Detection of instants of glottal closure using characteristics of excitation source

S. Guruprasad [1], *B. Yegnanarayana* [2], *and K. Sri Rama Murty*[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Madras, India
[2]International Institute of Information Technology Hyderabad, India
guru@cs.iitm.ernet.in, yegna@iiit.ac.in, ksrm@cs.iitm.ernet.in

Abstract

In this paper, we propose a method for detection of glottal closure instants (GCI) in the voiced regions of speech signals. The method is based on periodicity of significant excitations of the vocal tract system. The key idea is the computation of coherent covariance sequence, which overcomes the effect of dynamic range of the excitation source signal, while preserving the locations of significant excitations. The Hilbert envelope of linear prediction residual is used as an estimate of the source of excitation of the vocal tract system. Performance of the proposed method is evaluated in terms of the deviation between true GCIs and hypothesized GCIs, using clean speech and degraded speech signals. The signal-to-noise ratio (SNR) of speech signals in the vicinity of GCIs has significant bearing on the performance of the proposed method. The proposed method is accurate and robust for detection of GCIs, even in the presence of degradations.

Index Terms: glottal closure instants, excitation source, periodicity, coherent covariance sequence

1. Introduction

Voiced speech is produced by exciting a time-varying vocal tract system with a sequence of impulse-like excitations. The impulse-like excitation is due to the closure of glottis during the vibration of vocal folds. The time instant at which the closure is achieved (called glottal closure instant or GCI) is an important feature for analysis of speech signals. Detection of GCIs enables the identification of region of closed glottis within a pitch period. Analysis of short segments of speech signals over such regions helps in accurate estimation of vocal tract parameters such as formants [1], and also in the extraction of characteristics of voice source. In text-to-speech synthesis, accurate detection of GCIs is necessary for prosodic manipulation of speech sounds [2]. Moreover, speech signal in the vicinity of GCIs has relatively high signal-to-noise ratio (SNR), due to impulse-like excitation and damped sinusoid-like impulse response of the vocal tract system. These regions of high SNR are likely to preserve features specific to sound and speaker, even under the influence of degradations. Hence, methods for robust detection of GCIs in speech signals are necessary.

Some of the methods proposed for the detection of GCI assume a linear source-system model for the production of speech signal. These methods identify GCI with the time instant of strongest excitation which will be around the region with least predictability [3, 4, 5]. Normally, linear prediction (LP) residual is used as an estimate of source of excitation [5]. Another class of methods for detection of GCIs is based on the properties of minimum phase signals and group delay functions. In [6], the average slope of the unwrapped phase spectrum of speech signal is computed as a function of time, and the positive zero crossings of the phase slope function are hypothesized as the instants of glottal closure. In [7], the phase spectrum is computed from the LP residual instead of speech signal to reduce the effects of truncation. Robustness of the group delay based methods against noise and distortion is studied in [8]. In [9], properties of the phase slope function are used to hypothesize candidates for GCI, which are validated using a dynamic programming approach. Energy-weighted group delay is proposed as a measure for the detection of GCIs in [10].

In this paper, we propose a method for detection of instants of glottal closure using the periodicity of significant excitations in speech signals. In Section 2, we describe the representation of excitation source in terms of the Hilbert envelope of the linear prediction residual. The section describes the proposed method for detection of GCI, and also the issues involved in the choice of parameters used in the method. Section 3 discusses the experiments conducted for evaluating the performance of the proposed method, and the results of these studies. Conclusions are given in Section 4.

2. Proposed method for detection of GCI

We first describe a method to extract the significant excitations in speech signal. Then, an algorithm for determining the time instants of these excitations is proposed.

