Glottal Closure Instant detection from a glottal shape estimate

Gilles Degottex, Axel Roebel, Xavier Rodet

Glottal Closure Instant detection from a glottal shape estimate

Gilles Degottex, Axel Roebel, Xavier Rodet

IRCAM - CNRS - UMR9912-STMS, Analysis-Synthesis Team
1, place Igor-Stravinsky, 75004 Paris
gilles.degottex@ircam.fr

Abstract

The GCI detection is a common problem in voice analysis used for voice transformation and synthesis. The proposed innovative idea is to use a glottal shape estimate and a standard lips radiation model instead of the common pre-emphasis when computing the vocal-tract filter estimate. The time-derivative glottal source is then computed from the division in frequency of the speech spectrum by the vocal-tract filter. Indeed, prominent peaks are easy to locate in the time-derivative glottal source. A whole process recovering all GCIs in a speech segment is therefore proposed taking advantage of this. The GCI estimator is finally evaluated with synthetic signals and Electro-Glotto-Graphic signals.

1. Introduction

The source-filter model (eq. 1) is used in this paper to represent the voice production. This model decompose the voice production in three main components: the glottal source, the vocal-tract and the lips radiation. The glottal source is assumed to be produced by the periodic opening and closing of the glottis (the space between the vocal folds). Then, the vocal-tract transform this source like a filter. Finally, the lips radiate this transformed source outside of the mouth adding one more filter effect. Accordingly, this model represent a periodic excitation of two consecutive filters by a glottal source. Analyzing a recorded speech segment, we try in this paper to temporally synchronize a glottal model (a shape model of the glottal source), with a speech signal period. This time synchronization can be reduced to the detection of a maximum excitation instant. Physiologically, this instant corresponds more or less to the closure of the glottis and it is the reason to call it Glottal Closure Instant (GCI).

Numerous GCI detection methods already exist [1, 2, 3, 4, 5]. The source model is often seen as a Dirac and thus, one of the best approaches is to flatten the phases of a residual spectrum like in the DYPSA method [1, 3]. In the time domain, the dual solution is the localization of a maximum of energy [4]. It is also possible to use the Frobenius norm to locate such an instant [5]. Additionally, the error of an ARX model can be minimized using a full glottal shape instead of a Dirac [2].

Some GCI detection methods assume the glottal source to be a minimum-phase signal, like the vocal-tract impulse response. The Linear Prediction (LP) residual is thus used to recover an impulse train which should correspond to GCIs. However, the glottal source is a mixed-phase signal [6]. Therefore, in the Z-plane, roots exist outside of the unit circle in the glottal source as in the speech signal. Computing a minimum-phase envelope of the speech spectrum (like with the LP), theses roots will be mirrored into the unit circle. However, the phase contributions are not the same for a stable or an unstable root. Consequently, this phase difference remains in the residual. When detecting a GCI, this phase difference will create a bias. Therefore, computing the vocal-tract filter by LP (or any other minimum-phase envelope), it is very important to remove first the contributions of the glottal source and the lips radiation from the speech signal. Usually, the speech signal is pre-emphasized to compensate these two contributions [7]. Instead, to retrieve the vocal-tract filter, we use an estimate of a glottal model and a common lips radiation model.

Then, by removing the vocal-tract effects from the speech signal by deconvolution, the glottal source and the lips radiation remains. We call this residual the *radiated glottal source* (fig. 1). In one period, the glottal source decreases fast enough to create a prominent negative peak on his time-derivative. This peak corresponds to the GCI. Consequently, since the lips radiation is a time-derivative, we will see that the GCI is easy to recover from the radiated glottal source (fig. 1).

Our GCI detection method is the following: First, we assume the fundamental frequency f_0 of the glottal source to be known thanks to numerous methods which are able to extract such a feature directly from the speech signal [8, 9, 10]. Secondly, for each period, we are looking for one particular sample among the sampled signal which indicate the glottal model position. Finally, using the f_0 estimate again a subdivision algorithm is also proposed to recover all GCIs in a speech segment.

Section 2 presents the different spectral relationships obtained from the source-filter model: Thanks to an estimate of a glottal model, the vocal-tract filter can be retrieved. Then, the radiated glottal source is obtained. The main sources of errors disturbing these computations are also discussed at the end of this section. Section 3 propose a whole GCI detection process. Finally, in section 4, this GCI estimator is evaluated with synthetic signals and compared to Electro-Glotto-Graphic signals.

