Robustness of Group-Delay-Based Method for Extraction of Significant Instants of Excitation from Speech Signals

P. Satyanarayana Murthy and B. Yegnanarayana, *Senior Member, IEEE*

***Abstract*—In this paper, we study the robustness of a group-delay-based method for determining the instants of significant excitation in speech signals. These instants correspond to the instants of glottal closure for voiced speech. The method uses the properties of the global phase characteristics of minimum phase signals. Robustness of the method against noise and distortion is due to the fact that the average phase characteristics of a signal is determined mainly by the strength of the excitation impulse. The strength of excitation is determined by the energy of the residual error signal around the instant of excitation. We propose a measure for the strength of the excitation based on Frobenius norm of the differenced signal. The robustness of the group-delay-based method is illustrated for speech under different types of degradations and for speech from different speakers.**

***Index Terms*—Glottal pulse, group-delay, instants of excitation, residual signal.**

I. Introduction

SPEECH is produced as a result of excitation of a time-varying vocal tract system. In the case of voiced speech, the excitation is due to the quasiperiodic airflow resulting from the opening and closing of the glottis in each glottal cycle. Within a glottal cycle, the vocal tract system is strongly excited around the instant of glottal closure. We refer to this instant as the *significant instant* in this paper. Strong excitations such as at the release of unvoiced or voiced stops can also be considered as significant instants.

Instants of significant excitation are useful in several situations, for example, for accurate analysis and synthesis of speech [1]–[3]. For noisy speech, knowledge of the significant instants helps in performing robust spectrum analysis. This is because a short (2–4 ms) segment in the voiced speech signal immediately after the significant instant usually corresponds to a high signal-to-noise ratio (SNR) portion of the speech within a glottal cycle [4]. Hence, analysis of these short segments may yield better estimates of the characteristics of the vocal tract system.

Manuscript received February 14, 1997; revised August 1, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

P. S. Murthy is with the Department of Electrical Engineering, Indian Institute of Technology, Madras 600 036, India.

B. Yegnanarayana is with the Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India (e-mail: yegna@iitm.ernet.in).

Publisher Item Identifier S 1063-6676(99)08075-X.

Determination of the instants of significant excitation is difficult even for clean speech. In the case of strong voicing, due to sharp glottal closure in the voiced speech, the instant of significant excitation can be perceived even in the presence of noise. But in the case of voiced sounds where the glottal closure is gradual, the instant of glottal closure is difficult to perceive or identify, especially if the signal is corrupted by noise. Reliable identification of the instant of significant excitation depends on the strength of the excitation.

Several methods have been proposed in the literature for determining the instants of significant excitation [4]–[8]. Most of them depend on either the short-time energy of the speech signal or on the linear prediction (LP) residual signal. These methods are based on block-data processing, and hence there is some ambiguity in the locations of the instants. Moreover, the performance of these methods generally deteriorates when the speech signal is corrupted by noise and distortion.

In [9], a method was proposed for the extraction of the instants of significant excitation. The method is based on the fact that the average value of the group-delay function of a signal within an analysis frame corresponds to the location of the significant excitation within the frame. An improved method based on the computation of the group-delay function directly from the speech signal was proposed in [10]. In this paper, we propose further refinements of the method and then discuss the robustness of the group-delay-based method. Even though it was mentioned in [9] that the method would be sensitive to additive noise, the studies in this paper show that the group-delay-based method is indeed robust against additive random noise and channel distortions. This is because it is the strength of the excitation that determines the robustness of the method against noise.

In Section II, the modified group-delay-based method for the extraction of the instants of significant excitation is briefly reviewed. Some refinements of the method are also discussed. Since the robustness of the method is due to the strength of the excitation, we discuss in Section III the need for a measure of the strength of excitation, and propose a measure based on the Frobenius norm of the prediction matrix of the differenced speech signal. In Section IV, the robustness of the group-delay-based method is discussed for speech signals corrupted by additive noise and reverberation. In Section V, we study the performance of the method for different types of speech data with natural degradations.

II. Determination of Instants of Significant Excitation

In this section, we briefly present the group-delay-based method proposed in [9] and [10] for determining the instants of significant excitation from speech signals, and propose some refinements to the method. The method is based on the global phase characteristics of minimum phase signals. Since the average group-delay of a minimum phase system is zero [11], the average slope of the phase spectrum of the impulse response of the system corresponds to the location of the excitation impulse within the analysis frame [9]. In practice, the computed phase spectrum or the group-delay function depends on the window function used for analysis. To reduce the effects of the window function on the estimated group-delay function, it is preferable to compute the group-delay function from the LP residual signal. The residual signal is also preferable because some characteristics of the glottal source can be seen better in the residual error signal than in the speech signal. The average slope of the phase spectrum of the speech signal is the same for the residual signal also, because the inverse filter of the LP analysis is a minimum phase system [12]. The residual signal is derived by inverse filtering the speech signal, and the inverse filter is obtained using LP analysis. For LP analysis, a frame size of about 25 ms for every 10 ms may be chosen [9], [10]. The instants of significant excitation can be derived from the LP residual signal as follows [10]. Around each sampling instant a 10 ms segment of the LP residual signal is considered and the group-delay function is computed using the formula [13]

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)} \tag{1}$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ are the Fourier transforms of the windowed residual signal $x(n)$ and $y(n) = nx(n)$, respectively. The group-delay function is smoothed using a three-point median filter to remove any discontinuities in the group-delay function. The negative of the average of the smoothed group-delay function is called *phase slope*. The phase slope value is computed at each sampling instant to obtain the *phase slope function*. If the instant of significant excitation within a frame is at the midpoint of the frame, then the phase slope is zero. Therefore the positive zero-crossings of the phase slope function correspond to the instants of significant excitation. Short-time (1–3 ms) energy of the LP residual signal around the instant can be used to represent the strength of excitation associated with the instant [9], [10]. Fig. 1(a)–(d) show a segment of speech signal, the LP residual signal, the phase slope function and the extracted instants with estimated strengths, respectively. The speech signal shown corresponds to the utterance $/dzua/$, where $/dz/$ is a voiced palatal fricative as in *Julie*.

Sometimes the LP residual signal may contain some spurious impulses which may result in wrong estimation of the instants of significant excitation, as can be seen in Fig. 1(d), where the strengths are computed using the short-time energy of the residual signal centered around the estimated instants of significant excitation. The effect of these spurious impulses can be reduced by enhancing the region around the instants of significant excitation relative to other regions in the LP residual signal. This can be accomplished by deriving a weight function for the LP residual signal. The weight function is derived here by smoothing the LP residual signal with a Hamming window of duration 0.75 ms (eight samples at 11 kHz sampling rate). This smoothing reduces the noise fluctuations in the residual signal. The short-time energy of the smoothed residual signal is computed at every sample using a frame size of 1.4 ms (15 samples at 11 kHz sampling rate). The short-time energy curve will have large amplitudes around the significant instants. The short-time energy is normalized to a maximum value of one and is used as a weight function for the residual signal to enhance the regions in the residual signal around the significant excitations. The weighted residual signal is used to derive the instants of significant excitation. The phase slope function is smoothed using a five-point Hamming window. Positive zero-crossings of the smoothed phase slope function are used as the instants of significant excitation. Fig. 1(e) shows the plot of the instants derived after these

