pole/zero modeling of speech. Of the two techniques, homomorphic prediction with inverse LPC seems to provide better estimates of the antiresonance frequencies for the same number of coefficients. With either method a satisfactory representation usually includes several zeros in addition to the "true" antiresonances of the speech spectrum.

The basic idea of homomorphic prediction is to perform pole/zero analysis of the homomorphically estimated vocal-tract impulse response. In this paper only two particular identification methods were considered. Since homomorphic prediction avoids the problem of pitch-synchronization it is possible that some other previously rejected technique might be more effective. Specifically, there is a compelling reason to consider algorithms in which the poles and zeros are estimated simultaneously rather than in two separate

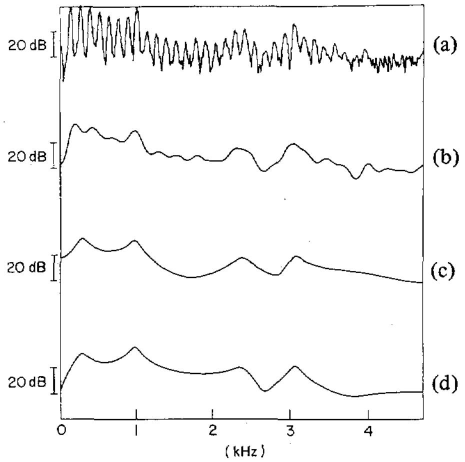

Fig. 8. Example 6: natural nasalized vowel. (a) Log spectrum of preemphasized |u| from "moon". (b) Log spectral envelope obtained by homomorphic filtering. (c) 10 pole/6 zero model with zeros estimated by Shanks' method. (d) 10 pole/6 zero model with zeros estimated by inverse LPC.

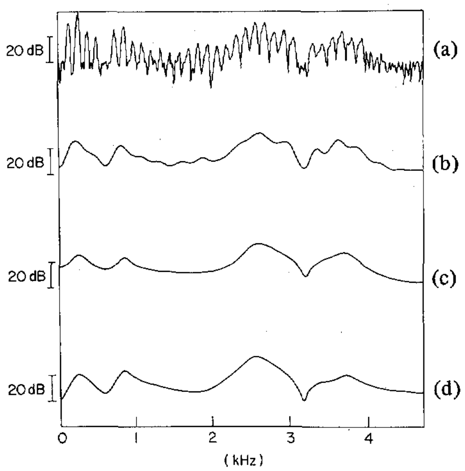

Fig. 9. Example 7: natural nasal consonant. (a) Log spectrum of preemphasized intervocalic |m| from "momo". (b) Log spectral envelope obtained by homomorphic filtering. (c) 12 pole/10 zero model with zeros estimated by Shanks' method. (d) 12 pole/10 zero model with zeros estimated by inverse LPC.

TABLE VII
SUMMARY OF EXAMPLE 7—ANALYSIS OF NATURAL NASAL CONSONANT (|m|)

est. poles (LPC, P=12) freq/bw (Hz)	est. zeros (Shanks', R=10)	est. zeros (inv. LPC, R=10)
233/162	-	-
834/142	-	-
2545/265	-	-
2666/583	-	-
3283/486	-	-
3738/256	-	-
-	0/3462*	-
-	0/160*	0/32*
-	695/353	596/161
-	2206/584	1901/494
-	3220/86	3180/51
-	-	3616/594
-	4309/1021	4704/879*

Note: Analysis interval = 54 ms (512 samples), cepstral cutoff = 4.3 ms (40 samples).

steps. In case of a nearly coincident pole-zero pair, the speech spectrum will exhibit neither a strong peak nor a sharp dip, even when the pole and zero are each narrow-band. Thus, if linear prediction is used to identify the poles, it is unlikely that the partially cancelled pole will be detected. As a result, the LPC error spectrum will be relatively featureless and the zero will be missed also. This may not be a problem in applications such as analysis/synthesis telephony where the goal is simply to reproduce the speech spectral envelope. However, in some situations, such as the acoustic analysis of nasalization, it is necessary to resolve interfering pole-zero pairs and a two-step approach would be inadequate.

A second limitation of the present study is that the experiments are based primarily on visual comparisons between the actual and model spectra of short speech segments. A more thorough investigation of homomorphic prediction would involve listening tests in which entire words or sentences are resynthesized from their pole/zero representation and evaluated.

Even within the context of homomorphic prediction defined in this paper, there are numerous unanswered questions. The algorithm is relatively complicated and many variations are possible. No serious attempt has been made to optimize any of the design decisions. For example, the low-time cepstral window is rectangular simply for convenience. It is possible that some other shape may lead to better pole-zero estimates because it smooths the speech spectrum differently.

