Exploring Bessel Features for Detection of Glottal Closure Instants

Chetana prakash[1], *Dhananjaya N*[2], *and Suryakanth V. Gangashetty*[1]

[1] International Institute of Information Technology, Hyderabad, India
[2] Indian Institute of Technology Chennai, India
chetana@research.iiit.ac.in, dhanu@cse.iitm.ac.in, svg@iiit.ac.in

Abstract

For voiced speech, the most significant excitation takes place around the instant of glottal closure. 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 by exploring Bessel feature, and the use of 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 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). The performance of the method is demonstrated using CMU-Arctic database. The corresponding electro-glottograph (EGG) signals are used as a reference for the validation of the detected GCI locations.
Index Terms: GCIs, Bessel expansion, AM-FM signal model, DESA.

1. Introduction

The instant of significant excitation of the vocal-tract system is referred to as the Glottal Closure Instants (GCI). An excitation is termed as significant if it is impulse-like with strength substantially larger than the strength of impulses in the neighborhood. In the context of speech most of the significant excitation takes place due to glottal vibration. The exceptions are strong burst releases of very short duration. 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.

Voiced speech analysis consists of determining the frequency response of the vocal-tract system and the glottal pulse representing the excitation source. Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal-tract system is within a glottal pulse. The significant excitation can be considered to occur at the instant of glottal closure. Many speech analysis situation depend on the accurate estimation of GCI locations within a glottal pulse. The glottal airflow is zero soon after the glottal closure, as a result the supralaryngeal vocal-tract is accoustically decoupled from the trachea. Hence, the speech signal in the closed phase region represents the free resonances of the supralaryngeal vocal-tract system. Analysis of the speech signal in the closed phase regions provides an accurate estimate of the frequency response of the supralaryngeal vocal-tract system [1]. With the knowledge of the GCIs, it is possible to determine the characteristics of the voice source by analysis of the signal within a glottal pulse. The GCIs can be used as pitch markers for prosody manipulation, which is useful for applications like text-to-speech synthesis, voice conversion and speech rate conversion. Knowledge of the GCIs locations is used for estimating the time-delay between speech signals collected over a pair of spatially distributed microphones. The segmental signal-to-noise ratio (SNR) of the speech signal is high in the regions around the GCIs, and hence, it is possible to enhance the speech by exploiting the characteristics of speech signals around the GCIs. The excitation features derived from the regions around the GCI location provide complementary speaker-specific information to the existing spectral features.

This paper concerns with the Fourier-Bessel (FB) representation of speech signal and its application to detecting the GCIs in speech signal. The FB representation of the speech signal is obtained using Bessel function as a basis set. In the linear speech production model, a typical glottal volume velocity signal of the steady-state non-nasalised voiced speech sound is considered as a quasi-periodic sequence of pulses. Each pulse is the result of glottal opening and closure. The resulting speech pressure wave signal has the form of a freely decaying oscillations. Such a speech waveform reveals the amplitude-decaying and the non-uniform zero-crossing nature of the voiced sound waveform during each pitch period, or more precisely each period of the glottal opening and closure. The Bessel function also displays amplitude-decaying and non-uniform zero-crossing characteristics. The use of the Bessel function as basis function for speech signal decomposition therefore seems logical and natural. 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 [2]. The discrete energy separation (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 2, we describe the Bessel expansion. The signal modeling based on AM-FM and its analysis using DESA is described in Section 3. Then we propose an approach for detection of GCIs using AE function from AM-FM model in Section 4. In Section 5, we compare the proposed method of GCI detection with the other existing methods.

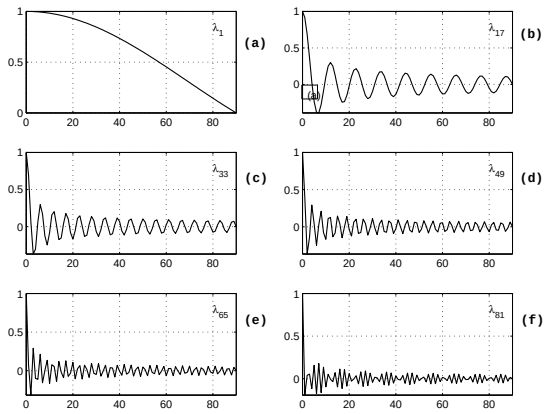

Figure 1: *The positive Bessel roots (a)* λ_1*, (b)* λ_{17}*, (c)* λ_{33}*, (d)* λ_{49}*, (e)* λ_{65}*, and (f)* λ_{81} *of zeroth order.*

2. 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 $J_0(\lambda_m t)$ and $J_1(\lambda_m t)$ as basis functions of representation, where $\lambda_m = \frac{t_m}{a}$, 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 [2]:

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

Where $\{\lambda_m,\ m = 1, 2, ...\}$ are the ascending order positive roots of $J_0(\lambda) = 0$, Fig. 1 shows few roots, and $J_0\left(\frac{\lambda_m}{a} t\right)$ are the zeroth-order Bessel functions. 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:

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

with $1 \leq m \leq M$, where *m* is the order of the Bessel expansion, and $J_1(\lambda_m)$ are the first-order Bessel functions. 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.

