Shape parameter estimate for a glottal model without time position

Gilles Degottex, Axel Roebel, Xavier Rodet

Shape parameter estimate for a glottal model without time position

Gilles Degottex, Axel Roebel, Xavier Rodet

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

Abstract

From a recorded speech signal, we propose to estimate a shape parameter of a glottal model without estimating his time position. Indeed, the literature usually propose to estimate the time position first (ex. by detecting Glottal Closure Instants). The vocal-tract filter estimate is expressed as a minimum-phase envelope estimation after removing the glottal model and a standard lips radiation model. Since this filter is mainly biased in low frequencies by the glottal model, an estimation method of a shape parameter is proposed. The evaluation of the results of such an estimator is still difficult. Therefore, this estimator is evaluated with synthetic signals. Such an estimate is useful for voice analysis (ex. glottal source estimation), voice transformation and synthesis.

1. Introduction

The source-filter model (eq. 1) is used in this paper to represent the voice production. This model decomposes the voice production in three main components: the glottal source, the vocal-tract filtering 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. Analyzing a recorded speech period, the shape of a glottal model (ex. Liljencrants-Fant model [1]) has a time position. In the literature, this time position is usually estimated by detection of Glottal Closure Instants (GCI) [2, 3]. Then, the shape parameters are estimated [4, 5, 6]. Thanks to an ARX model, Vincent *et al.* presented a joint estimate of the shape and the position [7]. We propose to obtain first a rough estimate of the shape, like presented in this paper. Then, we make the detection of GCIs [8]. As a refinement method, a joint estimate should conclude our glottal source estimation procedure.

To estimate the vocal-tract filter from the speech signal, the contributions of the glottal source and the lips radiation have to be compensated. The speech signal is usually pre-emphasized to compensate these two contributions [9]. Instead, we use a glottal model and a lips radiation model. This lips radiation is usually associated to a time-derivative. Consequently, we will see that the vocal-tract estimate is mainly biased by the shape parameter of the glottal model. Assuming the vocal-tract effects are negligible on low frequencies compared to the shape of the glottal source, we propose a mean to estimate the shape parameter of the glottal model.

Section 2 will present the derivation of the vocal-tract filter by means of a glottal model and a lips radiation model. Then, thanks to the previous assumption, an error function is proposed to estimate a shape parameter of the glottal model. Section 3 adds technical details about the estimation procedure and our implementation. The results of such an estimator are still difficult to validate. Indeed, under strong assumptions, the glottal flow can be assimilated to the source of the source-filter model. A measurement of this flow could be then compared to the glottal model estimates, but the measurement of such a flow is not yet possible *in vivo*. Therefore, in section 4, we evaluate the proposed estimator with synthetic signals. The comparison with Electro-Glotto-Graphic signals is also discussed.

2. Theoretical aspects: speech model, vocal-tract filter derivation and shape parameter estimate

This section will first present the source-filter model. Then, from a glottal model and a lips radiation model, the vocal-tract filter derivation is developed. Thanks to this derivation, the shape parameter estimate is presented in the last section.

2.1. Speech model

In the frequency domain, the source-filter model is expressed as:

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

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 a 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 minimum-phase property is denoted by the negative sign). In speech analysis, this filter is usually constrained to a stable all-pole filter corresponding to resonances [9]. 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 [9, 10] and therefore $L(\omega) = j\omega$.

2.2. Vocal-tract filter derivation

Our first goal is to show the relation between the vocal-tract estimate an the glottal model. First, we define $E_{-}(.)$ as a function computing the minimum-phase envelope spectrum of the argument. There are two means to do it: 1) The minimum phases of the argument can be retrieved trough the real cepstrum [11] 2) A minimum-phase envelope estimator can be used such as the *Cepstral envelope*[12], the *Linear Prediction* or the *Discrete All-Pole*[13]. Thanks to $E_{-}(.)$, by deconvolution in time, division in frequency, one can express the relation between the vocal-tract filter estimate C_{-}^{θ} and a given glottal model G^{θ}

parametrized by θ and the lips radiation model:

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

Compared to the ARX methods [7, 14], the main advantage of computing the vocal-tract filter by this mean offers the possibility to choose the envelope estimator (*Cepstral Envelope*, *LP*, *DAP*, etc.). Moreover, an articulatory model can be used to fit the argument of $E_-(.)$.

