Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform

DAVID Y. WONG, MEMBER, IEEE, JOHN D. MARKEL, SENIOR MEMBER, IEEE, AND AUGUSTINE H. GRAY, JR., SENIOR MEMBER, IEEE

Abstract—**Covariance analysis as a least squares approach for accurately performing glottal inverse filtering from the acoustic speech waveform is discussed. Best results are obtained by situating the analysis window within a stable closed glottis interval. Based on a linear model of speech production, it is shown that both the moment of glottal closure and opening can be determined from the normalized total squared error with proper choices of analysis window length and filter order. Results from actual speech are presented to illustrate the technique.**

I. Introduction

THE problem of estimating the glottal volume velocity waveform directly from the acoustical speech waveform has interested speech researchers for a number of years [1]-[7]. There has been much speculation about the possible influence of the glottal pulse shape on the perception of synthetic speech [8], [9]. Holmes [9] has pursued this area and shown that under certain listening conditions, the use of glottal pulses derived from speech significantly improves the naturalness over fixed glottal models. Wong and Markel have recently shown that retaining the phase characteristics of a typical glottal pulse can improve LPC synthesis quality [16].

By adopting the linear model of speech production [10], [11], accurate derivation of the glottal pulse shape from the speech waveforms also directly benefits another active area of speech research, namely, estimation of the vocal tract shape [11], [12]. If either the glottal waveform or vocal tract transfer function is accurately specified, then the other one can be obtained within the limits of the assumed model.

Considerable literature exists on analog glottal inverse filtering experiments and their interpretations [1], [2], [5]. However, two aspects of the technique have not been discussed in detail.

First, a physically meaningful mathematical basis for glottal inverse filtering has not been explicitly applied. The criterion used is usually defined qualitatively as "choose the settings which give minimum ripple during the interval of expected closure" [3], [7], [9], [13, p. 245]. Strube [14] has proposed a more mathematical approach of obtaining the inverse filter by linear prediction analysis over the closed glottis interval, and suggested using a log determinant measure for locating the instant of glottal closure. In this paper we will examine the covariance analysis technique more closely, and show that with a careful choice of analysis conditions, it provides a least squares approach to glottal inverse filtering. In addition, we will show that the minimum (rather than the maximum) [14] of the normalized total squared error offers a theoretically more precise measure for locating the instant of glottal closure and opening. This theoretical result is demonstrated in detail with an actual speech sample.

Second, the inverse filtering results are also very much dependent on various practical factors, such as the quality of the input speech recording. More specifically, ambient room noise, a low-frequency bias, and tape distortion have been found to lead to serious degradation to the results. Therefore, a brief discussion of these topics, along with a solution to the low-frequency bias problem, is presented.

Manuscript received January 13, 1977; revised December 30, 1977, July 21, 1978, and February 7, 1979. This work was supported by the Department of Defense under Contract MDA904-77-C-0197. Part of this work was performed while the authors were at the Speech Communications Research Laboratory, Inc., Santa Barbara, CA 93109.

D. Y. Wong and J. D. Markel are with Signal Technology, Inc., Santa Barbara, CA 93101.

A. H. Gray, Jr. is with the Department of Electrical Engineering and Computer Science, University of California, Santa Barbara, CA 93106, and with Signal Technology, Inc., Santa Barbara, CA 93101.

II. Linear Models and Analysis

A. Basic Models

A discrete time formulation of the linear speech production model for steady-state nonnasalized voiced speech, as illustrated in Fig. 1(a), is first assumed. The pertinent time sequences and their z-transforms are defined by

$$\begin{aligned} E(z) &\longleftrightarrow e(n) && \text{glottal excitation model signal} \\ U_G(z) &\longleftrightarrow u_G(n) && \text{glottal volume velocity signal} \\ U_L(z) &\longleftrightarrow u_L(n) && \text{lip volume velocity signal} \\ S(z) &\longleftrightarrow s(n) && \text{speech pressure wave signal.} \end{aligned} \tag{1}$$

The glottal excitation model signal $e(n)$ does not represent a physical signal (velocity or pressure), but rather, it is used as a mathematical input to a glottal model filter $G(z)$ in order to generate the glottal volume velocity signal $u_G(n)$. For a voiced sound, $e(n)$ is taken to be a periodic train of impulses.

