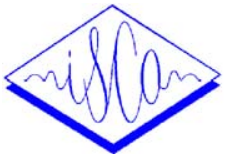
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A Robust Glottal Source Model Estimation Technique *

Qiang Fu †*and Peter Murphy*

Department of Electronic and Computer Engineering
University of Limerick, Limerick, Ireland
{qiang.fu, peter.murphy}@ul.ie

Abstract

This paper describes a robust glottal source estimation method based on a joint source-filter separation technique. In this method, the glottal flow derivative is modelled as the Liljencrants-Fant (LF) model and the vocal tract is described as a time-varying ARX model. Since the joint estimation problem is a multi-parameter nonlinear optimization procedure, we separate the optimization procedure into two passes. The first pass initializes the glottal source and vocal tract models providing robust initial parameters to the following joint optimization procedure. The joint estimation determines the accuracy of model estimation, which is implemented with a trust-region descent optimization algorithm. Experiments with synthetic and real voices show the proposed method is a robust glottal source parameter estimation method with a considerable degree of accuracy.

10.21437/Interspeech.2004-86

1. Introduction

Glottal source estimation can potentially be used for clinical diagnoses, speech analysis, synthesis, coding, front-end processing for speaker verification and identification, etc.. Traditionally, the glottal source is derived using *inverse filtering* techniques [1, 2, 3, 4]. For many years, a number of methods based on joint source-filter separation have also received much attention [5, 6, 7, 8, 9]. In such methods, the glottal source and vocal tract are modelled and estimated simultaneously under a *joint optimization* criterion. Compared with the time-invariant source-filter model based inverse filtering, it allows us to assume the vocal tract to be non time-invariant, i.e., the source is independent of the vocal tract [3]. Hence, the glottal source is able to be modelled by the glottal flow model without formant ripple, such as the Rosenberg [10] and Liljencrants-Fant (LF) [11] model, etc., which are all time domain multi-parameter models with the capacity for describing the glottal flow signal with sufficient temporal details.

In this paper, we introduce an accurate and robust glottal source joint estimation method. The LF model is chosen for glottal source modelling due to its widespread application and flexibility [7, 9, 12]. The time-varying ARX model is used for vocal tract modelling. A Kalman filtering [13] process is used here to adaptively estimate the vocal tract filter coefficients.

The joint estimation is essentially a multi-dimensional nonlinear optimization problem, whose error function is not simply convex [14]. The optimization procedure can quite possibly get stuck in local minimizations [12]. An adequate initialization of a given optimization procedure is of major importance if the global minimum is to be found [9]. The strategy of this paper is to separate the optimization procedure into two passes. The first pass initializes the glottal source and vocal tract models. This initial pass ensures that robust initial parameters are provided to the following joint optimization. The following joint estimation concerns the accuracy of model estimation.

We start from presenting the time-varying ARX speech production model and the formulation of the joint estimation problem in Section 2. The initialization for both the glottal flow derivative and the vocal tract models is described in Section 3, and the implementation of the joint estimation procedure is described in Section 4. Section 5 presents the experimental results for both synthesis and real speech data, while conclusions are given in Section 6.

2. Formulation of Joint Estimation Problem

The first step of this algorithm is to determinate the location of the beginning and end of each pitch period. Among all the events that characterize the pitch period the Glottal Closure Instant (GCI), which coincides with the instant where all formants are excited simultaneously and leads to a discontinuity in the glottal flow derivative as shown in Fig. 1, seems the most desirable one. We choose the Pitch Tracking Algorithm (PiTA) [2] as the GCI determination algorithm.

2.1. Time-varying ARX Speech Production Model

In the proposed algorithm, the speech production process is modelled as a time-varying IIR system as follows [8]

$$s[n] + \sum_{i=1}^{p} a_i[n]s[n-i] = \sum_{j=1}^{q} b_j[n]v_g[n-j] + v_g[n] \quad (1)$$
$$(0 \le n \le N_0 - 1)$$

where $s[n]$ and $v_g[n]$ denote an observed speech signal and an unknown input glottal flow derivative waveform within one pitch period, respectively. N_0 is the sample length of the pitch period. $a_i[n]$ and $b_j[n]$ are time-varying auto-regressive (AR) and moving average (MA) coefficients of the IIR filter, respectively. p and q are the filter order, i.e. the number of the poles and zeros of the model, respectively.

