Excitation-Synchronous Modeling of Voiced Speech

S. PARTHASARATHY, MEMBER, IEEE, AND DONALD W. TUFTS, FELLOW, IEEE

Abstract—**A new modeling technique for voiced speech is introduced. Salient features are detailed modeling of speech waveforms and the use of improved parameter estimation techniques. The ideas of pitch-synchronous analysis are extended to make two subintervals synchronous with regions of approximately closed and approximately open glottis. Two LPC models are used in each pitch period, and the model parameters are changed at estimated times of transition from open-to-closed and closed-to-open glottis. The excitation is provided by changing initial conditions at these transition instants. Experiments with real, connected speech indicate that the speech waveforms can be accurately represented using the analysis–synthesis approach presented here.**

I. INTRODUCTION

A simple and effective model for speech production is the linear source-system model [1]. In this model, voiced speech is represented as the response of the vocal tract to a sequence of glottal pulses. Many parameterizations for the glottal source and the vocal tract system have been considered [1] in an attempt to represent the important features of speech or to synthesize speech waveforms with a small number of parameters.

The decomposition of the speech signal into a source and a system appears to be a major problem in source-system modeling. Some speech sounds, such as steady vowels, can be reproduced with good quality even with a somewhat idealized glottal source consisting of pulses lasting a fraction of a pitch period and zero excitation for the rest of the pitch period [2]. In such cases, the glottis is assumed to be completely closed for a part of a pitch period, and the speech signal in this interval can be considered to be the zero-input response of the vocal tract system. The vocal tract parameters can then be obtained by analyzing the closed glottis periods [3], [4]. This approach, however, has serious drawbacks. In many instances, the glottis is never completely closed or is closed only for very short durations [25] and, therefore, parameter estimation is difficult. The use of improved linear prediction [5], [7] and maximum-likelihood [8], [9] techniques can alleviate the latter problem to some extent.

The properties of the vocal tract change even within a pitch period because of the opening and closing of the glottis [10], [11]. Coupling between the subglottal cavities and the upper vocal tract, and its effect on the vocal tract resonances, has been considered in some studies and there is some evidence that this may be important in good quality speech synthesis [10], [12]. These effects have been ignored in most conventional speech analysis methods. The model investigated in this study attempts to obtain a first-order approximation of these variations in the vocal tract characteristics within a pitch period.

Characterization of the glottal excitation has also proved to be a difficult problem. The vocal tract parameter estimates from an analysis of the closed glottis period is often used to estimate the glottal excitation by inverse filtering the speech signal [13]. Attempts at parameterizing the glottal excitation have not, in general, been very successful. It has also been found that the results can be misleading and require careful interpretation [13]. The conventional linear predictive coding (LPC) techniques [14], [15] model voiced speech signals as the output of an all-pole system to a quasi-periodic train of impulses, the period chosen to be the pitch period. This model assumes that there is a single instant of excitation in a pitch period. This assumption is valid in a very approximate way since the major excitation of the vocal tract occurs at glottal closure. The improvements provided by pitch-synchronous analysis methods [16] demonstrated this fact. The LPC excitation model is, however, too idealized for accurate speech representation.

An approach that circumvents the problem of explicit modeling of the glottal pulse is the multipulse excitation LPC model [17]. In this approach, the excitation is modeled as a sequence of impulses whose amplitudes and positions are determined such that the speech waveform is approximated as closely as possible in the least squares sense. The success of this model appears to indicate the merits of the waveform approximation approach. There are, however, some drawbacks in this model. It has been found that the pulse sequence does not have a close relationship to the glottal pulse shape and that the periodicity of the speech waveform is not reflected in the excitation sequence. In this paper, the system and excitation models are chosen to correspond more closely to the speech production process than in the multipulse LPC model, while using the waveform approximation error as the criteria for parameter estimation.

The knowledge about the features of the glottal source that has been acquired from the experimental studies by a number of researchers can be used in efficient modeling of the source. It has been observed that the vocal tract resonances are excited mainly at glottal closure [18], [19]. Secondary excitation has been observed at glottal opening although less significant excitation persists throughout the

Manuscript received November 30, 1985; revised July 12, 1986. This work was supported in part by the Office of Naval Research and by the National Science Foundation.

S. Parthasarathy was with the Department of Electrical Engineering, University of Rhode Island, Kingston, RI. He is now with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

D. W. Tufts is with the Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881.

IEEE Log Number 8715561.

0096-3518/87/0900-1241$01.00 © 1987 IEEE

pitch period [20]. It seems reasonable that each pitch period be represented by two steady states separated by short intervals during which changes take place at the glottis. The excitation regions are then assumed to be restricted to short durations around glottal closure and opening. This is, in a way, an extreme case of the method that uses multiple pulses and initial condition changes to represent the speech waveform [24]. By using initial condition changes alone at appropriate instants, the excitation can be modeled to conform more closely to the speech production process.

Fig. 1 shows the results of an experiment conducted to demonstrate the properties of the speech waveform discussed so far. The tracks of the formant frequencies obtained from real speech by an improved covariance analysis using a short, sliding rectangular window, the corresponding segment of the speech waveform, and the approximation error are shown. The formant frequencies and the approximation error at time instant t are obtained by an analysis of a 3.5 ms segment of the speech waveform beginning at time t. The tracks illustrate the changes in the estimated formant frequencies within a pitch period. The blank regions in the formant tracks indicate that reasonable estimates of the formant poles could not be obtained from the data in the corresponding analysis interval. It can be seen that when the analysis interval straddles two pitch periods, the covariance method does not yield reasonable pole estimates. This is to be expected since the speech waveform in this interval cannot be represented as a single segment that is approximately the zero-input response of an all-pole system. A more interesting observation is that a speech segment approximately centered in the pitch period does not yield good estimates of the formant poles, particularly the second formant in this example. The normalized waveform approximation error also shows periodic variation with large errors in regions marked $R2$ (corresponding to analysis intervals that include excitation regions) and small errors in regions marked $R1$ (corresponding to analysis intervals where the data can be approximately represented as the zero-input response of an all-pole system). This seems to indicate the presence of a secondary excitation within a pitch period, and this result appears to corroborate experiments by other researchers. The excitation at approximately the instants marked C and O are attributed to the glottal closure and opening, respectively. All the formant frequency estimates show perceptible change as the analysis interval is shifted from the beginning to the end of a pitch period. These changes are attributed to the interaction between the subglottal cavities and the upper vocal tract.