2. Theoretical aspects: Speech model, vocal-tract filter derivation and radiated glottal source

In this section, we will first present the speech model in the frequency domain. Then, given a glottal model and a lips radiation model, the vocal-tract filter is obtained. Finally, by deconvolution of the speech signal by the vocal-tract filter, the radiated glottal source is expressed.

2.1. Speech model

In the frequency domain, the source-filter model of a voiced speech segment is expressed as:

$$S = H^{f_0} \cdot e^{j\omega\phi} \cdot G \cdot C_{-} \cdot L \qquad (1)$$

For reading convenience, since this equation holds for all frequencies, the frequency arguments have been removed ($X \equiv$

$X(\omega)$) for all terms which are not of pure linear phase. H^{f_0} is a harmonic structure modeling a periodic Dirac of fundamental frequency f_0. In one period, $e^{j\omega\phi}$ define the time position ϕ of the glottal shape. G is a mixed-phase spectrum defining the shape of the glottal source. C_- is a minimum-phase filter corresponding to the vocal-tract filter (the property of minimum-phase is denoted by the negative sign). In speech analysis, this filter is usually constrained to a stable all-pole filter corresponding to resonances. The minimum-phase assumption is more general, it implies only stability. Roots of the Z-transform (poles and/or zeros) have to be inside the unit circle. Finally, L is the filter corresponding to the lips radiation. This filter is usually associated to a time derivative [7, 11] and therefore, $L(\omega) = j\omega$.

2.2. Vocal-tract filter derivation

The following process is fully described in a simultaneous publication [12]. We will summarize the main ideas in this section. Thanks to a shape parameter estimate of a glottal model (like the one presented in section 3.1), it is possible to retrieve an approximation of the vocal-tract filter C_-^θ by division in the frequency domain (deconvolution in time):

$$C_-^\theta = E_-(\frac{S}{G^\theta \cdot L}) = E_-\left(\frac{H^{f_0} \cdot e^{j\omega\phi} \cdot G \cdot C_- \cdot L}{G^\theta \cdot L}\right) \quad (2)$$

G^θ is the glottal model parametrized by θ (*ex.* the Liljencrants-Fant model [13] parametrized by Rd [14]). $E_-(.)$ is a smooth minimum-phase envelope estimate of the argument like the *Cepstral Envelope*[15], the *LP* or the *Discrete All-Pole*[16]. Compared to the ARX methods [2, 17], computing the vocal-tract filter by this mean offers the choice of the envelope estimator.

Because $E_-(.)$ is computed from the amplitudes, it has the property of distributivity on spectrum multiplication. We can thus express the estimate of the vocal-tract filter:

$$C_-^\theta = \frac{E_-(H^{f_0} \cdot e^{j\omega\phi} \cdot G \cdot C_-)}{E_-(G^\theta)} = \frac{E_-(S/L)}{G_-^\theta} \quad (3)$$

We chose the *Cepstral Envelope* to compute the numerator. It is a minimum-phase envelope estimate of the speech spectrum after removing the lips radiation effect. The denominator is the glottal model replacing the mixed phases by the minimum phases. G_-^θ is thus retrieved from the real cepstrum.

Finally, from the two previous equations, focusing on the result of the computation of C_-^θ: $E_-(.)$ is computed from the amplitudes, $e^{j\omega\phi}$ is thus ignored. The order of $E_-(.)$ is limited to avoid the modelization of the harmonics H^{f_0}. Therefore, we assume the sampling of C_- by f_0 to be sufficient and the envelope estimator precise enough to neglect H^{f_0}:

$$C_-^\theta = \frac{E_-(H^{f_0} \cdot e^{j\omega\phi} \cdot G \cdot C_-)}{G_-^\theta} = \frac{G_- \cdot C_-}{G_-^\theta} \quad (4)$$

Finally, end of equation 3 gives the mean to compute C_-^θ and end of equation 4 gives the result of this computation.

2.3. Radiated glottal source

To estimate the time position of our glottal model, we are looking in the time domain, to the maximum negative peak in the radiated glottal source $\tilde{G}'$. Therefore, we will focus now on the deconvolution of the speech signal by the vocal-tract filter.

