Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 98–109

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter

Prasanta Kumar Ghosh *, Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA

Received 25 December 2009; received in revised form 29 June 2010; accepted 19 July 2010

Abstract

We propose a glottal source estimation method robust to shimmer and jitter in the glottal flow. The proposed estimation method is based on a joint source-filter optimization technique. The glottal source is modeled by the Liljencrants–Fant (LF) model and the vocal-tract filter is modeled by an auto-regressive filter, which is common in the source-filter approach to speech production. The optimization estimates the parameters of the LF model, the amplitudes of the glottal flow in each pitch period, and the vocal-tract filter coefficients so that the speech production model best describes the observed speech samples. Experiments with synthetic and real speech data show that the proposed estimation method is robust to different phonation types with varying shimmer and jitter characteristics.

Keywords: Glottal flow derivative; Shimmer; Jitter; Glottal source estimation

1. Introduction

Estimation of glottal flow from the acoustic speech signal can be useful for many potential applications, such as speech analysis, modeling, synthesis, coding, and speaker verification/identification, as well as for noninvasive diagnosis of voice disorders (Rosenberg, 1971; Plumpe et al., 1999; Strik, 1998; Moore et al., 2003; Airas and Alku, 2006). Although glottal flow can be assessed accurately through direct, invasive measures within specific scientific or diagnostic setups, in practice, it is usually estimated from a signal which is recorded noninvasively (Frohlich et al., 2001).

Voiced speech is typically modeled as the output of a linear time-invariant (LTI) filter with glottal flow at its input. Under such a model, it is straightforward to derive the glottal flow derivative from the output speech signal using glottal inverse filtering (Quatieri, 2001; Hess, 1983). In glottal inverse filtering, the vocal-tract filter is first estimated from the output speech signal using linear prediction (LP) (Rabiner and Schafer, 2010), and then the output speech is filtered through the inverse of the estimated vocal-tract filter to obtain an estimate of the glottal flow derivative. The main problem in glottal inverse filtering is that the estimate of the vocal-tract filter is influenced by the glottal flow and, hence, may not be accurate. Pitch-synchronous LP (PSLP) is a more widely used approach for glottal inverse filtering for avoiding the effect of the harmonic structure of the speech spectrum on the LP analysis (Rabiner and Schafer, 2010). To avoid the influence of the glottal flow while estimating the vocal-tract filter, a common approach is to perform LP analysis only during the closed phase, i.e. the period during which the glottis is closed and there is no glottal flow (Krishnamurthy et al., 1986). For example, Wong et al. presented a classical pitch synchronous closed phase covariance linear prediction algorithm (Wong et al., 1979). However, a sufficiently long closed phase is necessary for estimating the vocal-tract filter accurately; unfortunately, this is not always the case, particularly for the speech of females and children due to the shorter glottal time periods. As an alternative, Alku (1992) proposed a low-order FIR filter for modeling the glottal source and used it to eliminate its effect on the out-

* Corresponding author. Tel.: +1 213 821 2433; fax: +1 213 740 4651.
E-mail address: prasantg@usc.edu (P.K. Ghosh).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.07.004

put speech and then performed a PSLP over the whole pitch period.

In natural speech production, there are more complex interactions between the glottal excitation and the vocal-tract filter beyond what is represented by the simple LTI filtering assumption (Carre, 1981). For example, as pointed out by Miller (1959), the coupling to the subglottal system causes appreciable damping of the formant oscillation during the open glottis interval. To capture the source-tract interaction, a common approach is to assume a model of glottal source and estimate the source and filter jointly. Almost all available glottal flow models are time domain models, such as the Rosenberg (Rosenberg, 1971), KLGLOTT88 (Klatt et al., 1990), Rosenberg++ (Veldhuis, 1998), Liljencrants–Fant (LF) (Fant and Lin, 1985) models, all of which have the capability of describing the glottal flow signal with sufficient temporal details. For example, in (Krishnamurthy, 1992) the glottal source is described using the LF model and the vocal tract is modeled as a pole-zero system with different sets of pole and zero locations in the closed phase (CP) and open phase (OP) to model the source-tract interaction. However, the estimates of the CPs and OPs from natural speech are not guaranteed to be always accurate, which may lead to wrong estimates of the LF model parameters. To reduce such error propagation due to wrong estimates of CP and OP, it is desirable to incorporate CP and OP estimation in the optimization framework itself. Frohlich et al. (2001) have presented a pitch-asynchronous simultaneous inverse filtering and model matching (SIM) method. A simplified LF model for the glottal source was incorporated within a discrete all-pole (DAP) modeling technique. The SIM method was proposed for a speech segment of 10 pitch cycles, and it assumes that the amplitudes of the glottal flow derivatives in the 10 cycles are constant (i.e., no shimmer); however, in practice, such an assumption does not hold often. Ding et al. (1995) adopted a completely time-varying autoregressive with exogenous input (ARX) model for the vocal tract, and the KLGLOTT88 glottal source model acts as its source. A simulated annealing optimization was used to identify the ARX model parameters in a pitch-synchronous fashion. More recently, Fu et al. (2006) proposed a pitch-synchronous method for jointly estimating source and filter parameters using the LF model for glottal source. A Kalman filtering process was embedded in the joint optimization process for adaptively identifying the vocal-tract parameters. In the pitch-synchronous method, Fu et al. considered signal segments between two consecutive glottal closure instants (GCIs) for analysis and assumed that the amplitudes of glottal flow derivative in both pitch cycles are identical; thus, the effect of shimmer was not directly incorporated in their optimization.