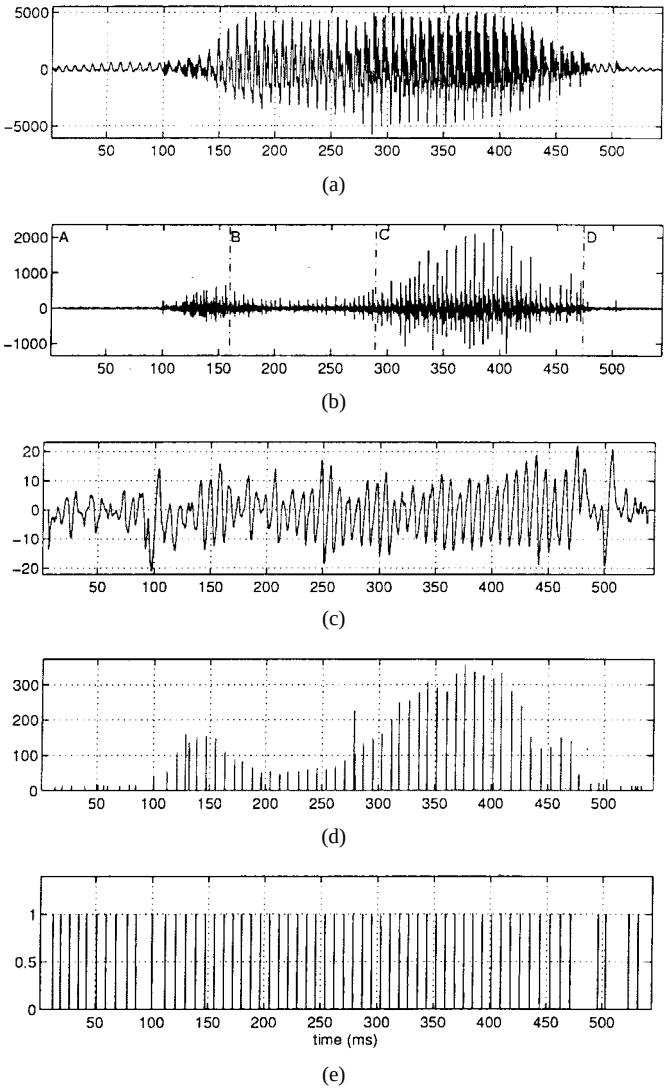


Fig. 1. (a) Clean speech for the utterance $/dzua/$. (b) LP residual signal derived from the signal in (a). (c) Phase slope function. (d) Significant instants, weighted by their strengths, derived from the signal in (a). (e) Significant instants, derived from the signal in (a) using the proposed algorithm.

refinements. Some of the errors in the estimation of instants in Fig. 1(d) are corrected in Fig. 1(e). The different steps in the algorithm for the computation of the instants of significant excitation are given in Fig. 9.

III. Measure of Strength of Excitation

Reliability of the extracted instants depends on the strength of excitation around the instants. In [9], [10] the short-time energy of the LP residual signal was used to represent the strength of excitation at each instant. In some cases it is difficult to use the short-time energy around the instant as a measure of the strength, especially when the residual signal is noisy, as in the region BC in Fig. 1(b). Moreover, the derived residual signal energy depends on the effectiveness of the LP analysis for these segments.

We propose an alternative measure for the strength of excitation, which is based on the use of the Frobenius norm. In [8] the Frobenius norm of a signal prediction matrix, formed by using the samples in a frame of about 3 ms, was proposed to locate the instants of glottal closure. The Frobenius norm was computed at each sampling instant. The locations of the peaks in the plot of the Frobenius norm as a function of time were considered as the desired instants. In this section we propose that the Frobenius norm [14] of the signal prediction matrix [8] formed by using the samples in a 3-ms frame of differenced speech centered around the identified instant of significant excitation can be used to represent the strength of excitation at that instant.

Consider a frame of the differenced speech signal with N samples, s_1, $s_2, \cdots, s_N$. Assuming a linear prediction order of p, the following prediction error vector can be formed:

$$\boldsymbol{e} = \boldsymbol{S}\boldsymbol{a} \tag{2}$$

where $\boldsymbol{S}$ is the Toeplitz signal prediction matrix of dimension $(N-p)\times(p+1)$

$$\boldsymbol{S} = \begin{bmatrix} s_{p+1} & s_p & \cdots & s_1 \\ s_{p+2} & s_{p+1} & \cdots & s_2 \\ & & \ddots & \vdots \\ \vdots & \vdots & & s_{p+1} \\ & & & \vdots \\ s_N & s_{N-1} & \cdots & s_{N-p} \end{bmatrix} \tag{3}$$

and $\boldsymbol{a}$ is the augmented vector of LPC's $[1\ a_1\ a_2 \cdots a_p]^T$. Assuming s_n, $n = 1, \cdots, N$ are the samples of a signal at the output of an all-pole system excited by a periodic impulse train, there is a linear dependence between the column vectors of $\boldsymbol{S}$, when the instant of excitation is not included in the analysis frame [8]. The error vector is then zero. But when the instant of excitation is included, the norm of the error vector goes up. The amplitudes of signal samples in the signal prediction matrix also go up, because of the excitation. Thus, the Frobenius norm of the signal prediction matrix, computed as the square root of the sum of all squared elements of the matrix, also goes up. The square of the Euclidean norm of $\boldsymbol{e}$, which is a measure of the energy (strength) of excitation, is given by

$$\begin{aligned} ||\boldsymbol{e}||_2^2 &= ||\boldsymbol{S}\boldsymbol{a}||_2^2 \\ &\le ||\boldsymbol{S}||_F^2 \cdot ||\boldsymbol{a}||_2^2 \end{aligned} \tag{4}$$

where $||\boldsymbol{S}||_F$ is the Frobenius norm of $\boldsymbol{S}$. The ratio $||\boldsymbol{e}||_2^2/||\boldsymbol{a}||_2^2$ is upper bounded by $||\boldsymbol{S}||_F^2$. Ignoring the variation in $||\boldsymbol{a}||_2^2$ compared to $||\boldsymbol{S}||_F^2$, we can use $||\boldsymbol{e}||_2^2/||\boldsymbol{a}||_2^2$ as a measure of the strength of excitation. Computing the Euclidean norm of $\boldsymbol{e}$ from (2), we get

$$\frac{||\boldsymbol{e}||_2^2}{||\boldsymbol{a}||_2^2} = \frac{\boldsymbol{a}^T(\boldsymbol{S}^T\boldsymbol{S})\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} \tag{5}$$

$$= \rho(\boldsymbol{a}) \tag{6}$$

where $\rho(\boldsymbol{a})$ is the Rayleigh quotient of $(\boldsymbol{S}^T\boldsymbol{S})$ [14]. It is shown in Appendix A [see (A.8)] that

$$\sigma_{p+1}^2 \le \rho(\boldsymbol{a}) \le \sigma_1^2 \tag{7}$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{p+1} > 0$ are the singular values of $\boldsymbol{S}$, and are also the eigenvalues of $\boldsymbol{S}^T\boldsymbol{S}$. It is also known that the square of the Frobenius norm is the sum of squared singular values [15]. So we have the inequality

$$\sigma_{p+1}^2 \le \frac{1}{(p+1)}\,||\boldsymbol{S}||_F^2 \le \sigma_1^2 \tag{8}$$

since $(p+1)^{-1}||\boldsymbol{S}||_F^2$ is the arithmetic mean of squared singular values. It is known that all the singular values rise in magnitude when there is an excitation within the analysis frame and fall when there is no excitation [8]. Therefore, both $\rho(\boldsymbol{a})$ in (7) and $(p+1)^{-1}||\boldsymbol{S}||_F^2$ in (8) track these changes. Therefore $(p+1)^{-1}||\boldsymbol{S}||_F^2$ can be used as a measure of the strength of excitation. We note that though this is a measure of energy of the residual signal, it is computed directly from the speech signal.

It is to be noted that since the square of the Frobenius norm of the signal prediction matrix is the sum of squares of all samples in the matrix, it is nothing but the short-time energy of the speech signal computed from the weighted samples of the speech signal.