A speech representation that incorporates the primary properties of the speech signal, and analysis techniques that provide solutions to some of the problems mentioned in the preceding discussion, are introduced in the following sections. Examples of real speech analysis and synthesis are presented to illustrate the performance of the new modeling technique.

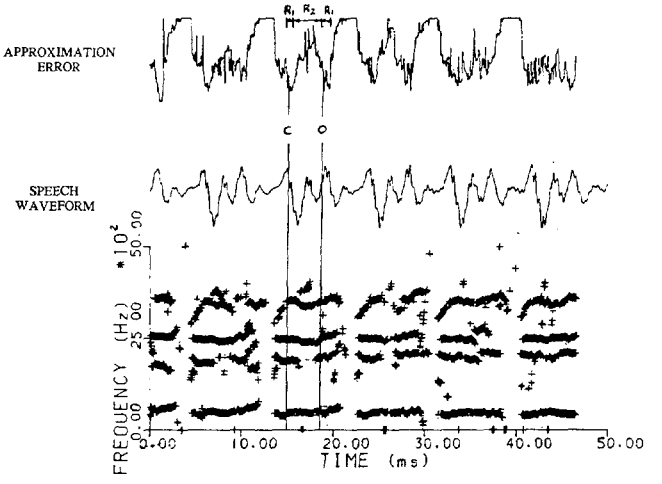

Fig. 1. Sliding window covariance analysis.

II. The Excitation-Synchronous Representation

The new speech representation is based on the following hypotheses.

1) The vocal tract is significantly excited only during short intervals around glottal closure and opening.

2) The interaction between the glottal source and the vocal tract can be effectively modeled by varying the vocal tract parameters at glottal closure and opening.

The vocal tract characteristics are represented by two steady states in each pitch period; one during an interval when the glottis is predominantly closed, and another when the glottis is almost open. Since the glottal closure and opening are transient phenomena lasting a finite duration, the steady-state regions are separated by short (compared to a pitch period) transition intervals. The speech waveform is, therefore, represented by concatenating zero-input responses of all-pole systems with distinct poles, these response segments being synchronous with closed and open glottis periods and separated by transition regions. Since the speech segments are modeled synchronously with the short duration excitation, this representation is called the excitation-synchronous representation. The system representing the vocal tract is chosen to be all-pole with linear prediction coefficients $\{a_k\}_{k=0}^{P}$ or poles $\{z_k\}_{k=1}^{P}$. The parameters of the all-pole system change abruptly at glottal closure $\{\tau_{ci}\}$ and opening $\{\tau_{oi}\}$. The transition interval width takes on the following values:

$$W = \begin{cases} W_c, & \text{for the open-to-closed transition of the glottis} \\ W_o, & \text{for the closed-to-open transition of the glottis.} \end{cases}$$

Subscripts c and o are used to indicate the parameters in the closed and open glottis regions, respectively. In each interval (labeled i) between glottal closures, the speech waveform in the closed glottis intervals $[\tau_{ci} : \tau_{oi} - W_o]$, and the open glottis intervals $[\tau_{oi} : \tau_{ci} - W_c]$, can be represented as a linear combination of exponentially damped sinusoids. The initial conditions that determine the zero-input responses are allowed to change at time instants

$\{\tau_{ci}\}$ and $\{\tau_{oi}\}$. There are a number of ways in which the transitions could be made from one set of model parameters to another. The initial conditions could be changed abruptly or be induced by an impulse-stream exciting the system during the transition region. In addition, the excitation pulse during the transition region could be parametrically modeled. Since the major emphasis of this paper is to study the basic excitation-synchronous analysis approach, a simple model for the changes in the initial conditions and the signal in the transition region will be assumed. The P samples preceding the start of a zero-input response segment of a system with P poles will be assumed to be noisy observations of the initial conditions for that segment. The transition region width of less than typical choices of P has been found to be adequate by experimentation. Therefore, the signal in the transition region can also be represented as a sum of exponentially damped sinusoids.

Modeling speech waveforms by concatenating exponential segments has been considered before [21]. The novelty in the excitation-synchronous approach is the use of two segments in a pitch period and the introduction of the transition regions. Further, noise-resistant parameter estimation techniques that have been introduced recently [5]–[9] are used to estimate the speech parameters more accurately than in the earlier experiments. Since the speech waveform is segmented so as to minimize the effects of glottal excitation in the determination of the vocal tract system, the order of the system is chosen to be equal to twice the number of formants that is expected. In some cases, additional poles may be required, the reasons for which are discussed later. An important feature is that extraneous poles that are usually required to model the combined effects of the glottal source and the vocal tract (and not required to adequately characterize the vocal tract alone) are avoided. The next section focuses on the estimation of the transition instants and the poles of the system in the closed and open glottis periods. The transition widths, W_c and W_o, are determined by experimentation and are kept fixed.

III. Estimation of System Poles and Transition Instants

The excitation-synchronous model is a detailed model of the speech signal that preserves changes in the vocal tract parameters not only over adjacent pitch periods but also within a pitch period. Therefore, a coarse estimate of the formants, such as those obtained by conventional LPC analysis using data from multiple pitch periods, is unacceptable. Simultaneous estimation of the system parameters and the transition instants is a computationally tedious procedure. The approach followed here leads to a suboptimal four-step algorithm. The first two steps consist of obtaining approximate estimates of the speech parameters, and the following steps refine these estimates. The details of the algorithms are presented in the rest of this section.

a) Approximate Estimation of the Instant of Glottal Closure: The first formant is excited mainly at glottal closure [15]. Define the quantity denoted by f_I to be the energy in a frequency band containing the first formant. The value of f_I estimated using a short (about half a pitch period) sliding window of speech is likely to peak when the speech segment is chosen to be near glottal closure [22]. This fact is used in obtaining the approximate estimate of glottal closure. The algorithm for approximate estimation of glottal closure is given below.

