Determination of Instants of Significant Excitation in Speech Using Hilbert Envelope and Group Delay Function

K. Sreenivasa Rao, *Member, IEEE*, S. R. Mahadeva Prasanna, *Member, IEEE*, and B. Yegnanarayana, *Senior Member, IEEE*

***Abstract*—This letter proposes a time-effective method for determining the instants of significant excitation in speech signals. The instants of significant excitation correspond to the instants of glottal closure (epochs) in the case of voiced speech, and to some random excitations like onset of burst in the case of nonvoiced speech. The proposed method consists of two phases: the first phase determines the approximate epoch locations using the Hilbert envelope of the linear prediction residual of the speech signal. The second phase determines the accurate locations of the instants of significant excitation by computing the group delay around the approximate epoch locations derived from the first phase. The accuracy in determining the instants of significant excitation and the time complexity of the proposed method is compared with the group delay based approach.**

***Index Terms*—Group delay function, Hilbert envelope, instants of significant excitation, linear prediction residual.**

I. Introduction

VOICED speech is produced as a result of excitation of the vocal tract system by a quasiperiodic sequence of glottal pulses. The major excitation of the vocal tract system within a pitch period takes place at the instant of glottal closure (GC) [1]. These instants are termed as instants of significant excitation (epochs) and can be automatically determined from a speech signal using the negative derivative of the unwrapped phase (group delay) function of the short-time Fourier transform of the signal [2], [3]. Though group delay based approach provides the accurate epoch locations, the approach is computationally intensive. In this letter, we propose a time-effective approach to determine the instants of significant excitation using Hilbert envelope (HE) of the linear prediction (LP) residual and group delay function.

Many of the speech analysis techniques depend on the accurate estimation of the instant of GC within a pitch period. For example, if such instances are known, the closed glottis region can be identified, and the vocal tract parameters such as formants may be derived accurately by confining the analysis to only those regions [4]. It is also possible to determine the characteristics of the voice source by careful analysis of the signal with the help of GC instants [1]. For some of the real-time applications such as text-to-speech (TTS) synthesis, voice conversion, and varying speech rate, we need to compute the information about the instants of significant excitation at a faster rate and use it for the specific applications [5]–[7]. For instance, in the TTS application, it is necessary to modify the durations and pitch contours of the basic units and words in order to incorporate the suprasegmental knowledge of an utterance containing a sequence of basic units [8].

The approximate locations of the instants of significant excitation can be derived by exploiting the unipolar nature of HE [9]. The strength of excitation in voiced speech is high around the GC instant. The impulse-like excitation results in large error in the LP residual around the GC instant. The region around the GC instant corresponds to the high energy portion of the excitation within a pitch period. However, it is difficult to determine the location of the GC instant due to bipolar fluctuations of the amplitudes of the residual samples around the instant of GC. Ideally, it is desirable to derive an impulse-like signal at the GC instant. A close approximation to this is possible by using the HE of the LP residual [1]. The accurate instants can be further determined by computing the group delay for the samples around the approximate locations given by HE of the LP residual, hence the motivation for the work.

In this letter, we present a method for determining the instants of significant excitation using the properties of HE and group delay function. In Section II, we discuss the method based on group delay function to compute the instants of significant excitation in speech. The proposed method to determine the instants of significant excitation is described in Section III. In Section IV, the accuracy in determining the locations of the epochs and time complexity of the proposed method is evaluated against the group delay based approach. In Section V, summary and possible extensions of the work are given.

II. Group Delay Based Method for Determining the Instants of Significant Excitation

The method is based on the global phase characteristics of minimum phase signals [2], [3]. Since the average group delay of a minimum phase system is zero, 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 [2]. The speech signal need not be a minimum phase

Manuscript received September 1, 2006; revised February 19, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Frederic Bimbot.

K. Sreenivasa Rao and S. R. Mahadeva Prasanna are with the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781 039, Assam, India (e-mail: ksrao@iitg.ernet.in; prasanna@iitg.ernet.in).

B. Yegnanarayana is with the International Institute of Information Technology (IIIT), Gachibowli, Hyderabad-500 032, Andhra Pradesh, India (e-mail: yegna@iiit.ac.in).