The speech spectrum S is computed with a window. Therefore, the following equation has to be studied:

$$\tilde{G}' = \frac{W \circledast S}{C_-^\theta} \quad (5)$$

However, if the main lobe of the window fall fast enough, one can assume:

$$\frac{W \circledast S}{C_-^\theta} \approx W \circledast \frac{S}{C_-^\theta} \quad (6)$$

The bigger is the variation of the amplitude spectrum of C_-^θ the bigger is the difference between the two sides of this equation. Since C_-^θ has a relatively smooth amplitude spectrum, the assumption will not introduce a large error. Consequently, we will focus on S/C_-^θ and consider the window effect remains in the final result.

From the speech model (eq. 1) and equation 4:

$$\tilde{G}' = \frac{S}{C_-^\theta} = H^{f_0} \cdot e^{j\omega\phi} \cdot \frac{G \cdot C_-}{G_- \cdot C_-/G_-^\theta} \cdot L \quad (7)$$

One can assume the envelope estimator is sufficiently precise [18]. Therefore, $|G_-| = |G|$ and the ratio of these two terms is an all-pass filter. We call this ratio the ***all-pass residual spectrum of*** G and we write it, for any variable, $X_{/-} = X/X_-$

$$\tilde{G}' = H^{f_0} \cdot e^{j\omega\phi} \cdot G_-^\theta \cdot G_{/-} \cdot L \quad (8)$$

In this equation, one can see the following results:

- *The pure linear phase term of* $\tilde{G}'$ is the one of the real glottal source $H^{f_0} \cdot e^{j\omega\phi} \cdot G$. Accordingly, the position of the glottal pulse is kept. This is the most important consequence for the detection of GCIs.
- *The amplitudes of* $\tilde{G}'$ are completely defined by the glottal model and the lips radiation $|G^\theta \cdot L|$.
- Conversely, the phases of the source model do not appear because they are replaced by the minimum-phase of the glottal model. Moreover, the all-pass residual of the real glottal source remains. Therefore, if the amplitudes of the glottal model are correct, *the phases of* $\tilde{G}'$ are the phases of the real source.

Finally, if we assume the real shape of the glottal source can be correctly represented by our chosen glottal model and there is only an error of parametrization $\Delta\theta$, from last equation:

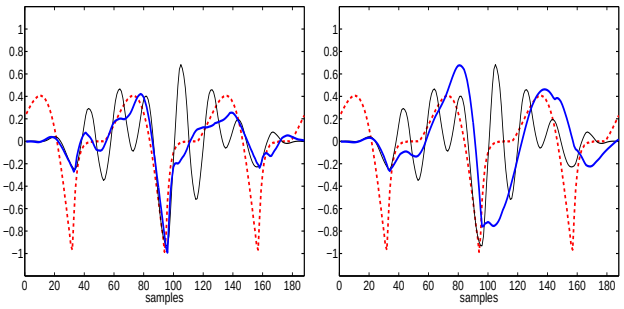
$$\tilde{G}' = H^{f_0} \cdot e^{j\omega\phi} \cdot G_-^\theta \cdot \frac{G^{\theta+\Delta\theta}}{G_-^{\theta+\Delta\theta}} \cdot L \quad (9)$$

defining $X^{\Delta\theta} = X^{\theta+\Delta\theta}/X^\theta$

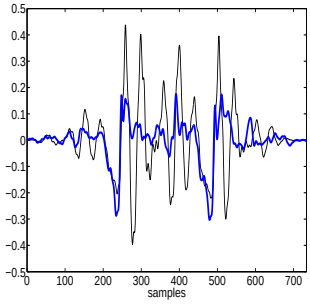
$$\tilde{G}' = H^{f_0} \cdot e^{j\omega\phi} \cdot G^\theta \cdot G_{/-}^{\Delta\theta} \cdot L \quad (10)$$

Like previously, the amplitudes are always defined by $|G^\theta \cdot L|$. Consequently, only the phases of $\tilde{G}'$ express the error of parametrization $\Delta\theta$.

