Estimation of Glottal Closure Instants from Telephone Speech using a Group Delay-Based Approach that Considers Speech Signal as a Spectrum

G. Anushiya Rachel, P. Vijayalakshmi, T. Nagarajan

Speech Lab, SSN College of Engineering, Chennai, India
vijayalakshmip@ssn.edu.in, nagarajant@ssn.edu.in

Abstract

Glottal closure instants (GCIs) are characterized by a strong negative valley in the speech signal and an abrupt change in the amplitude. In this paper, an algorithm that exploits these two properties of a GCI is proposed to estimate the location of GCIs, specifically from telephone speech. The algorithm considers a symmetrized voiced segment as the Fourier transform of an even signal. In such a case, the negative valleys in the spectrum correspond to zeros that lie outside the unit circle in the z-plane. The angular location of these zeros indicate the location of the GCIs. The estimated GCIs are identified from the group delay spectrum of the even signal, since a phase change of $2\pi$, between adjacent frequency bins, occurs at the location of a zero that lies outside the unit circle. The performance of the algorithm is evaluated on a simulated speech corpora derived from CMU and CSTR databases and the NTIMIT database, in terms of identification, false alarm, and miss rates. The proposed algorithm is compared with DYPSA, YAGA, and SEDREAMS, and is found to outperform all the algorithms when used on telephone speech.

Index Terms: glottal closure instants, group delay, telephone speech, chirp-z transform

1. Introduction

Voiced speech is produced by the excitation of the vocal tract by periodic vibrations of the glottis [1]. Owing to these vibrations, the glottis opens and closes once within a pitch period or glottal cycle. When the glottis is fully open or when it is just about to close, the strength of the excitation to the vocal tract is maximum within this cycle. This is reflected as a high amplitude region within a pitch cycle. In other words, the time instants at which there is an abrupt change in amplitude are identified. The proposed algorithm assumes that the excitation can be detected within a pitch period is maximum and negative. Estimation of GCIs plays a vital role in a number of applications. It is primarily used to identify pitch marks and for pitch tracking. The time domain pitch synchronous overlap and add (TD-PSOLA) technique, which is used to modify the prosody of speech, requires an estimate of pitch marks. The authors of [2] observed that preserving the regions of the speech signal around the GCIs results in more natural sounding speech and therefore the accurate location of these instants are vital. Estimation of GCIs also plays an important role in speech dereverberation, glottal source modelling, speech enhancement, and speech synthesis [3].

Several methods exist to estimate GCIs from a speech signal. These are typically designed to work on microphone speech. However, when applied to telephone speech, which is band-limited, noisy, and degraded, there might be a decrease in the performance of some of these algorithms. Some of the successful approaches to estimating the location of GCIs are the following: (i) ZFF, (ii) DYPSA, (iii) YAGA, and (iv) SEDREAMS.

Zero Frequency Filtering (ZFF) approach [4, 5] works on the principle that since the excitation to the vocal tract is impulse-like, information about the excitation exists at all frequencies, including 0 Hz. In this regard, the algorithm finds the speech signal through a 0 Hz resonator, to avoid the influence of the vocal tract information. The zero crossings of the resultant signal, which is a smoothed version of the speech signal that oscillates at the pitch frequency, indicates the locations of the GCIs. However, as mentioned by the authors in [6], since the algorithm relies solely on the zero frequency, it might not be appropriate for use on telephone speech.

Two algorithms that use dynamic programming to estimate the correct GCIs from an initial set of candidate GCIs, estimated from the group delay of the LP residue and multiscale analysis, are the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [7, 8] and Yet Another GCI Algorithm (YAGA) [9] respectively. The cost functions used in dynamic programming take into account the deviation in pitch period, correlation between the speech signal around successive GCIs, and energy of signal at the GCIs. Therefore, although spurious instants may be detected in the initial stage of these algorithms, dynamic programming discards the incorrect instants, thereby making it possible for these algorithms to perform reasonably well on telephone speech as well.

Speech Event Detection using Residual Excitation And a Mean-based Signal (SEDREAMS) [10] involves deriving a mean-based signal, which oscillates at the pitch frequency, using a window of size equal to one to two pitch periods. The zero crossings in this signal provides an estimate of the interval within which the GCIs would lie. The accurate GCIs are located by identifying the maxima within this interval in the LP residual of the speech signal. Since the algorithm relies on an estimate of the pitch period, its performance on telephone speech might not be as high as that obtained for microphone speech.