It has been shown in Table 1 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. 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 [3]. 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.

Table 1: 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
1000	40
2000	80
3000	120

3. 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 [4]. 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 NLE operator $\Psi(.)$ defined for the discrete signal $x[n]$ as

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

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 [4].

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

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. Only AM part of AM-FM decomposition is used in the proposed method.

4. 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

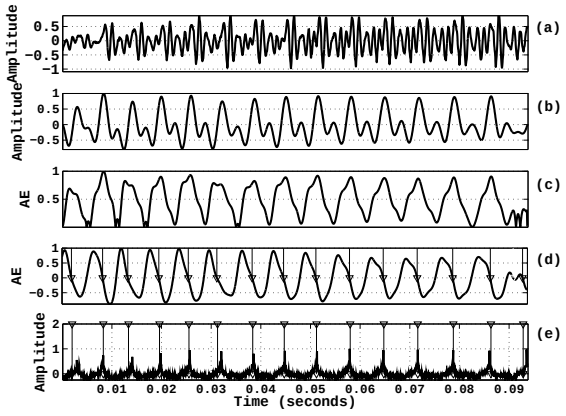

Figure 2: *A sample segment of (a) Speech waveform of male speaker, (b) Band-limited signal using Bessel expansion, (c) Amplitude envelope of band-limited signal, (d) Amplitude envelope convolved with the differenced Gaussian window, (e) Differenced EGG signal.*

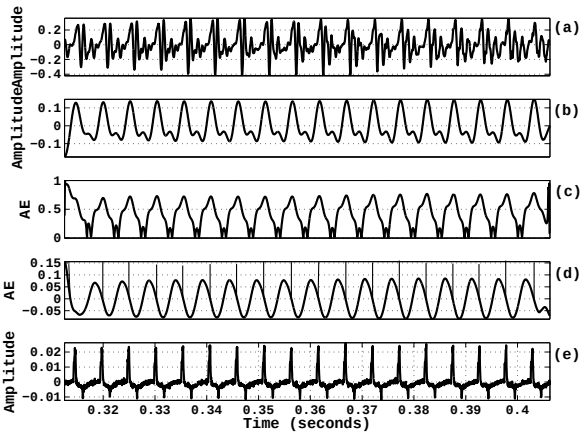

Figure 3: *A sample segment of (a) Speech waveform of female speaker, (b) Band-limited signal using Bessel expansion, (c) Amplitude envelope of band-limited signal, (d) Amplitude envelope convolved with the differenced Gaussian window, (e) Differenced EGG signal.*

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 [3]. 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 Hz. When we band limited the signal to 300 Hz, the residual effect of the lower resonances can be seen in the reconstructed signal as shown in Fig. 2(b). But this does not affect the GCI detection severely as can be seen from the amplitude envelope shown in Fig. 2(c). 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. The peaks of the amplitude envelope gives the GCI detection. When the amplitude envelope convolved with the differenced Gaussian window of suitable length, the negative zero crossing of the resultant signal corresponds to locations of hypothesized GCIs. For illustration, we consider a sample segment of speech signal uttered by male speaker whose waveform is shown in Fig. 2(a). Fig. 2(b) is a band limited signal of frequency band 0-300 Hz obtained by taking the appropriate Bessel coefficient. Fig. 2(c) is the amplitude envelope (AE) of the band-limited signal. Amplitude envelope of the reconstructed signal is convolved with the differenced Gaussian window of suitable length and is shown in Fig. 2(d). The negative zero crossings of Fig. 2(d) and positive peaks of amplitude envelope of Fig. 2(c) corresponds to the GCI locations. It is seen that negative zero crossings of the AE convolved with Gaussian window which is shown in Fig. 2(d) are agreeing in most of the cases with the peaks in the differenced EGG signal of the Fig. 2(e). Similar observations are also seen for the speech utterance of female speaker as in Fig. 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 utterence. It is also seen from Figs. 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.