Figure 1 shows examples of $\tilde{G}'$ for two synthetic signals and a real signal. The first synthetic signal (a) is computed without parametrization error, $\Delta\theta = 0 \Rightarrow G_{/-}^{\Delta\theta} = 1$. The second one (b) is computed with a parametrization error corresponding to $\approx 50\%$ of the parameter range. Both for synthesis and analysis, the glottal model is a Liljencrants-Fant (LF) model [14]. A *hanning* window is used to compute S and thus the windowing effect is also visible on $\tilde{G}'$. A few theoretical elements can be seen in this figure:



(a) Synthetic signals (b) Synthetic signals with error $\Delta\theta$



(c) Real signals

Figure 1: Examples of radiated glottal source $\tilde{G}'$ in the time domain: Synthetic signals (a,b): the waveform in thin black, the synthetic source in dashed red and $\tilde{G}'$ in thick blue. Real signals (c): the waveform in thin black and $\tilde{G}'$ in thick blue.

- In the synthetic examples (a,b): C_- is not perfectly reconstructed because the vocal-tract filter response is sampled by f_0. Consequently, ripples appears all along $\tilde{G}'$. However, the negative peak (in time domain) of $\tilde{G}'$ is close to the one of the real source.
- In the synthetic example with parametrization error (b): the negative peak of $\tilde{G}'$ is still prominent but slightly blurred. This result is very important for GCI detection because this peak position is hardly contested by other ripples. Consequently, a rough estimate of θ seems sufficient for such a detection (see section 4.1.1 for a quantitative evaluation).
- About the real example: the LF model has a strong negative peak and a relative smooth positive bump. However, positive peaks concentrated on one instant like at samples $\approx 400, \approx 230, \approx 490$ cannot be synthesized with such a model. More investigations should attempt to explain theses residues.

2.4. Main sources of errors

Consequence of the plane-wave hypothesis for GCI detection: The first limit of the source-filter model is the plane-wave hypothesis: Above $\approx 4000Hz$ the waves propagating inside the vocal-tract are not supposed to travel perpendicularly to the traveling axis [11]. Over such a frequency, the phases and the amplitudes of the glottal source are not properly propagated up to the mouth. Consequently, the source phases cannot be retrieved and thus there is no way to take advantage of them when trying to synchronize a source model with a recorded speech signal (ie. detecting GCIs). Therefore, the speech signal should be sampled at $8000Hz$. We are looking to more quantitative descriptions of this problem.

Preserved phases and polarity: The phases of the speech signal has to be preserved in the speech recording. Indeed, compression algorithms may disregard the phase information to improve the compression rate at the expense of quality. Additionally, the polarity of the signal has to be known. Indeed, in the time domain, the minimum of the time-derivative glottal shape is assumed to correspond to the GCI and the proposed method takes advantage of this. If the polarity is false, the proposed method may be confused with another peak.

3. Method

Using the theoretical results of the previous section, this section will present the complete proposed method with a few technical details.

In our implementation, the glottal model is the Liljencrants-Fant model [13]. The shape of this model is controlled by 3 parameters $\theta = (O_q, \alpha_m, t_a)$, the fundamental frequency f_0 and the excitation amplitude E_e. First, the relaxing parameter Rd is used to control a meaningful curve in the 3 shape parameter space [14, 19]. This shape parameter Rd is estimated thanks to an hypothesis made on the phases of the vocal-tract filter (see section 3.1). Secondly, we suppose f_0 to be known *a priori*. Numerous methods can be used to compute f_0 from the speech signal like *YIN*[8], *Swipep*[9] or by harmonic matching[10]. Finally, in the time domain, for each period, the minimum of the radiated glottal source $\tilde{G}'$ is assumed to correspond to the GCI, the prominent negative peak. Therefore, in each period, the proposed method looks for this minimum. Additionally, it is not necessary to estimate E_e, only the position of the peak is recovered.

The estimation of the radiated glottal source is computed following equations 3 and 7:

$$\tilde{G}' = S \cdot \frac{G_-^{Rd}}{E_-(S/L)} \tag{11}$$

where L is supposed to be a time derivative $L(\omega) = j\omega$. To compute the minimum-phase envelope estimate $E_-(.)$, we use the *Cepstral Envelope* because of his precision, robustness and the simple control of the parameters [18].