From the previous equation, since L is supposed to be known as $L(\omega) = j\omega$, it can be eliminated. Moreover, because $E_-(.)$ is computed from the amplitudes, this operator has the property of distributivity on spectrums multiplication. Therefore:

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

The numerator is equal to $E_-(S/L)$. It is the minimum-phase envelope estimate of the speech spectrum after removing the lips radiation effect. Concerning the denominator, G^θ is a model and is thus expressed analytically. Consequently, an envelope estimator is not necessary. However, the analytical derivation of G_-^θ, replacing the mixed phases of G^θ by the minimum phases, is not obvious. In a first approximation, G_-^θ can be computed from the real cepstrum. The computation of the vocal-tract filter is therefore:

$$C_-^\theta = \frac{E_-(S/L)}{G_-^\theta} \quad (4)$$

The numerator represents a common problem in the estimate of a filter: the filter response is sampled. In our case, the sampling is caused by the harmonic structure H^{f_0}. This problem will be discussed in section 2.4. We use the *Cepstral Envelope* to compute the numerator with a sufficiently low order to avoid the modelization of the harmonic structure H^{f_0}. Indeed, in our context, the *Cepstral Envelope* provides the best smooth minimum-phase envelope estimate compared to *LP* and *DAP* [15].

Focusing on the result of the computation of C_-^θ: the minimum-phase envelope estimate $E_-(.)$ has a limited order and is computed from amplitudes of the argument only. Therefore, according to the discussions above, H^{f_0} and $e^{j\omega\phi}$ are ignored. Finally, if C_- is assumed to be sufficiently well reconstructed, one can conclude with the following approximation:

$$C_-^\theta = \frac{E_-(H^{f_0} \cdot e^{j\omega\phi} \cdot G \cdot C_-)}{G_-^\theta} \approx \frac{G_- \cdot C_-}{G_-^\theta} \quad (5)$$

Therefore, the estimate of the vocal-tract filter is influenced by the difference between the real glottal source shape G and the glottal model G^θ. Moreover, these influences are linked to θ, the shape parameter.

2.3. Shape parameter estimate

We are looking for an estimate of θ and one can see the following result from equation 5: the estimation of C_- is mainly biased by the shape parameter of the glottal model. Therefore, we suggest in this section a hypothesis on C_- to optimize θ regarding to the ratio G_-/G_-^θ.

Considering the lips radiation effect with the glottal source $G \cdot L$ (like in the Liljencrants-Fant model definition [1]), one can see a bump dominating the amplitude spectrum in the result, like a formant. Since this bump is related to the glottal source it is usually called the glottal formant. We make 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 model*

$$\forall\omega \in [l,h] \qquad |\angle C_-(\omega)| \ll |\angle G_-^\theta(\omega)| \quad (6)$$

We will now evaluate this hypothesis for a simple case: The fundamental frequency f_0 is equal to $128Hz$. figure 1 shows the phase variation around the glottal formant of the vocal-tract filter and the minimum-phase Liljencrants-Fant glottal model. The ratio of the mean values is equal to $\approx$ 17 and the ratio of the standard deviation is $\approx$ 2. The distribution estimate of $|\angle G_-^\theta(\omega)|$ is precise since the shape parameter θ can be sampled as finely as necessary. However, the distribution estimate of $|\angle C_-(\omega)|$ is under estimated since we have only a limited number of different vocal-tract filters (5 filters in our case).