The vocal tract model $V(z)$ is assumed to be an all-pole model of the form [13]

$$V(z) = \left[1 + \sum_{i=1}^{K} c_i z^{-i}\right]^{-1}. \tag{2}$$

Taking K to be an even integer, $K/2$ formants or vocal tract

0096-3518/79/0800-0350$00.75

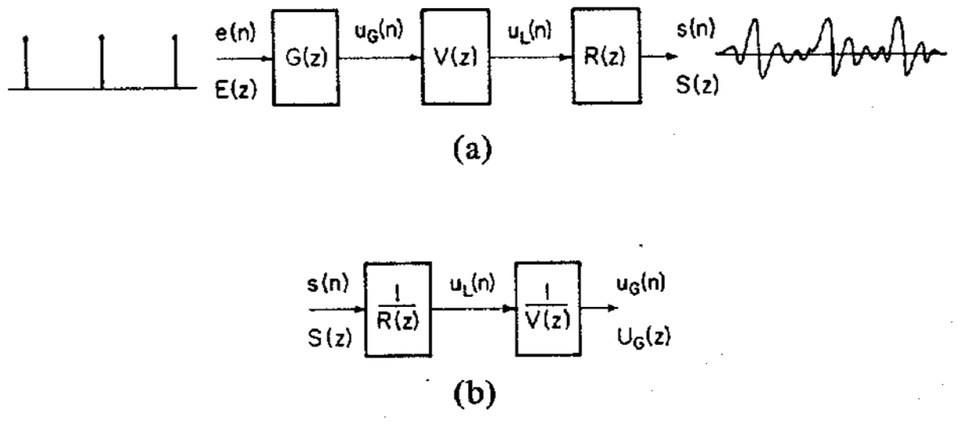

Fig. 1. (a) Block diagram representation of linear speech production model. (b) Block diagram of conceptualized glottal inverse filtering model.

resonances are included in $V(z)$ below the folding frequency.

The speech signal pressure wave is related to the volume velocity at the lips through a radiation impedance $R(z)$. For frequencies below about 4000 Hz, the sound pressure signal at a distance of l_1 from the lips is proportional to the time derivative of the unnormalized volume velocity at the lips with a time delay of l_1/c where $c = 35.3$ cm/ms [13, pp. 35, 38–41]. Excluding the proportionality constant and the time delay, a low-frequency approximation to the radiation impedance $R(z)$ is given by a differencing filter

$$R(z) = 1 - z^{-1}. \tag{3}$$

Based on the linear model just described, glottal inverse filtering is conceptually defined as solving for $U_G(z)$ by the equation

$$U_G(z) = \frac{S(z)}{V(z)\,R(z)}. \tag{4}$$

The relationships going from the speech pressure waveform to the glottal velocity waveform, in the form of an analysis model, are indicated in Fig. 1(b).

Since $R(z)$ is the same for different speech sounds, the fundamental problem in the estimation of the glottal volume velocity waveform is to determine the parameters of the inverse filter $1/V(z)$.

Before beginning the discussion on the glottal inverse filtering techniques, it must be noted that based upon analysis starting from the speech waveform, there are several important observations relating to the estimated and actual glottal volume velocity waveforms. An absolute dc value cannot be obtained for U_G since $R(z)$ has a zero at zero frequency. However, polarity is maintained from the glottal volume velocity wave to the speech pressure wave as long as the recording equipment does not introduce polarity reversals. The propagation delays from the glottis to the lips and then the transducer microphone are not included in the model here so that the timing of the various waveforms can be directly compared. Delays must be included if it is desired to make timing comparisons between the acoustically estimated volume velocity waveform and actual physical measurements made at the glottis.

B. Analysis

We will now show that covariance analysis provides a least squares estimate of $V(z)$ and that the normalized error energy can be used to determine the instants of glottal closure and

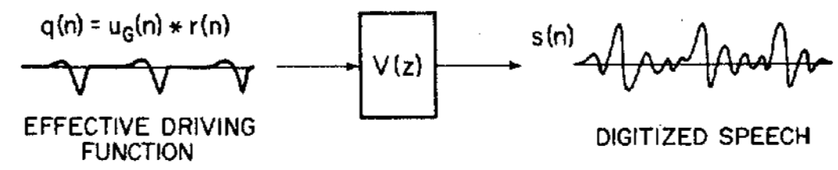

Fig. 2. Equivalent representation of a linear speech production model in terms of an effective driving function and vocal tract model.

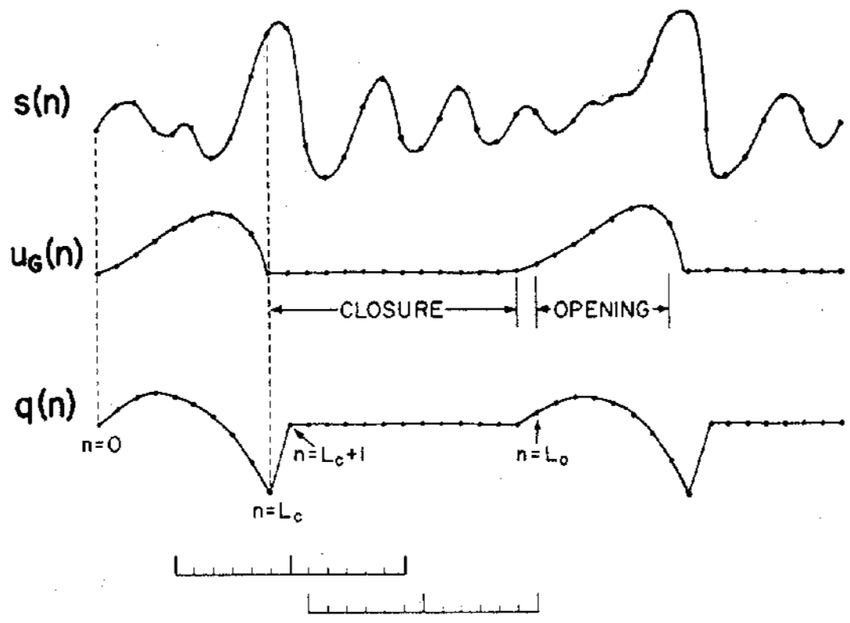

Fig. 3. Timing relationships among speech, glottal volume velocity, and effective driving function.