3.1. Rd estimate

We propose a simple way of computing a rough estimate of Rd without time synchronization. It seems sufficient for the proposed GCI detection method (sec. 4.1.1). We use the following hypothesis: *the phases of the vocal-tract filter around the glottal formant are negligible compared to the phases of the minimum-phase glottal source*:

$$\forall\omega \in [l,h] \qquad |\angle C_-(\omega)| \ll |\angle G_-(\omega)| \tag{12}$$

where the frequency band $[l,h]$ is chosen to contain all possible glottal formant frequencies ($\approx [1\cdot f_0, 3\cdot f_0]$). Consequently, C_- can be neglected in equation 4 when computing the phases of C_-^{Rd}:

$$\forall\omega \in [l,h] \qquad \angle C_-^{Rd}(\omega) = \angle G_-(\omega) - \angle G_-^{Rd}(\omega) \tag{13}$$

Finally, around the glottal formant, the phases of the vocal-tract filter estimate are biased by the shape parameter Rd of the glot-

tal model. Therefore, to estimate this shape parameter, the following error is minimized thanks to a Brent algorithm:

$$\epsilon(Rd) = \frac{1}{h-l}\int_l^h |\angle C_-^{Rd}(\omega)|\, d\omega \qquad (14)$$

More details are published in a simultaneous publication dedicated to this problem [12].

3.2. One-period detection

The minimum of $\tilde{G}'$ seems easy to locate, but when computing the spectrum, the windowing effect can displace it (fig. 1). From an arbitrary starting position, we propose an iterative method to converge to the nearest GCI:

1. Estimate of Rd and synthesis of G_-^{Rd}
2. Select $s : n$ periods of windowed speech signal with the starting position in the middle of the window
3. Compute S : the Discrete Fourier Transform (DFT) of s
4. Compute $(S/j\omega)_-$: the minimum-phase envelope estimate of S after removing the lips radiation effect
5. Compute $\tilde{G}' = S \cdot \frac{G_-^{Rd}}{(S/L)_-}$
6. Locate $GCI = argmin(DFT^{-1}(\tilde{G}'))$: locate the minimum of the radiated glottal source in the sampled time domain in a one-period interval around the starting position
7. Convergence test: stop if the GCI is already the sample of the middle of the window, else continue
8. Re-positioning: the window is moved to put the GCI in the middle of the window
9. back to 2

Usually, after 3 or 4 iterations the windowing effect is negligible and the method stops. In our implementation, we used a *hanning* window with a length of 3 periods. The vocal-tract filter is supposed to be stationary in such a window. Hence, the window cannot be arbitrary long.

3.3. Multiple-period detection

A method detecting different GCIs in a complete voiced speech segment has to take care of different aspects: 1) No GCI should be missed. 2) To minimize computation time, one GCI should not be detected twice. 3) An error of one GCI detection should not be propagated to detection of other GCIs. The main algorithm idea is to recursively subdivide a segment into two smaller segments if his duration is longer than a period. We take again advantage of the known f_0 and assume that the fundamental period $T_0(t)$ is known at any instants. As we will see, a high precision of T_0 is not necessary, the robustness against octave errors is more important.

1. First, put the start and end time of the speech segment into the top element of a stack
2. Select a starting and an ending position $[t_s, t_e]$ from the stack
3. From the middle position $t_m = (t_s + t_e)/2$, converge to the nearest GCI with the *One-period detection* method giving t_{GCI}
4. If $\alpha \cdot (t_{GCI} - t_s) > T_0(t_{GCI})$, put the time segment $[t_s, t_{GCI}]$ in the stack
5. and do the same for the segment $[t_{GCI}, t_e]$
6. If the stack is not empty, back to step 2

Doing so: 1) The algorithm subdivide the initial time interval into sub-intervals smaller than a period. Consequently, no GCI should be missed. 2) The search range is one period between two GCIs. Consequently, no one should be detected twice. 3) The subdivision process uses two different GCIs t_s and t_e, both should be erroneous to maximize the probability of propagating the error inside the time segment.

The α parameter control the minimum recursion size where a GCI is supposed to exist. Ideally, in a speech signal with constant f_0, α should be equal to 1. However, especially in speech, the f_0 variations are obviously not negligible. Moreover, a tolerance on the f_0 estimate should be accorded. Therefore, in our implementation, this parameter is fixed to $2/3$. Consequently, if the T_0 estimate is smaller than $1/3 \cdot T_0^*$, one third of the real period, the method creates a false alarm. Conversely, if the T_0 estimate is bigger than $1/3 \cdot 2T_0^*$, one third of two real periods, the method misses a GCI. However, creating a false alarm is better than missing a GCI. Indeed, the method always converge to the nearest GCI and so removing duplicated GCI is still possible but at the cost of useless computing time. An improved recursion test should dynamically set the α parameter from the f_0 standard deviation. Figure 2 shows an example of GCI detection on a *fry voice* segment, an especially aperiodic voice mode.

