# Automatic Voice-Source Parameterization of Natural Speech

*Javier Pérez and Antonio Bonafonte*

Department of Signal Theory and Communication
TALP Research Center
Technical University of Catalonia (UPC), Barcelona, Spain
{javierp,antonio}@gps.tsc.upc.edu

## Abstract

We present here our work in automatic parameterization of natural speech by means of a pitch synchronous source-filter decomposition algorithm. The derivative glottal source is modelled using the Liljencrants-Fant (LF) model. The model parameters are obtained simultaneously with the coefficients of an all-pole filter representing the vocal tract response by means of a quadratic programming algorithm. Synthetic data has been created and analyzed in order to show the appropriate function of the estimation method. The parameterization results in high quality synthesized speech for voiced frames. Voice quality extraction is performed on basis to the LF source representation. The inherent modelling of the voice source makes it suitable for voice modification tasks. Work is in progress to add this speech representation to emotional speech synthesis and voice conversion algorithms.

## 1. Introduction

Estimation of voice source parameters is an important part of many applications that would benefit of a correct source parameterization, speech synthesis being the most obvious. Speech analysis, voice conversion or emotional synthesis, are other applications requiring accurate knowledge about the voice source. One method to obtain this representation is to use inverse filtering techniques together with a parameterization of the estimated glottal source. This is a complex problem that has been studied for years with the goal of developing an automatic method of parameterization ([1], [2] or [3] among others).

We are developing an analysis tool that would allow us to obtain a physically relevant representation of the speech signal. Our goal is to be able to add voice quality to the speech generation system, since it is a profitable knowledge for applications requiring large voice modifications (synthesis of emotional speech of voice conversion among others).

In this article we report the current work performed in this direction. The parameterization algorithm is presented in the next section, together with some specific parts dealing with the glottal epochs detection (sec. 2.1), initial source modelling and the optimization algorithm (sec. 2.2). A description of the LF derivative glottal source is then presented in sec. 2.3. The results of the experiments carried out during this work are reported in sec. 3. We end the report by explaining future directions for the research and some improvements to the current version of the algorithm (sec. 4).

## 2. Voice-source parameterization algorithm

The estimation algorithm requires pitch-synchronization, since the parametric model for a glottal period needs be matched exactly. Both the instants of opening and closing of the glottis are needed, and to obtain that we will use the simultaneously recorded signal from the laryngograph (EGG signal). The joint estimation of the voice source and the vocal tract is performed using a simpler model for the source (KLGLOTT88) proposed by Klatt [4], followed by a parameterization using the final LF model.

We start with a description of the algorithm used for the glottal epoch detection (both opening and closure instants) in sec. 2.1 and continue by giving the details of the KLGLOTT88 representation and optimization algorithm in 2.2.

### 2.1. EGG based glottal closure/opening instant detection

Electroglottography is a technique used to obtain an indirect knowledge of the laryngeal behavior during speaking by measuring the variation in electrical impedance across the throat. The EGG waveform and its relation to the events occurring at the glottis is well-known and has been reported extensively.

We will obtain the glottal closure instants (gcis) as the instants of occurrence of the minimum of derivative. As can be seen in figure 1, closure points are easily identifiable, since closure of the glottis is often abrupt. However, the opening of the glottis is not so easily obtained from the derivative of the EGG (as seen in the first pulses, trying to use the maximum of the derivative between to closure points is rather difficult). Instead, we will use a thresholding method that has proved successful (e.g. [1]). The glottal opening instant is defined as the instant where the EGG waveform reaches a threshold of $35\%$ of the difference between the maximum and the minimum of the EGG in that period.

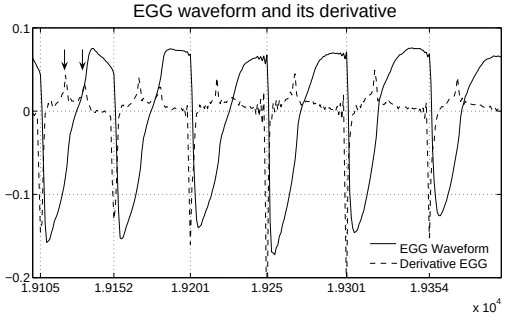


Figure 1: *Several pitch periods of EGG and differential EGG waveforms, showing the glottal closure instants occurring at the minimum of the derivative. The glottal opening instants, obtained by searching for the differential EGG maxima, are sometimes ill-defined, as marked by the two arrows in the left.*

