Detection of Glottal Closure Instants from Bessel Features using AM-FM Signal

Chetana Prakash
Speech and Vision Laboratory
International Institute
of Information Technology
Hyderabad, India
Email: chetana@research.iiit.ac.in

Dhananjaya N
Speech and Vision Laboratory
Indian Institute of Technology
Chennai, India
Email: dhanu@cse.iitm.ac.in

Suryakanth V. Gangashetty
Speech and Vision Laboratory
International Institute
of Information Technology
Hyderabad, India
Email: svg@iiit.ac.in

***Abstract*—Glottal closure instants (GCI) information is useful for accurate speech analysis. In particular accurate spectrum analysis is performed by considering the speech in the intervals of glottal closure. In this paper we propose an approach for detection of GCI based on Bessel features, amplitude and frequency modulated (AM-FM) signal. Using appropriate range of Bessel coefficients, the narrow band, band limited signal is obtained for the given signal. The bandlimited signal is considered as a AM-FM signal. The signal is band limited for 0-300 Hz to remove effect of formants. Amplitude Envelope (AE) function of the AM-FM signal model has been estimated by the discrete energy separation algorithm (DESA). We experimentally evaluated our approach to detect GCI on CMU-Arctic database. The corresponding electro-glottograph (EGG) signals are used as a reference for the validation of the detected GCI locations.**

I. Introduction

The instant of significant excitation of the vocal-tract system is referred to as the Glottal Closure Instants (GCI). Detection of GCI in voicd speech signal is important for various speech processing applications like pitch tracking, prosodic modification, modeling of voice source, speech dereverberation, and so on [1], [2]. The GCI is defined as an instant where there is a most significant excitation which is due to glottal vibrations in the vocal tract. The vocal folds vibrate during the production of voiced sound and speech signal is characterized by a substantial instantaneous increase in signal energy due to closure of the vocal folds at the end of each glottal cycle. This instantaneous change can be represented as an impulse-train like excitation signal [2].

Presently, there are many techniques for the detection of GCIs in the literature. The various quantitative measures for the performance of detection of GCI methods based on group delay is presented in [3]. A dynamic programming projected phase-slope algorithm (DYPSA) for automatic estimation of glottal closure instants in voiced speech was presented in [4]. The zero-frequency resonator based methods for extraction of GCIs is proposed in [5]. The positive zero-crossings of the filtered signal are corresponding to GCIs. The time-frequency domain based methods have been applied for GCI detection from the speech utterences in [6], [7].

In this paper, we propose an approach for detection of GCIs from the speech signal. The method is based on the Bessel expansion and the AM-FM model. The inherent filtering property of the Bessel expansion is used to weaken the effect of formants in the speech utterences. The Bessel coefficients are unique for a given signal in the same way that Fourier series coefficients are unique for a given signal. Bessel series expansion suitable for analysis of speech signal which is of non-stationary in nature, because basis function of Bessel transformation are of aperiodic and decay over time [8]. The discrete energy separation algorithm (DESA) method has been used to estimate amplitude envelope (AE) function of the AM-FM model due to its good time resolution. This feature is advantageous for detection of GCI as they are well localized in time-domain.

The paper is organized as follows: In Section II, we describe the Bessel expansion. The signal modeling based on AM-FM and its analysis using DESA is described in Section III. The proposed approach for detection of GCIs using AE function from AM-FM model is explained in Section IV. The Section V, gives experimental study of the performance of proposed GCI detection method.

II. Bessel Expansion

Rather than using a classical Fourier basis, a damped exponential set of functions is employed which corresponds more closely to the waveforms that occur in voiced speech. Bessel expansion of the speech signal is achieved by using zeroth-order $J_0(\lambda_m t)$ and first-order $J_1(\lambda_m t)$ of the Bessel function of the first kind as the basis functions of representation. For the zeroth-order Bessel function of the first kind $J_0(\lambda_m t)$, $\lambda_m = \frac{t_m}{a}$, and t_m is the m^{th} root of $J_0(t) = 0$, and a is the time frame of the analysis. The decomposition describes a speech signal as a linear combination of the orthogonal basis functions. The zeroth-order Bessel series expansion of a signal $x(t)$ considered over some arbitrary interval $(0, a)$ is expressed as [8]:

$$x(t) = \sum_{m=1}^{\infty} C_m J_0\left(\frac{\lambda_m}{a} t\right) \tag{1}$$

Where $\{\lambda_m,\ m = 1, 2, ..M.., \infty\}$ are the ascending order positive roots of $J_0(\lambda) = 0$, and Figure 1 shows few roots of