Digital Object Identifier 10.1109/LSP.2007.896454

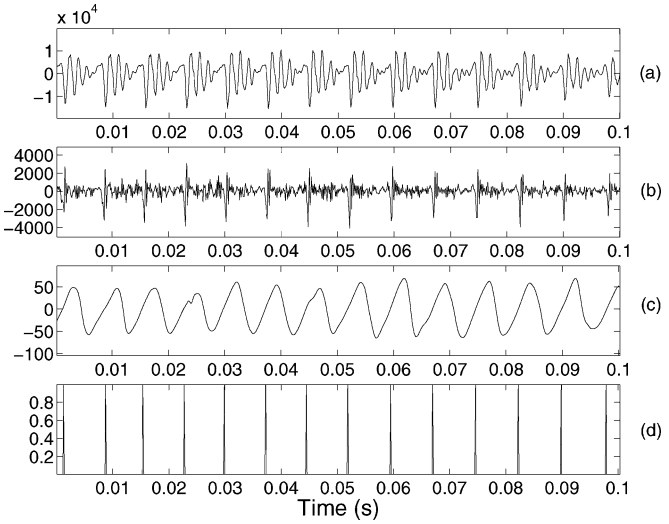

Fig. 1. (a) Segment of voiced speech, (b) LP residual, (c) phase slope function, and (d) instants of significant excitation.

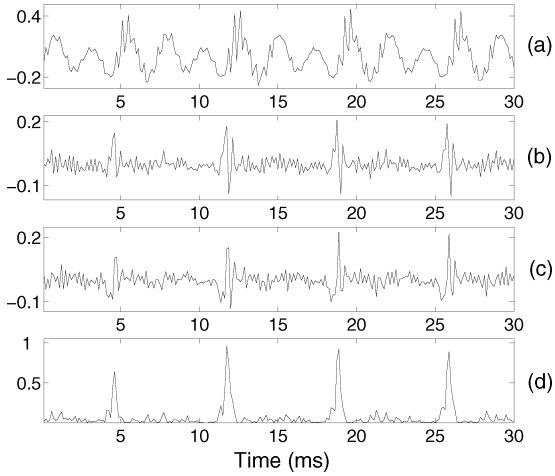

Fig. 2. (a) Segment of voiced speech, (b) LP residual, (c) Hilbert transform of the LP residual, and (d) HE of the LP residual.

signal always, whereas LP residual signal will be a minimum phase signal [2]. Hence, 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 residual signal is derived by inverse filtering the speech signal, and the inverse filter is obtained using LP analysis. LP analysis is performed using 10th order, with a frame size of 20 ms and frame shift of 10 ms. The instants of significant excitation can be derived from the LP residual signal as follows: Around each sample, a 10-ms segment of the LP residual signal is considered, and the group delay function $\tau(\omega)$ is computed using [9]

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

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ are the Fourier transforms of the windowed residual signal $x(n)$ and $nx(n)$, 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* [2]. 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. Fig. 1 shows a segment of voiced speech, the LP residual, the phase slope function, and the instants of significant excitation.

III. Proposed Method for Determining the Instants of Significant Excitation

Determining the instants of significant excitation using the group delay based method is a computationally intensive process, since the group delay is computed for every sample shift. The computational complexity can be reduced by computing the group delay only for few samples around the instants of GC. This is achieved by first detecting the approximate locations of the GC instants. The peaks in the HE of the linear prediction residual indicate the approximate locations of the GC instants [1], [10]. Even though the real and imaginary parts of an analytic signal (related through the Hilbert transform) have positive and negative samples, the HE of the signal is a positive function, giving the envelope of the signal [9]. Thus, the properties of HE can be exploited to derive the impulse-like characteristics of the GC events. The HE $h_e(n)$ of the LP residual $e(n)$ is defined as follows [9]:

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$ and is given by [1]

$$e_h(n) = IDFT[E_h(k)], \text{ where}$$
$$E_h(k) = \begin{cases} -jE(k), & k = 0, 1, \ldots, \left(\frac{N}{2}\right) - 1 \\ jE(k), & k = \left(\frac{N}{2}\right), \left(\frac{N}{2}\right) + 1, \ldots, (N-1). \end{cases}$$