The LF model is used to represent the glottal flow derivative, since it ensures an overall fit to commonly encountered wave shapes with a minimum number of parameters and is flexible in its ability to match extreme phonation [11]. Assume the discrete timing parameters, the instant of maximum glottal flow, maximum negative value of glottal flow derivative and glottal closure, are denoted by N_p, N_e and N_c, respectively, the

* This work is supported by Enterprise Ireland Research Innovation Fund 2002/0037.

† The author is now with the Institute of Acoustics, Chinese Academy of Sciences, P. R. China.

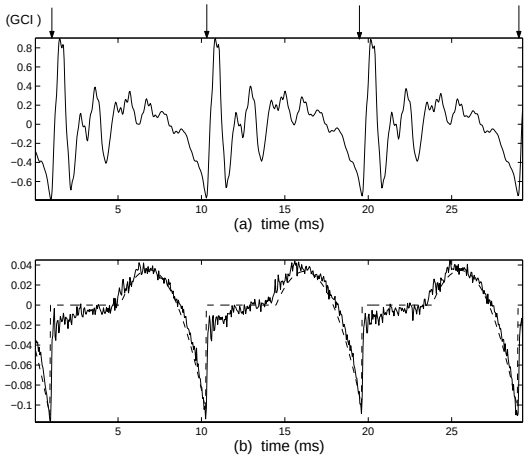


Figure 1: (a) Original speech data with GCI marks; (b) Estimated glottal flow derivative waveform using pitch-synchronous inverse filtering (the dash line is the best-fitted Rosenberg model using the method described in Section 3.1)

discrete-time version of LF model is then can be defined as

$$v_g^{LF}[n] = \begin{cases} E_o e^{\alpha n}\sin(\omega_g n), 0 \le n < N_e \\ \dfrac{-E_e}{\varepsilon N_a}[e^{-\varepsilon(n-N_e)} - e^{-\varepsilon(N_c-N_e)}], N_e \le n < N_c \\ 0, N_c \le n \le N_0 - 1 \end{cases} \tag{2}$$

Along with the magnitude of the glottal closure excitation E_e, the LF model may be specified by two independent sets of parameters: the direct synthesis parameters $(E_0, \alpha, \omega_g, \varepsilon)$, and the timing parameters (N_p, N_e, N_a, N_c). It has to be pointed out that, in our method, the starting and ending points of each pitch period are the GCIs, which is different with those of v_g^{LF}. There is a timing shift relation between the LF model expression v_g^{lf} and v_g^{LF}, which can be expressed as

$$v_g^{lf}[n] = \begin{cases} v_g^{LF}[n + N_e + 1], & 0 \le n \le N_0 - N_e - 2 \\ v_g^{LF}[n - (N_0 - N_e - 1)], & N_0 - N_e - 1 \le \\ & n \le N_0 - 1 \end{cases} \tag{3}$$

2.2. Formulation of Joint Estimation Problem

$\mathbf{X}[n] = \{a_1[n], \cdots, a_p[n], b_1[n], \cdots, b_q[n]\}^T$ and $\mathbf{H}[n] = \{-s[n-1], \cdots, -s[n-p], v_g^{lf}[n-1], \cdots, v_g^{lf}[n-q]\}^T$ are defined as the *state* and *measurement* vectors of the ARX model, respectively. To jointly estimate both the glottal source and vocal tract models from $s[n]$, a least-squares criterion is used here, i.e.,

$$\begin{aligned} \min_\theta E(\theta) &= \min_\theta \frac{1}{N_0}\sum_{n=0}^{N_0-1}\{e[n]\}^2 \\ &= \min_\theta \|e[n]\|^2 \\ &= \min_\theta \|s[n] - \mathbf{H}[n]^T\mathbf{X}[n] - v_g^{lf}[n]\|^2 \end{aligned} \tag{4}$$