opening. To begin, an *effective driving function* $Q(z) \Leftrightarrow q(n)$ is defined by

$$Q(z) = U_G(z)\,R(z) \tag{5a}$$

or

$$q(n) = u_G(n) * r(n) \tag{5b}$$

where $*$ denotes convolution, so that the linear model of Fig. 1(a) is equivalently described by the model in Fig. 2.

Since the radiation term has a zero at zero frequency, $q(n)$ must be a zero-mean signal. Assuming stable closed glottis conditions, the theoretical timing relationships among the digitized speech, the glottal volume velocity, and the effective driving function are illustrated in Fig. 3. The locations $n = L_c$ and $n = L_o$ define the first points of glottal closure and opening, respectively. If $u_G(n) = 0$ for $L_c \leqslant n < L_o$, then $q(n) = 0$ over $L_c + 1 \leqslant n < L_o$, assuming the lip radiation of the form in (3). The locations of the windows will be explained shortly.

In terms of the effective driving function $q(n)$, the speech production model is of the form

$$s(n) = \sum_{i=1}^{K} c_i s(n - i) + q(n). \tag{6}$$

The constants K and $\{c_i\}$ are as defined in (2). Since the glottis is assumed to be closed for $L_c \leqslant n < L_o$, $q(n) = 0$ for $L_c + 1 \leqslant n < L_o$, so that

$$s(n) = -\sum_{i=1}^{K} c_i s(n - i) \qquad (L_c + 1 \leqslant n < L_o). \tag{7}$$

Thus, one sample after glottal closure, the speech waveform becomes a freely decaying oscillation (sum of complex exponentials) that is strictly a function of the vocal tract resonances specified by $c_1, \cdots, c_K$ and the initial conditions $s(L_c), \cdots, s(L_c - K + 1)$. In fact, this result holds over the entire interval $L_c + 1 \leqslant n < L_o$.

Assume that an Mth-order analysis filter of the form

$$A(z) = \sum_{i=0}^{M} a_i z^{-1} \qquad (a_o = 1) \tag{8}$$

where $M \geqslant K$ is to be obtained. If $s(n)$ is passed through this filter, the output is defined as the residue or error signal, and is given by

$$\epsilon(n) = s(n) + \sum_{i=1}^{M} a_i s(n - i). \tag{9}$$

For $L_c + 1 \leqslant n < L_o$, the expression given by (7) applies, so $\epsilon(n)$ is also given by

$$\epsilon(n) = \sum_{i=1}^{M} (a_i - c_i)\, s(n - i). \tag{10}$$

Therefore, if the conditions

$$a_i = \begin{cases} c_i & \text{for } i = 1, \cdots, K \\ 0 & \text{for } i = K+1, \cdots, M \end{cases} \tag{11}$$

are satisfied, then $\epsilon(n) = 0$ for $L_c + 1 \leqslant n < L_o$. In obtaining $A(z)$ by the covariance method of linear prediction for an analysis window from $s(n - M)$ to $s(n + N - M - 1)$, the total squared error $\alpha_M(n)$ is computed as

$$\alpha_M(n) = \sum_{j=n}^{n+N-M-1} \epsilon^2(j) \tag{12}$$

when $\epsilon(j)$ is defined by (9), and the analysis filter coefficients are obtained by minimizing $\alpha_M(n)$. If the conditions $n \geqslant L_c + 1$ and $n + N - M < L_o$ are satisfied, $\epsilon(j)$ is equal to zero over the entire range of the summation in (12). This means that the covariance algorithm will generate $\{a_i\}$ according to (11), and $\alpha_M(n)$ can be theoretically reduced to zero.

The relationship between L_c, L_o and the analysis window just discussed are illustrated by the two windows in Fig. 3 for the case $M = 6$ and $N = 13$. For the first analysis window in Fig. 3, the first point of the window is exactly $L_c - 5$, and the last point is less than L_o. Assuming that $s(n)$ is given by (6) and $K \leqslant M$, then (7) is satisfied and $\alpha_M(n)$ is reduced to zero by selecting $\{a_i\}$ according to (11). The first six points of this window correspond to the initial conditions, and the last seven points correspond to the interval over which the total squared error of (12) is minimized.