2.1. Representation of excitation source

Linear prediction (LP) residual can be used as an estimate of the source of excitation. Linear prediction analysis [11] of voiced sounds results in multiple peaks of either polarity in the LP residual, around the instants of glottal closure. This is because the resulting digital inverse filter does not exactly compensate the phase of the vocal tract system. The difficulty in the detection of significant excitations from the LP residual, arising out of peaks of either polarity, can be resolved by deriving the Hilbert envelope of the LP residual [5]. The Hilbert envelope emphasizes abrupt positive-to-negative and negative-to-positive excursions in the LP residual. This approach is more accurate in preserving significant excitations, compared to the computation of the residual energy in a short interval of time. The Hilbert envelope $x[n]$ of the LP residual $r[n]$ is given by

$$x[n] = \sqrt{r^2[n] + r_H^2[n]}, \quad (1)$$

where $r_H[n]$ denotes the Hilbert transform of $r[n]$.

The emphasis of significant excitations in the Hilbert envelope relative to LP residual is shown in Fig. 1. It has been shown that the locations of the significant excitations, as observed from the Hilbert envelope of the LP residual, are close to GCIs [5]. While the significant excitations are prominent in

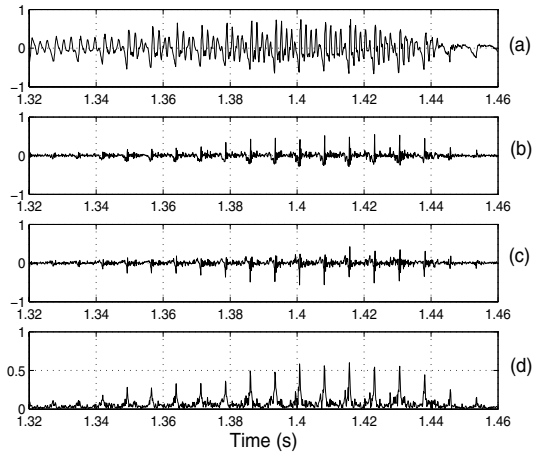

Figure 1: *Emphasis of significant excitations. (a) Waveform of a segment of voiced speech, (b) its linear prediction (LP) residual, (c) Hilbert transform of LP residual and (d) Hilbert envelope of LP residual. The LP residual was derived from 10th order LP analysis.*

the Hilbert envelope, automatic detection of their locations is a nontrivial task.

2.2. Computation of coherent covariance sequence

Periodicity of significant excitations can be exploited to derive a signal which is invariant to the dynamic range of the Hilbert envelope. This is described below. Let us consider a segment of the Hilbert envelope of the LP residual, of length N samples, starting at time index n. The mean of the samples in the segment is removed from each sample. Such a sequence is denoted as $X = \{x[n], x[n+1], \ldots, x[n+N-1]\}$. Let a vector $\mathbf{x}_{n,k}$ of dimension L be defined as

$$\mathbf{x}_{n,k} = [x[n+k] \quad \ldots \quad x[n+k+L-1]]^{\mathsf{T}}, \qquad (2)$$

where $0 \leq k \leq N-L$ and T denotes the transpose operator. Thus, contiguous subsequences of length L within the sequence X can be viewed as vectors. For the sequence X, the covariance sequence (vector) is defined as

$$\begin{aligned} \mathbf{\Phi}_n &= [\phi[n;0] \quad \phi[n;1] \quad \ldots \quad \phi[n;N-L]]^{\mathsf{T}}, \text{ where} \\ \phi[n;k] &= \frac{\mathbf{x}_{n,0}{}^{\mathsf{T}}\mathbf{x}_{n,k}}{||\mathbf{x}_{n,0}||||\mathbf{x}_{n,k}||}, \qquad 0 \leq k \leq N-L. \end{aligned} \qquad (3)$$