In the current work, a group delay-based algorithm that considers speech signal to be a spectrum is proposed, to estimate GCIs from telephone speech. While ZFF assumes that information of the excitation exists at zero frequency as well, the proposed algorithm assumes that the excitation can be detected at the maximum frequency. In other words, the time instants at which there is an abrupt change in amplitude are identified. The proposed technique involves symmetrizing the Feager energy of a voiced segment of speech and assuming it to be the Fourier transform of an even signal. The group delay spectrum derived from this even signal, which is obtained by taking the inverse Fourier transform of the assumed spectrum, is used to estimate the GCIs. The proposed algorithm is applied to tele-
phone speech and its performance is compared with the existing algorithms in terms of identification, false alarm, and miss rates.

The paper is organized as follows: Section 2 describes the proposed technique in detail. Section 3 provides details about the speech corpora used in the current work. Section 4 analyses the performance of the existing and proposed algorithms on telephone speech, and Section 5 concludes the paper.

2. The Proposed Algorithm

The current work uses the method proposed in [11] to estimate glottal closure instants from a speech signal by assuming the signal to be the Fourier transform of an even signal. To suit the properties of telephone speech, appropriate modifications, specifically, low pass filtering followed by the application of a Teager energy operator, are incorporated at the pre-processing stage. The basic theory of the algorithm is repeated here for clarity. The algorithm exploits the fact that the negative valleys in the Fourier transform correspond to zeros that lie outside the unit circle in the z-plane. Further, the group delay spectrum has a value of around $2\pi$ at these negative values, that is, a phase change of $2\pi$ exists between adjacent frequency bins when a zero lies outside the unit circle. Therefore, the samples in the group delay spectrum that have a value of $2\pi$ are identified as GCIs.

Consider a voiced speech segment, $x(n)$, of length $N$. It is symmetrized to obtain $y(n)$ as follows:

$$y(n) = \begin{cases} x(n), & n = 0 \text{ to } N - 1 \\ x(-n + 2N - 1), & n = N \text{ to } 2N - 1 \end{cases}$$

(1)

Since the Fourier transform of a real and even signal is real and symmetric, $y(n)$ can be considered to be the spectrum of an even signal. In a speech signal, there is bound to be an abrupt change in amplitude at the regions around the GCIs. Further, a strong negative valley exists at each GCI. Since the speech signal is assumed to be a spectrum, these valleys correspond to zeros lying outside the unit circle in the z-plane, as evident from the following discussion:

Consider a system, $H(z)$, with a single zero at an angular location, $w_0$, as given below.

$$H(z) = 1 - az^{-1}$$

(2)

Here $a = re^{j\theta_0}$, where $r$ is the radius of the zero in the z-plane. The corresponding Fourier transform can be expressed as

$$H(e^{j\omega}) = 1 - r\cos(w - w_0) + j\sin(w - w_0)$$

(3)

At the angular location of the zero, $w = w_0$,

$$H(e^{j\omega_0}) = 1 - r$$

(4)

When $r > 1$, $H(e^{j\omega})$ becomes negative, implying that when a zero lies outside the unit circle, the Fourier transform at that angular location possesses a negative value.

From the Fourier transform magnitude spectrum, it would not be possible to determine if a singularity lies outside or inside the unit circle. However, this could be discriminated using the phase spectrum and in turn the group delay spectrum, as described below.

Consider again the system with a single zero at the angular location $w_0$ with radius $r$. Let $w_1$ and $w_2$ be the frequency bins just below and above $w_0$, that is $w_1 = w_0 - \delta$ and $w_2 = w_0 + \delta$, where $\delta$ is very small. Using (3), the Fourier transform of the system at $w_1$ is given by

$$H(e^{j\omega_1}) = 1 - r\cos\delta - j\sin\delta$$

(5)

From the above equation, the phase at $w_1$ can be expressed as

$$\theta(e^{j\omega_1}) = \tan^{-1}\left(-\frac{\sin\delta}{1 - r\cos\delta}\right)$$

(6)

Similarly, the phase at $w_2$ is

$$\theta(e^{j\omega_2}) = \tan^{-1}\left(\frac{\sin\delta}{1 - r\cos\delta}\right)$$

(7)

where the phase is a four quadrant inverse tangent function.