### 2.2. Convex optimization

In order to obtain the different parameters of the system, we will formulate the problem in terms of the glottal wave. We will minimize the error between the glottal wave as modelled by the KLGLOTT88 model, and the estimated glottal wave, that would be obtained by inverse-filtering the speech waveform with the filter parameters to be estimated simultaneously. The KLGLOTT88 model [4] can be parameterized with a second order polynomial:

$$g_{KL}(t) = \begin{cases} b \cdot t \cdot (2\,t_{gci} - 3\,t) & ,0 \le t < t_{gci}, \\ 0 & ,t_{gci} \le t < T_0, \end{cases} \quad (1)$$

where $F_s$ is the sampling frequency, $T_0$ the pitch period of the voice, $t_{gci}$ is the glottal closure instant and $b$ is a parameter controlling the amplitude of the waveform. We have slightly modified the original notation to remark the fact that only one independent parameter controls the amplitude.

In order to minimize the error between the estimated and parameterized glottal waveforms, we need to choose which error norm we will use. If we choose $L_1$ or $L_{\inf}$ the problem can be solved using linear programming techniques [5]. However, we will use the $L_2$ norm, so the error minimization is transformed into a quadratic programming (QP) problem. There are several ways to solve a QP problem (a large collection of available software packages can be found in [6]); we will use the Sequential Unconstrained Minimization Method (SUMT) (see [7] for a mathematical description).

Let the filter coefficients be denoted as $[\hat{a}_1 \cdots \hat{a}_{N+1}]$, the estimated derivative glottal waveform as $g_{if}(n)$ and the (known) speech signal as $s(n)$:

$$g_{if}(n) = s(n) - \sum_{k=1}^{N+1} \hat{a}_k s(n-k). \quad (2)$$

The error between the parametric glottal wave (that we assume to be a KLGLOTT88 waveform) and the estimated glottal wave can then be formulated as:

$$g_{KL}(n) - g_{if}(n) = b \cdot n \cdot (2\,n_{gci} - 3\,n) + \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n), \quad (3)$$

in the open phase, and as:

$$g_{KL}(n) - g_{if}(n) = 0 + \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n), \quad (4)$$

when the glottis is closed. If we write down the error for the whole cycle length, we have (in matrix notation):

$$\mathbf{e} = \begin{bmatrix} (2\,n_c - 3) & s(0) & \cdots & s(-N) \\ \vdots & \ldots & \vdots & \vdots \\ -n_c^2 & s(n_c - 1) & \cdots & s(n_c - N - 1) \\ 0 & s(n_c) & \cdots & s(n_c - N) \\ \vdots & \ldots & \vdots & \vdots \\ 0 & s(M-1) & \cdots & s(M - N - 1) \end{bmatrix} \mathbf{x} - \begin{bmatrix} s(1) \\ \vdots \\ s(n_c) \\ s(n_c + 1) \\ \vdots \\ s(M) \end{bmatrix} = \mathbf{F}\mathbf{x} - \mathbf{y}, \quad (5)$$

where $\mathbf{x} = [b\ \hat{a}_1 \cdots \hat{a}_{N+1}]$ is a vector containing the parameters to be estimated. The equation error to minimize is then, in matrix notation:

$$\min_{\mathbf{x}} ||\mathbf{e}||^2 = \min_{\mathbf{x}} ||\mathbf{F}\mathbf{x} - \mathbf{y}||^2 = \min_{\mathbf{x}} \mathbf{x}^T \mathbf{F}^T \mathbf{F}\mathbf{x} - 2\mathbf{y}^T \mathbf{F}\mathbf{x}.$$

This minimization is a convex optimization problem [7], thus guaranteed to have only one (local) minimum (i.e. the optimal solution). We want to impose a low-pas characteristic for the final pole, since it is the spectral tilt characteristic of the KLGLOTT88 model. This means that the pole itself must be positive. Since it is the product of all the other poles, occurring in complex conjugate pairs due to the resonator characteristic of the filter, constraining the coefficient $a_{N+1}$ to be positive guarantees the low-pass characteristic. Furthermore, since this last coefficient is the product of all the pole magnitudes, we place an upper bound on it in order to obtain stable filters. The two additional constraints of the problem are then $0 < a_{N+1} \le 0.9 \cdot 0.985^N$ and $b > 0$. The values $0.9$ for the glottal spectral tilt and $0.985$ for the $N$ vocal tract poles were suggested by Lu [5] and successfully used in other research projects (e.g. [8]). As a result of this optimization stage, we obtain the estimated coefficients of the vocal tract filter, and an initial estimation of the glottal excitation, modelled as a KLGLOTT88 waveform.