Subject to $0 < N_p < N_e < N_c \le N_0$. Here θ denotes the discrete LF model parameter set, i.e., $\theta = \{N_p, N_e, N_a, N_c, Ee\}^T$. Once a set of θ is chosen, the $v_g^{lf}[n]$ is determined. Eq. (4) then reduces to

$$\min_{\mathbf{X}[n]} E(\mathbf{X}[n]) = \min_{\mathbf{X}[n]} \|z[n] - \mathbf{H}[n]^T\mathbf{X}[n]\|^2 \tag{5}$$

where$z[n] = s[n] - v_g^{lf}[n] (0 \le n \le N_0 - 1)$.

The time-varying state vector $\mathbf{X}[n]$ can be uniquely identified by Kalman filtering (Section 4.1)). Every θ within the feasible field will give a unique $E(\theta)$, thus forming a multi-dimensional error function of θ. Conclusively, from the view of optimization theory, Eq. (4) shows an unconstrained general nonlinear minimization problem subject to bound. Apparently, there is no closed-form solution for $E(\theta)$. Moreover, by observing Eq. (2) and (4), it's easy to find out that the error surface of $E(\theta)$ is non-convex, which means no efficient iterative algorithm can be guaranteed to converge to the global optimum solution [15]. Our strategy to solve this complex nonlinear optimization problem is to firstly construct an approximate problem that is relatively easier to locate the global solution; the global solution of this approximate problem is then taken as the initial starting point of a descent algorithm to finally solve the problem of Eq. (4). In this way, a robust and accurate solution can be achieved.

3. Initialization

Instead of accuracy, convergence and robustness are our concerns in the initialization pass. A two-parameter glottal waveform fitting procedure is used here for this purpose, which differs from the joint estimation in two aspects: firstly, to simplify the problem, the vocal tract is assumed to be an all-pole LTI system; secondly, the glottal flow derivative waveform is now generated by a Rosenberg model, instead of the LF model, because of its simplicity.

3.1. Construction of the Approximate Problem

Under the LTI speech production process assumption, the vocal tract transfer function is typically modelled as $V(z) = 1/(1 - \sum_{i=1}^{p} a_i' z^{-i})$. The output speech is therefore presented as

$$s[n] = \sum_{i=1}^{p} a_i' s[n-i] + v_g'[n], \quad 0 \le n \le N_0 - 1 \tag{6}$$

here the $v_g'[n]$ is represented by a Rosenberg model [10], which is a simple second order polynomial in the open phase and zero in the closed phase, as shown in Fig. 1,

$$v_g^{RS}(n) = \begin{cases} 2cn - 3dn^2, & 0 \le n \le OQ \cdot N_0 \\ 0, & OQ \cdot N_0 < n \le N_0 - 1 \end{cases} \tag{7}$$

$$c = \frac{27 \cdot AV}{4 \cdot (OQ^2 \cdot N_0)} \tag{8}$$

$$d = \frac{27 \cdot AV}{4 \cdot (OQ^3 \cdot N_0^2)} \tag{9}$$

We denote the parameter vector $\theta' = [OQ, AV]^T$. $OQ (0 \le OQ \le 1)$ is the open quotient of the glottal source, i.e. the ratio of open time to total period duration and AV is the ratio of the peak amplitude of the glottal flow to N_0. For the same reason as described in Section 2.1, there is a timing shift relation between the Rosenberg model expression v_g^{rs} and v_g^{RS}, i.e.,

$$v_g^{rs}[n] = \begin{cases} v_g^{RS}[n + OQ \cdot N_0 + 1], 0 \le n \le N_0 - OQ \cdot N_0 - 2 \\ v_g^{RS}[n - (N_0 - OQ \cdot N_0 - 1)], N_0 - OQ \cdot N_0 - 1 \\ \qquad \le n \le N_0 - 1 \end{cases} \tag{10}$$