In homomorphic prediction with inverse LPC, the zeros are obtained from the inverse error signal, $\{\tilde{e}^{-1}(n)\}$, rather than from $\{v^{-1}(n)\}$, to allow comparison with homomorphic prediction using Shanks' method. Whether this is the better approach in terms of identification accuracy remains to be determined. Furthermore, since $\tilde{E}^{-1}(z)$ is ideally all-pole, if the zeros are estimated from $\{\tilde{e}^{-1}(n)\}$ it may be better to use the autocorrelation formulation of LPC rather than the covariance method.

Finally, using the proper preemphasis strategy, it may be possible to reduce the number of coefficients required to represent any particular speech segment. For example, it was pointed out that there always seems to be a strong zero at dc. By first-differencing $\{\tilde{e}^{-1}(n)\}$, this can be removed when using homomorphic prediction with inverse LPC so that one less parameter is needed.

References

[1] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," *Proc. IEEE*, vol. 63, pp. 662-677, April 1975.

[2] J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum*, vol. 7, pp. 22-45, Oct. 1970.

[3] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442-449, Aug. 1969.

[4] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio and Electroacoustics*, vol. AU-20, Dec. 1972.

[5] B. Gold, "Note on buzz-hiss detection," *J. Acoust. Soc. Amer.*, vol. 36, pp. 1659-1661, 1964.

[6] J. I. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561-580, Apr. 1975.

[7] J. L. Shanks, "Recursion filters for digital processing," *Geophysics*, vol. 32, pp. 33-51, 1967.

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

[9] J. Durbin, "The fitting of time-series models," *Rev. Inst. Int. Statist.*, vol. 28, pp. 233-243, 1960.

[10] A. V. Oppenheim, G. E. Kopec, and J. Tribolet, "Signal analysis by homomorphic prediction," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-24, pp. 327-332, Aug. 1976.

[11] J. Tribolet, "Identification of linear discrete systems with applications to speech processing," M.S. thesis, Dep. Elec. Eng., Mass. Inst. of Tech., Cambridge, MA, Jan. 1974.

[12] F. D. Galiana, "Review of basic principles and of available techniques in system identification," Mass. Inst. of Tech., Cambridge, MA, 1969.

[13] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "Non-linear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264-1291, Aug. 1968.

[14] 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.

[15] A. V. Oppenheim, "A speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Amer.*, vol. 45, pp. 458-465, Feb. 1969.

[16] J. Makhoul and J. Wolf, "Linear prediction and the spectral analysis of speech," *BBN Rep. 2304*, Bolt, Beranek, and Newman, Inc., Boston, MA, Aug. 31, 1972.

[17] L. C. Wood and S. Treitel, "Seismic signal processing," *Proc. IEEE*, vol. 63, pp. 649-661, Apr. 1975.

[18] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, ch. 10, 1975.

[19] R. W. Schafer, "Echo removal by discrete generalized linear filtering," Mass. Inst. of Tech. Research Lab. of Electron., Cambridge, MA, *Tech. Rep. 466*, Feb. 1969.

[20] J. L. Goldstein, "Auditory spectral filtering and monaural phase perception," *J. Acoust. Soc. Amer.*, vol. 41, pp. 458-479, 1967.

[21] R. E. Kalman, "Design of a self-optimizing control system," *Trans. ASME*, pp. 468-478, Feb. 1958.

[22] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50 (part 2), pp. 637-655, 1971.

[23] V. N. Faddeeva, *Computational Methods of Linear Algebra* (Transl. by C. D. Benster). New York: Dover, 1959.

[24] E. M. Hofstetter, "An introduction to the mathematics of linear predictive filtering as applied to speech analysis and synthesis," Mass. Inst. of Tech. Lincoln Lab., Lexington, MA, *Tech. Note 1973-36*, Rev. 1, Apr. 12, 1974.

[25] J. D. Markel and A. H. Gray, "A linear prediction vocoder simulation based on the autocorrelation method," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-22, pp. 124-134, Apr. 1974.

[26] D. H. Klatt, "Acoustic theory of terminal analog speech synthesis," in *Conf. Rec., 1972 Conf. on Speech Communication and Processing*, IEEE-AFCRL, Newton, MA, IEEE Cat. 72 CHO 596-7 AE, Apr. 24-26, 1972.

[27] O. Fujimura, "Spectra of nasalized vowels," *Quarterly Progress Report*, M.I.T. Res. Lab. of Electron., Cambridge, MA, no. 58, pp. 214-218, July 1960.

[28] O. Fujimura, "Analysis of nasal consonants," *J. Acoust. Soc. Amer.*, vol. 34, pp. 1865-1875, Dec. 1962.

[29] G. Kopec, "Speech analysis by homomorphic prediction," *S.M. thesis*, Dep. Elec. Eng. and Comput. Sci., Mass. Inst. of Tech., Cambridge, MA, Feb. 1976.

[30] J. Makhoul, "Spectral linear prediction: properties and applications," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-23, pp. 283-296, June 1975.

[31] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, June 1968.

[32] C. J. Weinstein and A. V. Oppenheim, "Predictive coding in a homomorphic vocoder," *IEEE Trans. Audio and Electroacoustics*, Sept. 1971.