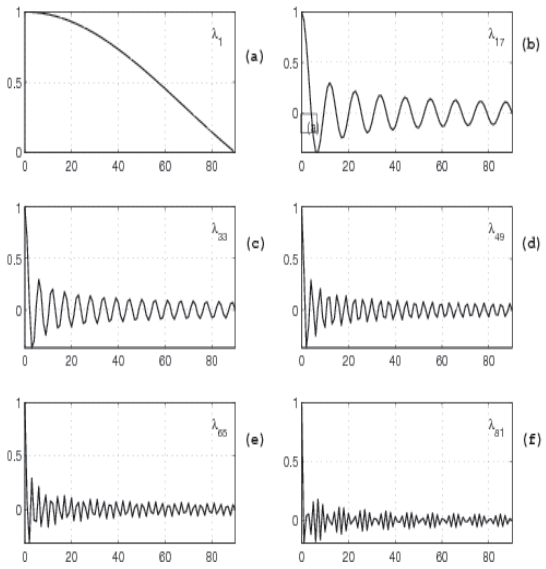

Fig. 1. *The positive Bessel roots (a) λ_1, (b) λ_{17}, (c) λ_{33}, (d) λ_{49}, (e) λ_{65}, and (f) $\lambda_{81}$of zeroth order Bessel function of first kind.*

the zeroth-order Bessel function. Using the orthogonality of zeroth-order Bessel functions $J_0\left(\frac{\lambda_m}{a}t\right)$, the Bessel coefficients C_m are computed by using the following equation:

$$C_m = \frac{2}{a^2[J_1(\lambda_m)]^2}\int_0^a tx(t)J_0\left(\frac{\lambda_m}{a}t\right)dt \tag{2}$$

with $1 \leq m \leq M$, where *m* is the order of the Bessel expansion, and $J_1(\lambda_m)$ is the first-order Bessel function of the first kind. The Bessel expansion order *M* must be known a priori. The interval between successive zero-crossings of the Bessel function $J_0(\lambda)$ increases slowly with time and approaches π in the limit. If order *M* is unknown, then in order to cover full signal bandwidth, the half of the sampling frequency, *M* must be equal to the length of the signal.

TABLE I
C_m coefficients against frequency

Frequency F_{max} (Hz)	**Ideal value m-index**
a = 20 ms and sampling frequency 8 kHz	
200	8
500	20
700	28
1000	40
2000	80
2500	100
3000	120
3400	136
3600	144

It has been shown in Table I that there is one-to-one correspondence between the frequency content of the signal and the order (m) at which the coefficient attains peak magnitude [9]. If the AM-FM component or formant of the speech signal are well separated in the frequency domain, the speech signal components will be associated with various distinct clusters of non-overlapping Bessel coefficients [10]. Each component of the speech signal can be reconstructed by identifying and separating the corresponding Bessel coefficients. In this paper, the inherent band pass filtering property of the Bessel expansion is used to separate the low-frequency band of the speech signal.

III. AM-FM Signal and DESA Method

A general monocomponent continuous time nonstationary signal has the form of a modulated signal defined as:

$$x(t) = A(t)\cos[\omega(t) + \phi(t)] = A(t)\cos[\Phi(t)] \tag{3}$$

Where $A(t)$ is the time-varying amplitude envelope (AE) of $x(t)$ with instantaneous frequency (IF) $\Omega(t)$ given by:

$$\Omega(t) = \frac{d\Phi(t)}{dt} = \omega + \frac{d\phi(t)}{dt} \tag{4}$$

Equation (3) has both amplitude modulation (AM) and frequency modulation (FM). These signals have been used in speech processing applications for modeling of speech resonances [11]. The discrete-time version of a monocomponent signal $x[n]$ is given by:

$$x[n] = A[n]\cos(\Phi([n]) \tag{5}$$

Both the instantaneous frequency and the amplitude envelope of the signal $x[n]$ can be derived from the Teager's nonlinear energy (NLE) operator. The Teager's NLE operator $\Psi(.)$ defined for the discrete signal $x[n]$ as [12]:

$$\Psi(x[n]) = x^2[n] - x[n-1]x[n+1] \tag{6}$$

It is applied to the AM-FM signal $x[n]$ and the difference signal $y[n] = x[n] - x[n-1]$. The amplitude envelope function $|A[n]|$ of the signal $x[n]$ is estimated by the discrete energy separation algorithm (DESA) as [11].

$$|A[n]| \approx \sqrt{\frac{\Psi[x[n]]}{1 - \left[1 - \frac{\Psi[y[n]] + \Psi[y[n+1]]}{4\Psi[x[n]]}\right]^2}} \tag{7}$$

The amplitude envelope function of the AM-FM signal thus estimated exhibit ripples and therefore requires smoothing using a filter. The performance of the energy operator/DESA approach is vastly improved if the signal is first filtered through a bank of band pass filters, and at each instant analyzed (via Ψ and DESA) using the dominant local channel response.

IV. Approach for Detection of GCIs

In order to detect GCIs we emphasize the low frequency contents of the speech signal in the range of 0 to 300 Hz. This is achieved by using the appropriate order *M* of the Bessel expansion. Since the resultant band-limited signal is a narrow band signal, it can be modeled by using AM-FM model. The reason for choosing 0 to 300 Hz band is that the characteristics of the time-varying vocal-tract system will not affect the location of the GCIs. This is because the vocal-tract system has resonances at higher frequencies than 300

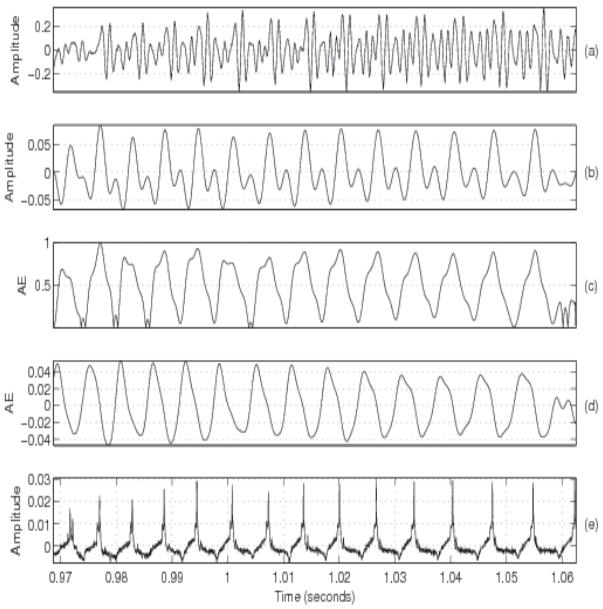

Fig. 2. *A 3000 sample segment of (a) Speech waveform of the male speaker, (b) Band-limited signal using Bessel expansion, (c) Amplitude Envelope of band-limited signal, (d) Smoothened Amplitude Envelope of band-limited signal, (e) Differenced EGG signal.*