To illustrate the need for a measure for the strength of excitation, let us consider the differentiated glottal pulses [Fig. 2(a)] generated using the LF model [16]. All the parameters of the model are kept constant except the time constant of the return phase and the instant of peak positive excitation. To vary the rate of closure, the time constant of the return phase is increased from 0.05–1.5 ms from left to right. The amplitudes of the pulses are progressively scaled up (from left to right) so that all the pulses have equal negative peak amplitudes. These differentiated glottal pulses are used to excite an all-pole model to obtain a synthetic voiced sound shown in Fig. 2(c). It should be noted that, in the first 40 ms of the speech signal, the signal components due to higher formants can be clearly seen. This is due to the sharp closing phase, which results in a magnitude spectrum of the excitation pulses that is less steep. This feature is not seen in the latter portion of the signal in Fig. 2(c) due to the gradual closing phase. The second derivative of the glottal pulse and the twelfth-order LP residual signal are shown in Fig. 2(b) and (d), respectively. From these figures it is evident

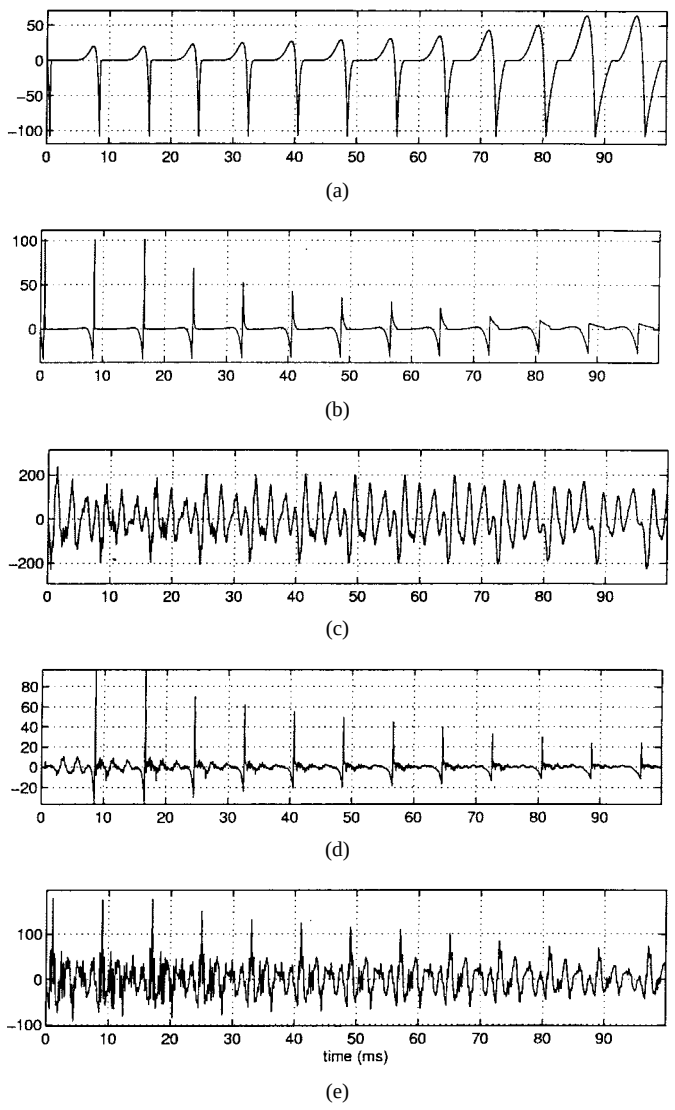


Fig. 2. (a) Differentiated glottal pulses. (b) Second derivative of glottal pulses. (c) Synthetic signal. (d) Residual signal derived from the signal in (c). (e) First-order difference of the signal in (c).

that the amplitudes of the excitation impulses are higher for the glottal pulses with sharper closure. The strength of excitation is higher for sharper closure, although the amplitude and energy of the speech signal in Fig. 2(c) is nearly the same throughout for all the glottal pulse shapes. It should be noted in Fig. 2(a) that the energy concentration is higher for the pulses in the initial portion than in the latter portion of the signal.

If we consider the differenced signal of Fig. 2(c), as shown in Fig. 2(e), we notice that the strength of excitation is also evident in the differenced signal. It can also be seen by considering a difference operation $(1 - z^{-1})$ on the z-transform of the signal, $S(z) = E(z)H(z)$, where $E(z)$ corresponds to the differentiated glottal pulse excitation, and $H(z)$ corresponds to the vocal tract system. We have

$$(1 - z^{-1})S(z) = (1 - z^{-1})E(z)H(z). \tag{9}$$

Thus, the differenced signal can be viewed as a signal that results due to the excitation of the vocal tract system with the second derivative of the glottal pulse. The second derivative of the glottal pulse in Fig. 2(b) and the differenced signal in Fig. 2(e) both show the characteristics of the strength of excitation. These figures suggest that the Frobenius norm of the differenced signal can be used as a measure of the strength of excitation around the instant of significant excitation.

IV. Robustness of the Group-Delay-Based Method

In this section we shall examine the robustness of the group-delay-based method for two types of degradations, namely, additive random noise and echo/reverberation.

A. Robustness Against Additive Noise

Let us consider an excitation signal $x(n)$ consisting of an impulse of amplitude $\mathcal{A}$ at time $n = L$ and a zero-mean additive white Gaussian noise $v(n)$.

$$x(n) = \mathcal{A}\delta(n - L) + v(n), \qquad n = 0, 1, \cdots, N - 1. \tag{10}$$

The Fourier transform of $x(n)$ is

$$X(\omega) = \mathcal{A}\exp(-j\omega L) + V(\omega) \tag{11}$$

where

$$\begin{aligned} V(\omega) &= \sum_{n=0}^{N-1} v(n)\exp(-j\omega n) \\ &= |V(\omega)|\exp(j\phi_v(\omega)). \end{aligned} \tag{12}$$

$|V(\omega)|$ and $\phi_v(\omega)$ are random variables corresponding to the magnitude and phase of $V(\omega)$, respectively. Without loss of generality, the phase spectrum $\phi_v(\omega)$ can be assumed to have a uniform probability density function over the range $[-\pi, \pi]$ [17]. Let $|X(\omega)|$ and $\phi_x(\omega)$ be the magnitude and phase of $X(\omega)$, respectively.

$$\begin{aligned} &\log[|X(\omega)|] + j\phi_x(\omega) \\ &\quad = \log(\mathcal{A}) - j\omega L \\ &\qquad + \log\left[1 + \frac{|V(\omega)|}{\mathcal{A}}\exp(j(\phi_v(\omega) + \omega L))\right]. \end{aligned} \tag{13}$$

It is shown in Appendix-B [see (B.4)] that

$$\frac{\mathcal{E}[|V(\omega)|]}{\mathcal{A}} < 10^{-(E_s/20)} \tag{14}$$

where $\mathcal{E}$ denotes ensemble average and E_s is the excitation SNR, defined as the logarithm of the ratio of average excitation signal power per sample $(\mathcal{A}^2/N)$ to the average noise power per sample (σ_v^2)

$$E_s = 10\log_{10}\left(\frac{\mathcal{A}^2}{N\sigma_v^2}\right) \quad \text{dB}. \tag{15}$$

For $E_s = 0$ dB, the upper bound on the expected value of the magnitude of $[(|V(\omega)|/\mathcal{A})\exp(j(\phi_v(\omega) + \omega L))]$ is one. If the Fourier transform in (11) is evaluated using an N-point discrete Fourier transform (DFT), the magnitude of the DFT $|V(\omega_k)|$ can be shown to be less than $\mathcal{A}$ with 99% confidence when $E_s \geq 6.6$ dB [see App. B, (B.7)]. Expanding the third

term on the right hand side of (13) by Taylor series expansion, the phase term of (13) can be approximated as

$$\phi_x(\omega) = -\omega L + \frac{|V(\omega)|}{\mathcal{A}} \sin(\phi_v(\omega) + \omega L). \tag{16}$$

The group-delay function $(\tau_x(\omega))$ is given by

$$\tau_x(\omega) = -\frac{d\phi_x(\omega)}{d\omega}. \tag{17}$$

Hence, the average value of the group-delay function is given by

$$\begin{aligned} \overline{\tau}_x &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \tau_x(\omega)\, d\omega \\ &= -\frac{1}{2\pi} [\phi_x(\pi) - \phi_x(-\pi)]. \end{aligned} \tag{18}$$

Substituting (16) in (18) and noting that $\phi_x(\omega)$ is an odd function of ω and that the second term in (16) vanishes at π, we have

$$\overline{\tau}_x = L \tag{19}$$

i.e., the average value of the group-delay function gives the location of the impulse.