Algorithm GCDFT: Compute the N-point discrete Fourier transform (DFT) of a Hamming windowed, zero-padded segment of duration less than one pitch period of the speech signal (typically about 3–4 ms for a male speaker).

$$D_t(m) = \frac{1}{N}\sum_{k=0}^{N_1-1} s(t+k)\, e^{-j\omega_m k}; \qquad m = 1, 2, \cdots, N \tag{1a}$$

where $\{s(k)\}_{k=t}^{t+N_1}$ is a segment of the speech waveform, t denotes the beginning of the analysis interval, and N is typically chosen to be 128. Find the energy in the frequency band containing the first formant—the frequency range is chosen to be 300 Hz (corresponding to bin m_1) to 1000 Hz (corresponding to bin m_2) since the first formant usually lies in this range.

$$e'(t) = \sum_{k=m_1}^{m_2} \left|D_t(m)\right|^2. \tag{1b}$$

Scale $e'(t)$ by the short time energy (averaged over 20 ms) of the speech signal. The scaled $e'(t)$ is referred to as f_I.

$$f_I(t) = \frac{e'(t)}{\sum_{k=t}^{t+t_1} s^2(k)} \tag{1c}$$

where $t_1 = 20 \times 10^{-3} \times f_s$; f_s is the sampling frequency. Evaluate $e(t)$ for the range of t over which the instant of glottal closures are to be determined. This waveform exhibits quasi-periodic peaks and a simple algorithm is used to pick the appropriate peaks of this waveform. Scaling by the short time energy ensures that the peaks in the low amplitude regions of the speech waveform are not obscured. The locations of these peaks indicate glottal closure.

The plot of $\{f_I(t)\}$ computed for the speech waveform in Fig. 3(a) is shown in Fig. 3(c). It can be seen that the peaks in $f_I(t)$ correspond approximately to the visual estimate of the instant of maximum excitation of the formants.

b) Approximate Estimation of the System Poles in the Closed and Open Glottis Periods: The previous step of the algorithm yielded approximate estimates of glottal closure. For the purposes of this paper, the instant of glottal closure is assumed to be the beginning of the pitch period and the period between two glottal closures is cho-

sen to be the pitch period. The glottis is closed or nearly closed for a fraction (generally 0.2–0.5) of the pitch period. Therefore, approximate estimates of the system poles in the closed and open glottis regions can be obtained by restricting the analysis intervals near the beginning or the end of the pitch period as indicated in Fig. 2. Once the approximate closed or the open glottis periods have been determined, the problem becomes one of estimating the poles of an all-pole system from observations (perhaps noisy) of its zero-input response. Conventional covariance methods can be used to obtain estimates of pole locations, but the performance of these methods deteriorates rapidly in the presence of noise. In addition, the number of data samples in the open or closed glottis periods is usually small and this makes accurate estimation of the parameters difficult. The problem of estimating the exponential parameters from short, noisy observations has received a lot of attention recently [5]–[9]. The algorithm outlined below uses the idea of increasing the predictor length beyond the minimum required (i.e., the true order of the system) [5], [7] and also singular value decomposition (SVD) based solutions [5], to obtain accurate estimates with reduced variance.

Algorithm POLE-LP: The observations $\{y(t)\}$ in the closed or open glottis intervals can be written as

$$y(t) = \sum_{k=1}^{P} c(k) z^t(k) + n(t); \qquad t = 0, 2, \cdots, N-1 \tag{2}$$

where N is the number of data points in the interval, P is the order of the system, $\{c_k\}$ are complex amplitudes of the corresponding exponential, and $\{n(t)\}$ are perturbations assumed to be white, Gaussian noise in this case. Set up a system of linear prediction equations (backward or forward) [5].

$$\begin{bmatrix} y(1) & y(2) & \cdots & y(L) \\ y(2) & y(3) & \cdots & y(L+1) \\ \vdots & \vdots & & \vdots \\ y(N-L) & y(N-2) & \cdots & y(N-1) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \vdots \\ a(L) \end{bmatrix} = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(N-L-1) \end{bmatrix}. \tag{3}$$

$a(k)$ are the predictor coefficients. The predictor length L is chosen to be larger than the true system order P. The least-squares solution of the linear prediction vector $\boldsymbol{a}$ is obtained as [5]

$$\boldsymbol{a} = \sum_{k=1}^{R} \frac{1}{\sigma(k)} \boldsymbol{u}_k^H \boldsymbol{y} \boldsymbol{v}_k \tag{4}$$

where $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are left and right singular vectors, and σ_k are the singular values in the SVD of the data matrix $\boldsymbol{Y}$,

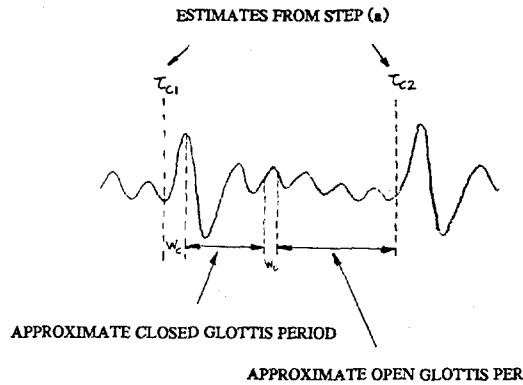

Fig. 2. Choice of analysis intervals for approximate estimation of system poles [step b)].

and R specifies the desired rank in the low-rank approximation. The roots of the polynomial

$$A(z^{-1}) = 1 + a(1) z^{-1} + \cdots + a(L) z^{-L}$$

are denoted by $\tilde{z}(k)$. The linear prediction pole estimates obtained by the Lth-order analysis are then

$$z(k) = \frac{1}{\tilde{z}^*(k)}$$

where * denotes complex conjugate.

