Estimation of LF glottal source parameters based on an ARX model

Damien Vincent[(1)], *Olivier Rosec*[(1)] *and Thierry Chonavel*[(2)]

[(1)] France Telecom, R&D Division
[(2)] ENST Bretagne, Signal & Communication Department
{damien.vincent,olivier.rosec}@francetelecom.com, thierry.chonavel@enst-bretagne.fr

Abstract

We propose a method to estimate the glottal flow based on the ARX model of speech production and on the LF model of glottal flow. This method splits the analysis in two stages: a low frequency analysis to estimate the glottal source parameters which have mainly a low pass effect and a second step to refine the parameters which have also a high pass effect. Along with this new analysis scheme, we introduce a new algorithm to efficiently minimize the nonlinear function resulting from the least square criterion applied to the ARX model. Results on synthetic and natural speech signals prove the effectiveness of the proposed method.

1. Introduction

Several studies ([1, 2]) have highlighted the link between vocal quality and the glottal source signal. This link allows to consider some interesting applications targeting voice characterization or voice transformation.

Direct measurement of the glottal flow can only be carried out using some intrusive methods and therefore is incompatible with the aforementioned applications. To solve this issue, the glottal source signal must be estimated using only the speech signal. To deal with this deconvolution problem, numerous solutions have been proposed including time domain inverse filtering [3, 4], as well as separation of glottal source and vocal tract in the z-plane [5].

Interestingly, in [6], an ARX (Auto-Regressive eXogenous) model has been proposed as a representation of the speech production process and combined with the RK (Rosenberg-Klatt) glottal source model. However, this method has to cope with a very difficult optimization problem for which no practical satisfactory solution has been given. In this article we propose a new estimation scheme based on an ARX model excited by a LF (Liljencrant-Fant) model [7] of the glottal source. Our contributions lie in the split of the analysis in two stages: a low frequency analysis to estimate the parameters whose effects are mainly located in the low part of the spectrum followed by a full band analysis to refine parameters which also have high frequency effects. This paper is organized as follows. Section 2 presents the ARX model as well as the LF model. The low band and full band analysis are described in section 3 and 4 while section 5 describes the experiments led on synthetic as well as on natural speech.

2. Speech production model

2.1. The ARX model

To extract the glottal source, we have to make some assumptions on the physical process of speech production. In its simplified form, 3 components have an effect in the speech production: the glottis, the vocal tract and the lips. Neglecting the nonlinear effects such as the interaction between the vocal folds and the vocal tract and modeling the lip radiation as a differentiator, the speech production can be represented by the following ARX model:

$$s(n) = -\sum_{k=1}^{p} a_k(n)s(n-k) + b_0 u(n) + e(n) \quad , \qquad (1)$$

where $s(n)$ denotes the speech signal, $u(n)$ the glottal flow derivative, $a_k(n)$ the time-varying AR coefficients modeling the vocal tract and $e(n)$ the residual.

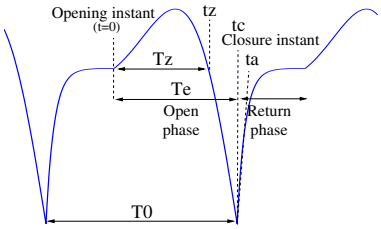

Figure 1: *One period of the glottal flow derivative.*

2.2. The LF glottal flow model

In the ARX model, some constraints can be added to the source component of the AR model which by itself gives only a poor estimation of the glottal source for voiced speech. A particularly convenient parametrization can be obtained by using the LF model, shown in figure 1, which enables the characterization of the glottal source signal with 5 parameters: one for the location of the glottal source (the reference is usually the glottal closure instant), one for the amplitude (already in the ARX model in the b_0 coefficient) and three to define the shape of the glottal flow. Among the possible parameter sets to define the shape, the vector $\theta = (O_q, \alpha_m, Q_a)$ has been chosen: O_q corresponds to the open quotient ($O_q = \frac{T_e}{T_0}$), α_m to the asymmetry coefficient ($\alpha_m = \frac{T_z}{T_e}$) and Q_a to the return phase quotient ($Q_a = \frac{t_a - t_c}{(1-O_q)T_0}$). Θ denotes the space of shape parameters. The explicit expression of the model for one fundamental period is given by:

$$\begin{aligned} u(t) &= E_1 e^{at}\sin(wt) && 0 \le t \le T_e \\ u(t) &= -E_2\left[e^{-b(t-T_e)} - e^{-b(T_0-T_e)}\right] && T_e \le t \le T_0 \end{aligned} \qquad (2)$$

where the parameters a, b and w are implicitly connected to θ.