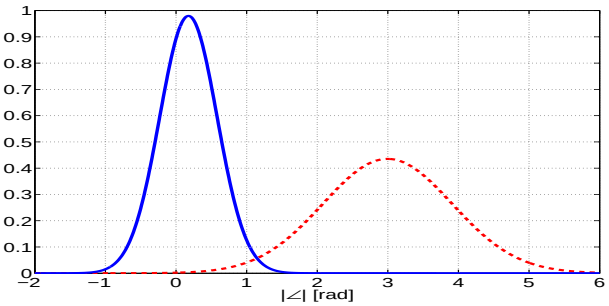


Figure 1: Phase variation around the glottal formant: the vocal-tract filter (solid blue line) and the minimum-phase glottal model (dashed red line).

In the general case, the phases of the vocal-tract filter are never zero in $[l,h]$. Moreover, the phase values depend on the formants position. The estimator is thus sensible to the variations of the vocal-tract filter. This variance can be seen in section 4. Moreover, a fundamental frequency $128Hz$ is a convenient assumption for a male speaker. However, as we will see in section 4.1.1, the bias of the estimator is proportional to the fundamental frequency. Consequently, the limits of application of such a hypothesis have to be kept in mind.

Using the hypothesis of equations 5 and 6, we find that the phases of the vocal tract estimate C_-^θ is approximately given by:

$$\forall\omega \in [l,h] \qquad \angle C_-^\theta(\omega) = \angle G_-(\omega) - \angle G_-^\theta(\omega) \quad (7)$$

Consequently, the shape parameter θ can be directly estimated by minimizing the phases of C_-^θ. The problem is solved in a minimization context and we use the following error function related to the shape parameter:

$$\epsilon(\theta) = \frac{1}{h-l}\int_l^h |\angle C_-^\theta(\omega)|\, d\omega \quad (8)$$

The frequency band $[l,h]$ is chosen to contain all possible glottal formant frequencies.

The vocal-tract filter reconstruction is based on the harmonics. Therefore, the confidence on phase values should be more important on harmonic frequencies. Moreover, as we will see in section 3, only a small number of harmonics exists in $[l,h]$ (actually $\approx$ 3). A few investigations should attempt to improve the computation of the error function.

This error function is computed on the vocal-tract filter which is computed from amplitudes since it is a minimum-phase filter. Therefore, the estimate of θ is indirectly based on the amplitudes. The phases offer a different conditioning of the problem.

2.4. Main sources of errors

Consequence of the plane-wave hypothesis: 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 [10]. Above this limit, the waves may not respect the minimum-phase property of the vocal-tract filter C_- and the all development above may not hold (the situation is the same if C is supposed to be an all-pole stable filter since it is a minimum-phase filter too). Therefore, the problem has to be solved for frequencies below $4000Hz$ and so, the speech signal has to be sampled at $8000Hz$. We are looking to more quantitative descriptions of this problem.

Vocal-tract filter reconstruction: The vocal-tract filter frequency response can be measured only if a source excites it. Therefore, the reconstruction of this filter is subject to the following conditions: i) To retrieve a correct vocal-tract filter response C_-, the sampling of this filter by f_0 has to be big enough compared to the shape of the filter. In the cepstral domain, the significant vocal-tract cepstral coefficients have to be lower than T_0. ii) Harmonics have to be present up to the Nyquist frequency. For example, in high frequencies, the noise level exceeds the harmonic level. The envelope estimator $E_-(.)$ will model this noise level. Consequently, when computing equation (4), this will affect the phases of C_-^θ over all frequencies by the minimum-phase property linking the phases to the amplitudes. To minimize this drawback, the problem as thus to be constrained to the smallest sufficient frequency band. iii) In this paper, we used a very precise filter model [15], but maybe flexible too much for the possible vocal-tract filters. Indeed, this filter is usually modelized only with poles, like the LP and DAP methods do, which are less flexible models. Further research should attempt to strike a balance between precision and flexibility.