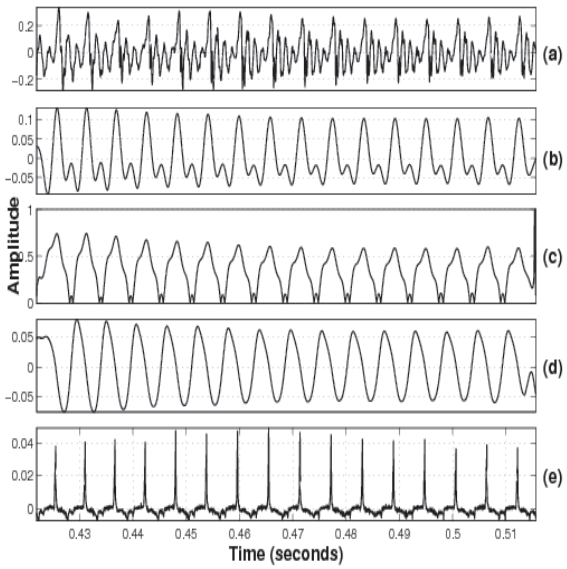

Fig. 3. *A 3000 sample segment of (a) Speech waveform of the female speaker, (b) Band-limited signal using Bessel expansion, (c) Amplitude Envelope of band-limited signal, (d) Smoothened Amplitude Envelope of band-limited signal. (e) Differenced EGG signal.*

Hz. Therefore, we propose that the characteristics of peaks due to GCIs can be extracted by reconstructing the speech signal using the Bessel expansion of order $M = 56$ in (1) for the sampling frequency 32000 Hz (3000 sample segment is taken for analysis, 16000 Hz corresponds to 3000^{th} sample since the sampling frequency is 32 kHz) . The DESA technique is applied on this band-limited signal to separate AM-FM component of the signal. Convolving AM component with differenced Gaussian window of suitable length to smoothen the AE. The zero crossings of resultant signal obtained from the smoothed AE corresponds to locations of hypothesized GCIs.

For illustration, we consider a 3000 sample segment of speech signal uttered by male speaker whose waveform is shown in Figure 2(a). Figure 2(b) is a band limited signal of frequency band 0-300 Hz obtained by taking the appropriate Bessel coefficient. Amplitude Envelope (AE) of the band limited signal is shown in Figure 2(c). AE of the band-limited signal is smoothed by convolving result of Figure 2(c) with the differenced Gaussian window of suitable length and is shown in Figure 2(d). The negative zero crossings of Figure 2(d) and positive peaks of amplitude envelope of Figure 2(c) corresponds to the GCI locations. It is seen that negative zero crossings of the smoothed AE which is shown in Figure 2(d) are agreeing in most of the cases with the peaks in the differenced EGG signal of the Figure 2(e). Similar observations are also seen for the speech utterance of female speaker as in Figure 3. This enable us to identify the locations of GCIs from the peaks of the amplitude envelope of the band-limited AM-FM signal of the given speech utterance. It is also seen from Figures 2 and 3 that the number of GCIs for female speaker are more than the male speaker for the same duration of speech segment. This is due to the fact that generally the fundamental frequency (reciprocal of the difference between successive GCIs) of female speakers is higher than the male speakers.

V. Studies on Detection of GCIs

We study the performance of the GCI detection approach for the speech utterances of CMU-Arctic database [13]. This database has EGG signals associated with each utterance. The peaks in the differenced EGG signals are used as references GCIs. The Arctic database consists of various phonetically balanced English sentences spoken by male and female speakers. We consider male speaker data from bdl directory (consists of 1131 speech utterances) and female speaker data from slt directory (consists of 1132 speech utterances). The duration of each utterance is approximately 3 seconds, which makes the duration of the entire database to be around 2 hours 40 minutes. The database was collected in a soundproof booth and digitized at a sampling frequency of 32 kHz. In addition to the speech signals, the Arctic database contains the simultaneously recordings of EGG signals collected using a laryngograph. The speech and EGG signals were time-aligned to compensate for the larynx-to-microphone delay, determined to be approximately 0.7 ms. Reference locations of the GCIs were extracted from the 3000 samples voiced segments of the EGG signals by finding peaks in the differenced EGG signals. The performance of the proposed approach is evaluated by comparing the detected GCI locations with reference GCIs (differenced EGG signal).

TABLE II

The Performance of dectecting GCIs in continuous speech. The performance is given as a total number of GCIs in the continuous speech utterance for the matching, missing and spurious hypotheses.

Gender	Speech Utterance Ids	Reference # of GCIs	GCI detection performance for		
			Matching	Missing	Spurious
Male	10501	12	12	0	0
	10368	11	11	0	1
	10882	9	9	0	0
	11007	11	11	0	0
	11131	11	11	0	0
Female	30401	16	16	0	0
	31128	15	15	0	0
	30890	17	17	0	0
	30747	17	17	0	0
	31016	14	14	0	0

The performance is measured in terms of the number of matching, missing and spurious GCIs of speech utterances. For testing we consider 5 male and 5 female speech utterances selected at random from 1131 and 1132 sentences respectively. The performance of the GCI detection approach for each of the sentence is given in Table II.