For the second window in Fig. 3, the first point is past $L_c + 1$, and the last point is exactly L_o. The error signal samples $\epsilon(j)$ corresponding to the last seven points are no longer within the range $[L_c + 1, L_o - 1]$, so (7) is not satisfied and $\alpha_M(n)$ cannot be reduced to zero. The two window locations n_1 and n_2 therefore delimit the range over which $\alpha_M(n)$ can be minimized to zero. Conversely, these window locations can be obtained by examining the $\alpha_M(n)$ as follows if the speech signal is correctly represented by (6).

Assume that the total squared error is computed on a sequential basis, that is, $\alpha_M(n)$ is computed by moving the N-length analysis window one sample at a time. Then, at the first sample n_1, where $\alpha_M = 0$, glottal closure is defined by

$$L_c = n_1 - 1 \text{ (closure)}. \tag{13a}$$

At the next sample where nonzero error occurs, n_2, the glottal opening location is defined by n_2 plus the number of samples over which error is minimized $(N - M)$ minus one, i.e.,

$$L_o = n_2 + N - M - 1 \text{ (opening)}. \tag{13b}$$

The window locations $n = n_1$ and $n = n_2$ are those illustrated in Fig. 3.

The condition $\alpha_M = 0$ can only occur theoretically as discussed above. For actual speech data, it defines a squared error measure for judging how well the digitized speech matches the model of speech production during glottal closure. This assumes that sufficient computer accuracy is used so that numerical roundoff errors are masked by the modeling and recording condition noises. In addition, it assumes that M equals two times the possible number of formants in the sampled speech signal.

To ensure that results are not a function of absolute system gain (such as recording or voice level), it is preferable to use the normalized total squared error as the measure of goodness, i.e.,

$$\eta(n) = \alpha_M(n)/\alpha_0(n) \tag{14}$$

where $\alpha_0(n)$ is the input signal energy. Therefore, if η is sufficiently small, the corresponding filter coefficients may represent the vocal tract model $V(z)$ without any influence of the glottal or radiation terms. Under these conditions, the volume velocity waveform is estimated from

$$\hat{U}_G(z) = \hat{Q}(z)/(1 - z^{-1}) \tag{15a}$$

where

$$\hat{Q}(z) = S(z)/\hat{V}(z) = S(z)\, A(z) \tag{15b}$$

and $A(z)$ is the all-zero filter defining the inverse vocal tract transfer function estimate. The details for constructing $A(z)$ based upon practical considerations are presented below.

III. Analysis Procedure and Observations

A. Procedure and Example

A structural procedure for glottal volume velocity waveform estimation, based on the analysis developed in the last section, is shown in Fig. 4. It is, therefore, generally for the case when stable glottal closure intervals of adequate duration are present.

Before analysis the digitized speech samples are first passed through a linear-phase high-pass filter to remove any low-frequency energy due to the recording conditions. Since the actual speech signal may not be exactly represented as complex exponentials, as in (7), the criterion based upon $\alpha_M(n) = 0$ is operationally reformulated as follows. First, sequential covariance method analysis is performed, that is, the analysis window is sequentially moved one sample at a time throughout the incoming speech data. For each frame, the normalized squared error η is computed, corresponding to the filter $A(z)$. A threshold value η_{th} is then defined to estimate both the moments of glottal closure and opening. The locations where $\eta < \eta_{th}$ correspond to the interval $[n_1, n_2 - 1]$ are discussed in Section II, from which the points of glottal closure and opening can be determined.

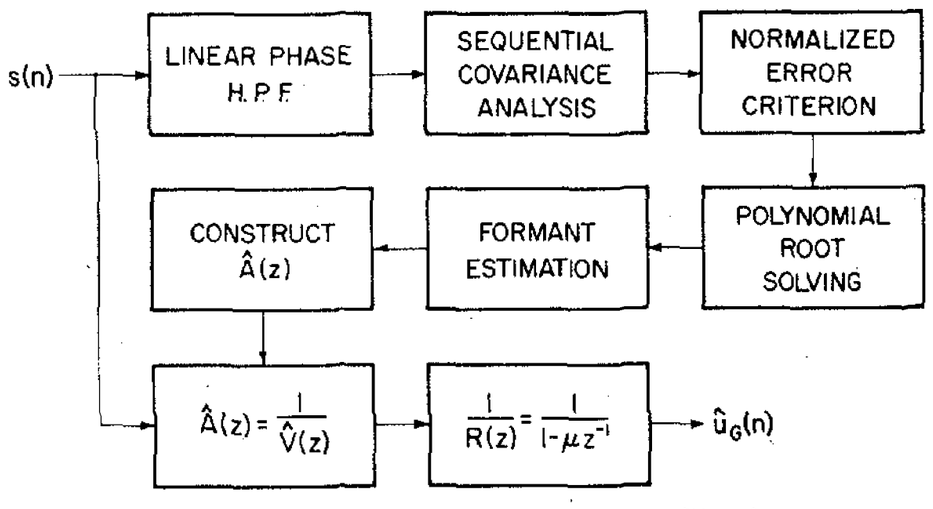

Fig. 4. Block diagram of the glottal inverse filtering system.