In case of an abrupt return of the glottis ($Q_a = 0$), the glottal source spectrum has an asymptotic behaviour similar to

a second order filter (see figure 2): this resonance is also called glottal formant by analogy with the frequency resonances of the vocal tract. Adding a return phase increases the spectral slope by -6dB/oct on frequencies above the cut-off frequency $F_a = \frac{1}{2\pi}\frac{1}{Q_a(1-O_q)T_0}$, but has almost no influence on the glottal formant frequency which mainly depends on the open quotient O_q and the asymmetry coefficient α_m.

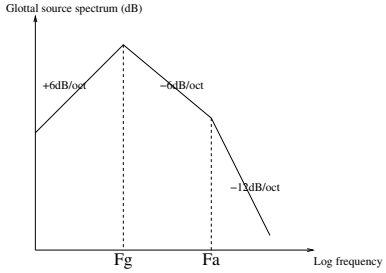

Figure 2: *Asymptotic behaviour of the glottal flow derivative spectrum.*

3. Estimation of the low frequency component of the glottal source

3.1. Analysis

3.1.1. Overview

In the previous section, we have seen that the glottal formant is a low frequency component of the glottal source (F_g is below 1kHz). Therefore, an analysis over [0-1kHz] is able to capture the glottal formant and thus should lead to a good estimation of the shape parameters O_q and α_m. In the following, $s_l(n)$ and $u_l(n)$ denote respectively the low pass component ([0-1kHz]) of the speech signal $s(n)$ and the glottal source signal $u(n)$. The ARX model can be rewritten using the low pass speech signal:

$$s_l(n) = -\sum_{k=1}^{p} a_k(n) s_l(n-dk) + b_0 u_l(n) + e(n) \qquad (3)$$

where d is the decimation coefficient ($d = 4$ for $f_s = 8kHz$).

In the analysis scheme presented thereafter, the fundamental frequency f_0 is assumed to be known (eg estimated by the YIN method [8]). For a given glottal closure instant t_c, the speech signal is analysed over a two fundamental period interval using a Hanning window centered on t_c. Under these conditions, the ARX model can be written:

$$S = M_u A + E \quad , \qquad (4)$$

where:
$A = (a_1, \cdots, a_p, b_0)^T$
$M_u = [-D|U]$
$D = \begin{pmatrix} w(ta_1)s_l(ta_1-d) & \cdots & w(ta_1)s_l(ta_1-pd) \\ \cdots & \cdots & \cdots \\ w(ta_N)s_l(ta_N-d) & \cdots & w(ta_N)s_l(ta_N-pd) \end{pmatrix}$
$S = (w(ta_1)s_l(ta_1), \cdots, w(ta_N)s_l(ta_N))^T$
$U = (w(ta_1)u_l(ta_1), \cdots, w(ta_N)u_l(ta_N))^T$
and where $\{ta_k\} = [t_c - T_0, t_c + T_0] \bigcap (t_c + d\mathbb{Z})$ denote the time analysis instants.

The glottal source parameters are estimated by minimizing the least square error criterion $\|E\|^2$ over the source parameters and the vocal tract parameters. When the source parameters are given, the system is linear and the partial minimization with respect to A can be computed easily; the partial minimum will be noted $E_u(t_c, \theta) = \|S - M_u(M_u^T M_u)^{-1} M_u^T S\|^2$. The minimization with respect to the source parameters is much more difficult: in section 3.2, we will introduce an efficient method to optimize E_u with respect to θ. Regarding the glottal closure instant, the minimization can be restricted to the vicinity of an initial estimate obtained either by a group delay algorithm ([9]) or by EGG (Electro-GlottoGraph) measurements if available. We also consider an AR order selection scheme based on a new cost function $\bar{E}_u$ (defined in the next section) instead of the function E_u. Algorithm 1 summarizes the global low pass estimation process.