The P system poles can be chosen out of $\{z(k)\}_{k=1}^{L}$ by a subset selection procedure [6]. In the case of speech signals where the poles are not too closely spaced in frequency, the simpler algorithm given below suffices.

Algorithm SELECT-POLES: Select the L_1 poles whose radii lie between ρ_{in} and ρ_{out}. Compute the amplitudes $\{c(k)\}$ in (2). The least squares solution for the complex amplitude vector $\boldsymbol{c} = (c(1)\, c(2) \cdots c(L_1))^H$ can be obtained by solving the set of linear equations

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ z(1) & z(2) & \cdots & z(L_1) \\ z^2(1) & z^2(2) & \cdots & z^2(L_1) \\ \vdots & \vdots & \vdots & \vdots \\ z^{N-1}(1) & z^{N-1}(2) & \cdots & z^{N-1}(L_1) \end{bmatrix} \begin{bmatrix} c(1) \\ c(2) \\ \vdots \\ c(L_1) \end{bmatrix} = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(N-1) \end{bmatrix}. \tag{5}$$

Compute the energy contribution of each real pole or pole pair in the signal reconstruction as

$$E_i = \begin{cases} c^2(i) \left(\dfrac{1 - z(i)^{2N}}{1 - z(i)^2} \right), & \text{for real } z(i), c(i); \\ 2|c(i)|^2 \left(\dfrac{1 - |z(i)|^{2N}}{1 - |z(i)|^2} \right) \\ \quad + 2\Re\left(c(i) \dfrac{1 - z(i)^N}{1 - z(i)} \right), & \text{for complex pairs} \\ \quad (z(i), z^*(i))), (c(i), c^*(i))). \end{cases} \tag{6}$$

where $\Re(\cdot)$ denotes the real part of $(\cdot)$.

Choose the system poles to be the P poles that contribute the largest energies in the reconstruction.

Choices of various parameters such as the predictor order and the rank in forming the solution given in (6) are determined based on the properties of the speech signal. These considerations are discussed in Section IV.

c) Least-Squares Estimation of the Transition Instants: This step of the speech analysis procedure depends on the appropriate choice of the duration and location of the analysis interval. The analysis interval is chosen to be smaller than a pitch period and positioned such that it contains only one transition. The transition instant estimation problem can be stated as follows.

Segment a given interval of the speech signal into two homogeneous regions (approximately satisfying a homogeneous difference equation) separated by an excitation region of width W to minimize the square reconstruction error. The observations are written as

$$y(t) = \begin{cases} y_1(t); & 0 \le t \le \tau - W \\ y_3(t - \tau + W + 1); & \tau - W < t < \tau \\ y_2(t - \tau); & \tau \le t \le N - 1 \end{cases} \quad (7a)$$

where

$$y_1(t) = \sum_{k=1}^{P_1} c_1(k) z_1^t(k) + n_1(t)$$

$$y_2(t) = \sum_{k=1}^{P_2} c_2(k) z_2^t(k) + n_2(t). \quad (7b)$$

$y_3(t)$ is the signal in the transition region. In the speech examples in Section IV,

$$y_3(t) = \sum_{k=1}^{P_2} c_2(k) z_2^{-t}(k) + n_3(t). \quad (7c)$$

The negative index for t has been chosen only to indicate that the signal values in the transition region are in fact the initial conditions for the following segment, which in this case is an extrapolation of the signal $y_3(t)$ for negative values of the time index. Therefore, the data in each segment can be written in the following form:

$$y_i = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_i(1) & z_i(2) & \cdots & z_i(P_i) \\ z_i^2(2) & z_i^2(2) & \cdots & z_i^2(P_i) \\ \vdots & \vdots & \vdots & \vdots \\ z_i^{m-1}(1) & z_i^{m-1}(2) & \cdots & z_i^{m-1}(P_i) \end{bmatrix} \begin{bmatrix} c_i(1) \\ c_i(2) \\ \vdots \\ c_i(P_i) \end{bmatrix} + \begin{bmatrix} n_i(0) \\ n_i(1) \\ \vdots \\ n_i(m-1) \end{bmatrix} \quad (8)$$

where the dimensions of the matrices depend on the transition time τ. The approximation of y_i is given by

$$\hat{y} = P_{Z_i} y_i; \quad P_{Z_i} = Z_i (Z_i^H Z_i)^{-1} Z_i^H y_i. \quad (9)$$

Z_i now depends on τ. The problem is to find the value of τ that best segments the observations into two zero-input responses and a transition region of width W such that the norm $\Sigma_{i=1}^{3} \| y_i - \hat{y}_i \|^2$ is minimized. The pole locations are assumed to be known.

Algorithm TRANS: Evaluate

$$f(\hat{\tau}) = \sum_{i=1}^{3} \| y_i - P_{Z_i} y_i \|^2 \quad (10)$$

for $\hat{\tau} \in \{P_1, P_1 + 1, \cdots, N - P_2\}$ and pick τ_{LS} to be the value of τ that corresponds to the minimum of $f(\hat{\tau})$.

Efficient, recursive methods for computing $f(\hat{\tau})$ are given in the Appendix. To estimate the instant of glottal closure and opening, proceed as follows.

Algorithm GCGO: Select a speech segment of duration less than a pitch period (typically about 0.7 times pitch period) approximately centered around the estimate of glottal closure obtained in step a). Using the approximate estimates of the poles in the open and closed glottis regions obtained in step b), estimate the transition instant using algorithm TRANS. This yields an estimate of the instant of open-to-closed transition of the glottis.