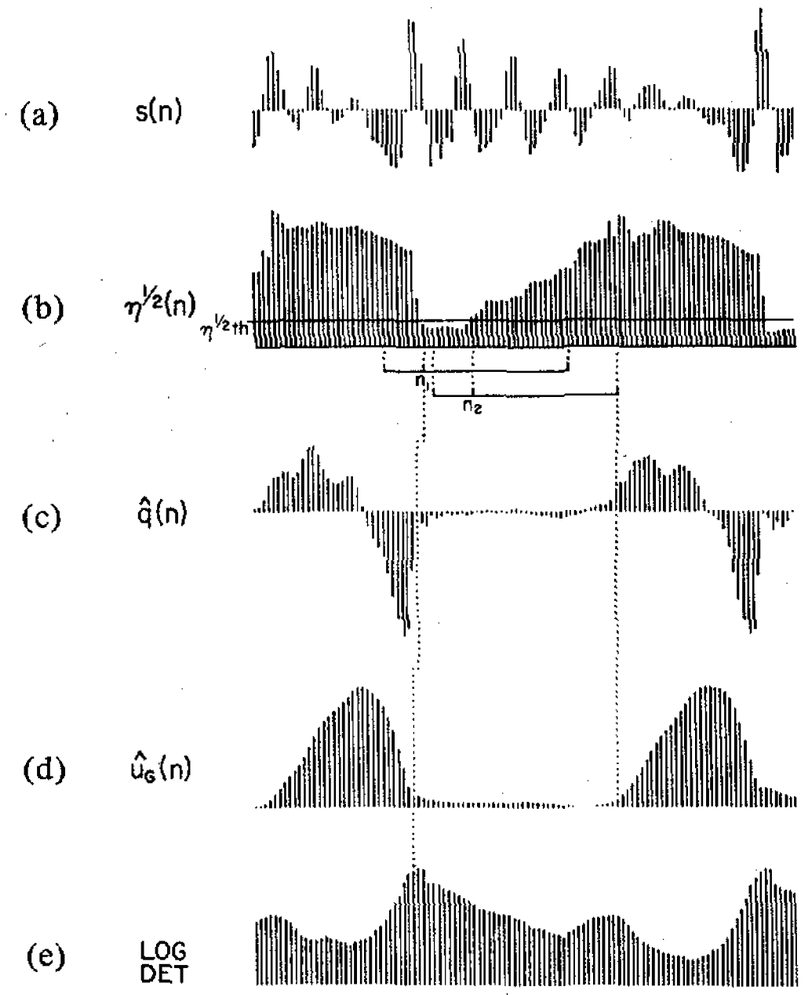

Fig. 5. Analysis Results: (a) Speech waveform. (b) Normalized error signal. (c) Effective driving function. (d) Glottal volume velocity waveform. (e) Strube's log determinant measure.

Experiments have shown that the point of closure is much more accurately determined than the point of opening by a threshold η_{th}. This is because the closure phase is normally much more abrupt than the opening phase so that η changes much more rapidly at the point of closure. However, by preemphasizing the speech data before computing the covariance terms in the analysis, the slope of the opening phase becomes steeper, and is more accurately determined from η. With this added preemphasis, following the same argument and notations as for Fig. 3, the point of glottal closure L_c is given by $n_1 - 2$ where n_1 is the first point when $\eta < \eta_{th}$. The point of opening is still given by (13) where n_2 is the first point when $\eta > \eta_{th}$. We will illustrate the technique with an actual speech sample below.

An example of the digitized speech waveform for /a/ and the square root of the normalized squared error for slightly more than one waveform period is shown in Fig. 5(a) and (b). The square root of the normalized squared error is shown in the figure as an explicit function of the sample index n. A preemphasis factor of 0.95 is used in the analysis. The location of each vertical line in the normalized squared error corresponds to the location of the first error sample used for computing $\alpha_M(n)$ as given by (12). The error sample $\epsilon(j)$ is obtained from (9). The fast rate of change at n_1 is a strong relative indicator of glottal closure. The rate of change corresponding to the glottal opening phase ($n > n_2$) is much more gradual, which is indicative of a slow glottal pulse rise during opening. The two analysis windows that correspond to n_1 and n_2 are also shown in Fig. 5(b). The window length and filter order are $N = 38$ and $M = 8$, respectively. There is no theoretical criterion for defining η_{th}. Operationally, η_{th} is obtained here as a value just exceeding a sample $\eta(n)$ that is abruptly lower than the preceding sample, as is the case here for $\eta(n_1)$ in Fig. 5. Once η_{th} is defined according to this behavior (corresponding to an abrupt glottal closure), n_2 is also operationally defined.

To estimate the actual volume velocity waveform, the filter $A(z)$ for a single period is chosen as that corresponding to η_{min} (the minimum value obtained between n_1 and n_2). There is, however, little difference between the filters obtained at any of the points between n_1 and n_2.