Algorithm 1: Global optimization algorithm

forall $t_c \in$ *list of a priori locations* **do**
  Generation of the analysis window ;
  forall $p \in$ *list of AR orders to be tested* **do**
    Minimization of $\bar{E}_u$ with respect to θ
    $\Rightarrow \hat{\theta}(t_c, p)$;
$(\hat{t}_c, \hat{p}) = \text{argmin}_{t_c,p}\, \bar{E}_u(t_c, \hat{\theta}(t_c, p))$;
Shape parameters given by: $\hat{\theta}(\hat{t_c}, \hat{p})$;

3.1.2. AR order selection

The cost function $E_u(t_c, \theta)$ is not convenient to select the optimal AR order p as the error E_u is always decreasing as the order p is increasing. To get a better measure of how well the LF glottal source models the speech signal, we introduce a normalized prediction error $\bar{E}_u(t_c, \theta)$ defined by:

$$\bar{E}_u(t_c, \theta) = \frac{E_u(t_c, \theta)}{E_0(t_c)} = \frac{\|S - M_u(M_u^T M_u)^{-1} M_u^T S\|^2}{\|S - D(D^T D)^{-1} D^T S\|^2} \qquad (5)$$

where $E_0(t_c)$ corresponds to the usual LPC prediction error using the same analysis window centered on t_c. $E_0(t_c)$ is always greater than the prediction error $E_u(t_c, \theta)$ based on a source model. Thus, the new cost function is normalized between 0 and 1: it tends towards 0 if the speech is strictly voiced and the source model perfectly fits the real glottal source and tends towards 1 if either the speech is unvoiced or if the LF source model is far away from the real glottal signal of the current frame.

This new cost function is not decreasing anymore as the order p is increasing: the error E_u is decreasing but E_0 is also decreasing. For an analysis on the frequency range [0-1kHz], orders from 2 to 5 are used. Another advantage of the cost function $\bar{E}_u$ is its independance to the signal amplitude.

3.2. Efficient cost minimization

In algorithm 1, we did not mention how to optimize the function $\bar{E}_u(t_c, \theta)$ with respect to θ. This kind of nonlinear minimization problem could be solved by a simulated annealing algorithm but would be very time consuming. Instead, we use a two step algorithm: θ is first estimated by an exhaustive evaluation of $\bar{E}_u(t_c, \theta)$ over a finite subspace of Θ; then this first estimate is further refined using a simplex algorithm detailed in [10].

3.2.1. LF space quantization

To quantize the glottal source shape, the correlation function ρ between two glottal waveforms $u_{\theta 1}$ and $u_{\theta 2}$ (where $\theta 1$ and $\theta 2$ are shape vectors) is used as a similarity measure. Thus, given a minimum correlation coefficient ρ_m, the corresponding finite subspace $\tilde{\Theta}_{\rho_m}$ is built so as to satisfy the following two conditions:

1. covering condition:
   $\forall \theta \in \Theta : \exists \tilde{\theta} \in \tilde{\Theta}_{\rho_m}$ such that $\rho(u_\theta, u_{\tilde{\theta}}) \geq \rho_m$;
2. condition to prevent redundancy:
   $\forall \tilde{\theta}_1, \tilde{\theta}_2 \in \tilde{\Theta}_{\rho_m} : \rho(u_{\tilde{\theta}_1}, u_{\tilde{\theta}_2}) < \rho_m$.

Algorithm 2 allows to create such a subspace whose cardinal is denoted thereafter by L.

Algorithm 2: Algorithm for building the subspace $\tilde{\Theta}$

```
Θ̃ ← ∅ and Δ ← 0 ;
repeat
    Vector θ generated by uniform sampling over Θ;
    if max_{θ̃∈Θ̃} ρ_{u_θ̃, u_θ} < ρ_m then
        Θ̃ ← Θ̃ ⋃ {θ};
        Δ ← 0;
    else
        Δ ← Δ + 1;
until Δ < Δ_max;
```

The number of shape vectors increases rapidly as the correlation coefficient ρ_m tends toward 1, as shown in figure 3. The coefficient ρ_m must be high enough so that the estimated shape is close to the optimal shape. However, a space $\tilde{\Theta}_{\rho_m}$ too large would lead to a high complexity without necessarily improving the estimation since in the second step, the simplex algorithm compensates for the discrete structure of the subspace $\tilde{\Theta}_{\rho_m}$. Taking $\rho_m = 0.99$ proved to be a good compromise between these two opposite considerations.