An estimate of the instant of closed-to-open transition of the glottis can be obtained using the same procedure as above. The analysis interval in this case starts at the estimate of glottal closure obtained above.

d) Accurate Estimation of the System Poles in the Closed and Open Glottis Regions: Step c) provided estimates of glottal closure and opening and hence the closed glottis and open glottis periods. Using the pole estimation algorithm of step b) and analyzing the signal in the closed and open glottis periods alone, the corresponding pole locations can be obtained. Iterative maximum-likelihood techniques can be used to improve upon the linear prediction estimates [8]. The least-squares or maximum-likelihood (if the additive noise is assumed to be Gaussian) estimates of the poles can be obtained by maximizing with respect to $\{z_k\}_{k=1}^{P}$, where P_Z has the same definition as in (9). An efficient algorithm developed by Golub and Pereyra [23] based on a quasi-Newton method is used. To ensure that the pole estimates lie in a ring enclosed by circles with radii ρ_{in} and ρ_{out} (usually chosen to be 1), the optimization is performed with the modified variable α where

$$r = \rho_{in} + (\rho_{out} - \rho_{in}) \sin^2(\alpha)$$

instead of the radius r. If the initial linear prediction estimates yielded poles outside the unit circle, these poles are reflected inside the unit circle before the maximum-likelihood iterations are begun.

IV. Examples of Speech Analysis and Synthesis

Voiced segments extracted from continuous speech is analyzed using the algorithm presented in Section III. The intermediate results of each step of the algorithm are presented to provide intuitive feel for the performance of the algorithm. Parameters such as the predictor length in steps b) and d) of the algorithm, the analysis intervals, and the transition region width are chosen based on experimentation and are somewhat arbitrary. Noisy speech is ana-

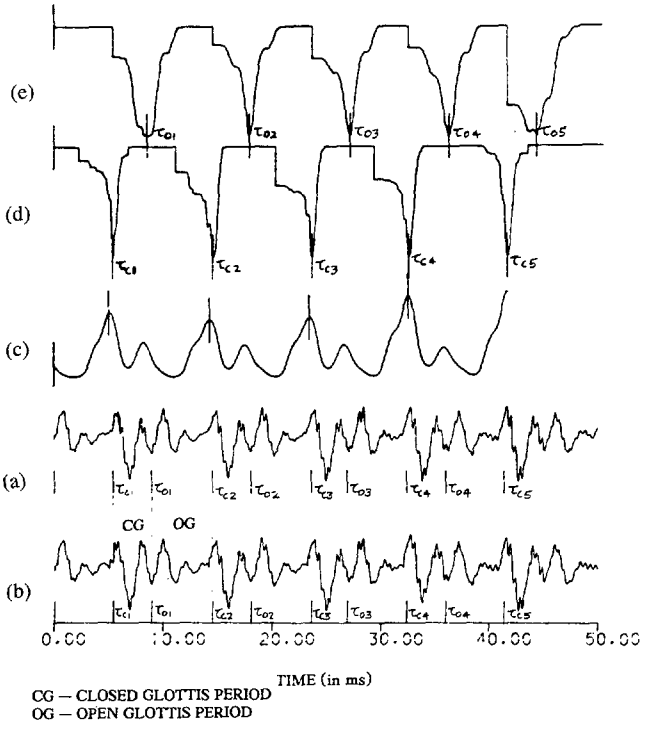

Fig. 3. Excitation-synchronous analysis and reconstruction of voiced speech. (a) Speech waveform. (b) Reconstruction. (c) Approximate glottal closure (f_l). (d) Estimation of glottal closure. (e) Estimation of glottal opening.

lyzed using the same algorithm and the results are presented (Fig. 4). A 50 ms segment of voiced speech sampled at 20 kKHz is shown in Fig. 3(a). A synchronized plot of $f_l(t)$ computed using algorithm GCDFT is shown in Fig. 3(c). The analysis interval for computing each value of $f_l(t)$ is chosen to be 3.5 ms corresponding to 70 samples. Although there are peaks at times other than the instant of glottal closure, it is obvious from Fig. 3(c) that the appropriate peaks can be selected. A simple algorithm that selects peaks locally and also looks for approximate long time periodicity is used to perform the peak picking and peak selection tasks. Approximate estimates of the poles in the open and closed glottis periods are obtained using algorithm POLE–LP. The linear predictor length L is chosen to be around 30 corresponding to a time duration of 1.5 ms. The system order is chosen to be 10. Four pole pairs (8 poles) corresponding to the first four formants are estimated. An additional pole pair around pitch frequency is often estimated. This accouns for the source periodicity and is essential for obtaining a low approximation error. In some applications such as formant estimation, a preemphasized speech waveform is used. In such cases, the low-frequency pole is not estimated and the system order can be reduced to 8. The rank R is chosen to be 14. This is because the speech waveform is strictly not a rank 10 signal. Using a rank lower than about 14 provides poor estimates of the formant bandwidths. The approximate pole estimates obtained in step b) are in most cases good estimates. Therefore, for the purposes of transition region estimation, it can be assumed that the locations of the system poles in the closed and open glottis poles are known. The contribution to the approximation error [$f(\tau)$ in (10)] due to errors in the pole locations are inherently different from those due to errors in the location of the excitation. In practice, small errors in the pole locations do not seem to significantly affect the transition estimates. The function $f(\tau)$ evaluated for estimating each transition is normalized and concatenated for convenient plotting. The waveforms in Fig. 3(d) and (e) show the error functions where the location of the minima indicate the instants of glottal closure or opening. The transition widths for glottal closure and opening are assumed to be 0.35 ms (7 samples) and 0.15 ms (3 samples), respectively. If the maximum-to-minimum value ratio of the approximation error function is below a certain threshold, it is assumed that a transition does not exist in the analysis interval. In such cases, the analysis interval is shifted forward or backward until a transition is estimated. The glottal opening and closure are marked in Fig. 3. The algorithm described in step d) is used to estimate the poles in the closed and open glottis regions. The iteration limit is set to 10 although the first few iterations appear to yield most of the improvement. The speech waveform in the two regions can be reconstructed as in (7)–(9). The signal in the transition region is reconstructed by extrapolating the exponential waveform of the following segment backwards. Better models for the transition region are being investigated. The reconstructed speech is shown in Fig. 3(b). It can be seen that there is a close match between the original speech