In practice, the group-delay function is computed at discrete frequencies, and hence the computed average deviates from (19). Random fluctuations and spikes appear in the group-delay function [18]. These spikes may bias the mean value of the group-delay function. Therefore, it is preferable to use median smoothing of the computed group-delay function before computing the average.

So far we have considered an excitation signal corrupted by additive noise. Let us now consider a noisy speech signal $y(n)$

$$y(n) = s(n) + w(n) \tag{20}$$

where $s(n)$ is the speech signal and $w(n)$ is the additive white noise. To derive the instants of significant excitation, let us consider the LP residual signal. The frequency response of the inverse filter obtained from the LP analysis is given by

$$\begin{aligned} A(\omega) &= |A(\omega)| \exp(j\phi_A(\omega)) \\ &= \sum_{k=0}^{p} a_k \exp(-j\omega k) \end{aligned} \tag{21}$$

where $a_0 = 1$ and $a_1, \cdots, a_p$ are the LP coefficients (LPC's). The residual error signal obtained after inverse filtering is given by

$$x(n) = r(n) + v(n) \tag{22}$$

where $r(n)$ is the component at the output of the inverse filter due to the speech signal $s(n)$ and $v(n)$ is the colored noise due to filtering of the white noise $w(n)$. Note that even though the speech signal is assumed to be the output of an all-pole system, the noisy signal $y(n)$ corresponds to a pole-zero system [19]. The power spectrum of the colored noise component $v(n)$ is given by

$$\begin{aligned} P_v(\omega) &= \frac{1}{N}\, \mathcal{E}[|V(\omega)|^2] \\ &= |A(\omega)|^2 \sigma_w^2. \end{aligned} \tag{23}$$

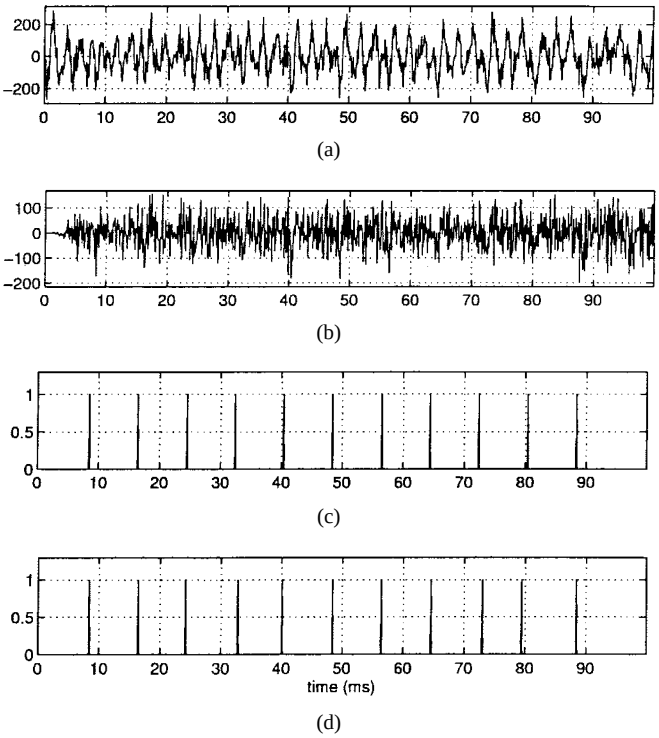


Fig. 3. (a) Synthetic speech of Fig. 2(c) at an average SNR of 5 dB. (b) LP residual signal derived from the signal in (a). (c) The true locations of the instants of significant excitation. (d) The instants of significant excitation derived from the noisy signal in (a).

The second moment $\mathcal{E}[|V(\omega)|^2]$ depends on the frequency ω. Let us consider the worst case situation, i.e., the maximum value of $(1/N)\mathcal{E}[|V(\omega)|^2]$. Let

$$\begin{aligned} \sigma_{v\max}^2 &= \max_{\omega} \frac{1}{N}\, \mathcal{E}[|V(\omega)|^2] \\ &= A_{\max}^2 \sigma_w^2 \end{aligned} \tag{24}$$

where $A_{\max}$ is the maximum value of $|A(\omega)|$ given by

$$\begin{aligned} A_{\max} &= \max_{\omega} |A(\omega)| \\ &= \max_{\omega} \left| 1 + \sum_{k=1}^{p} a_k \exp(-j\omega k) \right|. \end{aligned} \tag{25}$$

In the expression for E_s in (15), the σ_v^2 is replaced by $\sigma_{v\max}^2$. Assuming that $A_{\max} > 1$, the effective E_s for the residual signal is reduced.

The above analysis is valid even when the speech is corrupted by additive colored random noise, except that $A_{\max}$ now also depends on the maximum value of the power spectrum of the colored noise.

The robustness of estimation of the instant of excitation depends on the excitation SNR (E_s). For a constant additive noise, E_s will decrease as the strength of the excitation decreases. This is illustrated in Fig. 3 for a noisy case of the synthetic signal generated by exciting an all-pole filter with the differentiated glottal pulses of Fig. 2(a). The overall SNR of the speech signal is 5 dB. Note that the periodicity cannot be immediately seen from the noise corrupted speech signal. Since it is a synthetic case, the strength of excitation can be approximated to the amplitude of the second derivative

of the glottal pulse shown in Fig. 2(b). Fig. 3(c) shows the actual instants of significant excitation. Fig. 3(d) shows the instants of significant excitation estimated from the noisy speech signal. The figure shows that the accuracy of the extracted instants depends on the excitation signal-to-noise ratio. Reliability of the extracted instants decreases with a decrease in the excitation SNR, as can be seen from the deviation of the instants in Fig. 3(d) relative to the instants in Fig. 3(c). The excitation SNR (E_s) is defined as the ratio of the square of the amplitude of the impulse and the noise power. Note that even though the average SNR of the speech signal is nearly constant, i.e., 5 dB, the excitation SNR is decreasing from left to right on the time scale.

B. Robustness Against Echo and Reverberation

Let us consider the following reverberant signal $x(n)$ for an impulse of strength $\mathcal{A}$ and delayed by $n = L$ samples.

$$x(n) = \mathcal{A}\delta(n-L) + \lambda\mathcal{A}\delta(n-L-D) + \lambda^2\mathcal{A}\delta(n-L-2D) + \cdots \tag{26}$$

where λ is the attenuation factor ($0 < \lambda < 1$) and D is the delay due to reverberation. The Fourier transformation of (26) yields

$$|X(\omega)|\exp(j\phi_x(\omega)) = \frac{\mathcal{A}\exp(-j\omega L)}{(1-\lambda\exp(-j\omega D))} \tag{27}$$

where $|X(\omega)|$ and $\phi_x(\omega)$ are the magnitude and phase of the Fourier transform of $x(n)$, respectively. Taking natural logarithm on both sides of (27), we get [20]

$$\log[|X(\omega)|] + j\phi_x(\omega) = \log(\mathcal{A}) - j\omega L - \log(1-\lambda\exp(-j\omega D)). \tag{28}$$

Neglecting the higher order terms in the Taylor series expansion of the last term above, the phase component is given by

$$\phi_x(\omega) = -\omega L - \lambda\sin(\omega D). \tag{29}$$

The group-delay is

$$\tau_x(\omega) = L + \lambda D\cos(\omega D). \tag{30}$$

The mean value of the group-delay $\tau_x(\omega)$ is L. For a single echo, the term $-\log[1-\lambda\exp(-j\omega D)]$ in (28) can be replaced by $\log[1+\lambda\exp(-j\omega D)]$. The expression for the phase is same as in (29) and hence the group-delay for the case of echo is same as in (30).

It should be noted that the above analysis is valid only under mild echo and reverberant conditions ($\lambda \ll 1$). We have also assumed that the signal characteristics are stationary. Due to nonstationarity of speech signals, the model of reverberation in (26) may not be valid in real situations.

C. Robustness Due to Weighting of the LP Residual Signal

In this section, we show that suitable weighting of the LP residual signal improves the robustness of the algorithm for extraction of the instants of significant excitation. This is because the excitation SNR E_s can be improved by weighting, as shown below.