When the sequence X is periodic, peaks in the covariance sequence $\mathbf{\Phi}_n$ occur at an interval which is equal to the interval between the peaks in X. However, the peaks in $\mathbf{\Phi}_n$ are not aligned with those in X, as shown in Fig. 2(b). Hence, another covariance sequence, denoted by $\mathbf{\Psi}_n$, is computed using the sequences $\mathbf{\Phi}_n$ and X as follows:

$$\begin{aligned} \mathbf{\Psi}_n &= [\psi[n;0] \quad \psi[n;1] \quad \ldots \quad \psi[n;N-L]]^{\mathsf{T}}, \text{ where} \\ \psi[n;k] &= \frac{\mathbf{\Phi}_n{}^{\mathsf{T}}\mathbf{x}_{n,k}}{||\mathbf{\Phi}_n||||\mathbf{x}_{n,k}||}, \qquad 0 \leq k \leq N-L. \end{aligned} \qquad (4)$$

For the innerproduct in (4), the dimensions of $\mathbf{\Phi}_n$ and $\mathbf{x}_{n,k}$ should be same, i.e., $N-L = L-1$. We have chosen $N = 239$

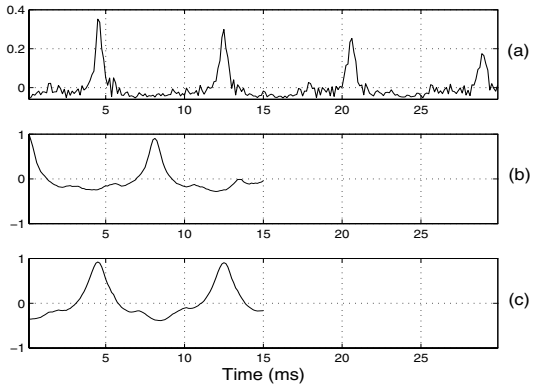

Figure 2: *(a) A sequence X of Hilbert envelope signal, (b) covariance sequence $\mathbf{\Phi}_n$ computed from the samples of X, and (c) coherent covariance sequence $\mathbf{\Psi}_n$ computed from X and $\mathbf{\Phi}_n$. The peaks in (c) are aligned with those in (a). $\mathbf{\Phi}_n$ and $\mathbf{\Psi}_n$ have durations of 15 ms only.*

and $L = 120$, corresponding to 30 ms and 15 ms, respectively, at a sampling frequency of 8 kHz. Fig. 2(c) shows the sequence $\mathbf{\Psi}_n$ in which the significant peaks of $\mathbf{\Phi}_n$ are aligned with those of X. Hence we call $\mathbf{\Psi}_n$ as *coherent covariance sequence*. The sequence $\mathbf{\Psi}_n$ can be updated over successive segments of speech, as shown in Fig. 3(c).

The computation of coherent covariance sequence using unit normalized vectors helps in overcoming the dynamic range of the Hilbert envelope. The resulting coherent covariance sequence is a smooth function which enables the detection of significant peaks more easily than the sequence X. The effect of spurious peaks in the Hilbert envelope is reduced, since these spurious peaks do not contribute to the periodicity. The peaks in the coherent covariance sequence are hypothesized as the instants of glottal closure. Fig. 3 shows a segment of voiced speech, where the locations of peaks in the coherent covariance sequence (Fig. 3(c)) are in agreement with the true GCIs obtained from the differential of the electroglottograph (EGG) signal (Fig. 3(d)).

2.3. Choice of analysis parameters

A 10$^{\text{th}}$ order linear prediction analysis is performed on speech segments of 10 ms duration, with an overlap of 5 ms between successive segments. The Hilbert envelope of the LP residual is weighted using the function derived by coherent addition of segments of speech signal. This will emphasize the regions of high SNR while reducing the effect of noise. The weighted Hilbert envelope is then used for computation of the coherent covariance sequence. The length N of the segment used for computation of the coherent covariance sequence is chosen so as to contain at least two complete pitch periods. The choice of N corresponding to 30 ms is based on the assumption that the maximum pitch period does not exceed 15 ms. For female speakers, a window of 30 ms may contain several pitch periods. This may lead to errors in the locations of the GCIs due to excessive averaging. Hence, a segment of smaller length may be chosen for analysis of female voices. While updating the coherent covariance sequence, a shift of 5 ms is used to track the changes in the gross envelope of the Hilbert envelope.