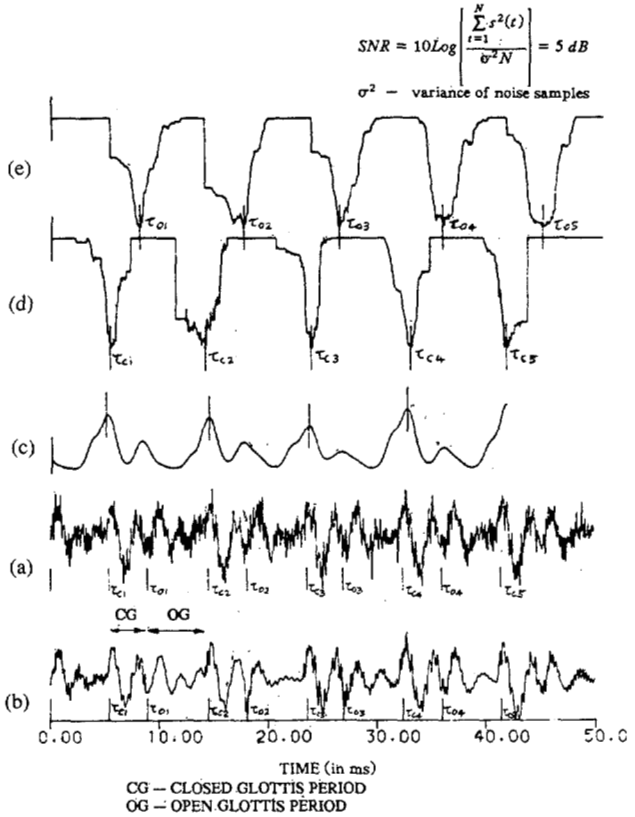

Fig. 4. Excitation-synchronous analysis of noisy speech. (a) Noisy speech. (b) Reconstruction. (c) Approximate glottal closure (f_i). (d) Estimation of glottal closure. (e) Estimation of glottal opening.

waveform and the reconstruction. The voiced segments of the sentence "A lathe is a big tool" were analyzed using the procedure outlined above, and the results were similar to the results presented for the 50 ms segment. Finally, noisy speech with a signal-to-noise ratio of 5 dB was analyzed, and the results of each step of the algorithm are indicated in Fig. 4. The reconstruction looks much less noisy than the original. However, in high noise situations some detail may have to be sacrificed for robustness. For example, the pole estimates in each closed or open period could be estimated using data from the corresponding intervals in adjacent pitch periods. This means that the resolution of the tracks of the system poles will no longer be one pitch period, but it is often acceptable since such averaging results in improved noise resistance.

Some general conclusions regarding noisy speech processing are: i) the DFT based coarse glottal closure estimator is extremely resistant to noise; ii) the improved linear production techniques, particularly the low-rank approximations [5], and the maximum-likelihood methods contribute to better estimates of the system poles, but below about 5 dB SNR some kind of averaging as suggested above will be required; and iii) the least squares transition estimator is noise-resistant provided the initial pole estimates are good.

V. Conclusions

A new speech modeling technique that incorporates the primary features of the acoustic speech production process, and also noise-resistant parameter estimation techniques, is introduced. The focus of this paper is a study of the applicability of the excitation-synchronous model and the improved parameter estimation techniques. It is seen that short segments of voiced speech, if appropriately chosen, can be represented to a good approximation as a linear combination of exponentials. The speech parameters can be estimated accurately using the improved pole and transition estimation techniques. Experiments on real, connected speech indicate that speech waveforms can be accurately represented using this model. The results of the analysis of noisy speech are encouraging. Algorithms for the estimation of the various parameters are presented. Potential applications in speech coding and recognition are the subjects of future research.

Appendix

The least-squares error function can be written as the sum of the reconstruction error norms for the individual segments as in (10). The first error term e_1 for a particular value of $\hat{\tau} = k$ can be written as

$$e_1(k) = \boldsymbol{d}_k^H \boldsymbol{R}_k^{-1} \boldsymbol{d}_k \qquad \text{(A.1)}$$

where

$$\boldsymbol{d}_k = \boldsymbol{Z}_1^H(k)\, \boldsymbol{y}_1(k)$$

$$\boldsymbol{R}_k = \boldsymbol{Z}_1^H(k)\, \boldsymbol{Z}_1(k).$$

Observe that the number of rows in $\boldsymbol{Z}_1(\hat{\tau})$ increases by

one for every increment in $\hat{\tau}$. That is,

$$\boldsymbol{Z}_1(k+1) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1(1) & z_1(2) & \cdots & z_1(P_1) \\ z_1^2(1) & z_1^2(2) & \cdots & z_1^2(P_1) \\ \vdots & \vdots & & \vdots \\ z_1^{k-1}(1) & z_1^{k-1}(2) & \cdots & z_1^{k-1}(P_1) \\ z_1^k(1) & z_1^k(2) & \cdots & z_1^k(P_1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_1(k) \\ \boldsymbol{g}_k \end{bmatrix} \tag{A.2}$$

and

$$\boldsymbol{R}_{k+1} = \boldsymbol{Z}_1^H(k)\,\boldsymbol{Z}_1(k) + \boldsymbol{g}_k\boldsymbol{g}_k^H. \tag{A.3}$$

Using (A.2), (A.3), and the matrix inversion lemma [26], the recursions for $\boldsymbol{d}_k$ and $\boldsymbol{R}_k^{-1}$ are

$$\begin{aligned} \boldsymbol{d}_{k+1} &= \boldsymbol{d}_k + y(k)\,\boldsymbol{g}_k \\ \boldsymbol{R}_{k+1}^{-1} &= \boldsymbol{R}_k^{-1} - \frac{\boldsymbol{R}_k^{-1}\boldsymbol{g}_k\boldsymbol{g}_k^H\boldsymbol{R}_k^{-1}}{1 + \boldsymbol{g}_k^H\boldsymbol{R}_k^{-1}\boldsymbol{g}_k}. \end{aligned} \tag{A.4}$$