Consider the impulse-in-noise sequence $x(n)$ in (10). Let $\gamma(n)$, $n = -(N-1)/2, \cdots, 0, \cdots, (N-1)/2$ be a positive window function such that $\gamma(0) > \gamma(n)$, $n \neq 0$. Let

$$\begin{aligned} x_\gamma(n) &= x(n)\gamma(n-L) \\ &= \mathcal{A}\gamma(0)\delta(n-L) + \gamma(n-L)v(n), \\ &\qquad n = 0, 1, \cdots, (N-1) \end{aligned} \tag{31}$$

be the weighted excitation signal, such that the impulse at $n = L$ is given the maximal weight of $\gamma(0)$. By following the steps in the analysis presented in Section IV-A, we have

$$\phi_{x\gamma}(\omega) = -\omega L + \frac{|V_\gamma(\omega)|}{\mathcal{A}\gamma(0)}\sin(\phi_{v\gamma}(\omega) + \omega L) \tag{32}$$

where

$$\begin{aligned} V_\gamma(\omega) &= \sum_{n=0}^{N-1} v(n)\gamma(n-L)\exp(-j\omega n) \\ &= |V_\gamma(\omega)|\exp(j\phi_{v\gamma}(\omega)) \end{aligned} \tag{33}$$

and $\phi_{x\gamma}(\omega)$ is the phase of the Fourier transform of the weighted excitation sequence $x_\gamma(n)$. The approximation in (32) is justified provided that

$$\frac{\mathcal{E}[|V_\gamma(\omega)|]}{\mathcal{A}\gamma(0)} < 1.$$

Assuming that $\{v(n)\}$ are zero-mean Gaussian random variables with variance σ_v^2, we have from (33)

$$\begin{aligned} \mathcal{E}[|V_\gamma(\omega)|^2] &= \sigma_v^2 \sum_{n=0}^{N-1} \gamma^2(n-L) \\ &= \gamma^2(0)\sigma_v^2 S_\gamma \end{aligned} \tag{34}$$

where

$$S_\gamma = \sum_{n=0}^{N-1} \left[\frac{\gamma(n-L)}{\gamma(0)}\right]^2. \tag{35}$$

Following the steps in the analysis presented in Appendix B, we define the weighted excitation SNR as

$$\begin{aligned} E_w &= 10\log_{10}\left(\frac{\mathcal{A}^2\gamma^2(0)}{\mathcal{E}[|V_\gamma(\omega)|^2]}\right) \text{ dB} \\ &= 10\log_{10}\left(\frac{\mathcal{A}^2}{N\sigma_v^2}\,\frac{N}{S_\gamma}\right) \text{ dB} \end{aligned} \tag{36}$$

Using (15), we get

$$E_w = E_s + 10\log_{10}\left(\frac{N}{S_\gamma}\right). \tag{37}$$

Note that for the case without weighting of the LP residual signal, $\gamma(n) = 1$. Therefore, from (35) and (37), $E_w = E_s$. For any other window function with a broad peak around the location of the impulse i.e., $n = L$, $S_\gamma < N$. Thus, there is some gain in the excitation SNR. For the limiting case of a weight function with a narrow peak at $n = L$, the gain in the excitation SNR tends to $10\log_{10}(N)$.

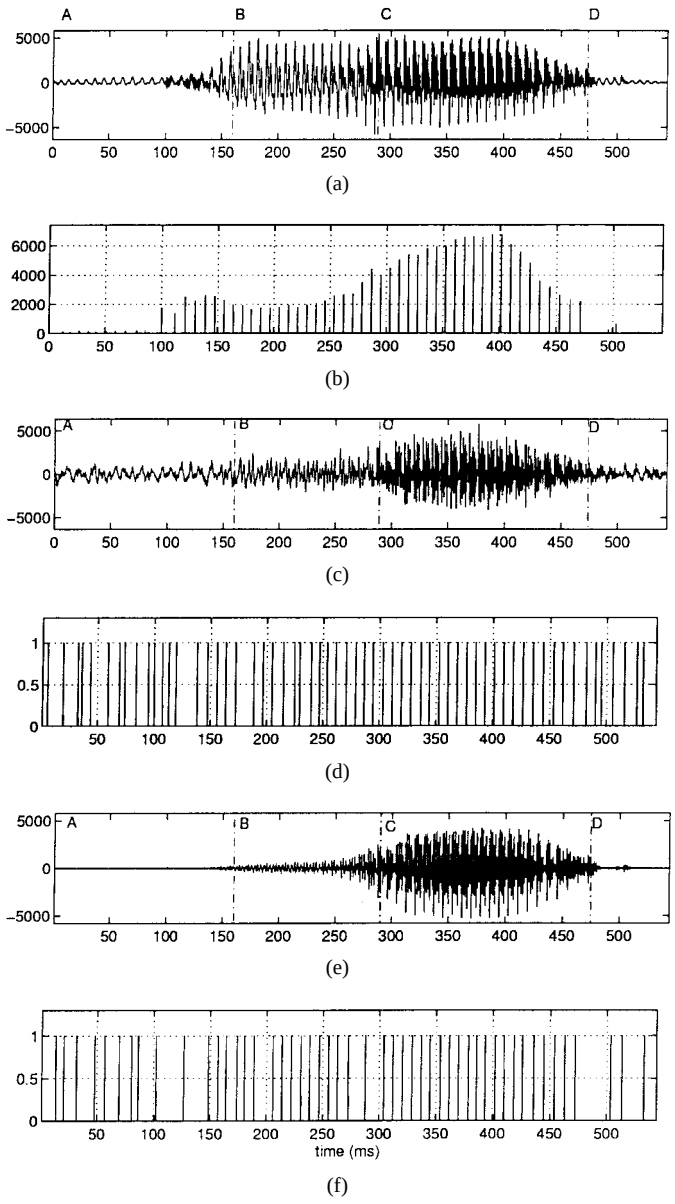


Fig. 4. (a) Clean speech for the utterance /dzua/. (b) Strengths of excitation based on the Frobenius norm. (c) Speech degraded by ambient noise. (d) Significant instants derived from the signal in (c). (e) Telephone speech. (f) Significant instants derived from the signal in (e).

V. Performance Evaluation of the Group-Delay-Based Method

In this section, we consider some examples of speech data under actual conditions of degradation, and examine the performance of the group-delay-based method for extraction of the instants of significant excitation. Since we do not have a method for estimating the SNR of the strength of excitation for signals with natural degradations, the results can only be interpreted from our *a priori* knowledge of the characteristics of the excitation for different categories of sounds. Wherever appropriate, the Frobenius norm of the differenced speech signal can be used as a measure of the strength of excitation.

Fig. 4 shows the performance of the algorithm for noise and telephone channel degradations for the segment of speech given in Fig. 1(a). The strengths of excitation at the extracted instants computed using the Frobenius norm are shown in Fig. 4(b). For this signal, the strength of excitation is lower for the segment /u/ in the region BC compared to the region CD. The noisy speech signal in Fig. 4(c) corresponds to the same speech as in Fig. 4(a), but recorded by a microphone placed 50 cm away from the speaker. The signal in the region AB is affected by the additive noise more than the signal in the region CD due to lower signal amplitudes. Hence the instants extracted for the signal in region AB are not reliable. Most of the extracted instants [Fig. 4(d)] for the signal in the region BC are correct, even though in Fig. 4(c) there appears to be no visible periodicity in the signal in the region BC. From Fig. 4(b) and (d), it can be seen that the instants are correctly extracted for the signal in the region CD. The results are similar for the case of telephone speech shown in Fig. 4(e) and (f). In the telephone speech shown in Fig. 4(e), the signal in the region AB is lost and it is significantly attenuated in the region BC. This is because the low first formant of the vowel /u/, is severely attenuated due to the bandpass nature of the telephone channel characteristics. The errors in the region AB are due to low levels of the signal itself in that region. It is important to note that although the signal level is high in the region BC for the clean speech, the strength of excitation is low for the instants in that region. Hence, the extracted instants in this region are more prone to errors compared to the extracted instants in the region CD.