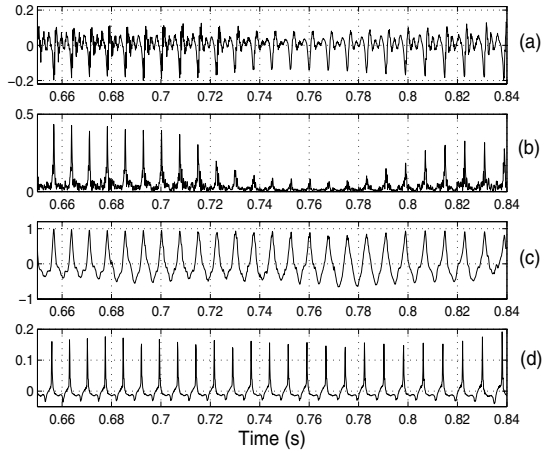

Figure 3: *(a) A segment of voiced speech, (b) Hilbert envelope of LP residual, (c) coherent covariance sequence and (d) differential of EGG signal.*

3. Experimental studies

3.1. Speech data

The accuracy of the GCIs hypothesized by the proposed method is evaluated using speech signals and their corresponding EGG signals from ARCTIC speech corpus [12]. Speech signals were sampled at 8 kHz. The chosen set contains 300 sentences from two male speakers and one female speaker. Only voiced regions longer than five pitch cycles were considered throughout the study, resulting in a total of 65094 true GCIs which were detected using EGG signals. This set is called Set I. Another set, called Set II, is obtained by retaining only those GCIs from Set I for which the short-term energy of the corresponding speech signals is above a threshold. The performance of the proposed method is evaluated for clean speech, and for speech signals corrupted by additive white Gaussian noise to obtain signals with different levels of overall signal-to-noise ratio (SNR). Also, speech signals were played through a loudspeaker, and the signals were collected by placing microphones at distances of 3 ft, 4 ft, 5 ft and 6 ft. The recordings were done in a laboratory environment, and the collected signals were time-synchronized with the played signals. The objective in this case is to observe the performance of the method in a practical case.

3.2. Performance measures

The method described in Section 2.2 is used to compute the coherent covariance sequence, the peaks of which are hypothesized as GCIs. For each true GCI, if the closest hypothesized GCI lies within a time interval of 3 ms from the true GCI, then the two are associated with each other. The true GCIs which have not been associated with any hypothesized GCI are assigned to the set $\mathcal{S}_m$ of missing GCIs. Similarly, those hypothesized GCIs which have not been associated with any true GCIs are assigned to the set $\mathcal{S}_f$ of false alarms. Let N_t denote the number of true GCIs. The missing rate η_m and the false alarm rate η_f are given by

$$\eta_m = \frac{|\mathcal{S}_m|}{N_t} \qquad \text{and} \qquad \eta_f = \frac{|\mathcal{S}_f|}{N_t}, \quad (5)$$

where $|\mathcal{S}|$ denotes the cardinality of set $\mathcal{S}$. The error rates η_m and η_f are dependent on the threshold that is applied on the peaks of the coherent covariance sequence. The error rates η_m and η_f are computed as functions of the threshold, and equal error rate (EER) is achieved at a threshold when η_m and η_f are equal.