The following recursions are the result of using (A.2) and (A.4) and straightforward algebra.

$$\begin{aligned} e_1(k+1) &= e_1(k) - \frac{|K_2|^2}{1+K_1} \\ &\quad + 2y(k+1)\frac{\Re K_2}{1+K_1} + y^2(k+1)\frac{K_1}{1+K_1} \\ \boldsymbol{R}_{k+1}^{-1} &= \boldsymbol{R}_k^{-1} - \frac{\boldsymbol{w}(k+1)\,\boldsymbol{w}^H(k+1)}{1+K_1} \end{aligned} \tag{A.5}$$

where

$$\begin{aligned} \boldsymbol{w}(k+1) &= \boldsymbol{R}_k^{-1}\boldsymbol{g}_{k+1} \\ \boldsymbol{u}(k+1) &= \boldsymbol{R}_k^{-1}\boldsymbol{d}_k \\ K_1 &= \boldsymbol{g}_{k+1}^H\boldsymbol{w}_{k+1} \\ K_2 &= \boldsymbol{g}_{k+1}^H\boldsymbol{u}_{k+1}. \end{aligned}$$

A similar recursion is applicable for the evaluation of the error function for the second segment.

This method still requires the inversion of a $P_i \times P_i$ matrix to initiate the recursions. This overhead might be unacceptable if the number of observations is small. The method developed below is useful in such cases.

The error terms of (10) can be rewritten in terms of the linear prediction coefficients $a(k)$. This leads to efficient recursions for the error functions that are more efficient in some cases than the recursions given above. Details of the derivations follow.

The data vector $\boldsymbol{y}_1$ of length $k \geq P_1$, generated as in (8), can be represented as a linear combination of k linearly independent vectors, P_1 of them forming a basis for the signal subspace, and $k - P_1$ vectors defining the noise subspace.

$$\begin{aligned} \boldsymbol{y}_1 &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1(1) & z_1(2) & \cdots & z_1(P_1) \\ z_1^2(1) & z_1^2(2) & \cdots & z_1^2(P_1) \\ \vdots & \vdots & & \vdots \\ z_1^k(1) & z_1^k(2) & \cdots & z_1^k(P_1) \end{bmatrix} \begin{bmatrix} c_1(1) \\ c_1(2) \\ \vdots \\ c_1(P_1) \end{bmatrix} \\ &\quad + \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{k-P_1} \end{bmatrix} \begin{bmatrix} c'(1) \\ c'(2) \\ \vdots \\ c'(k-P_1) \end{bmatrix}. \end{aligned} \tag{A.6}$$

The vectors $\boldsymbol{x}_i$ are chosen such that

- they form a linearly independent set, and
- each $\boldsymbol{x}_i$ is orthogonal to every column of $\boldsymbol{Z}_1(k)$.

A simple choice of the noise subspace basis vectors can be obtained by recalling the fact that the $z(k)$'s are roots of the prediction error filter polynomial $\Sigma_{i=0}^{P_1} a(k)\,z^{-i}$. Let the $\boldsymbol{x}_i$'s be chosen as follows:

$$\boldsymbol{x}_i = \begin{bmatrix} \underbrace{0 \;\cdots\; 0}_{i-1} & a_1(P_1) & a_1(P_1-1) & \cdots & a_1(0) & \underbrace{0 \;\cdots\; 0}_{k-i-P_1} \end{bmatrix}. \tag{A.7}$$

It is obvious that the vectors in (A.7) have the desired properties to be used as basis functions for the noise subspace. The matrix $\boldsymbol{X}(k)$ therefore takes the form

$$\boldsymbol{X}(k) = \begin{array}{c} \\ k \end{array} \overset{\overset{\qquad\qquad\qquad\qquad\qquad\qquad k-P_1}{\downarrow}}{\begin{bmatrix} a_1(P_1) & 0 & \cdots & 0 \\ a_1(P_1-1) & a_1(P_1) & & \vdots \\ \ddots & \ddots & \ddots & 0 \\ \ddots & & & a_1(P_1) \\ a_1(0) & & & a_1(P_1-1) \\ 0 & a_1(0) & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & a_1(0) \end{bmatrix}}.$$

Since the columns of $\boldsymbol{Z}(k)$ and $\boldsymbol{X}(k)$ span mutually orthogonal subspaces, the least-squares solutions for $\boldsymbol{c}$ and $\boldsymbol{c}'$ are

$$\begin{aligned} \boldsymbol{c} &= \left(\boldsymbol{Z}_1^H(k)\,\boldsymbol{Z}_1(k)\right)^{-1}\boldsymbol{Z}_1^H(k)\,\boldsymbol{y}; \\ \boldsymbol{c}' &= \left(\boldsymbol{X}_1^T(k)\,\boldsymbol{X}_1(k)\right)^{-1}\boldsymbol{X}_1^T(k)\,\boldsymbol{y}. \end{aligned} \tag{A.8}$$

Projection matrices that project the observed data onto the

signal subspace spanned by the columns of $\boldsymbol{Z}_1(k)$ and the noise subspace spanned by the columns of $\boldsymbol{X}_1(k)$ can be defined to be

$$\boldsymbol{P}_{Z_1} = \boldsymbol{Z}_1\left(\boldsymbol{Z}_1^H(k)\,\boldsymbol{Z}_1(k)\right)^{-1}\boldsymbol{Z}_1^H(k);$$
$$\boldsymbol{P}_{X_1} = \boldsymbol{X}_1\left(\boldsymbol{X}_1^T(k)\,\boldsymbol{X}_1(k)\right)^{-1}\boldsymbol{X}_1^T(k). \qquad \text{(A.9)}$$

Using these definitions

$$\boldsymbol{y}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{P}_{Z_1}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{P}_{X_1}\boldsymbol{y}. \qquad \text{(A.10)}$$