Having obtained a filter $A(z)$, minor adjustments are necessary to ensure that the inverse filter will only remove the formant poles from the speech signal. This step is needed because the estimated vocal tract model $1/A(z)$ may have real poles at either zero frequency or the half-sampling frequency. Formants for the vocal tract, however, are always defined from only complex pole pairs. The real pole at zero frequency will typically occur due to low-frequency recording noise or a nonzero mean in the short analysis window. Real poles may also occur when the required filter order is over specified. The first effect is avoided by high-pass filtering the speech data, but the second cause may still lead to a real zero. If it is not removed, "jags" at the points of glottal closure will occur. A real pole may also occur at the half-sampling frequency. When it is of narrow bandwidth, it generally indicates a formant location nearby, and thus should be retained. If a real pole occurs due to spectral shaping requirements in the analysis, without there being a nearby resonance, it will generally be of wide bandwidth. Including such a pole in the inverse filter will have a minimal effect on the results. Therefore, as a practical matter, we do not remove poles at the half-sampling frequency, if they occur. After eliminating any real roots near $f = 0$ Hz, a polynomial $\hat{A}(z)$ (of possibly reduced order) is reconstructed as the final inverse vocal tract model estimate.

After an estimate of the filter $\hat{A}(z)$ is obtained, the digitized speech is processed through it to obtain the effective driving function estimate $\hat{q}(n)$. The glottal volume velocity estimate is then obtained by integrating $\hat{q}(n)$ from (15a). To avoid possible overflow on some machines, the pure difference function in (15a) can be replaced with $1 - \mu z^{-1}$, where $0.98 \leqslant \mu < 1.0$. For steady-state vowels, several periods of the glottal waveform may be satisfactorily obtained using one estimate $\hat{A}(z)$. Otherwise, $\hat{A}(z)$ has to be updated as the formant frequencies show noticeable change.

The two waveforms obtained from the procedure are shown in Fig. 5(c) and (d). Note the time synchronization of the instants of opening and closure of the estimated glottal waveform with those predicted by η.

The effective driving function estimate $\hat{q}(n)$ shown in Fig. 5(c), as stated earlier, is a zero-mean signal whose absolute reference $\hat{q}(n) = 0$ is shown. Furthermore, $\hat{q}(n)$ is the actual linear prediction error or residual signal corresponding to $\eta_{\min}$ over the analysis window.

The resultant glottal volume velocity waveform estimate, obtained by integrating the effective driving function, is shown in Fig. 5(d). A relatively slow rise time at the onset of the glottal opening is seen with a small amount of superimposed ripple. The fall time is somewhat more rapid during glottal closure. This behavior for speech production with moderate intensity is typical and has been observed by most researchers in this area for moderate intensity vowel phonation [1], [2], [5], [15].

Due to the lip radiation impedance which introduces a zero at dc, the glottal volume velocity baseline cannot be recovered theoretically. A baseline can be estimated by connecting the minima for two periods as is done in Fig. 5(d).

In Fig. 5(e), the log determinant of the $M+1$ by $M+1$ covariance matrix is plotted for comparison. This measure has been suggested by Strube [14] for locating the instant of glottal closure. It is seen here that its peaks are within a few points from the point of closure predicted by the η measure. However, theoretically there is no precise mathematical relationship between the point of closure and the log determinant peak. Our experiments showed that to obtain well-defined peaks for the log determinant measure, such as the one shown here, very short analysis windows must be used. Longer analysis windows produce log determinant values with flat tops and therefore ill-defined peaks. A 2 ms window was suggested by Strube. The window used for Fig. 5(e) is slightly shorter than 2 ms long. We find the accuracy of the measure to be dependent on strong peaks in the speech wave generated by sharp glottal closures, and not so much on how well the speech fits the linear all-pole model. As a result, gain normalization for the log determinant measure is difficult, as Strube has pointed out.

The measure was also examined by Strube for locating the instant of glottal closure. Instead of searching for the minimum, he chose the location of the *maximum* of the total squared error $\eta(n)$ as the defining point for glottal closure. It is easily seen from Fig. 5(b) that the location of the maxima has no relationship to the point of glottal closure whatsoever.

B. Observations

From the example illustrated in Fig. 5, several observations on the estimation of formant parameters can be made.

The formant frequencies and bandwidths computed during the closure interval are different from those estimated during glottal opening, or those estimated over several periods (as in the autocorrelation method) because of effective vocal tract length increases and energy dissipation into the subglottal region [14]. The estimated bandwidths tend to be larger and the formant frequency estimates tend to be lower than those for closed glottis conditions. The frequency difference is barely visible on the glottal estimate of Fig. 5(d) (during glottal opening), but is easily seen as an oscillation on the effective driving function during the corresponding interval in Fig. 5(c).