3. Complete method

This section will present the complete method with a few technical details. In this paper, the Liljencrants-Fant model is used [1]. This model is controlled by 3 shape parameters (O_q, α_m, t_a), the fundamental frequency f_0 and the excitation amplitude E_e. We suppose f_0 to be known *a priori*. Numerous methods can be used to compute f_0 from the speech signal (*YIN*[16], *Swipep*[17] or by harmonic matching[18]). Since the proposed method works on phases, it is not necessary to estimate E_e. Finally, the relaxing parameter Rd is used to control a meaningful curve in the 3 shape parameter space [19, 5]. When Rd tends to big values, the time-derivative glottal model approaches to a period of a sinusoid. If Rd tends to small values, it approaches roughly to a negative Dirac. This Rd parameter control well the glottal formant. The literature describe the f_0-normalized glottal formant frequency by: $F_g = \frac{1}{2O_q\alpha_m}$. This value fit the LF-model amplitude spectrum by a triangle in a mel frequency scale [20]. However, since the term "formant" refer to a similarity to the vocal-tract formants, we prefer to focus on the frequency of the amplitude spectrum maximum. We call thus $F_{gm} = argmax(|G^{Rd}(\omega)|)$. For low Rd values, F_{gm} is high and inversely, like F_g. The fundamental frequency f_0 apply a frequency scaling to the glottal source spectrum. Therefore, F_{gm} is proportional to f_0. We use thus the frequency band $[l, h] = [f_0, 3 \cdot f_0]$, because $F_{gm}/f_0 = 1$ for the biggest Rd value and $F_{gm}/f_0 \approx 3$ for the smallest Rd value.

Our proposed method is the following: from equation 4, the term C_-^{Rd} is computed:

$$C_-^{Rd} = \frac{E_-(S/L)}{G_-^{Rd}} \tag{9}$$

To compute $E_-(S/L)$: as already discussed in the previous section, we chose the *Cepstral Envelope*. The error function (eq. 8) focuses on low frequencies. Therefore, it is very important to take care of the behavior of the minimum-phase envelope estimator around frequency zero. In the Z-plane, the derivative effect of the lips radiation is a zero on the unit circle. Such a zero creates a spectral shape which is difficult to imitate by a smooth envelope estimator and even more disturbing for an all-pole model like LP and DAP. To compute $E_-(S/L)$, our method is the following:

- Remove the lips radiation effect by integration of the speech signal, in frequency: divide S by $j\omega$
- Around the frequency zero, dividing by small numbers, the spectrum values can degenerate . Therefore, the amplitudes of S/L between 0 and $f_0/2$ have to be set to zero.
- Apply the *Cepstral Envelope* estimator[12] on the resulting spectrum.

Especially for low frequencies, this method provides a robust estimate of the minimum-phase envelope of (S/L).

According to equation 8, we need only the phases of C_-^{Rd} to compute $\epsilon(\theta)$. The phase operator can thus be distributed to the terms $E_-(S/L)$ and G_-^{Rd}. Therefore, we use the following error function which has the same minimum as the one of equation 8:

$$\epsilon'(Rd) = 1 - \frac{1}{h-l}\int_l^h \cos(\angle E_-(S/L) - \angle G_-^{Rd})\, d\omega \tag{10}$$

A Brent's method [21] is used to solve this minimization problem. This error function is very efficient since $\angle E_-(S/L)$ can be computed a single time. Such a computation is therefore possible in real-time.

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

Because the results of such an estimator are difficult to validate with real measurements, we evaluate them with synthetic signals. The correlates with the Electro-Glotto-Graphy (EGG) will be discussed.

4.1. Evaluation with Synthetic signals

In this section, the estimator is used on a synthetic signal controlled by Rd (the shape parameter to estimate) and f_0. Additionally, five different vocal-tract filters C_- are used to model five different vowels: /a/, /e/, /i/, /o/, /u/. Figure 2 (left plot) shows a comparison between known Rd values and the estimated Rd values for the 5 vocal-tract filters while keeping $f_0 = 128$.