This expression can be substituted in (A.9) to obtain the following expression for the LS error:

$$e(k) = \boldsymbol{y}^T\boldsymbol{P}_{Z_1}\boldsymbol{y}. \qquad \text{(A.11)}$$

Defining

$$\boldsymbol{d}_k = \boldsymbol{X}_1(k)\,\boldsymbol{y}$$
$$\boldsymbol{R}_k = \boldsymbol{X}_1^T(k)\,\boldsymbol{X}_1(k)$$
$$\boldsymbol{w}_k = \boldsymbol{R}_k^{-1}\boldsymbol{d}(k),$$

the error function can be rewritten as an inner product.

$$e_1(k) = \boldsymbol{d}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{d}(k) = \boldsymbol{d}_k^T\boldsymbol{w}(k). \qquad \text{(A.11)}$$

$\boldsymbol{R}_k$ is Toeplitz.

Just as in the first method, the structure of the matrix $\boldsymbol{R}_k$ can be used to obtain efficient recursions for the computation of the error function for various values of k. For each increment in k, the matrix $\boldsymbol{R}_k$ is extended by a row and a column and

$$\boldsymbol{d}_{k+1} = \left(\boldsymbol{d}_k \sum_{i=0}^{P_1} a(i)\,y(k+i)\right).$$

The vector $v(k)$ can therefore be computed for $k = (P_1 + 1, P_1 + 2, \cdots, k - P_1)$ by using the Levinson recursions. For $k = P_1 + 1$, $\boldsymbol{R}_k$ is a scalar. Therefore, there is no need to invert a $P_1 \times P_1$ matrix to start the recursions as in method 1. The disadvantage with this method is that the dimensions of the matrix $\boldsymbol{R}_k$ grows with the size of the observation vector. For large orders, $\boldsymbol{R}_k$ tends to become ill conditioned. This method is most applicable for short observations.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[2] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *JASA*, vol. 49, pp. 583-590, 1971.

[3] E. N. Pinson, "Pitch-synchronous time-domain estimation of formant frequencies and bandwidths," *JASA*, Aug. 1963.

[4] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, Aug. 1979.

[5] R. Kumaresan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 833-840, Dec. 1982.

[6] R. Kumaresan, D. W. Tufts, and L. L. Scharf, "A Prony method for noisy data: Choosing the signal components and selecting the order in exponential signal models," *Proc. IEEE*, vol. 72, Feb. 1984.

[7] T. L. Henderson, "Geometric methods for determining system poles from transient response," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, Oct. 1981.

[8] S. Parthasarathy and D. W. Tufts, "Maximum-likelihood estimation of the parameters of exponentially damped sinusoids," *Proc. IEEE*, vol. 73, pp. 1528-1530, Oct. 1985.

[9] R. Kumaresan and A. K. Shaw, "An algorithm for pole-zero modeling and spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 637-640, June 1986.

[10] G. Fant, "Glottal source and excitation analysis," STL-QPSR, KTH, no. 1, Stockholm, Sweden, pp. 85-107, 1979.

[11] B. Cranen and L. Boves, "The effects of glottal termination impedance on the formants of speech signals," in *Signal Processing II: Theories and Applications*, H. W. Schussler, Ed. Amsterdam, The Netherlands: Elsevier Science, 1983.

[12] D. R. Allen and W. J. Strong, "A model for the synthesis of natural sounding vowels," *JASA*, vol. 78, no. 1, pp. 58-69, July 1985.

[13] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least-squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 350-355, Aug. 1979.

[14] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *JASA*, vol. 50, no. 2, pp. 637-655, 1971.

[15] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[16] M. V. Mathews, J. E. Miller, and E. E. David, Jr., "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Amer.*, vol. 33, pp. 179-186, 1961.

[17] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. ICASSP 1982*, pp. 614-617.

[18] R. L. Miller, "Nature of the vocal cord wave," *JASA*, vol. 31, pp. 667-677, 1959.

[19] J. N. Holmes, "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter," in *Proc. IV Int. Congr. Acoust.*, Aug. 1962, pp. 1-4.

[20] ——, "Formant excitation before and after glottal closure," in *Conf. Rec. 1976 IEEE Conf. Acoust., Speech, Signal Processing*, Apr. 1976, pp. 39-42.

[21] H. J. Manley, "Analysis-synthesis of connected speech in terms of orthogonalized exponentially damped sinusoids," *JASA*, vol. 35, no. 4, Apr. 1963.

[22] D. W. Tufts, S. E. Levinson, and R. Rao, "Measuring pitch and formant frequencies for a speech understanding system," in *Proc. ICASSP*, Apr. 1976, pp. 314-317.

[23] G. H. Golub and V. Pereyra, "The differentiation of pseudo-inverses and non-linear least-squares problem whose variables separate," *SIAM J. Numer. Anal.*, vol. 10, no. 2, pp. 413-432, Apr. 1973.

[24] H. J. Trussell and M. R. Civanlar, "Optimal initial conditions and pulse values for multipulse speech coding," in *Proc. IEEE Conf. ASSP*, 1985, pp. 264-267.

[25] H. Fletcher, *Speech and Hearing in Communication*. New York: Van Nostrand, 1953.

[26] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: The Johns Hopkins University Press, 1983.

S. Parthasarathy (S'81-M'86) received the B.E. (honours) degree in electronics and communication engineering from the University of Madras, Madras, India, and the M.S. and Ph.D. degrees in electrical engineering from the University of Rhode Island, Kingston, in 1980, 1982, and 1986, respectively.

From 1980 to 1982 he was associated with the Robotics Research Group at the University of Rhode Island, and since then has been working on digital signal processing problems. He is now with AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include speech analysis and synthesis, signal modeling, and parameter estimation.

Donald W. Tufts (S'58-M'61-SM'78-F'82), for a photograph and biography, see p. 355 of the March 1987 issue of this TRANSACTIONS.