Following Eq. (4), a similar least-squares error minimization problem can be formulated as

$$\min_{\theta'} E(\theta') \quad = \quad \frac{1}{N_0} \sum_{n=N_0-OQ\cdot N_0-1}^{N_0-1} \{\hat{v'}_g[n] - v_g^{rs}[n]\}^2$$

where $\hat{v'}_g[n]$ is the estimated glottal flow derivative waveform via the pitch synchronous covariance LP method [3], $\mathbf{v}_g^{rs} = \{v_g^{rs}[0], v_g^{rs}[1], \cdots, v_g^{rs}[N_0-1]\}^T$. Although the Rosenberg glottal model is relatively less accurate and flexible than the LF model, it's simple mathematical expression ensures the corresponding glottal waveform fitting procedure easily converges to the global minimum solution. The computation is also reduced. This is a very important property since the main purpose of the initialization pass is to provide robust initial parameters to the following joint optimization.

Considering there is the possibility of error detection of GCI and the existence of noise and ripples in $\hat{v}'_g[n]$, the glottal fitting procedure is implemented using the Simplex algorithm [16], which is a direct search method that does not use numerical or analytic gradients. We denote the optimum solution in this phase as OQ^1, AV^1.

3.2. Direct Estimation

The glottal waveform fitting procedure also requires an initial value to get started. Since the estimated glottal flow derivative $\hat{v'}_g[n]$ often contains ripples, which are the effects of formants that were not inverse filtered, and a noise component, it is firstly low-pass filtered to reduce the measurement error. Specifically, $\hat{v'}_g[n]$ is convolved with a Blackman window with the length of $M = 7$. The rules used in the direct estimation are basically followed [17]. $U_m = \max(\hat{u'}_g^{lp})$: the maximal amplitude of $\hat{u'}_g^{lp}$, N_z: the instant of glottal closure, and N_o: the onset of glottal opening are determined. We set the initial values of the Rosenberg model as $OQ^0 = (N_0 - N_o)/N_0$, i.e., the initial value of open quotient; $AV^0 = U_m/N_0$, i.e., initial value of amplitude. Moreover, we also can obtain the duration of the return portion as $\Delta_{ce} = N_z + 1$.

3.3. Parameter Conversion from Rosenberg to LF model

Because it is the LF model that is actually used as the glottal source in the joint estimation, we have to convert the initial Rosenberg model parameter estimates (OQ^1, AV^1) into LF ones $(N_p^1, N_e^1, N_a^1, N_c^1, Ee^1)$. To avoid the staircase phenomenon of the error function of the joint estimation [12], we use the ratio of the timing parameters to N_0, i.e.,$(r_p^1 = N_p^1/N_0, r_e^1 = N_e^1/N_0, r_a^1 = N_a^1/N_0, r_c^1 = N_c^1/N_0)$, as the input parameters of the error function instead of the discrete-time ones. The equations are as follows:

$$r_e^1 = OQ^1 \tag{11}$$

$$r_p^1 = \frac{2r_e^1}{3} \tag{12}$$

$$r_c^1 = r_e^1 + \frac{\Delta_{ce}}{N_0} \tag{13}$$

$$r_a^1 = \frac{2\Delta_{ce}}{3} \tag{14}$$

$$E_e^1 = -v_g^{rs}[N_e^1] = -v_g^{rs}[int(r_e^1 \cdot N_0)] \tag{15}$$

4. Implementation of Joint Estimation

4.1. Identification of ARX model

The state-space model to solve the ARX identification problem in Eq. (5) is defined as

$$\begin{cases} \mathbf{X}[n] &= \mathbf{X}[n-1] \\ z[n] &= \mathbf{H}^T[n]\mathbf{X}[n] + e[n] \end{cases} \tag{16}$$

where Eq. (16) is the state transition equation and Equation 16 is the observation equation, respectively.

For each pitch period, the initial condition of $\mathbf{X}[n]$ is set as $\mathbf{X}[0] = \{\hat{a}'_0, \hat{a}'_1, \cdots, \hat{a}'_p, \bar{b}_0, \bar{b}_1, \cdots, \bar{b}_q\}^T$, where $\bar{b}_j$ represents the average of the jth MA coefficient of ARX model over the previous pitch period.