As stated before, this simpler model is useful for the mathematical formulation of the problem, but there are other models performing better for a wider range of voice types. The next step is then to reparameterize the derivated glottal inverse waveform with the LF model, extensively reported and studied for several phonations.

### 2.3. LF model

The LF model has been widely used and is by now well established. Its parameters have been correlated to physiological and acoustic parameters. We are first presenting the model, and then explaining how it is incorporated into our system and how can we extract common measures to characterize difference voice qualities.

The model is capable of characterize the shape of the derivative glottal wave for a wide range of voices, both in the open and closed phases. In figure 2 a glottal LF cycle is presented, ranging from 0 to the fundamental period $t_o$. The other time marks are: $t_p$, representing the maximum of the glottal flow (and thus a value of 0 for the derivative); $t_e$, the time instant of the minimum in the derivative; $t_a$, defined as the point where the tangent to the exponential return phase crosses 0; $t_c$ the moment when the return phase reaches 0; and $Ee$ as the absolute value of the minimum of the derivative. The mathematical description of the LF model is:

$$g_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t) & ,0 \le t \le t_e, \\ -\frac{E_e}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(tc-te)}] & ,t_e < t \le t_c, \\ 0 & ,t_c < t \le t_0, \end{cases} \quad (6)$$

The rest of the parameters ($\alpha$, $w_g$, $E_0$ and $\epsilon$) are computed from the temporal ones by fulfilling some requirements of area balance and continuity (for details refer to [9], [10] or [3]):

$$\int_0^{t_o} g_{LF}(t) = 0 \quad (7)$$

$$w_g = \frac{\pi}{t_p} \quad (8)$$

$$\epsilon t_a = 1 - e^{-\epsilon(tc-te)} \quad (9)$$

$$E_0 = -\frac{E_e}{e^{\alpha t_e} \sin w_g t_e}. \quad (10)$$

One of the main differences with respect to the KLGLOTT88 model seen in section 2.2 is the non-abrupt return

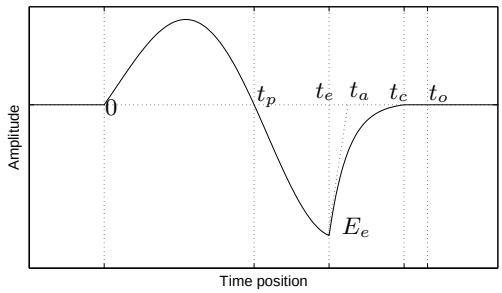


Figure 2: *The LF model. The figure shows a glottal period (from* 0 *until the period length* $t_o$*, and the parameters* $t_p$*,* $t_e$*,* $t_c$*,* $t_a$ *and* $Ee$*.*

phase. This is a more realistic modelling, since the glottis closure does not occur instantaneously. Thus, the LF model can result in a better approximation of the (derivative) glottal waveform and in an improved synthesis quality for several voice types.

#### *2.3.1. LF model fitting*

The fitting of an LF model to the inverse filtered speech signal is performed by means of non-linear optimization algorithms. Thus, a robust initial estimate needs to be computed in order to minimize the probability of ending with a local (non-optimal) minimum.

There are several methods to obtain the initial estimation. They can be extracted from the estimated glottal waveform obtained with the inverse filtering. In this case, it needs to be low-pass filtered prior to any computation since it normally contains aspiration noise. The time markers $t_p$, $t_e$, $t_c$, $t_a$ and the value of $E_e$ can then be obtained by identifying the minimum value of the signal and the zero-crossings (for details see [11]).

In our work, we have decided to take advantage of the similarities between the parametric KLGLOTT88 model obtained in the optimization step and the final LF model. $Ee$ is set to the minimum value of $g_{KL}$, $t_e$ to the time position of this minimum, and $t_p$ as the zero-crossing to the left of $t_e$. Since the glottal cycles are already available (as explained in 2.1), $t_c$ and $t_o$ are set to the end of the glottal cycle. Note that we are using here the common approach of not using a different value for $t_c$.

With this initial estimate, the parameters are further refined via constrained non-linear optimization methods (see [11] or [5]).