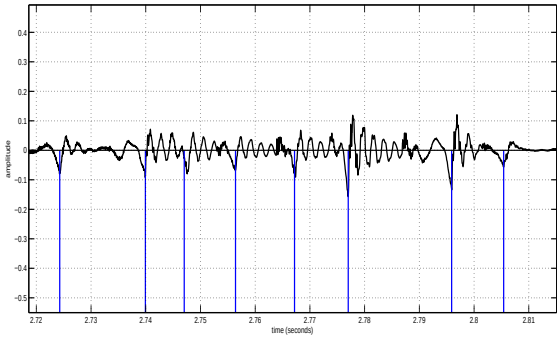


Figure 2: GCI detection on a *fry voice* segment

4. Comparison with synthetic signals and Electro-Glotto-Graphy

Validation of parameters estimate of glottal source is not obvious. Indeed, under strong assumptions, the glottal flow is usually associated to the source of a source-filter model [7], but the measurement *in vivo* of this flow is not yet possible. Nevertheless, correlates can be studied between a source estimate and physiological measurements like the Electro-Glotto-Graphy (EGG) [20].

In following sections, as a validation procedure, the GCI detection method is first evaluated with synthetic signals. Then, reference GCIs are computed from EGG signals and compared to the detected GCIs computed with the proposed method.

4.1. Evaluation with synthetic signals

In this section, the estimator is used on a synthetic signal with a known GCI with a shape model controlled by Rd and a known

f_0. Additionally, five different vocal-tract filters C_- are used to model five different vowels: /a/, /e/, /i/, /o/, /u/.

4.1.1. *Error related to Rd*

The used Rd estimate is not very accurate [12]. Consequently, the error of the GCI estimator related to the Rd error has to be evaluated. Thanks to equation 10, one can see that the term $G_{/-}^{\Delta\theta}$ has to be small enough to: 1) Do not challenge the gross position by keeping the global minimum of $\tilde{G}'$ close to the real GCI 2) Do not blur the linear-phase term $e^{j\omega\phi}$ and deteriorate the precision of this position. The error is computed for 13 f_0 values between $96 - 288Hz$, 23 Rd values between $0.3 - 2.5$ and the 5 different vocal-tract filters. The mean and standard deviation of the error is then computed. The results are shown in figure 3. For $|\Delta Rd| < 1$, the standard deviation is still below 10% of the period. Consequently, a rough estimate of Rd is sufficient.

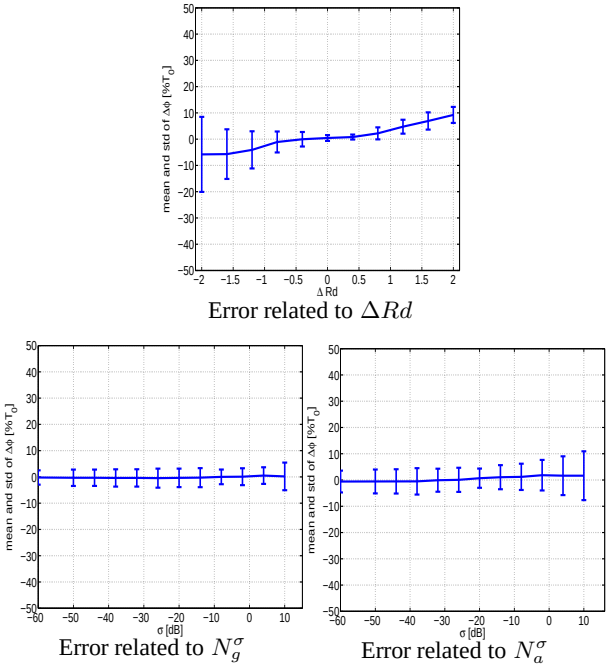
4.1.2. *Error related to the noise*

This test evaluate the estimator error related to two different white Gaussian noises of standard deviation σ: one is added to the signal while the other one is added to the the glottal source. Consequently, reference speech signals are synthesized with these two models:

$$S^{N_g^\sigma} = H^{f_0} \cdot e^{j\omega\phi} \cdot (G + N_g^\sigma) \cdot C_- \cdot L \qquad (15)$$

$$S^{N_a^\sigma} = H^{f_0} \cdot e^{j\omega\phi} \cdot G \cdot C_- \cdot L + N_a^\sigma \qquad (16)$$