4.2. Glottal Source Optimization

With the initialization procedure, we already have a good initial point $\theta^1 = \{r_p^1, r_e^1, r_a^1, r_c^1, E_e^1\}$ for the problem of Eq. (4), then the descent algorithm based on the gradient and Hessian matrix of the objective function $E(\theta)$ can be used to get a more accurate solution. Moreover, (4) is a nonlinear least-squares problem. The special structure of the gradient and Hessian matrix can be exploited to achieve more efficient convergence than in the general unconstrained problem. To implement the descent procedure, we choose the interior trust region algorithm described in [18] since it generates strictly feasible iterates and possesses local quadratic convergence properties. It can also deal with simple bounds as in Eq. (4). In trust region method, the increment $s_k = \theta_{k+1} - \theta_k$ is an approximate solution to a quadratic subproblem with a bound on the step:

$$\min_{s\in\Re^5}\{Q_k(s) = g_k^T s + \frac{1}{2}s^T M_k s : \|D_k s\| \leq \triangle_k\} \tag{17}$$

where

$$M_k = B_k + C_k \tag{18}$$

$$C_k = D_k diag(g_k) J_k^v D_k \tag{19}$$

Here $g_k \doteq \nabla E(\theta_k)$, B_k is a symmetric approximation to the Hessian matrix $\nabla^2 E(\theta_k)$, Δ_k is a positive scalar representing the trust region size, D_k is a scaling matrix that is defined as

$$D(\theta) = diag(|v(\theta)|^{-\frac{1}{2}}) \tag{20}$$

Here $v(\theta) \in \Re^5$ is defined as follows: for each component $1 \leq i \leq 5$, (i) if $g_i < 0$ then $v_i = \theta_i - 1$; (ii) if $g_i \geq 0$ then $v_i = \theta_i$. Thus, $J_k^v \in \Re^{5\times 5}$ in Eq. (19) is the Jacobian matrix of $|v(\theta)|$ whenever $|v(\theta)|$ is differentiable.

5. Experiments and Discussions

5.1. Synthesized Voices

The synthesized voices are generated according to the linear time-invariant source-filter model and using the LF model as the glottal flow derivative. We firstly assess the proposed method with the synthetic vowels /IY/ with 500 different LF timing parameter configurations that conceptually cover 5 typical voice types, i.e., modal, vocal fry, breathy, falsetto and harsh, as described in [19]. For each type of the generated vowel samples, we have 100 different LF timing parameter configurations which form a Gaussian distribution. An average relative difference δ_r [9] is calculated as

$$\delta_r = \frac{1}{100}\sum_{j=1}^{100}\frac{|es_j - or_j|}{or_j} \tag{21}$$

Table 1: Estimation Results for Synthesized Typical Voices

Voice	$\mathbf{t_p}(\%)$	$\mathbf{t_e}(\%)$	$\mathbf{t_a}(\%)$	$\mathbf{t_c}(\%)$	$\mathbf{E_e}(\%)$
Modal	09.97	09.56	28.39	10.80	10.33
Vocal fry	17.17	10.80	43.18	15.33	15.70
Breathy	04.68	06.23	11.97	06.99	23.65
Falsetto	05.36	07.36	10.42	04.63	13.15
Harsh	16.85	12.07	11.86	11.74	17.89

where es_j and or_j are the estimated and original LF timing parameters values, respectively. The results are shown in Table 1. From Table 1, a high degree of estimation accuracy is observed for t_p, t_e, t_a and t_c, while t_a and E_e are slightly worse than the others. The proposed method is also found to be robust against additive glottal noise. Different voice type has different noise levels ranging from $50dB$ (falsetto) to $10dB$ (harsh), the estimation results show the proposed method is insensitive to the different glottal noise levels.

5.2. Real Vowels

For the real speech material, we select the sentence "We were away a year ago" with a comfortable pitch and loudness and produced by a male speaker from the CD-ROM with [19].