3.3. Results and discussion

The error rates, and histograms of deviation between the correctly detected GCIs and the corresponding true GCIs are shown in Fig. 4. The deviation between the true GCIs and the hypothesized GCIs is primarily due to two reasons: (a) The locations of peaks in the Hilbert envelope of LP residual are not exactly same as the locations of GCIs. (b) Due to the averaging involved in the computation of coherent covariance sequence, the locations of hypothesized peaks are not exactly the same as those of the peaks in Hilbert envelope. Regions of weak voicing contribute to missing GCI, while secondary excitations have been observed as the main source of false alarms. For signals collected at a distance, the deviation is greater compared to that in signals corrupted by additive noise, may be due to the effect of reverberation and low SNR. In contrast, the instants of excitation are still preserved in the signals corrupted by additive noise. Hence, the spread of deviation around the mean value is lesser in the case of additive noise (Fig. 4(b)), compared to that in signals collected at different distances (Figs. 4(f), (h), (j) and (l)). The histograms also indicate that as the distance increases, the spread of the deviation increases too. Tables 1 and 2 list the performance of GCI detection for speech signals collected at different distances, and for speech signals corrupted by additive noise, respectively. In both the cases, EER and deviation are smaller for Set II than for Set I. This highlights the importance of SNR associated with the GCIs. Weak voiced sounds with low SNR in the vicinity of GCI may be missed even in clean speech, due to poor estimation of the source of excitation from the LP residual. In contrast, voiced regions with significant SNR near GCIs are detected, even when the speech signal is collected at a distance or is corrupted by additive noise. The EER and deviation do not increase appreciably over the range of SNR (0-20 dB), or distance (3-6 ft). Thus, the SNR of speech signal in the vicinity of GCIs, in conjunction with the periodicity of significant excitations, lends robustness to the proposed method.

4. Conclusion

In this paper, the characteristics of source of excitation of the vocal tract are exploited for detection of the glottal closure instants in voiced speech. The computation of coherent covariance sequence is proposed, which is more amenable to detection of peaks due to significant excitations than the Hilbert envelope of LP residual. Experimental results indicate that the proposed method is robust and accurate, even when speech signal is corrupted by degradations. The coherent covariance sequence can be processed further to derive a weighting function, which selectively emphasizes the regions of speech signals in the vicinity of significant excitations. Such an emphasis can be useful for (a) enhancement of speech for human listening, and (b) extraction of robust parameters from degraded speech.

Table 1: *Performance of the proposed method for detection of GCI, for speech signals collected at different distances.*

Distance (ft)	EER (%)		Deviation (ms)	
	Set I	Set II	Set I	Set II
6	12.8	10.4	1.35	1.09
5	10.1	8.3	1.22	1.04
4	9.1	7.4	1.10	0.93
3	8.8	7.2	1.16	0.98