#### *2.3.2. Glottal flow measurements*

Several measures can be extracted from the glottal flow waveform in order to characterize in a numerical (an thus comparable) form the voice. A part from the directly available LF set of parameters, the speed quotient and the open quotient, two of the most extensively used characteristics in the literature, can be extracted.

The open quotient measure defines the duration of the open phase in relation to total cycle length and is computed as: $\mathrm{OQ} = t_e/T_0$, where $T_0$ is the total cycle length. A small modification can be added to penalize the very small glottal flow values occurring when the term $t_e$ in eq. 6 is large (the glottis is not considered open until the flow surpasses a experimentally set threshold [2]). The relative duration of the closing phase, as a percentage of the cycle length, is defined by the closing quotient: $\mathrm{CQ} = (t_e - t_p)/T_0$. A last measure describing the temporal skewing between the closing and opening phase is the speed quotient: $\mathrm{SQ} = \frac{\mathrm{OQ}-\mathrm{CQ}}{\mathrm{CQ}}$. Depending on the point of view, the returning phase $t_a$ can or cannot be considered part of the preceding measures.

## 3. Results

In this section we present some of the results obtained in this work. We have experimented with analysis/synthesis of natural speech using simultaneously recorded laryngograph data, obtaining promising results in the terms of synthetic speech quality. The method is robust and has been used to analyze several utterances from different databases. We will start by presenting some results obtained used artificially created data, in order to validate the method. Then, the results of the analysis of real data are explained in sec.3.2.

### 3.1. Synthetic data validation

We have performed an initial evaluation with synthetic data in order to validate the algorithm. Following the work by Strik *et al.* [11], we performed the estimations on a series of LF pulses created with the parameters proposed there (all the units are milliseconds):

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_p$ | 4 | 4 | 6 | 6 | 6 | 6 | 4 | 4 | 5.2 | 5.2 | 5.2 |
| $t_e$ | 5.2 | 5.2 | 7.2 | 7.2 | 8.8 | 8.8 | 6 | 6 | 7.2 | 7.2 | 7.2 |
| $t_a$ | 0.4 | 1.6 | 0.4 | 1.6 | 0.4 | 0.8 | 0.4 | 1.6 | 0.4 | 1 | 1.6 |

The other parameters of the LF model were kept constant ($t_o = 0$, $t_c = 10$, $Ee = 1024$) since their influence on the estimation error is small. This pulses have been filtered with a LPC derived filter computed for a vowel *a*.

The error measure we will use is the Averaged Perceptual Error (APE) [11], computed by averaging the percentual error of the input and estimated parameters according to $PE = |\hat{P} - P|/P$, where $\hat{P}$ is the estimated value and $P$ is the original parameter used in synthesis.

The resulting APEs are smaller in comparison with the results reported before: $APE_{te} = 0.0186$, $APE_{ta} = 1.1402$ and $APE_{tp} = 0.2380$. This is due to the fact that in our case, the $t_e$ parameter is actually an input to the system, since it corresponds to the glottal closure instant and we are assuming it known. The algorithm actually performs a reestimation around this original value to account for small errors in the GCI detection from the EGG signal, that explains the APE for $t_e$ not being zero. Nevertheless, the results show that the method is performing as expected for synthetic data. The next section presents some experiments performed using natural speech.

### 3.2. Natural speech analysis

Figure 3 shows a sample of the fitting process for both models (KLGLOTT88 and LF). As it can be seen, the LF model achieves a better shape modelling than the KLGLOTT88 (the objective quality of the fitting in terms of segmental SNR also improves).

We did some experiments incorporating a weighting filtering to the LF model fitting, although no significant improvement improvement (in terms of synthetic speech quality) was observed (in some cases it slightly decreased). In this configuration, the error of the LF fitting procedure is computed in the speech domain. Prior to compute the $L_2$ norm, the error