The group delay function at the location of the zero can be expressed as

$$\tau = \theta(e^{j\omega_2}) - \theta(e^{j\omega_1})$$

When the order of the Fourier transform is high, the difference between adjacent frequency bins, $\delta$, is very small, making $\sin\delta = \delta, \cos\delta = 1$. Therefore, the group delay function is given by

$$\tau = -2\tan^{-1}\left(\frac{r\delta}{1 - r}\right) = \tau_c$$

(8)

The above group delay function can be referred to as “conditional group delay”, $\tau_c$, since it is conditioned by the order of the Fourier transform.

When a zero lies outside the unit circle, $1 - r$ is negative. This implies that the denominator of the fourth quadrant inverse tangent function in (8) is negative, while the numerator is positive. In such a case, the conditional group delay can now be written in terms of the two quadrant inverse tangent function as

$$\tau_c = -2\left[\tan^{-1}\left(\frac{r\delta}{1 - r}\right) + \pi\right]$$

(9)

For $\delta$ being very small and $r > 1$, $\tan^{-1}\left(\frac{r\delta}{1 - r}\right)$ is a very small and negative value. Therefore, $\tau_c \approx -2\pi$, at the angular location of the zero. In this regard, the angular locations at which the group delay function yields a value of $-2\pi$ are identified as the locations of the GCIs.

Further, while the Fourier transform is conventionally computed on the unit circle, applying a window to the signal would result in zeros on the unit circle. In such a case, the group delay function would exhibit several spikes. Therefore, as suggested by [12], the chip group delay can be computed on a circle of radius greater than one, to avoid these spikes. This is portrayed in Fig. 1.

In summary, following are the sequence of steps (portrayed in Fig. 2) in the proposed approach when applied to telephone speech:

1. Low pass filter the given voiced speech segment. In the current work, an FIR filter of order 25, with a cut off frequency of 400 Hz is used.
2. Apply the Teager energy operator to the filtered speech to suppress the low frequency components and emphasize the high frequency ones. The Teager energy function is given by

$$t_e(n) = x^2(n) - x(n - 1)x(n + 1)$$

(10)

3. Negate the resulting signal to obtain strong negative valleys.
4. Perform mean subtraction to remove DC-offset, if any.
5. Symmetrize the segment by repeating the samples, as in (1), and consider it to be the Fourier transform of an even signal.
6. Compute the inverse Fourier transform of the symmetrized signal and consider only the causal portion.
7. Compute the chirp-Z transform of this signal with a radius greater than one and derive the phase spectrum.
8. Compute group delay spectrum. The locations at which the group delay takes a value of $-2\pi$ are identified as the GCIs.

3. Speech Corpora

The speech corpora used in the current work are described as follows:

3.1. NTIMIT

The NTIMIT database [13] or network speech database is developed by NYNEX Science and Technology Speech Communication Group. It is a telephone bandwidth version of the TIMIT database [14]. The database is derived by passing all 6300 phonetically balanced TIMIT utterances (10 utterances each from 630 speakers), recorded at a sampling rate of 16 kHz, through an actual telephone channel. The utterances are time-aligned with those in the TIMIT database. Since the NTIMIT database does not contain EGG recordings, reference GCIs are manually marked for 10 sentences (one each from 10 speakers - five male and five female).

3.2. Simulated telephone speech corpus

To perform an extensive analysis, the CMU Arctic [15], KED, and RAB databases [16], which contain simultaneous speech and EGG recordings, are converted to resemble telephone speech by using the G.191 software tools for speech and audio coding standardization, available from ITU-T. The sequence of steps described in [17] are used to derive the simulated telephone speech corpus. The data from these databases are considered at a sampling frequency of 8 kHz. The CMU Arctic database consists of one hour of data (1132 utterances) each, from three American speaker, BDL (male), JMK (male), and SLT (female). The KED corpus is recorded from an American male speaker and consists of 450 TIMIT utterances, corresponding to about 20 minutes of speech data, while the RAB database consists of around 2000 non-sense words, corresponding to 28 minutes of data, from an English male speaker.