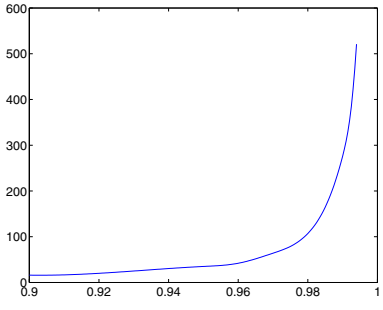

Figure 3: *Cardinal of the subspace $\tilde{\Theta}$ with respect to ρ_m.*

From a spectral point of view, the quantization of space Θ mainly captures the glottal formant shape as the glottal formant corresponds to the frequency range of highest energy, while the spectral slope is less well represented in the subspace $\tilde{\Theta}_{\rho_m}$.

3.2.2. Algorithmic considerations

As can be seen from the above description, the joint estimation of the source parameters and the AR filter coefficients requires the resolution of many linear systems ($280 + 30$ using $\rho_m = 0.99$). In order to further decrease the complexity of the system resolution $S = M_u A$, we can take advantage of the similarity of the matrix M_u when only the source shape parameters are modified: a modification of θ results only in a modification of the last column of the matrix M_u. Thus, the p first steps of the QR algorithm, needed to solve the linear system in a least square sense, can be computed once, then the evaluation of $\bar{E}_u$ for the $L-1$ remaining shape parameters corresponds only to an update of the QR algorithm. This leads to a global complexity of $O(Np^2) + O((L-1)Np)$ to evaluate $\bar{E}_u$ over $\tilde{\Theta}$ instead of $O(LNp^2)$ without using this similarity property.

The algorithm also requires an explicit generation of the LF source signal. Instead of using equation (2) to generate the signal, a faster way consists in a recursive generation of the source signal: the open phase can be computed using a second order filter and the return phase using only a first order filter.

4. Full band analysis

The low band analysis yields good estimates of O_q and α_m and a rather precise location of the glottal closure instant (the results will be given in section 5) but may lead to a poor estimation of Q_a especially if the cut-off frequency F_a is greater than 1kHz. The full band analysis aims at improving the Q_a estimation and getting a more accurate glottal closure instant location. This analysis is carried out in the same way as the low band analysis except that the least square criterion is minimized only with respect to Q_a and t_c. Thus, O_q and α_m remain unchanged while t_c is constrained to be close to the value estimated from the low band analysis. In this full band analysis, the decimation coefficient is set to 1 and the AR order is set to 14.

Once the vocal tract and source parameters have been estimated, if the LF model does not perfectly fit the deterministic part of the glottal source, the residual $e(n)$ holds a fraction of the deterministic part of the glottal source. This modeling error can be included either in the glottal source as a LF model deviation or in the AR filter modeling the vocal tract. When the speech is purely voiced, the modeling error can be included in the AR filter by estimating the AR frequency response from the speech and the glottal source spectral envelopes respectively denoted by $S_e(f)$ and $U_e(f)$: $\frac{G}{A(f)} = \frac{S_e(f)}{U_e(f)}$.

5. Experiments

5.1. Synthetic signals

Two synthetic tests have been designed: the first one to check the validity of the method, and the second one to analyse the behaviour of the method when a model perturbation is introduced to generate the synthetic speech signal. The results will be analysed using the standard deviations of the open quotient O_q, the asymmetry coefficient α_m, and of $Q_a(1-O_q)$ which is more physically significant than Q_a itself ($F_a = \frac{1}{2\pi}\frac{1}{Q_a(1-O_q)T_0}$) and by providing the average correlation coefficient ρ_{mean} between the estimated glottal flow derivative and the theoretical glottal flow derivative. As the estimation did not reveal any meaningful bias, this bias is not provided in the results.

In the first experiment, the speech signal is generated at $fs = 8$kHz using a glottal source whose shape parameters are uniform random variables such that $O_q \in [0.3, 0.9]$, $\alpha_m \in [0.66, 0.90]$ and $Q_a \in [0, 0.30]$. Vocal tract AR filters are uniformly sampled from a set of seven predetermined filters corresponding to various French vowels. A noise component (HNR=25dB) has been added to the glottal source component. Table 1 shows the results for two fundamental frequencies. When $f_0 = 100$Hz, the shape parameters O_q and α_m are well

estimated, the corresponding estimator variance is below the human hearing threshold which is $0.14O_q$ for O_q (the threshold depends linearly on O_q) and 0.022 for α_m as stated by Henrich in [1]. For $f_0 = 180$Hz, the glottal formant is not as well estimated, resulting in a estimation variance increase. However the variances are below the hearing thresholds of non-expert people [1].