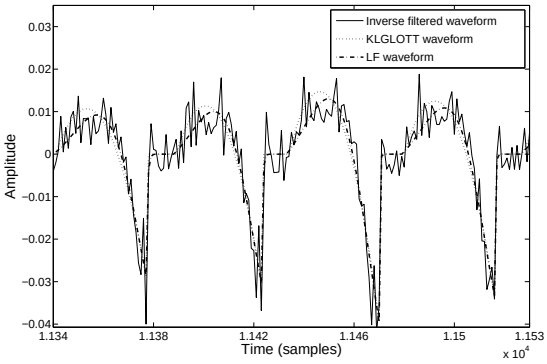


Figure 3: *Several glottal cycles obtained by inverse filtering the speech signal and the corresponding fitted KLGLOTT88 and LF models. The higher flexibility of the LF model results in a better shaping in terms of root mean square error.*

is passed though a perceptual weighting filter used to concentrate the error in the less noticeable parts of the spectrum. The weighting filter coefficients $b_k$ in eq. 11 are derived performing a LPC analysis using a window of 25 milliseconds centered on the pitch period. The weighting filter is then constructed as:

$$W(z) = \frac{1 - \sum_{k=1}^{K} b_k \gamma_1^k z^{-k}}{1 - \sum_{k=1}^{K} b_k \gamma_2^k z^{-k}}, \qquad (11)$$

where $\gamma_1 > \gamma_2$ are used to control the sharpness of the amplitude response of $W(z)$. This is currently used in speech coding schemes based on linear predictive analysis-by-synthesis (e.g. [12]).

In principle, it would be interesting to use the perceptual weighting filter in the joint (KLGLOTT88 source and all-pole filter) analysis. Unfortunately, the IIR characteristic of this filter would break the convex formulation of the problem, thus invalidating the actual algorithm used in our work. As a suboptimal approach, we develop the first joint estimation as before, and include the perceptual measure in the adjustment of the LF parameters. Further work is required to study the effect of incorporating perceptual weighting measures when computing the vocal tract estimation.

## 4. Conclusions

We have presented here a robust joint source-filter decomposition algorithm. It has been successfully used to analyze and re-synthesize whole utterances from our databases. The method presented here works reasonably well for voiced speech, but it needs to be extended to the unvoiced parts. In order to do this, the aspiration noise of the adopted human production system needs to be estimated together with the glottal excitation.

Several ways to obtain the noise estimations have been proposed. One possibility is to assume a noise residual model, and estimate the parameters using the glottal residual waveform. Lu [5] proposed a statistical method using wavelet denoising, to estimate a noise residual model with two components, one is an amplitude modulated pitch-synchronous Gaussian noise, and the other a zero mean, unit variance, white Gaussian noise. On the other hand, it is also possible to adopt a stochastic approach and train a codebook with the glottal derivative residuals (e.g. [8]).

From our point of view the first approach is more interesting since we want to obtain a production model related to the human production system. However, unless a good noise model is adopted, the second approach will probably result in better quality.

## 5. Acknowledgements

This work has been partially sponsored by the Spanish Government under grant TIC2002-04447-C02 (ALIADO project [13]) and by the European Union under grant FP6-506738 (TC-STAR project [14]).

## 6. References

[1] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. Acoust. Soc. Am.*, vol. 115, no. 3, March 2004.

[2] M. Frölich, D. Michaelis, and H. W. Strube, "SIM–simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *J. Acoust. Soc. Am.*, vol. 110, no. 1, July 2001.

[3] C. Gobl, "The voice source in speech communication," Ph.D. dissertation, KTH, 2003.

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

[5] H.-L. Lu, "Toward a high-quality singing synthesizer with vocal texture control," Ph.D. dissertation, Stanford University, 2002.

[6] "Neos guide: Optimization guide." [Online]. Available: http://www-fp.mcs.anl.gov/otc/Guide/SoftwareGuide

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[8] Y. E. Kim, "Singing voice analysis/synthesis," Ph.D. dissertation, MIT, 2003.

[9] G. Fant, "The lf-model revisited. transformations and frequency domain analysis." STL-QPSR, Tech. Rep. 2-3:119–156, 1995.

[10] Q. Lin, "Speech production theory and articulatory speech synthesis," Ph.D. dissertation, KTH, 1990.

[11] H. Strik, B. Cranen, and L. Boves, "Fitting a LF-model to inverse filter signals," in *3rd European Conference on Speech Communication and Technology, Proceedings of the*, 1993, pp. 103–106.

[12] K. Järvinen, J. Vainio, P. Kapanen, T. Honkanen, and P. Haavisto, "Gsm enhanced full rate speech codec," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*. IEEE, 1997, pp. 771–774.

[13] *ALIADO, Tecnologías del habla y el lenguaje para un asistente personal*, http://gps-tsc.upc.es/veu/aliado/main.html.

[14] *TC-STAR, Technology and Corpora for Speech to Speech Translation*, http://www.tc-star.org/.