Because there are no "correct" LF timing parameters known for real vowels, as a reference measurement, we use the open quotient OQ_{EGG} based on the electroglottographic (EGG) signal that is recorded synchronously with the speech signal. The calculation of the OQ_{EGG} follows the Glottal Opening Instants (GOIs) and GCIs determination algorithm described in [3]. Having the GOI and GCI, the OQ_{EGG} can be obtained easily.

The measured average relative difference δ_r of the parameter t_c, which has correspondence with the OQ_{EGG}, is 10.47%.

6. Conclusions

A pitch-synchronous glottal source estimation method requiring no closed phase information is described. Within a pitch period, the parametric glottal source and the vocal tract are estimated simultaneously - iteratively and adaptively. By using Kalman filtering, both the estimated glottal source and vocal tract are time variant, and more complex vocal tract properties can be described. The choice of the LF glottal source model also provides a flexible expression for different source types which provides the possibility of a wide range of applications. An important feature of the proposed method is that it separates the multi-dimensional joint estimation procedure into two passes, thus the convergence of the optimization algorithm can be guaranteed. Both the results for synthesized vowels and real speech experiments show the proposed method gives a robust and accurate performance for glottal source model estimation.

7. References

[1] Wong, D. Y., Markel. J. D., and Gray. Jr. A. H., "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 4, pp. 350-355, 1979.

[2] Hess, W., Pitch Determination of Speech Signals - Algorithms and Devices, Ed. by K-S. Fu, *et.al.*, Springer-Verlag, Berlin, 1983.

[3] Krishnamurthy. A. K., and Childers. D. G., "Two-channel speech analysis," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No. 4, pp. 730-743, 1986.

[4] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Communication, Vol. 11, pp. 109-118, 1992.

[5] Milenkovic, P., "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. on Acoustics, Speech and Signal prceossing, Vol. ASSP-34, No. 1, pp. 28-41, 1986.

[6] Fujisaki, H, and Ljungqvist, M., "Estimation of voice source and vocal tract parameters based on ARMA anlysis and a model for the glottal source waveform," Proc. IEEE ICASSP, pp. 637-640, 1987.

[7] Krishnamurthy, A. K., "Glottal source estimation using a sum-of-exponentials model," IEEE Trans. on Signal Processing, Vol. 40, No. 3, pp. 682-686, 1992.

[8] Ding, W., Kasuya, H., and Adachi, S., "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model," IEICE Trans. Inf. & Syst., Vol. E78-D, No. 6, 1995.

[9] Frohlich, M., Michaelis, D., and Strube, H. W., "SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," J. Acoust. Soc. Am., Vol. 110, No. 1, pp. 479-488, 2001.

[10] Rosenberg, A. E., "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am., Vol. 49, pp. 583-590, 1971.

[11] Fant, G., "A four-parameter model of glottal flow," Speech Transmission Lab. Quart. Progr. Rep. 4/85, KTH, pp. 1-3, 1985.

[12] Strik, H., "Automatic parammetrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses," J. Acoust. Soc. Am., Vol. 103, No. 5, pp. 2659-2669, 1998.

[13] Haykin, S., Adaptive Filter Theory, 4th ed., Prentice Hall, 2001.

[14] Lu, H-L., Smith III, J. O., "Joint estimation of vocal tract filter and glottal source waveform via convex optimization," in Proc. 1999 IEEE Workshop on Application of Signal Procecessing to Audio and Acoutics, New paltz, New York, Oct. pp17-20, 1999.

[15] Rao, S. S., "Engineering Optimization: Theory and Practice," 3rd Ed. Interscience, 1996.

[16] Nelder, J. A. and Mead, R., "A simplex method for function minimization," Comput. J. (Swizerland), Vol. 7, pp. 308-313, 1964.

[17] Alku. P., and Vilkman, E., "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering," J. Acoust. Soc. Am, Vol. 98, No. 2, pp. 763-767, 1995.

[18] Coleman, T. F., and Li, Y., "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds," SIAM Journal on Optimization, Vol. 6, pp. 418-445, 1996.

[19] Childers, D. G., Speech processing and synthesis toolboxes, John Wiley & Sons, Inc. 2000.