Here IDFT denotes the inverse discrete Fourier transform, and $E(k)$ is the discrete Fourier transform of $e(n)$. Fig. 2 shows a segment of voiced speech, its LP residual, Hilbert transform, and the HE. The major peaks in the HE indicate approximate locations of epochs. The evidence of GC instants is obtained by convolving the HE with a Gabor filter (modulated Gaussian pulse) given by $g(n) = (1/\sqrt{2\pi}\sigma)e^{(-(n-N/2)^2/2\sigma^2 + j\omega n)}$, where σ defines the spatial spread of the Gaussian, ω is the frequency of modulating sinusoid, n is the time index varying from 1 to N, and N is the length of the filter [11]. The Gabor filter used in this study is shown in Fig. 3. The Hilbert envelope of the LP residual is convolved with the Gabor filter shown in Fig. 3 to obtain the plot of evidence shown in Fig. 4, which is termed as *GC Evidence Plot* in Fig. 4(c). In the GC evidence plot, the instants of positive zero-crossings correspond to approximate locations of the instants of significant excitation. To determine the accurate locations of the GC instants, the phase slope function is computed for the residual samples around the

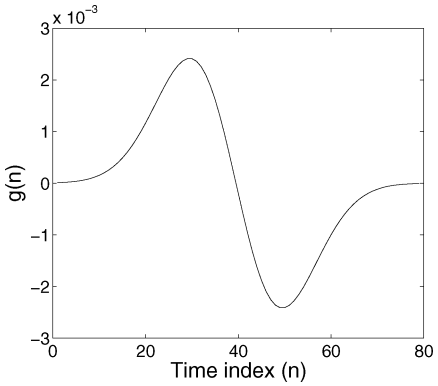

Fig. 3. Gabor window for $\sigma = 10$, $\omega = 0.1175$, and $N = 80$.

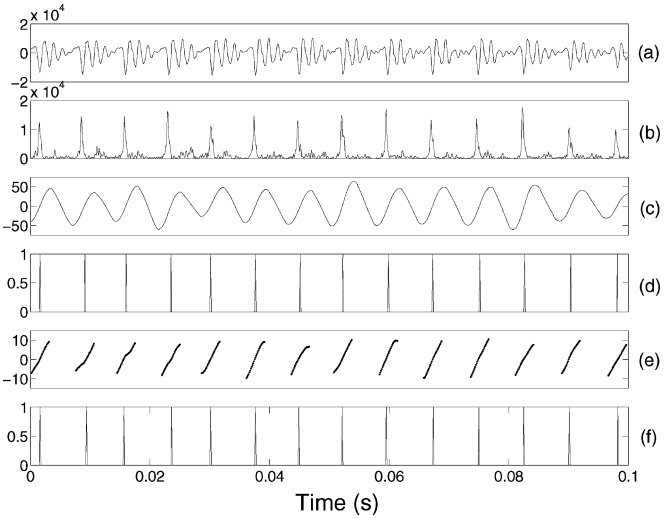

Fig. 4. (a) Segment of voiced speech, (b) HE of the LP residual, (c) GC instant evidence plot, (d) approximate GC instant locations, (e) phase slope function, and (f) accurate locations of the instants of significant excitation.

approximate GC instant locations. The positive zero-crossings of the phase slope function correspond to accurate locations of the instants of significant excitation. Fig. 4 shows a segment of voiced speech, the HE of the LP residual of a speech segment, the GC instant evidence plot, approximate locations of GC instants, phase slope function, and the locations of the instants of significant excitation. The proposed method is summarized in Table I.

IV. Evaluation of the Proposed Method

The computational efficiency of the proposed method depends on the number of approximate epoch locations derived from the HE of the LP residual and the number of samples considered around each GC instant. For evaluating the performance of the proposed method, 100 speech utterances, each of duration of 3 s, are considered. Among the utterances, 50 are uttered by male speakers and 50 are uttered by female speakers. For each utterance, the instants of significant excitation are computed by the proposed method using different window sizes (number of samples around the approximate instant location). The epochs determined by the standard group delay method are used as reference epochs [2]. Table II shows the number of instant locations derived by the proposed method for different

TABLE I
Steps for Automatic Detection of the Instants of Significant Excitation Using HE of LP Residual and Group Delay Function

1. Preemphasize the input speech.
2. Compute LP residual with 10^{th} order LP analysis, frame size of 20 ms and shift of 10 ms.
3. Compute Hilbert envelope of the LP residual,Enhance the peaks in Hilbert envelope by dividing each sample of the HE with the running mean around that sample. (Sample amplitude to running mean ratio is high around the GC instants (peaks of HE) and low at other places.)
4. Obtain the GC instant evidence plot by convolving the enhanced HE with the Gabor filter.
5. Find the positive zero-crossing locations in the GC instant evidence plot, which are hypothesized as the approximate locations of the instants of significant excitation.
6. Compute the group delay for the samples within the window around the approximate GC instant locations. Size of the window can be chosen based on the trade-off between computational efficiency and accuracy in determining the instant locations.
7. Derive the phase slope function from group delay values, and smooth it with a Hamming window.
8. Identify positive zero-crossings of the phase slope as the accurate instants of significant excitation.