For the next plots (fig. 2(right plot) and fig. 3). The mean and standard deviation of the estimator error is computed with the 5 different vocal-tract filters.

4.1.1. *Error related to f_0*

This test measure the error of the estimator related to f_0. A speech signal is synthesized with the base model of equation 1.

$$S = H^{f_0} \cdot G^{Rd} \cdot C_- \cdot j\omega \tag{11}$$

The results are shown in figure 2 (right plot). f_0 varies between $96Hz$ and $384Hz$ with a step of $1/8^{th}$ of an octave. The estimator variance increases with f_0 since the sampling by f_0 of the filter response C_- does not provide enough information to reconstruct C_- perfectly.

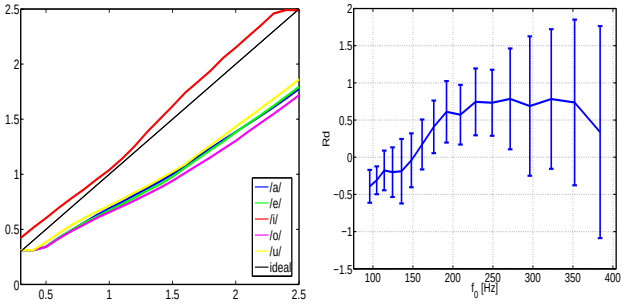


Figure 2: to the left: Synthetic Rd vs. estimated Rd for 5 different vowels. to the right: Rd mean and standard deviation of the error related to f_0.

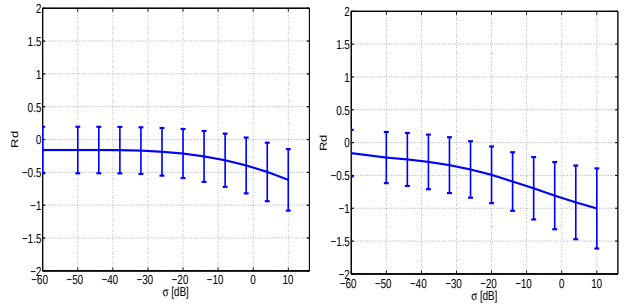
4.1.2. *Error related to the noise*

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

$$S^{N_g^\sigma} = (H^{f_0} \cdot G^{Rd} + N_g^\sigma) \cdot C_- \cdot j\omega \tag{12}$$

$$S^{N_a^\sigma} = H^{f_0} \cdot G^{Rd} \cdot C_- \cdot j\omega + N_a^\sigma \tag{13}$$

While keeping constant the amplitude parameter E_e, the mean and standard deviation of the estimator error is computed for values of σ between $-50dB$ and $10dB$ relative to E_e (fig. 3). For each σ value, the error is computed 8 times with f_0 values between $96Hz$ and $384Hz$ and the 5 different vocal-tract filters.



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

Figure 3: Mean and standard deviation of the estimator error related to glottal noise (to the left) and additive noise (to the right).

We will now analyze analytically the bias made by the noise terms. Therefore, Rd is considered to be known and H^{f_0} and $e^{j\omega\phi}$ are neglected:

For the glottal noise N_g^σ, from equations 2 and 12:

$$C_-^{N_g^\sigma} = E_- \left(\frac{(G + N_g^\sigma) \cdot C_- \cdot L}{G \cdot L} \right) = E_- \left(\frac{N_g^\sigma}{G} + 1 \right) \cdot C_-$$

For a given frequency, the larger is the amplitude of the glottal source $|G(\omega)|$, the smaller is the influence of the noise. Moreover, $|G(\omega)|$ decreases for increasing frequencies. Consequently, in low frequencies, N_g^σ does not disturb the estimate as we will see for the additive noise. However, in high frequencies, the noise influence is emphasized by $|G(\omega)|$. This is another reason to constrain the computation of C_-^θ to the smallest sufficient frequency band (*sec.* 2.4).