While keeping the excitation amplitude E_e constant, the mean and standard deviation of the error is computed for σ values between $-50dB$ and $10dB$ relative to E_e (fig. 3). For each σ value, the error is computed 8 times with 13 f_0 values, 23 Rd values and the 5 different vocal-tract filters.



Error related to ΔRd

Error related to N_g^σ Error related to N_a^σ

Figure 3: Error related to the parameter error ΔRd of the glottal model, glottal noise N_g^σ and additive noise N_a^σ: The disturbing parameter on the horizontal axis; mean and standard deviation of the GCI estimator error on the vertical axis.

The noise effects can be analyzed analytically. Focusing on the noise influence, the frequency sampling of the harmonics can be neglected. From the two previous signal models (15,16), the following equations are derived:

$$\tilde{G}'^{N_g^\sigma} = e^{j\omega\phi} \cdot G_-^\theta \cdot (G + N_g^\sigma)_{/-} \cdot L \qquad (17)$$

$$\tilde{G}'^{N_a^\sigma} = e^{j\omega\phi} \cdot G_-^\theta \cdot (G \cdot C_- + \frac{N_a^\sigma}{L})_{/-} \cdot L \qquad (18)$$

Comparing these equations with the unnoisy equation 8, we make these conclusions: The amplitudes are always fixed by $|G^\theta \cdot L|$ because the all-pass residual is an all-pass filter. For the glottal noise, like in equation 8, the vocal-tract filter C_- is correctly removed by the minimum-phase envelope $E_-(.)$. Hence, the influence of this kind of noise should not be dependent on the formant positions. Conversely, for additive noise, the noise term interacts with the vocal-tract filter. Additionally, the noise is emphasized in low frequencies by the lips radiation effect. Consequently, $\tilde{G}'$ has to be high-pass filtered as high as possible, just below f_0.

4.2. Comparison with Electro-Glotto-Graphy

The validation of GCI estimators is usually made with an EGG. The main assumptions are strong correlations between the vocal folds motion, the glottal air flow and the glottal source of a source-filter model. On an EGG signal, GCIs are detected by locating peaks on the time derivative. These maximums of derivative are usually correlated to the instants when the vocal folds touch each other. However, they can corresponds to the instant when they move the fastest [20]. Additionally, the sub-glottal pressure takes an important part in the glottal flow shape [11, 21]. Therefore, in the time domain, the relations between the EGG signal and the glottal source are time dependent. Keeping in mind these differences, which define the bounds of the comparison, a reference set of GCIs is created from the EGG [2]. Then, they are compared with the detected GCIs.

The standard deviation between detected and reference GCIs are computed on the three *CMU Arctic* databases [22]. Moreover, the standard deviation normalized by the period is computed. The rate of errors bigger than 10% of the period (Gross Error Rate (GER)) is also computed. To compute this error, the propagation delay between the glottis and the microphone has to be compensated. A delay of $0.6ms$ is used. Three methods are compared: the proposed one, the DYPSA method [3] and a Group-Delay (GD) method [23] (table 1). The methods are evaluated only on voiced segments, but the determination of such segments is not obvious. To minimize the influence of the voicing estimator on the GCI detection results, the voiced segments are computed from the EGG: For each instant in the EGG signal, this instant is defined voiced if there is a reference GCI closer than one-half a period.

The standard deviation in milliseconds or normalized by the period is always smaller with the proposed method. About the standard deviation in milliseconds, the error of the proposed method is $\approx 71\%$ of the error of the DYPSA method. Relatively to the period, it is $\approx 53\%$. About the GER, except for the *jmk* database, the proposed method offers excellent results compared to the state of the art.

