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 non-linear 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:

\[ s(n) = - \sum_{k=1}^{p} a_k(n)s(n-k) + b_0u(n) + e(n) \]  

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.

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 \( \theta = (O_q, \alpha_m, Q_o) \) has been chosen: \( O_q \) corresponds to the open quotient \( (O_q = \frac{T_e}{T_0}) \), \( \alpha_m \) to the asymmetry coefficient \( (\alpha_m = \frac{T_e}{T_0}) \) and \( Q_o \) to the return phase quotient \( (Q_o = \frac{T_e}{1-O_qT_0}) \). \( \Theta \) denotes the space of shape parameters. The explicit expression of the model for one fundamental period is given by:

\[
\begin{align*}
    u(t) &= E_1 e^{at} \sin(wt) \\
    u(t) &= -E_2 \left[ e^{-b(t-T_e)} - e^{-b(T_0-T_e)} \right] \quad 0 \leq t \leq T_e \\
    u(t) &= E_3 \left[ e^{-a(T_0-t)} - e^{-a(T_0-T_e)} \right] \quad T_e \leq t \leq T_0
\end{align*}
\]

where the parameters \( a, b \) and \( w \) are implicitly connected to \( \theta \).

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

\[ X \]
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
\( F_G = \frac{1}{2\pi Q_s(1-\alpha_m)T_0} \), but has almost no influence on the glottal
formant frequency which mainly depends on the open quotient
\( Q_s \) and the asymmetry coefficient \( \alpha_m \).

In the previous section, we have seen that the glottal formant is
3.1. Overview

spectrum.

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 two fundamental period interval
using a Hanning window centered on \( t_c \). Under these con-
ditions, the ARX model can be written:

\[
S = M_u A + E ,
\]

where:

\[
A = (a_1, \cdots, a_p, b_0)^T
\]

\[
M_u = [-D|U]
\]

\[
D = \begin{pmatrix}
w(ta_1)s(ta_1-d) & \cdots & w(ta_1)s(ta_1-pd) \\
\cdots & \cdots & \cdots \\
w(taN)s(taN-d) & \cdots & w(taN)s(taN-pd)
\end{pmatrix}
\]

\[
S = (w(ta_1)s(ta_1), \cdots, w(taN)s(taN))^T
\]

\[
U = (w(ta_1)u(ta_1), \cdots, w(taN)u(taN))^T
\]

and where \( \{ t_c \} = \{ t_c - T_0, t_c + T_0 \} \cap (t_c + d\mathbb{Z}) \) 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 re-
spect to \( A \) can be computed easily; the partial minimum will be
noted \( E_{0}(t_c, \theta) = \| S - M_u M_u^T M_u^{-1} M_u^T S \|^2 \). The mini-

\begin{algorithm}
\begin{algorithmic}
\caption{Global optimization algorithm}
\ForAll { \( t \in \text{list of a priori locations} \) }
\State Generation of the analysis window ;
\EndFor
\ForAll { \( p \in \text{list of AR orders to be tested} \) }
\State Minimization of \( E_u \) with respect to \( \theta \);
\EndFor
\State \( (t_c, \hat{\theta}) = \arg \min_{t_c, \theta} E_u(t_c, \theta(t_c, p)) \);
\State Shape parameters given by: \( \theta(t_c, \hat{\theta}) ; \)
\end{algorithmic}
\end{algorithm}

\begin{align}
\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} \tag{5}
\end{align}

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 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 or-
der \( p \) 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 \( \theta \). This kind of nonlinear minimiza-
tion problem could be solved by a simulated annealing algo-
rithm but would be very time consuming. Instead, we use a two
step algorithm: \( \theta \) is first estimated by an exhaustive evaluation of
\( \bar{E}_u(t_c, \theta) \) over a finite subspace of \( \theta \); then this first estimate is
further refined using a simplex algorithm detailed in [10].