A systematic investigation was carried out to study the accuracy of the extracted instants for synthetic and natural vowels. Histograms of the spread of the errors are shown in Figs. 5 and 6 for five synthetic and natural vowels (/a/, /e/, /i/, /o/ and /u/), respectively, for an overall SNR of 10 dB. All the synthetic vowels are generated by the same LF-model-based differentiated glottal pulses. The length of each pulse was chosen to be 80 samples. In the case of the natural vowels, the glottal cycle duration varied from 9 ms for vowel /a/ to 7 ms for vowel /u/. In Fig. 5, the histogram for each synthetic vowel is obtained by computing the histogram of deviations of the estimated instants of significant excitation from the true locations for 50 realizations of noise. There are ten glottal cycles in the signal for each vowel and hence we get 500 such deviations for each vowel. In Fig. 6, the deviations are obtained by subtracting the estimated locations from the locations extracted from the clean speech signal. Larger spread of the histograms indicates larger deviation of the extracted instants from the true locations of the instants. The errors are typically larger for the close vowels /u/ and /i/ than for the open vowels /a/, /e/, and /o/. For the synthetic case shown in Fig. 5, all the instants have the same strength and hence the spread of errors is less compared to the case of natural vowels. It is important to note that the variation in the spread of the errors for different vowels is also due to the artifacts of the LP analysis. For the synthetic case shown in Fig. 5, the spread is larger for the close vowels /u/ and /i/, despite the excitation strength being the same for all the five vowels, because of the dominance of the first formant in the LP analysis of the noise-corrupted signals for these close vowels. This is also true in the case of natural vowels shown in Fig. 6. There is a systematic bias in the estimated locations of the instants of excitation for

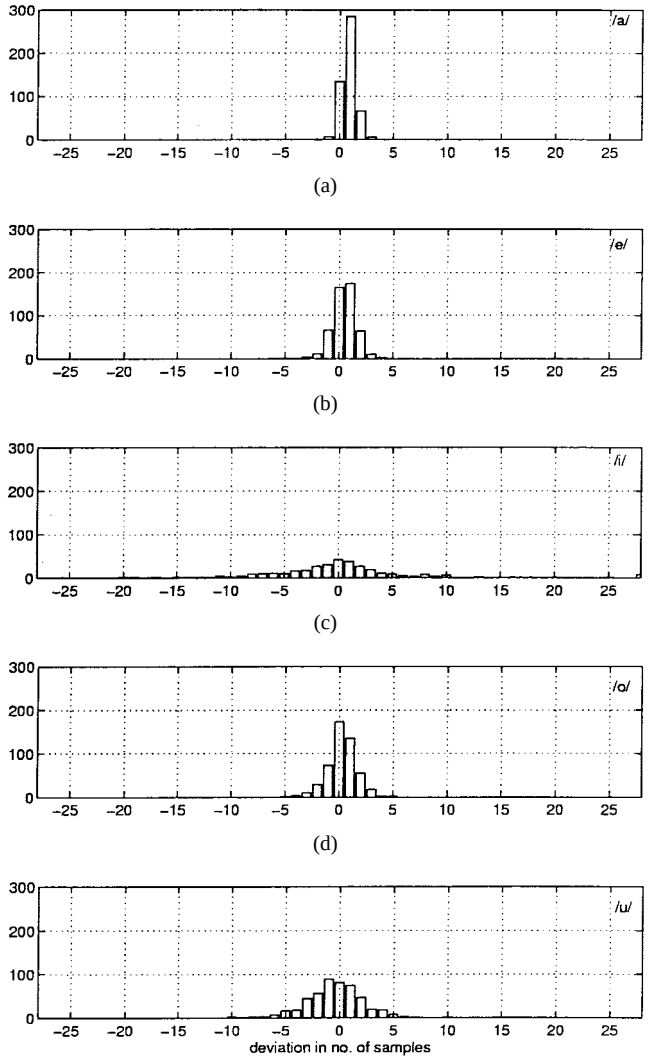


Fig. 5. Histogram of errors in the estimated instants for five synthetic vowels for SNR = 10 dB. (a) /a/, (b) /e/, (c) /i/, (d) /o/ ,(e) /u/.

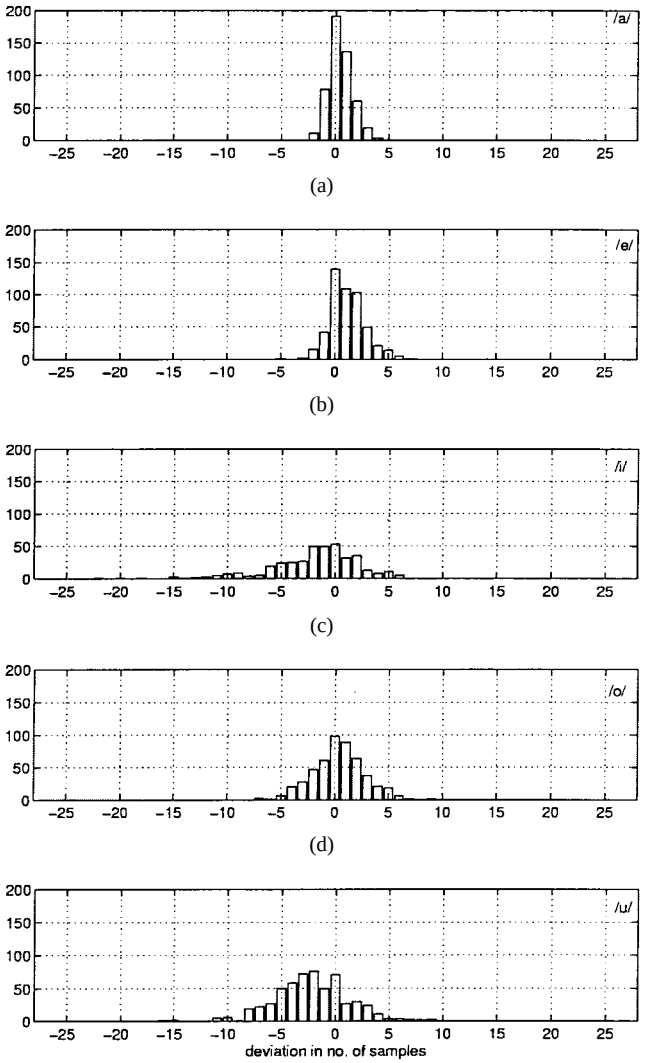


Fig. 6. Histogram of errors in the estimated instants for five natural vowels for SNR = 10 dB. (a) /a/, (b) /e/, (c) /i/, (d) /o/, (e) /u/.

the case of synthetic vowels. The bias is about two samples for the average glottal cycle length of 80 samples. That is, the bias is about 3%. The bias may have been caused due to weighting the LP residual signal before computing the instants of excitation. The weight function depends on the nature of the voiced sound, and the extent of degradation caused by noise. That is why the bias is positive in some cases and negative in some other cases.