For the additive noise N_a^σ, from equations 2 and 13:

$$C_-^{N_a^\sigma} = E_- \left(\frac{G \cdot C_- \cdot L + N_a^\sigma}{G \cdot L} \right) = E_- \left(\frac{N_a^\sigma}{G \cdot L} + C_- \right)$$

In that case, the denominator is $G \cdot L$ and not G. Below F_{gm}, the derivative effect of the lips radiation make the amplitudes of this denominator low and therefore emphasize the noise effect. This is a second reason to manage carefully the frequency zero when computing $E_-(S/L)$ (sec. 3). Conversely, above F_{gm}, $|(G \cdot L)(\omega)|$ decreases slower for increasing frequencies than $|G(\omega)|$ does for glottal noise. In high frequencies, the influence of additive noise is therefore less important than glottal noise.

4.2. Correlation with Electro-Glotto-Graph

In voice analysis, the glottal source estimates are often compared to EGG [22, 7]. The main assumptions are strong correlations between the vocal folds motion, the glottal air flow and the glottal source of a source-filter model. Two features are usually extracted from the EGG: the glottal closure instant and the open-quotient (EGG-O_q). As Rd control the shape parameter O_q of the LF-model, it should be correlated to EGG-O_q. However, the sub-glottal pressure P_{sub} has an important role in the glottal flow shape [10, 23]. Consequently, the relations between P_{sub}, EGG-O_q and Rd are far from simple linear equations. On speech sentences, our experiments show sometimes very strong correlations between Rd and EGG-O_q. Meanwhile, this is not the case at all in other situations. A correlation measure gives a result of only ≈ 0.5 on the Arctic-bdl database [24]. Maybe the correlation is stronger in singing voices since sub-glottal pressure is more sustained than in speech. More investigations are needed about comparisons of *in vivo* measurements and estimates [25].

5. Conclusion

Concerning the estimation of the glottal source, contrarily to the current literature, we proposed to obtain first a rough estimate of the shape of a glottal model, then an estimate of his time position like with a GCI detection.

In the theoretical part, we have seen that the vocal-tract filter can be estimated in the frequency domain with a glottal model and a standard lips radiation model instead of a pre-accentuation. Even if we chose the *Cepstral Envelope* in our implementation, this formulation offers the possibility to use any other envelope estimator like LP, DAP and even an articulatory model.

We have seen that the vocal-tract filter estimate is biased by the shape parameter of the glottal model. Thanks to this result, we proposed to neglect the phases of the vocal-tract filter around the glottal formant to estimate the shape parameter of the glottal model.

Finally, this shape parameter estimate is dependent on the formant positions and rough for high f_0 values. However, it is independent of the time position of the glottal model and especially robust against glottal noise. Moreover, the estimation process is fast enough to run in real-time.

6. References

[1] 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.

[2] 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.

[3] 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.

[4] Baris Bozkurt, Boris Doval, Christophe DAlessandro, and Thierry Dutoit, "Zeros of z-transform (zzt) decomposition of speech for source-tract separation," *ICSLP*, 2004.

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

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

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

[8] Gilles Degottex, Axel Roebel, and Xavier Rodet, "Glottal closure instant detection from a glottal shape estimate," in *Proc. 13th International Conference on Speech and Computer, SPECOM*, St. Petersburg, Russia, June 2009.

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

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

[11] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*, Prentice–Hall, 1975.

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

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

[14] 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.

[15] 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.

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

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

[18] 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.

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

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

[21] R. P. Brent, *Algorithms for Minimization without derivatives*, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

[22] Nathalie Henrich, *Etude de la source glottique en voix parlee et chantee*, Ph.D. thesis, UPMC, 2001.

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

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

[25] 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.