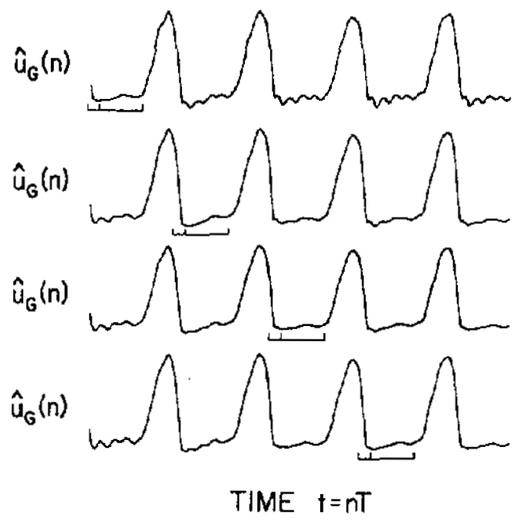

Fig. 6. Glottal volume velocity estimates for different analysis periods.

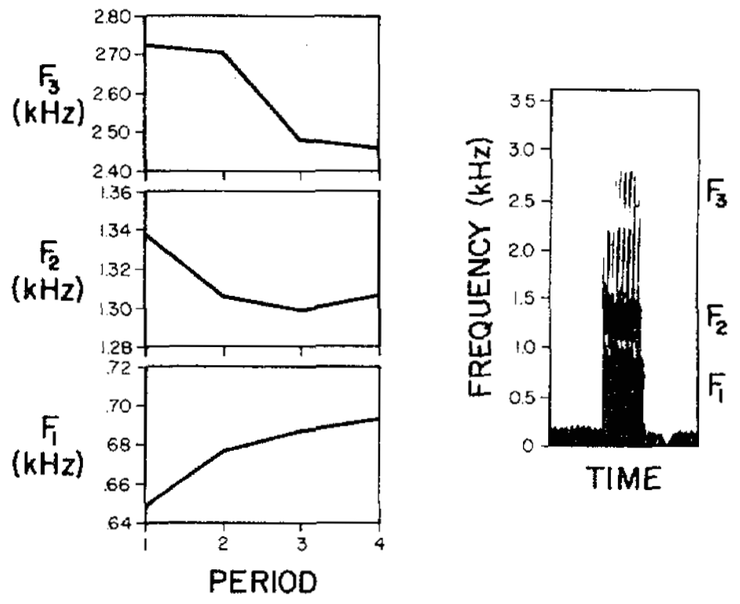

Fig. 7. The first three formant frequencies for four analysis periods along with the corresponding sonogram section.

This result shows one level of detail not easily seen in earlier studies.

To demonstrate the resolution of the inverse filter attainable by covariance analysis over the closed glottis, Fig. 6 shows the estimated glottal volume velocity waveforms based upon analysis window locations as shown for each of the periods. The glottal waveforms within each specific analysis period are remarkably similar, thus indicating that substantial resolution has been obtained in separating the glottal waveform structure from the supraglottal structure. To further demonstrate this point, Fig. 7 shows the first three formant frequencies obtained by solving for the roots of each polynomial $A(z)$, corresponding to $\eta_{\min}$ within each of the first four periods. A wideband sonogram for the same interval is also shown in Fig. 7 for comparison. The resolution of the analysis is significantly increased with respect to what can be reliably estimated from the sonogram, particularly for F_1 and F_2. It should be emphasized that these formant frequencies are also accurate in an absolute sense because the criterion of error between the linear model of speech production and the digitized speech over the intervals analyzed is on the order of $\eta = 0.003$. By comparing just the first two periods for the waveform at the top of Fig. 6, it can be seen that 30–40 Hz errors in formant frequency estimation for F_1 and F_2 can rather severely affect the details of the volume velocity waveform in the closure region.

IV. Discussion

For the procedures presented in the last section, it is necessary that the interval over which error minimization occurs is, at most, equal to the glottal closure interval, i.e., $L_o - L_c -$

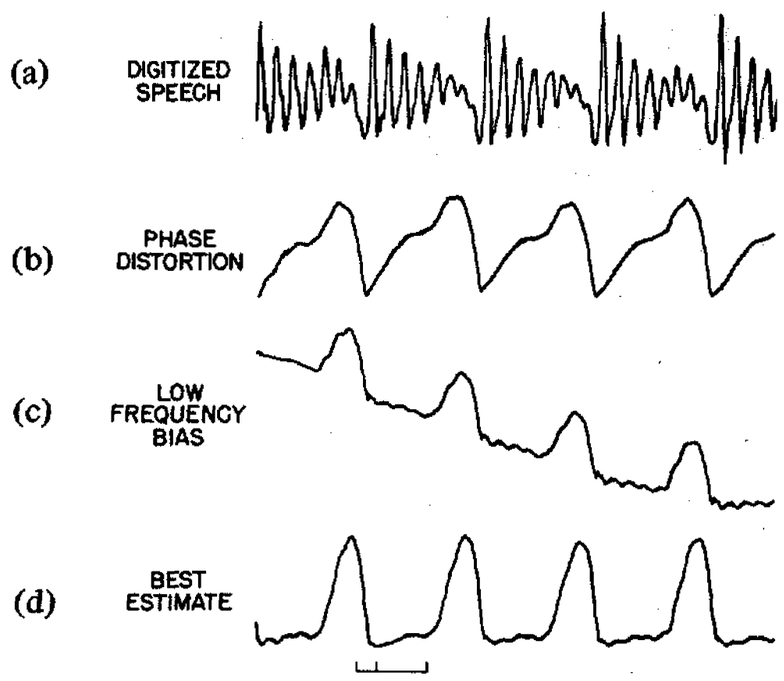

Fig. 8. Examples of the effects of recording distortion on the glottal wave estimate.