Errors in the extracted instants were also studied for utterances taken from the standard NTIMIT [21], [22] data for male and female speech. Since the TIMIT [23] data was available for reference, the spread was estimated using the deviations of the extracted instants for the NTIMIT data from those for the TIMIT data. The TIMIT and NTIMIT data taken for study were lowpass filtered and downsampled to 8 kHz before processing. The TIMIT and NTIMIT data was first time-aligned before the deviations were computed. The histograms of errors for one male voice and one female voice are shown in Figs. 7 and 8. The data for the male voice corresponds to the file: */test/dr2/mmdm2/sa1.wav* in the TIMIT/NTIMIT database. The data for the female voice corresponds to the file: */test/dr5/fjcs0/sa1.wav*. The instants of significant excitation were extracted only from the voiced regions, which were identified using the phonetic transcription files provided with the TIMIT database. From Figs. 7 and 8, it can be seen that there are more values of deviation in the histogram of deviations for female speech than for the male speech. This is because the average pitch of the female speaker is about 210 Hz and that of the male speaker is about 100 Hz. So there are more glottal cycles in the utterance of the female speaker than for the male speaker. The spread of errors is larger for these utterances compared to the errors for the vowels in Fig. 6, because the SNR is different for different segments in this case, whereas for vowels it was constant. The speech SNR varies over a range of 20–50 dB for the utterances taken from the TIMIT data and over a range of 5–40 dB for the utterances taken from the NTIMIT data for both male and female voices. The SNR for different segments was computed as the ratio of the energy of the signal samples to the energy of the noise samples in the silence regions. The bias and spread

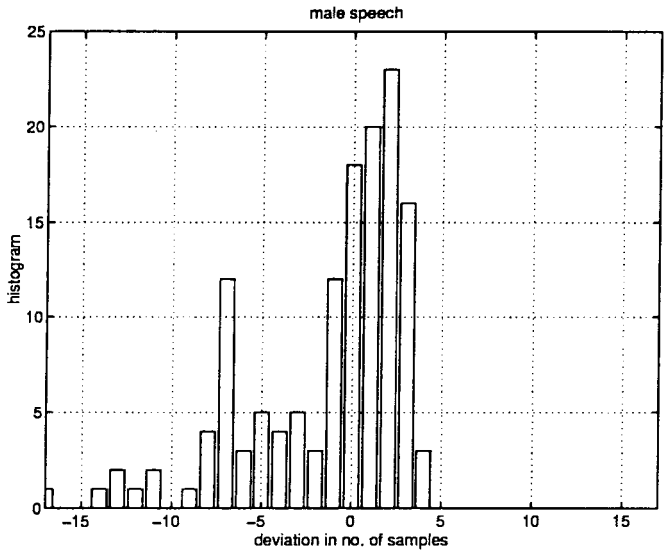


Fig. 7. Histogram of errors for the utterance "She had your dark suit in greasy wash water all year" uttered by a male speaker.

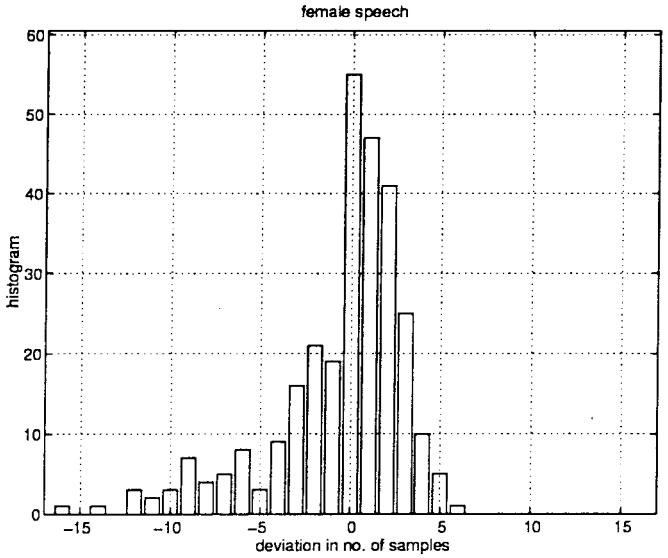


Fig. 8. Histogram of errors for the utterance "She had your dark suit in greasy wash water all year" uttered by a female speaker.

of the errors in Figs. 7 and 8 can be attributed not only to the variations of SNR for different segments, but also to the weight function used on the LP residual signal before computing the instants of excitation.

VI. Conclusions

In this paper, we have demonstrated that the group-delay-based method proposed in [9] and [10] is indeed robust against degradations in speech due to additive noise and channel distortion. The robustness is due to the fact that the energy of the signal is concentrated around the instant of significant excitation, which for voiced speech corresponds to the instant around glottal closure. We have discussed the importance of the strength of excitation, which cannot be directly inferred from the speech signal. We have shown that the errors in the extracted instants are small for many practical signals such as in the NTIMIT speech data.

Appendix A
Bounds on the Rayleigh Quotient

Let the singular value decomposition (SVD) [15] of $\boldsymbol{S}$ be

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{A.1}$$

where the columns of $\boldsymbol{U}_{(N-p)\times(N-p)}$ and $\boldsymbol{V}_{(p+1)\times(p+1)}$ are the left and right singular vectors of $\boldsymbol{S}$, respectively. $\boldsymbol{\Sigma}_{(N-p)\times(p+1)}$ is the matrix of singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{p+1} > 0$. Therefore

$$\boldsymbol{S}^T\boldsymbol{S} = \boldsymbol{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})\boldsymbol{V}^T. \tag{A.2}$$

So $\sigma_1^2 \cdots \sigma_{p+1}^2$ are the eigenvalues of $(\boldsymbol{S}^T\boldsymbol{S})$ and $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots \boldsymbol{v}_{p+1}\}$, the columns of $\boldsymbol{V}$, are its eigenvectors. The Rayleigh quotient of $(\boldsymbol{S}^T\boldsymbol{S})$ is defined as [14]

$$\rho(\boldsymbol{a}) = \frac{\boldsymbol{a}^T(\boldsymbol{S}^T\boldsymbol{S})\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} \tag{A.3}$$

where $\boldsymbol{a} \in R^{p+1}$. Assuming that the eigenvalues of $(\boldsymbol{S}^T\boldsymbol{S})$ are all distinct, its eigenvectors form an orthonormal basis in R^{p+1}. Hence, $\boldsymbol{a}$ can be expressed as

$$\boldsymbol{a} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_{p+1}\boldsymbol{v}_{p+1} \tag{A.4}$$

where $c_1, c_2, \cdots, c_{p+1}$ are the components of $\boldsymbol{a}$ w.r.t. the basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots \boldsymbol{v}_{p+1}\}$. Premultiplying both sides of (A.4) by $(\boldsymbol{S}^T\boldsymbol{S})$ and noting that σ_i^2 and $\boldsymbol{v}_i$, $i = 1 \cdots p+1$, are its eigenvalues and eigenvectors, respectively, we have

$$(\boldsymbol{S}^T\boldsymbol{S})\boldsymbol{a} = c_1\sigma_1^2\boldsymbol{v}_1 + c_2\sigma_2^2\boldsymbol{v}_2 + \cdots + c_{p+1}\sigma_{p+1}^2\boldsymbol{v}_{p+1}. \tag{A.5}$$

Premultiplying (A.5) by $\boldsymbol{a}^T$ and noting that the eigenvectors form an orthonormal set, we have

$$\boldsymbol{a}^T(\boldsymbol{S}^T\boldsymbol{S})\boldsymbol{a} = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_{p+1}^2\sigma_{p+1}^2. \tag{A.6}$$

From (A.3), (A.4), and (A.6), we have

$$\rho(\boldsymbol{a}) = \frac{(c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_{p+1}^2\sigma_{p+1}^2)}{(c_1^2 + c_2^2 + \cdots + c_{p+1}^2)}. \tag{A.7}$$

From (A.7), it is clear that

$$\sigma_{p+1}^2 \leq \rho(\boldsymbol{a}) \leq \sigma_1^2 \tag{A.8}$$

i.e., the Rayleigh quotient is bounded by the extreme eigenvalues of $(\boldsymbol{S}^T\boldsymbol{S})$.