Figure 2: Asymptotic behaviour of the glottal flow derivative
spectrum.
3.2.1. LF space quantization

To quantize the glottal source shape, the correlation function \( \rho \) 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 \( \rho_{m} \), the corresponding finite subspace \( \Theta_{m} \) is built so as to satisfy the following two conditions:

1. covering condition:
   \[ \forall \theta \in \Theta : \exists \theta_{m} \in \Theta_{m} \text{ such that } \rho(u_{\theta}, u_{\theta_{m}}) \geq \rho_{m} \]
2. condition to prevent redundancy:
   \[ \forall \theta_{1}, \theta_{2} \in \Theta_{m} : \rho(u_{\theta_{1}}, u_{\theta_{2}}) < \rho_{m} \]

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

**Algorithm 2**: Algorithm for building the subspace \( \Theta \)

\[
\Theta \leftarrow \emptyset \quad \text{and} \quad \Delta \leftarrow 0 ; \\
\text{repeat} \\
\quad \text{if max}_{\theta_{m} \in \Theta} \rho(u_{\theta_{m}}, u_{\theta}) < \rho_{m} \text{ then} \\
\quad \quad \Theta \leftarrow \Theta \cup \{ \theta \}; \\
\quad \quad \Delta \leftarrow 0; \\
\quad \text{else} \\
\quad \quad \Delta \leftarrow \Delta + 1; \\
\text{until } \Delta < \Delta_{\text{max}} \]

The number of shape vectors increases rapidly as the correlation coefficient \( \rho_{m} \) tends toward 1, as shown in figure 3. The coefficient \( \rho_{m} \) must be high enough so that the estimated shape is close to the optimal shape. However, a space \( \Theta_{m} \) 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 \( \Theta_{m} \). Taking \( \rho_{m} = 0.99 \) proved to be a good compromise between these two opposite considerations.

![Figure 3: Cardinal of the subspace \( \Theta \) with respect to \( \rho_{m} \).](image)

From a spectral point of view, the quantization of space \( \Theta \) 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 \( \Theta_{m} \).

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 \( S = M_{a}A \), we can take advantage of the similarity of the matrix \( M_{a} \) when only the source shape parameters are modified: a modification of \( \theta \) results only in a modification of the last column of the matrix \( M_{a} \). 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 \( \hat{E}_{\Theta} \) 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 \( \hat{E}_{\Theta} \) over \( \Theta \) 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 \( \alpha_{m} \) 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 \( f_{o} \) 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 \( Q_{o} \) and \( t_{c} \). Thus, \( O_{q} \) and \( \alpha_{m} \) 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 \( S_{o}(f) \) and \( U_{o}(f) \): \( \hat{Q}_{c} \equiv \frac{S_{o}(f)}{U_{o}(f)} \) and by providing the average correlation coefficient \( \rho_{m_{\text{mean}}} \) 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 \( fs = 8 \text{kHz} \) 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 \( f_{o} = 100\text{Hz} \), the shape parameters \( O_{q} \) and \( \alpha_{m} \) are well
estimated, the corresponding estimator variance is below the human hearing threshold which is 0.140q for Oq (the threshold depends linearly on Oq) and 0.022 for α0, as stated by Henrich in [1]. For f0 = 180Hz, 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].

<table>
<thead>
<tr>
<th>f0</th>
<th>σq0</th>
<th>σα0</th>
<th>σQq(1−Oq)</th>
<th>ρmean</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.023</td>
<td>0.013</td>
<td>0.005</td>
<td>0.996</td>
</tr>
<tr>
<td>180</td>
<td>0.044</td>
<td>0.033</td>
<td>0.019</td>
<td>0.979</td>
</tr>
</tbody>
</table>

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 Oq estimate. An increase of the fundamental frequency has the same effect as the previous synthetic test.

<table>
<thead>
<tr>
<th>f0</th>
<th>σq0</th>
<th>σα0</th>
<th>σQq(1−Oq)</th>
<th>ρmean</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.033</td>
<td>0.018</td>
<td>0.006</td>
<td>0.995</td>
</tr>
<tr>
<td>180</td>
<td>0.070</td>
<td>0.043</td>
<td>0.015</td>
<td>0.972</td>
</tr>
</tbody>
</table>

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

5.2. Natural speech

The French words “Médecine prédictev” 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-synthesized 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-synthesized 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 estimation variance increase. We also plan to apply this analysis scheme in speech modification.

7. References