f_0	σ_{O_q}	σ_{α_m}	$\sigma_{Q_a(1-O_q)}$	ρ_{mean}
100	0.023	0.013	0.005	0.996
180	0.044	0.033	0.019	0.979

Table 1: *Glottal source parameter estimation using a ideal source filter speech generation.*

The second set of experiments moves away from the perfect linear source filter model and introduces a basic modeling of the source/filter interaction. We define two AR filters, the first one models the vocal tract when the glottis is closed, the second one models the vocal tract when the glottal flow is highest; for intermediate glottal flow values, the AR filter is set by a linear interpolation between these two filters (interpolation done on the LSF coefficients). The filter on the open phase is supposed to have a first formant with a wider bandwidth than the AR filter of the closed phase. Table 2 shows only a slight degradation on the O_q estimate. An increase of the fundamental frequency has the same effect as the previous synthetic test.

f0	σ_{O_q}	σ_{α_m}	$\sigma_{Q_a(1-O_q)}$	ρ_{mean}
100	0.033	0.018	0.006	0.995
180	0.070	0.043	0.015	0.972

Table 2: *Glottal source parameter estimation when nonlinear effects are introduced.*

5.2. Natural speech

The French words *"Médecine prédictive"* pronounced by a female speaker have been analysed by the proposed method. Figure 4 shows the waveform and the spectrogram of the original and re-synthetized speech signals. The similarity in the time and frequency domain is confirmed by an informal listening test: although the synthesis using only the LF model is not completely transparent, the re-synthetized speech signal is very close to the original one.

6. Conclusion

We have described a new method based on a joint estimation of the vocal tract and source parameters which takes advantage of the source parameter influence in the spectral domain. It basically splits the estimation process in two steps: the first one to estimate the glottal formant by an analysis on [0-1kHz] and the second one to estimate the return phase quotient by a full band analysis. Along with this estimation process, a new normalized least square error criterion has been proposed in order to select the AR model order. Moreover, particular care was dedicated to algorithmic optimization so as to decrease the overall computational load. The whole process leads to good results on synthetic signals as well as on natural speech signals. However, we noticed a degradation of the results when the fundamental frequency increases; a future work is needed to ensure the continuity of source parameters over frames to compensate for this

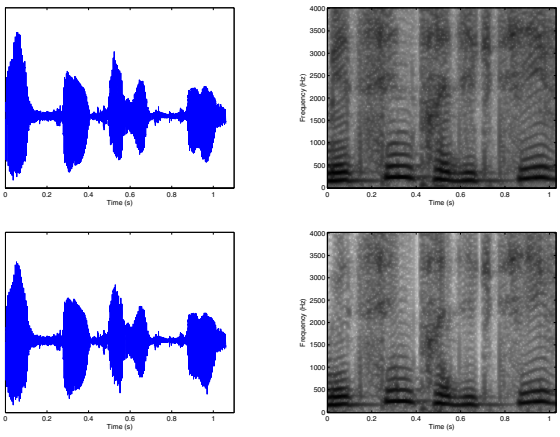

Figure 4: *Original speech signal in the time and frequency domain (top). Resynthesized speech signal (bottom).*

estimation variance increase. We also plan to apply this analysis scheme in speech modification.

7. References

[1] N. Henrich, "Etude de la source glottique en voix parlée et chantée," Ph.D. dissertation, Université de Paris 6, November 2001.

[2] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, February 1990.

[3] M. D. Plumpe, T. F. Quateri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 5, pp. 569–586, September 1999.

[4] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Com.*, vol. 11, pp. 109–118, 1992.

[5] B. Bozkurt, B. Doval, and C. D'Alessandro, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347, April 2005.

[6] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model," *IEICE Trans. Inf. Syst.*, vol. E78-D, no. 6, pp. 738–743, June 1995.

[7] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[8] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, April 2002.

[9] R. Smits and B. Yegnanarayana, "Determination of instants of significantg excitation in speech using group delay function," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 325–333, September 1995.

[10] J. Nelder and R. Mead, "A simplex method for function minimisation," *Computer Journal*, vol. 7, pp. 308–313, 1964.