Appendix B
Excitation Signal-to-Noise Ratio

For the zero-mean Gaussian distributed random variables $v(n)$, the Fourier transform $V(\omega)$ is a complex zero-mean Gaussian random variable. Therefore we have

$$\mathcal{E}[|V(\omega)|^2] = N\sigma_v^2. \tag{B.1}$$

Since the square of the mean is always less than the second moment, i.e.,

$$(\mathcal{E}[|V(\omega)|])^2 < \mathcal{E}[|V(\omega)|^2] \tag{B.2}$$

- Calculate the linear prediction residual signal using a frame of size 25 ms, Hamming window and a 10th order linear prediction analysis by autocorrelation method. The frame is shifted successively by 10 ms.
- Smooth the residual signal with a 0.75 ms Hamming window. Compute the short–time energy of the smoothed residual signal for every sample, over a frame of duration 1.4 ms. Normalise the short–time energy function to a maximum value of one. Multiply the residual signal obtained from the speech signal with the normalised short–time energy function, to obtain the *weighted residual signal.*
- Select a frame size between one to two periods of a glottal cycle, apply a Hamming window and compute the group–delay of the weighted residual signal at each sampling instant using the formula

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)}$$

  where $X(\omega) = X_R(\omega) + jX_I(\omega)$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ are the Fourier transforms of the weighted residual signal $x(n)$ and $y(n) = nx(n)$, respectively.
- Smooth the group–delay function computed at each sampling instant using a 3–point median filter.
- Compute the average of the smoothed group–delay function.
- The negative of the average of the group–delay function obtained at each sampling instant is plotted with time to obtain the *phase slope function.* Smooth the phase slope function using a Hamming window of length 5 samples.
- The positive zero–crossings in the smoothed phase slope function are identified as the significant instants.

Fig. 9. Algorithm for determination of instants of significant excitation.

we have

$$\mathcal{E}[|V(\omega)|] < (N\sigma_v^2)^{1/2}. \tag{B.3}$$

Hence

$$\frac{\mathcal{E}[|V(\omega)|]}{\mathcal{A}} < 10^{-(E_s/20)} \tag{B.4}$$

where E_s is the excitation SNR:

$$E_s = 10\log_{10}\left(\frac{\mathcal{A}^2}{N\sigma_v^2}\right) \quad \text{dB} \tag{B.5}$$

Let us consider an N-point discrete Fourier transform (DFT) of the sequence given in (10), computed at $\omega_k = (2\pi k/N)$, $k = 0, \cdots, N-1$. It can be shown [24] that the real and imaginary parts of the DFT of $v(n)$, $V_R(\omega_k)$ and $V_I(\omega_k)$, are (real) independent identically distributed (i.i.d.) Gaussian random variables for $k = 1, 2, \cdots, ((N/2)-1)$. Therefore, the vectors $\boldsymbol{v}_R = [V_R(\omega_1)\ V_R(\omega_2) \cdots V_R(\omega_{(N/2)-1})]^T$ and $\boldsymbol{v}_I = [V_I(\omega_1)\ V_I(\omega_2) \cdots V_I(\omega_{(N/2)-1})]^T$ are $\sim \mathcal{N}(\mathbf{0}, (N\sigma_v^2/2)\boldsymbol{I})$. Under these conditions the magnitude of the DFT of $v(n)$, $|V(\omega_k)| = [V_R^2(\omega_k) + V_I^2(\omega_k)]^{1/2}$, is Rayleigh distributed [24]. Since we have the knowledge of both the mean and variance of $|V(\omega_k)|$, we get

$$\begin{aligned}\frac{\mathcal{E}[|V(\omega_k)|]}{\mathcal{A}} &= \frac{\sqrt{\pi}}{2}10^{-(E_s/20)} \\ &\approx (0.9)10^{-(E_s/20)}\end{aligned} \tag{B.6}$$

which is indeed close to the upper bound $10^{-(E_s/20)}$ given in (B.4) above. From the cumulative distribution function of a Rayleigh distribution [25], we may write

$$\begin{aligned}P[|V(\omega_k)| < \mathcal{A}] &= 1 - \exp\left(-\frac{\mathcal{A}^2}{N\sigma_v^2}\right) \\ &= 1 - \exp(-10^{(E_s/10)})\end{aligned} \tag{B.7}$$

where $P[|V(\omega_k)| < \mathcal{A}]$ is the probability that $|V(\omega_k)|$ is less than $\mathcal{A}$. From (B.7), we note that $|V(\omega_k)| < \mathcal{A}$ with more than 99% confidence, when $E_s \geq 6.6$ dB.

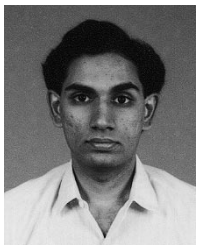
ACKNOWLEDGMENT

The authors would like to thank Dr. H. A. Murthy for providing the data required for some of the studies in this

paper, and the three anonymous reviewers for their critical comments, which greatly helped improve the presentation of the paper.

REFERENCES

[1] K. S. Nathan, Y.-T. Lee, and H. F. Silverman, "A time-varying analysis method for rapid transitions in speech," *IEEE Trans. Signal Processing*, vol. 39, pp. 815–824, Apr. 1991.
[2] A. K. Krishnamurthy, "Glottal source estimation using a sum-of-exponentials model," *IEEE Trans. Signal Processing*, vol. 40, pp. 682–686, Mar. 1992.
[3] C. Hamon, E. Moulines, and F. J. Charpentier, "A diphone synthesis system based on time domain prosodic modifications of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, U.K, May 1989, pp. 238–241.
[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, pp. 309–319, Aug. 1979.
[5] H. W. Strube, "Determination of the instant of glottal closure," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1625–1629, 1974.
[6] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 562–570, Dec. 1975.
[7] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1805–1814, Dec. 1989.
[8] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 258–265, Apr. 1994.
[9] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay functions," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 325–333, Sept. 1995.
[10] B. Yegnanarayana and R. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, MI, May 1995, pp. 776–779.
[11] E. A. Robinson, T. S. Durrani, and L. G. Peardon, *Geophysical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
[12] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[13] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1983.
[15] S. J. Leon, *Linear Algebra with Applications*. New York: Macmillan, 1990.
[16] G. Fant, "Glottal flow: Models and interaction," *J. Phonet.*, vol. 14, pp. 393–399, Oct.–Dec. 1986.
[17] X. Li and N. M. Bilgutay, "Wiener filter realization for target detection using group delay statistics," *IEEE Trans. Signal Processing*, vol. 41, pp. 2067–2074, June 1993.
[18] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Processing*, vol. 40, pp. 2281–2289, Sept. 1992.
[19] S. M. Kay, *Modern Spectral Estimation—Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[20] R. C. Kemeriat and D. G. Childers, "Signal detection and extraction by cepstrum techniques," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 745–759, Nov. 1972.
[21] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, Albuquerque, NM, Apr. 1990, pp. 109–112.
[22] C. Jankowski, "The NTIMIT speech database," from documentation accompanying the NTIMIT CD-ROM, Nynex Sci. Technol. Ctr., White Plains, NY, Jan. 1991.
[23] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, Feb. 1986, pp. 93–99.
[24] S. M. Kay, *Fundamentals of Statistical Signal Processing—Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.



P. Satyanarayana Murthy was born in Kakinada, India, in 1971. He received the B.E. degree in electronics and communication engineering from Chaitanya Bharathi Institute of Technology, Osmania University, Hyderabad, the M.Tech. and Ph.D. degrees in electrical engineering from the Indian Institute of Technology (IIT), Madras, in 1994 and 1999, respectively.

From January to July 1994, he was a Senior Project Officer in the Department of Computer Science and Engineering, IIT. He is currently a Manager with Speech and Software Technologies, Madras. His research interest is in speech signal processing.

B. Yegnanarayana (M'78–SM'84) was born in India on January 9, 1944. He received the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978 in the Department of Electrical Communication Engineering, Indian Institute of Science. From 1966 to 1971, he was engaged in the development of environmental test facilities for the Acoustic Laboratory, Indian Institute of Science. From 1977 to 1980, he was a visiting Associate Professor of computer science at Carnegie Mellon University, Pittsburgh, PA. He was a Visiting Scientist at ISRO Satellite Center, Bangalore, from July to December 1980. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology, Madras. He was a Visiting Professor at the Institute for Perception Research, Eindhoven Technical University, Eindhoven, The Netherlands, from July 1994 to January 1995. Since 1972, he has been working on problems in the area of speech signal processing. He is presently engaged in research activities in digital signal processing, speech recognition, and neural networks.

Dr. Yegnanarayana is a member of the Computer Society of India, a Fellow of the Institution of Electronics and Telecommunications Engineers of India, a Fellow of the Indian National Science Academy, and a Fellow of the Indian National Academy of Engineering.