4. Performance Analysis

The proposed algorithm is applied to the voiced segments in the corpora described in Section 3. The voiced segments are identified from the reference instants. For the NTIMIT database, as mentioned in Section 3, the reference instants are marked manually for one utterance each, from ten speakers. For the simulated telephone corpus, the reference GCIs are derived from the corresponding EGG signals (with the delay between the microphone and EGG input compensated), using the SIGMA [18] algorithm. The proposed algorithm can be applied to the voiced segment as a whole or to smaller frames of the segment. In the current work, each voiced segment is split into frames of size 25ms, irrespective of the speaker. The radius at which the group delay is measured is chosen to be 1.05.

DYPSSA, YAGA, and SEDREAMS are also applied to the speech corpora and the performance of these algorithms are compared against the proposed algorithm. Since the authors in [6] suggest that ZPF might not be appropriate for telephone speech, as mentioned in Section 1, the algorithm is not evaluated in the current work. The GCIs derived from a voiced segment of a male speaker from the NTIMIT database, using all the algorithms is compared in Fig. 3. The algorithms are evaluated in terms of the following three metrics:

- Identification rate (IDR) - Percentage of glottal cycles in which only one GCI is identified
- False alarm rate (FAR) - Percentage of glottal cycles that contain more than one GCI
- Miss rate (MR) - Percentage of glottal cycles in which no GCIs are identified

It is observed that the SIGMA algorithm results in spurious instants in certain voiced segments, particularly at the transition regions. Therefore, the YAGA algorithm, with a voiced/non-voiced classifier, is applied to the EGG signals (i.e. in YAGA, initially the source signal or the glottal flow derivative is estimated. In the current work, this source signal is replaced by the derivative of the EGG signal), to derive another set of reference GCIs. The IDR, FAR, and MR are calculated for all the algorithms with reference instants from SIGMA and YAGA, for the
The current work proposes a group delay-based method that considers speech to be a spectrum, to determine glottal closure instants from telephone speech. It exploits the fact that, the speech signal at the GCI exhibits an abrupt change in the amplitude and possesses a strong negative value. Assuming that the signal is the Fourier transform of an even signal enables the use of the properties that negative valleys in the Fourier transform correspond to zeros lying outside the unit circle in the z-plane and that the group delay at angular locations of these zeros has a value of $2\pi$. The performance of the proposed algorithm is evaluated on the NTIMIT database and a simulated telephone speech corpus derived from the CMU and CSTR databases. On an average the algorithm outperforms the existing algorithms in both corpora.

5. Conclusion

The authors would like to thank the speech group at IIT Guwahati and the authors of YAGA for providing the MATLAB codes to their algorithms.

6. Acknowledgements

The tables reveal that, on an average, the proposed algorithm outperforms the existing algorithms with an average IDR of 84.17% and 88.83% (with reference GCIs from SIGMA and YAGA, respectively) in the simulated telephone speech corpus and around 8% in NTIMIT. An improvement of around 4-5% in the simulated corpus and around 8% in NTIMIT is observed, when compared with the corresponding highest IDR attained by the existing algorithms. With regard to the performance of the algorithms on individual speakers, Table 1 reveals that the IDR obtained with the proposed algorithm is around 0.5% (for BDL) to 23% (for RAB) greater than the highest IDR obtained with the existing methods.

5. Conclusion

The current work proposes a group delay-based method that considers speech to be a spectrum, to determine glottal closure instants from telephone speech. It exploits the fact that, the speech signal at the GCI exhibits an abrupt change in the amplitude and possesses a strong negative value. Assuming that the signal is the Fourier transform of an even signal enables the use of the properties that negative valleys in the Fourier transform correspond to zeros lying outside the unit circle in the z-plane and that the group delay at angular locations of these zeros has a value of $2\pi$. The performance of the proposed algorithm is evaluated on the NTIMIT database and a simulated telephone speech corpus derived from the CMU and CSTR databases. On an average the algorithm outperforms the existing algorithms in both corpora.

6. Acknowledgements

The authors would like to thank the speech group at IIT Guwahati and the authors of YAGA for providing the MATLAB codes to their algorithms.
7. References


