Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 136–153

www.elsevier.com/locate/csl

Estimation of glottal closure instants from degraded speech using a phase-difference-based algorithm☆

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

Speech Lab, SSN College of Engineering, Chennai, India

Received 20 September 2016; received in revised form 17 March 2017; accepted 17 May 2017
Available online 23 May 2017

Abstract

The estimation of glottal closure instants (GCIs) plays a vital role in several glottal synchronous applications, and may not be restricted to clean speech. This necessitates the development of a GCI estimation algorithm that performs as well on degraded speech as on clean speech. Degradations in speech may be in the form of spectral or temporal perturbations. This could result in several spurious discontinuities, as in noisy speech, excitations that are not well-defined, as in band-limited speech, aperiodicity and variations in amplitude, as in pathological speech, thereby making the task of identifying GCIs more challenging. In this regard, a conditional group-delay/phase-difference-based (PD) algorithm that was initially proposed for use on clean speech is extended to operate on degraded speech, specifically telephone, noisy, and pathological speech. The performance of this algorithm is compared with six existing algorithms, in terms of identification, false alarm, and miss rates, and identification accuracy. It is observed that the PD algorithm is robust to degradations in speech and performs better than or on par with existing algorithms in all cases considered. Further, it is also observed that unlike existing algorithms, the PD algorithm scarcely estimates GCIs in non-voiced regions and this is verified in terms of a new metric proposed in the paper, namely, the spurious instants rate in non-voiced regions.

Keywords: Glottal closure instants; Phase difference; Group delay; Degraded speech; Telephone; Noisy; Pathological

1. Introduction

The estimation of glottal closure instants (GCIs) plays a vital role in glottal-synchronous processing of speech, in applications such as prosody modification (Thomas et al., 2008; Anushiya Rachel et al., 2014; 2015; Rao, 2012), speech dereverberation (Thomas et al., 2007; Gaubitch and Naylor, 2007), speech enhancement, glottal source modeling (Wong et al., 1979; Thomas et al., 2009), causal-anticausal deconvolution (Drugman et al., 2012), closed phase inverse filtering (Gudnason et al., 2014), and speech synthesis (Stylianou, 2001; Drugman et al., 2009). Vocal fold pathologies may be detected and diagnosed by performing pitch jitter and amplitude shimmer analyses on

☆ This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author.
E-mail address: anushiyarachel@yahoo.co.in (G. Anushiya Rachel), vijayalakshmip@ssn.edu.in (P. Vijayalakshmi), nagarajant@ssn.edu.in (T. Nagarajan).

http://dx.doi.org/10.1016/j.csl.2017.05.008
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

speech. These analyses also require an estimate of GCIs. Another application that is benefited by prior knowledge of GCIs is artificial bandwidth extension (ABWE) (Thomas, 2010; Thomas et al., 2010). With the advent of wide-band telephony, and the current co-existence of narrow-band and wide-band systems, it would be desirable to improve the quality of narrow-band telephone speech received on wide-band systems by ABWE. Further, ABWE is also useful in improving the intelligibility of telephone speech for cochlear implant users (Liu et al., 2009).

From the applications listed above, it is evident that the estimation of GCIs is not restricted to speech that is recorded in a noise-free environment and over a microphone. In other words, the speech signal may be degraded due to the presence of spectral and temporal perturbations, making the task of identifying GCIs more challenging. While GCIs are well-defined in clean speech, noisy speech is bound to contain additional rapid variations or discontinuities that make it difficult to distinguish GCIs from spurious fluctuations. Telephone speech, being band-limited, would contain excitations that are not well-defined, while pathological speech would contain a significant amount of amplitude shimmer and pitch jitter. This necessitates a GCI estimation algorithm that is not very sensitive to any degradation in speech.

Existing GCI estimation algorithms primarily focus either on identifying the fundamental frequency component or the maximum frequency component. Algorithms of the former kind, generally involve some kind of a low pass filtering operation to reduce the effect of the vocal tract/system information, which also tends to reduce the effect of noise components that may be present in the signal. Therefore, these algorithms would perform well on noisy speech. However, since band-limited/telephone speech do not contain the frequency components below 300 Hz, these algorithms may not perform well. On the other hand, the performance of algorithms that estimate GCIs by identifying the maximum frequency component or discontinuities in the signal, would not be affected to a greater extent on telephone speech. On noisy speech however, owing to the presence of discontinuities at instants other than the excitation instants, these algorithms may not perform very well.

Another means by which the existing algorithms may be categorized, is on the basis of whether they rely on an estimate of the pitch period and attempt to maintain periodicity or not. The algorithms that do rely on the pitch period, may not perform well when the speech signal possesses a significant amount of pitch jitter, as in pathological speech. Similarly, algorithms that are based on the energy of the speech signal will not be robust to amplitude shimmer. In this regard, it would be desirable that an algorithm be robust to any spectral or temporal perturbations that may be present in the signal.

1.1. Overview of existing algorithms

Commonly used algorithms to estimate GCIs from a speech signal, which are considered in this paper, are the group delay (GD)-based algorithm, dynamic programming projected phase slope algorithm (DYPSA), yet another GCI algorithm (YAGA), zero frequency filtering (ZFF) algorithm, speech event detection using the residual excitation and a mean-based signal (SEDREAMS) algorithm, and dynamic plosion index (DPI) algorithm. The GD algorithm (Smits and Yegnanarayana, 1995) attempts to identify GCIs from the linear prediction (LP) residual by using a group delay function, which contains a zero crossing at the location of the impulsive events in the signal. DYPSA (Naylor et al., 2007), which is an extension of the group-delay algorithm, aims at reducing the number of misses in the GD algorithm. In this regard, it initially uses a window size that is less than one pitch period to derive the group delay function and estimates all possible GCIs in the signal, including a number of spurious instants. These spurious instants are then eliminated by means of a dynamic programming-based candidate selection algorithm. YAGA (Thomas et al., 2012) is quite similar to DYPSA, and primarily varies in its use of iterative adaptive inverse filtering and wavelet transform to eliminate the vocal tract information from the speech signal, rather than using the LP residue. These three algorithms estimate GCIs by identifying the maximum frequency component and hence may be suitable for use on telephone speech, though not on noisy speech. Further, the GD algorithm requires that the GD function be calculated over a window of size equal to one to two pitch periods. Though DYPSA and YAGA do not rely on an exact estimate of the pitch period, the dynamic programming algorithm uses the local pitch period (pitch deviation cost defined in Naylor et al. (2007)) as a measure to accept or reject a candidate GCI. Therefore, these three algorithms may also be sensitive to pitch jitter.

The ZFF algorithm (Murty and Yegnanarayana, 2008) assumes that since the excitation is impulse-like, information about it must be available at all frequencies, including the zero frequency. It therefore derives a zero frequency filtered signal and attempts to identify GCIs from the zero crossings of this signal. The SEDREAMS algorithm