$1 \geqslant N - M$. Since N must at least be equal to $2M$, the absolute minimum value of $L_o - L_c$ is given by $M + 1$. For high fundamental frequency voices, the speech production model fails when used as an analysis structure even though the underlying events may still be represented by the model.

A number of factors in the recording procedure can significantly degrade the glottal inverse filtering results. They are: 1) ambient noise, 2) low-frequency bias due to breath burst on the microphone, 3) equipment and tape distortion, and 4) improper A/D conversion.

To minimize ambient noise, the example illustrated above was recorded in a soundproof room, and the microphone was placed only a few inches from the talker's lips. To ensure faithful frequency response, a Bruel and Kjer-type 2603 microphone amplifier and a type 4131 1 in condenser microphone was used. Combined, a flat response from 20 to 20 000 Hz was achieved. A speech segment based upon these conditions is shown in Fig. 8(a). However, the extremely good low-frequency response of this microphone and its close distance from the talker's mouth led to another serious problem as shown in Fig. 8(c) for the corresponding glottal estimate. Assume that a signal C (representing the additive signal bias) is applied to the general form of an integrator as

$$\frac{1}{R(z)} = 1/(1 - \mu z^{-1})$$

where $0 < \mu < 1$. The integrator output $y(n)$ is then

$$y(n) = C(1 - \mu^{n+1})/(1 - \mu).$$

This simple model describes the typical sloping form in the glottal volume velocity estimate as shown in Fig. 8(c). As n increases so that μ^{n+1} becomes small (for $\mu < 1$), the bias at the integrator output is seen to be multiplied by $1/(1 - \mu)$. Since practical values for μ are in the range $0.98 \leqslant \mu < 1.0$, the bias multiplication factor will typically be greater than 50.0. To remove the low-frequency bias, an additional filtering action is necessary to eliminate unpredictable, but very low-frequency bias). An appropriate solution is obtained by designing a linear-phase high-pass finite-impulse-response (FIR) filter [25] with a cutoff not exceeding $F_o/2$.

Analog tape recording introduces serious phase distortion in the glottal estimate as shown in Fig. 8(b). Suffice it to say that glottal inverse filtering should be attempted only with FM recorded signals [1] or with direct A/D conversion from a high-quality microphone amplifier into the computer. With proper care, the results as shown in Fig. 8(d) are obtained.

V. Summary

A methodology has been described for obtaining an accurate estimation of the glottal volume velocity waveform. A figure of merit was shown to be the normalized total squared error η. It was shown that if the digitized speech waveform analyzed approximately matches the linear speech production model, then both the moments of glottal opening and closure can be predicted. Experimental results have demonstrated this methodology to produce results with high resolution for real speech data with low fundamental frequency.

References

[1] R. L. Miller, "Nature of the vocal chord wave," *J. Acoust. Soc. Amer.*, vol. 31, pp. 667–677, 1959.
[2] J. N. Holmes, "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter," in *Proc. 4th Int. Congr. Acoust.*, Copenhagen, Denmark, 1962.
[3] P. B. Carr and D. Trill, "Long-term larynx-excitation spectra," *J. Acoust. Soc. Amer.*, vol. 36, pp. 2033–2040, 1964.
[4] T. Takasugi and J. Suzuki, "Speculation of glottal waveform from speech wave," *J. Radio Res. Lab. Jap.*, vol. 15, pp. 279–293, 1968.
[5] J. Lindqvist, "The voice source studied by means of inverse filtering," *STL-QPSR*, vol. 1, pp. 3–9, 1970.
[6] J. B. Allen and T. H. Curtis, "Automatic extraction of glottal pulses by linear estimation," in *86th Meet. Acoust. Soc.*, suppl., p. 36, 1973.
[7] J. N. Holmes, "Low-frequency phase distortion of speech recordings," *J. Acoust. Soc. Amer.*, vol. 58, pp. 747–749, 1975.
[8] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, 1971.
[9] J. N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 298–305, June 1973.
[10] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Heidelberg, New York: Springer-Verlag, 1976.
[11] G. C. M. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.
[12] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–427, Oct. 1973.
[13] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. Berlin, Heidelberg, New York: Springer-Verlag, 1972, p. 245.
[14] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1625–1629, 1974.
[15] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1632–1645, 1973.
[16] D. Y. Wong and J. D. Markel, "An excitation function for LPC synthesis which retains the human glottal phase characteristics," in *Conf. Rec. 1978 IEEE Int. Conf., Acoust., Speech, Signal Processing*, Apr. 1978, pp. 171–174.