5. Comparision of Proposed GCIs Extraction with Other Methods

In this section, the proposed method of GCI detection is compared with three existing methods in terms of identification accuracy. The three methods chosen for the comparision are the Hilbert envelope based (HE-based) method [5], the group-delay-based (GD-based) method [6], and the DYPSA algorithm [7]. The performance of the algorithm was evaluated on the clean data. CMU-Arctic database was employed to evaluate the proposed method of GCIs detection and to comapare the results with the existing methods [8]. The Arctic database consists of 1132 phonetically balanced English sentences spoken by two male and one female talkers. The duration of each utterence is approximately 3 s, which makes the duration of the entire database to be around 2 h 40 min. 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 simultaneous 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 voiced segments of the EGG signals by finding peaks in the differenced EGG signal. The performance of the algorithm was evaluated only in the voiced segments (detected from EGG signal) between the reference GCI location and the estimated GCI locations. The database contains a total of 792,249 GCIs in the voiced regions. The performance of the GCI detection method was evaluated using the measures defined in [7]. The following measures were defined to evaluate the performance of GCI detection algorithms.

- *Larynx cycle*: The range of samples $(1/2)(l_{r-1} + l_r) \leq n \leq (1/2)(l_r + l_{r+1})$, given an GCI reference at sam-

Table 2: *Performance comparision of GCI detection methods on CMU-Arctic database. IDR–Identification rate, MR–Miss rate, FAR–False alarm rate, IDA–Identification accuracy.*

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
HE-based	89.86	1.43	8.71	0.58
GD-based	92.8	4.01	3.18	0.67
DYPSA	96.66	1.76	1.58	0.59
Proposed	96.83	1.83	1.33	0.63

ple l_r with preceding and succeeding GCI references at samples l_{r-1} and l_{r+1}, respectively.

- *Identification rate (IDR)*: The percentage of larynx cycles for which exactly one GCI is detected.
- *Miss rate (MR)*: The percentage of larynx cycles for which no GCI is detected.
- *False alarm rate (FAR)*: The percentage of larynx cycles for which more than one GCI is detected.
- *Identification error (ζ)*: The timing error between the reference GCI location and the detected GCI location in larynx cycles for which exactly one GCI was detected.
- *Identification accuracy σ (IDA)*: The standard deviation of the identification error ζ. Small values of σ indicate high accuracy of identification.

The Fig. 4(a) shows the locations of the reference GCIs obtained from the differencing the EGG signal and taking the peaks of the Hilbert envelope, and the hypothesized GCIs detected from the proposed method. The Fig. 4(b) gives the deviation value of the hypothesized GCIs from the referenced GCIs.

Table 2 shows the performance results on Arctic database for identification rate, miss rate, false alaram rate, and identification accuracy for the three methods HE-based, GD-based, and DYPSA algorithm, as well as for the proposed method. The performance is measured in terms of the number of matching, missing and spurious GCIs of speech utterences. From Table 2 it can be concluded that the DYPSA algorithm performed best among the three existing techniques, with an identification rate of 96.66%. The proposed method of GCI detection gives slightly better identification rate as well as decrease in the False alarm rate compared to the results from the DYPSA algorithm.

6. Summary and Conclusions

In this paper we explore 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 differenced Gaussian window which is convolved with 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 utterence, 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 explores the Bessel features for the accurate detection of GCIs, so

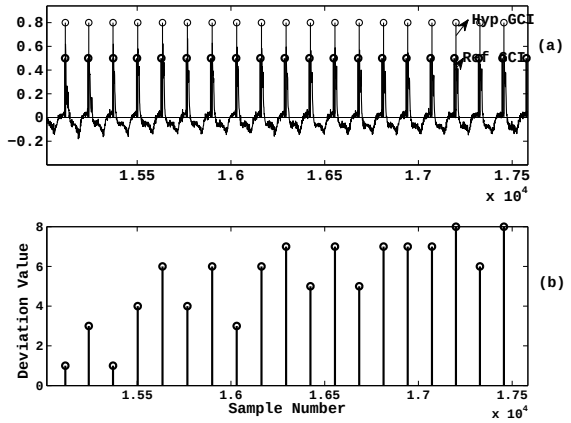

Figure 4: *(a) Plot of referenced GCIs with reference to differenced EGG signal with hypothesized GCIs obtained from proposed method, (b) Deviation of the hypothesized GCIs from the referenced GCIs location.*

the results are useful to develop methods for accurate estimation of fundamental frequency, formant estimation and other speech analysis applications.

7. References

[1] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *In Proc. IEEE Trans. Audio, Speech Lang. Processing*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[2] J. Schroeder, "Signal processing via Fourier-Bessel series expansion," *Digital Signal Processing*, vol. 3, pp. 112–124, 1993.

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

[4] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulation with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41 no. 10, pp. 3024–3051, Oct., 1993.

[5] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *IEEE signal Processing Lett.*, vol. 14, no. no. 10, pp. 762–765, Oct.,, YEAR =.

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

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

[8] J. Kominek and A. Black, "The CMU Arctic speech databases," *In Proc. 5th ISCA Speech Synthesis Workshop*, pp. 223–234, 2004.