TABLE II
Number of Instants Derived Using the Proposed Method for Different Window Sizes

Window size (ms)	Male speakers		Female speakers	
	# instants	% instants	# instants	% instants
0.5	7813	63.08	13510	67.17
1.0	11207	90.49	18792	93.43
1.5	11865	95.80	19644	97.67
2.0	12031	97.14	19775	98.32
2.5	12142	98.04	19883	98.86
3.0	12226	98.72	19940	99.14
3.5	12284	99.18	19974	99.31
4.0	12308	99.38	20020	99.54

window sizes. The total number of instants derived from the utterances of male speakers and female speakers are 12 385 and 20 113, respectively, by using the group delay method. The total number of approximate instant locations from the utterances of male speakers and female speakers, using the HE of the LP residual, is 12 867 and 20 926, respectively. The analysis shows that with a window size of 2 ms, about 97% of the GC instants are detected accurately for male speakers, and for female speakers, about 98% of the GC instants are detected accurately (see Table II). For instance, time complexity analysis in the case of male speakers indicates that for a window size of 2 ms, the proposed method determines the instants of significant excitation approximately in one fourth of the time compared to the group delay method (assuming that the average pitch period for male speakers is 8 ms). It is observed that when the size of the window is small, the computational efficiency is high, but at the same time, some of the epochs will be missing. As

TABLE III
NUMBER OF APPROXIMATE INSTANTS DERIVED FROM HE FOR DIFFERENT DEVIATIONS WITH RESPECT TO REFERENCE INSTANT LOCATIONS

Deviation	Male speakers		Female speakers	
# samples	# instants	% instants	# instants	% instants
0	2672	21.57	4574	22.74
1	3076	24.84	4745	23.59
2	2079	16.79	4198	20.87
3	2245	18.13	3260	16.21
4	1145	9.26	2037	10.13
5	537	4.34	526	2.62

the size of the window increases, the computational efficiency decreases, but the number of missing epochs also decreases.

The deviation in the approximate epoch locations with respect to their reference locations is computed. The results of the analysis are given in Table III. The entries in Table III indicate the number of approximate instants and their deviation in terms of number of samples with respect to reference instants. On the whole, the average deviation per instant is found to be 2.1 samples (0.26 ms) and 1.7 samples (0.21 ms) for male and female speakers utterances, respectively.

V. SUMMARY AND CONCLUSIONS

The proposed method for determining the instants of significant excitation provides the time-effective solution, which is more appropriate for real-time applications. The method first derives the approximate locations of the instants of significant excitation using HE of the LP residual of the speech signals, and then, the accurate locations of the instants of significant excitation are derived by using the group delay function of the samples around the approximate instant locations. The effect of the number of samples (window size) considered for group delay analysis in deriving the accurate locations of the instants of significant excitation is analyzed. The amount of deviation present in the approximate instant locations derived by HE of the LP residual is also discussed. Since we have a time-effective method for computing the instants of significant excitation, the effectiveness of the same may be verified in application like prosody modification for TTS synthesis and voice conversion.

REFERENCES

[1] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.

[2] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[3] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant excitation from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 609–619, Nov. 1999.

[4] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 313–327, Jul. 1998.

[5] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 453–467, Dec. 1990.

[6] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Commun.*, vol. 8, pp. 147–158, Jun. 1989.

[7] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 972–980, May 2006.

[8] B. Yegnanarayana, S. Rajendran, V. R. Ramachandran, and A. S. M. Kumar, "Significance of knowledge sources for TTS system for Indian languages," *SADHANA Acad. Proc. Eng. Sci.*, vol. 19, pp. 147–169, Feb. 1994.

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

[10] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Orlando, FL, May 2002, vol. 1, pp. 541–544.

[11] D. Gabor, "Theory of communication," *J. Inst. Elect. Eng.*, vol. 93, no. 2, pp. 429–457, 1946.