The GCIs detected with a deviation up to 2 ms are considered as the matching hypotheses. When the deviation of hypothesised GCI is more than 2 ms or there is no hypothesised GCI around the actual (referenced) GCI, the GCIs of such segments are considered as the missing hypotheses. When there are multiple hypotheses with in 2 ms around the actual GCI, such hypotheses are considered as spurious ones.

It is seen from table II that the proposed approach has detected the GCI locations accurately and number of GCIs detected is equal to number of referenced GCIs in most of the speech utterances.

VI. Summary and Conclusions

In this paper we propose Bessel based AM-FM signal model approach for the location of the golttal closure instants. The peak of the amplitude envelope (AE) and its corresponding zero crossing of the smoothened amplitude envelope provides the locations of GCIs. Since the method is based on the amplitude characteristics of the signal, it is possible to locate the GCIs locations from the speech signal with good time resolution. Since, the proposed method does not include any formants of the speech utterance, the peaks in the amplitude envelope are well manifested. This enables us to identify the impulsive nature of the GCIs. The CMU-Arctic database is considered for the analysing our studies since, the availability of speech signal and its corresponding EGG signals in the database. Since the proposed method provides accurate location of GCIs, the results are useful to develop methods for accurate estimation of fundamental frequency, formant estimation.

References

[1] Douglas O'Shaughnessy, *Speech Communications Human and Machine*, Wiley-IEEE Press, 2nd Edition 1999.

[2] A. K. Krishnamurthy, ''Glottal source estimation using a sum-of-exponential model", *IEEE Trans. Acoust. Speech Signal Process.*, vol. 40, no. 3, pp. 682-686, 1992.

[3] M. Brookes, P. A. Naylor, and J. Gundnason, ''Aquantitative assessment of group delay method for identifying glotal closure in voiced speech", *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 456-466, Mar. 2006.

[4] P. A. Naylor, A. Kounoudes, J. Gundnason, and M. Brookes, ''Estimation of glottal closure instants in voiced speech using the DYPSA algorithm", *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 34-43, 2007.

[5] K. S. R. Murthy and B. Yegnanarayana, ''Epoch extraction from speech signals", *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[6] J. I. Navarro-Mesa, E. Lleida-Solano, and A. Moreno-Bilbao, ''A new method for epoch detection based on the Cohen's class of time-frequency representations", *IEEE Signal Process. Letters*, vol. 8, no. 8, pp. 225-227, 2001.

[7] L. Kaushik and Douglas O'Saughnessy, ''A novel method for epoch extraction from speech signals", *Proc. Interspeech*, pp. 2883-2886, Sept. 2009.

[8] J. Schroeder, ''Signal processing via Fourier-Bessel series expansion", *Digital Signal Process.*, vol. 3, pp. 112-124, 1993.

[9] R. B. Pachori and P. Sircar, ''EEG signal analysis using FB expansion and second-order linear TVAR process", *Signal Process.*, vol. 88, no. 2, pp. 415-420, 2008.

[10] R. B. Pachori and P. Sircar, ''Analysis of multicomponent AM-FM signals using FB-DESA method", *Digital Signal Process.*, vol. 20, pp. 42-62, 2010.

[11] P. Maragos, J. F. Kaiser, and T. F. Quatieri, ''Energy separation in signal modulation with applications to speech analysis", *IEEE Trans. on Signal Process.*, vol. 41, no. 10, pp. 3024-3051, Oct. 1993.

[12] J. F. Kaiser, ''On a simple algorithm to calculate the 'energy' of a signal", *IEEE ICASSP*, pp. 381-384, 1990.

[13] J. Kominek and A. Black, ''The CMU Arctic speech database", *Proc. 5th ISCA Speech Synthesis Workshop*, pp. 223-234, 2004.