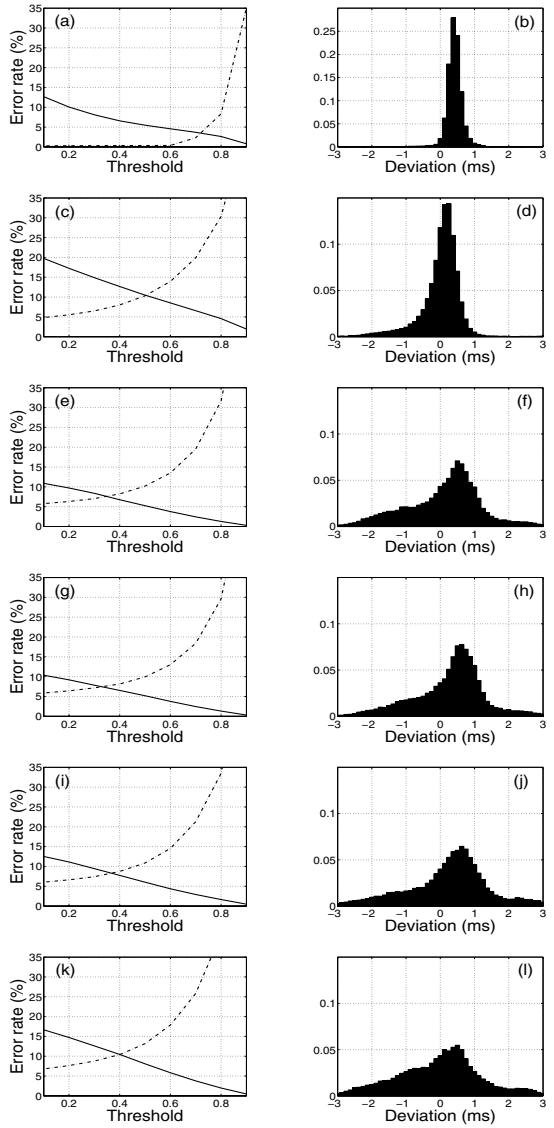

Figure 4: *Evaluation of GCI detection. (a) False alarm rate (solid curve) and missing rate (dashed curve) as functions of threshold and (b) normalized histogram of deviation between true and correctly hypothesized GCIs, for clean speech signals. The plots (c) and (d) correspond to speech signals with an overall SNR of 0 dB. Plots (e) and (f) correspond to speech signals collected at a distance of 3 ft, while (g), (h) correspond to 4 ft, (i), (j) to 5 ft and (k),(l) correspond to 6 ft. The evaluation was performed on Set II. Note the change in the vertical scale of the histogram in (b) and in the rest of the histograms.*

Table 2: *Performance of the proposed method for detection of GCI, for speech signals with different levels of overall SNR.*

SNR (dB)	EER (%)		Deviation (ms)	
	Set I	Set II	Set I	Set II
0	10.4	9.8	0.59	0.56
5	8.5	7.9	0.53	0.49
10	7.7	7.1	0.48	0.44
15	7.2	6.6	0.45	0.40
20	6.9	6.1	0.41	0.35
Clean speech	3.5	3.1	0.35	0.28

5. References

[1] Yegnanarayana, B., and Raymond Veldhuis, N. J., "Extraction of vocal-tract system characteristics from speech signals", IEEE Trans. Speech Audio Proc., 6(4):313–327, 1998.

[2] Sreenivasa Rao, K., and Yegnanarayana, B., "Prosody modification using instants of significant excitation", IEEE Trans. Audio Speech Lang. Proc., 14(3):972–980, 2006.

[3] Strube, H. W., "Determination of the instant of glottal closure from the speech wave", J. Acoust. Soc. Amer., Vol. 56, No. 5, 1625–1629, 1974.

[4] Changxue Ma, Yves Kamp, and Lei Willems, F., "A Frobenius norm approach to glottal closure detection from the speech signal", IEEE Trans. Speech Audio Proc., 2(2):258–265, 1994.

[5] Ananthapadmanabha, T. V., and Yegnanarayana, B., "Epoch extraction from linear prediction residual for identification of closed glottis interval", IEEE Trans. Acoust. Speech Signal Proc., 27(4):309–319, 1979.

[6] Yegnanarayana, B., and Smits, R. L. H. M., "A robust method for determining instants of major excitations in voiced speech", Proc. Int. Conf. Acoust. Speech Signal Process., 776–779, Detroit, USA, 1995.

[7] Smits, R. L. H. M., and Yegnanarayana, B., "Determination of instants of significant excitation in speech using group delay function", IEEE Trans. Speech Audio Proc., 3(5):325–333, 1995.

[8] Satyanarayana Murthy, P., and Yegnanarayana, B., "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals", IEEE Trans. Speech Audio Proc., 7(6):609–619, 1999.

[9] Kounoudes, A., Patrick Naylor, A., and Mike Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech", Proc. Int. Conf. Acoust. Speech Signal Process., 349–352, Orlando, USA, 2002.

[10] Mike Brookes, Patrick Naylor, A., and Jon Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech", IEEE Trans. Audio Speech Lang. Proc., 14(2):456–466, 2006.

[11] Makhoul, J., "Linear Prediction: A tutorial review", Proc. IEEE, 63(4):561–580, 1975.

[12] ARCTIC speech corpus, "http://festvox.org/cmu_arctic/".