5. Conclusion

The contributions of the glottal source and the lips radiation have to be compensated before computing the vocal-tract filter. Accordingly, instead of the common pre-emphasis, we use a

DataBase	Method	std[ms]	std[%T_0]	GER[%]
Arctic bdl	proposed	0.40	2.9	2.45
Arctic bdl	DYPSA	0.71	6.36	9.41
Arctic bdl	GD	0.63	10.34	33.06
Arctic slt	proposed	0.26	4.28	5.43
Arctic slt	DYPSA	0.48	8.91	25.19
Arctic slt	GD	0.53	10.52	31.65
Arctic jmk	proposed	0.72	3.76	9.17
Arctic jmk	DYPSA	0.75	5.30	8.56
Arctic jmk	GD	1.02	8.38	16.94

Table 1: *std*: standard deviation of duration between the reference and the detected GCIs. *GER*: Gross Error Rate: number of differences$> 0.1 \cdot T_0$ compared to the number of reference GCIs.

rough estimate of the shape of a glottal model and the standard lips radiation model. By deconvolution of the speech signal by this vocal-tract filter, the radiated glottal source can thus be retrieved. Even if the shape of the glottal model is roughly estimated, a negative peak is still prominent in the radiated glottal source. Therefore, a robust GCI detection method has been proposed in this paper taking advantage of this. The method seems robust even for an aperiodic voice mode.

From an analytical point of view, we have seen that the amplitudes of the radiated glottal source are always the one of the chosen model. Conversely, the all-pass residual is the one of the real glottal source.

Moreover, thanks to a statistical evaluation and an analytic examination, we conclude that the proposed method is robust against glottal noise. Additive noise doesn't seems to be a lot more disturbing. Finally, the method is precise compared to the state of the art.

6. References

[1] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 325–333, 1995.

[2] Damien Vincent, Olivier Rosec, and Thierry Chonavel, "Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints," *ICASSP*, 2006.

[3] Anastasis Kounoudes, Patrick A. Naylor, and Mike Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech," *ICASSP*, 2002, Department of Electrical and Electronic Engineering, Imperial College, London, UK.

[4] Geoffroy Peeters and Xavier Rodet, "Non-stationary analysis/synthesis using spectrum peak shape distortion, phase and reassignement," in *ICSPAT (DSP-World)*, Inc. Miller Freeman, Ed., Orlando, USA, Novembre 1999.

[5] C. Ma, Y. Kamp, and L. Willems, "A frobenius norm approach to glottal closure detection from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 258 – 265, 1994.

[6] Boris Doval, Christophe dAlessandro, and Nathalie Henrich, "The voice source as a causal/anticausal linear filter," *VOQUAL*, 2003.

[7] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer Verlag, Berlin, 1976.

[8] Alain de Cheveigne and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, April 2002.

[9] Camacho, *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*, Ph.D. thesis, University of Florida, USA, December 2007.

[10] C. Yeh and A. Roebel, "A new score function for joint evaluation of multiple f0 hypothesis," in *International Conf. on Digital Audio Effects (DAFx)*, Naples, Italy, Octobre 2004, pp. 234–239.

[11] J.L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer Verlag, 1972.

[12] Gilles Degottex, Axel Roebel, and Xavier Rodet, "Shape parameter estimate for a glottal model without time position," in *Proc. 13th International Conference on Speech and Computer, SPECOM*, St. Petersburg, Russia, June 2009.

[13] Gunnar Fant, Johan Liljencrants, and Qi guang Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[14] Gunnar Fant, "The lf-model revisited. transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[15] A. Roebel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," *DAFX*, 2005.

[16] Amro El-Jaroudi and John Makhoul, "Discrete all-pole modeling," *IEEETSP*, 1991.

[17] Qiang Fu and Peter Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 492–501, march 2006.

[18] A. Roebel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters*, vol. 28-11, pp. 1343–1350, 2007.

[19] Hui-Ling Lu, *Toward a High-quality Singing Synthesizer with Vocal Texture Control*, Ph.D. thesis, Stanford, 2002.

[20] G. Degottex, E. Bianco, and X. Rodet, "Usual to particular phonatory situations studied with high-speed videoendoscopy," in *The 6th International Conference on Voice Physiology and Biomechanics*, Tempere, Finland, Aug. 2008, pp. 19–26.

[21] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, 1982.

[22] John Kominek and Alan W Black, "Cmu arctic databases for speech synthesis," 2003.

[23] R. Fernandez, *A Computational Model for the Automatic Recognition of Affect in Speech*, Ph.D. thesis, Massachusetts Institute of Technology, 2004.