The shimmer and jitter (Yoshiyuki, 1982) in the glottal flow are two potential sources of perturbations in the parameters of the glottal flow waveform. Thus, the joint source-filter optimization should be formulated in a way to handle both shimmer and jitter. The amplitude of the glottal flow derivative signal can change from one pitch period to the next; this is known as shimmer. On the other hand, jitter occurs when the pitch period itself changes from one cycle to the next. Hence, the assumption of a fixed amplitude of the glottal flow in every pitch cycle may not be realistic. Similarly, assumption of fixed pitch periods while analyzing multiple pitch cycles also may cause errors in glottal source estimation. Thus, the joint source-filter optimization should be formulated in a way to handle both shimmer and jitter.

In this work, we present a joint source-filter optimization approach for estimating glottal flow using the LF model of the glottal flow derivative where the effects of shimmer and jitter are explicitly tackled. The vocal-tract filter is modeled by an auto-regressive filter. In this optimization approach, the amplitudes of the glottal flow derivative in each pitch cycle are estimated along with glottal flow and vocal-tract filter parameters. Experiments are conducted under a variety of shimmer and jitter conditions and the robustness of the proposed optimization method is demonstrated. The remainder of the paper begins with the details of the source and vocal-tract filter models used in the proposed optimization framework.

2. Source-filter model of speech production

2.1. AR speech production model

The proposed optimization method is developed based on the auto-regressive (AR) speech production model. In the AR speech production model, the speech signal $x[n]$ is considered to be the output of an all-pole linear time-invariant(LTI) filter[1] with input source signal $g[n]$ (Childers, 2000)

$$x[n] = -\sum_{p=1}^{P} a_p x[n-p] + g[n]. \tag{1}$$

The input source signal $g[n]$ is assumed to be the sum of the white gaussian noise $w[n]$ and samples of glottal flow derivative signal[2] $v_{T_0}[n] = v_{T_0}(nT_s)$, where $v_{T_0}(t)$ is the continuous-time glottal flow derivative signal with period T_0, and T_s is the sampling frequency. We assume that, at the operating sampling frequency $F_s = \frac{1}{T_s}$, the aliasing error due to sampling the non-bandlimited signal $v_{T_0}(t)$ is minimal. There can be cycle-to-cycle variations in the amplitude of the glottal derivative (shimmer) as well as in the period T_0 itself (jitter). Moreover, depending on the voice type, the

[1] In this paper, we have used the LTI AR model for simplicity. However, it can be easily extended to the time-varying (TV) AR model using an approach similar to (Hall et al., 1983).

[2] Since the speech production system is assumed to be LTI, the lip-radiation differentiator and the glottal flow $u_{T_0}[n]$ can be combined to result in the glottal flow derivative signal as the input to the filter. When the system is TV, such an operation is still valid assuming the vocal tract is slowly-varying in vowels (Fu et al., 2006).

amount of the additive noise $w[n]$ also changes. We incorporate all these effects in our proposed optimization framework.

2.2. Glottal flow derivative model

There are various models available for the glottal flow derivative waveform. We use the Liljencrants–Fant (LF) model in this paper because it provides a good fit for most of the commonly encountered glottal flow derivative waveform shapes and is flexible in its ability to match extreme phonations (Fant and Lin, 1985). The LF model is defined for continuous time glottal flow derivative signals as follows:

$$v_{T_0}(t) = \begin{cases} E_0 e^{\alpha t}\sin(\omega_g t) & \text{if } 0 \leqslant t \leqslant t_e; \\ -\frac{E_e}{\epsilon t_a}\left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}\right] & \text{if } t_e \leqslant t \leqslant t_c; \\ 0 & \text{if } t_c \leqslant t \leqslant T_0. \end{cases} \tag{2}$$

The LF model over one pitch period is composed of two sections: an open phase with an exponentially growing sinusoid followed by a decaying exponential return phase. Along with the magnitude of the glottal closure excitation E_e, the LF model for glottal flow derivative can be specified with two sets of independent parameters, namely the direct synthesis parameters $\{E_0, \alpha, \omega_g, \epsilon\}$ and the timing parameters $\{t_p, t_e, t_a, t_c\}$ (Fant and Lin, 1985). In this paper, we choose to work with the timing parameters. Given the timing parameters, the direct synthesis parameters can be obtained with the following constraints:

$$\begin{cases} \int_0^{T_0} v_{T_0}(t)dt = 0, \\ \omega_g = \frac{\pi}{t_p}, \\ \epsilon t_a = 1 - e^{-\epsilon(t_c-t_e)}, \\ E_e = -E_0 e^{\alpha t_e}\sin(\omega_g t_e). \end{cases} \tag{3}$$

As illustrated in Fig. 1, t_p, t_e, t_c represent the instants of the maximum glottal flow, maximum negative value of the glottal flow derivative, and glottal closure, respectively; t_a is the effective duration of the return phase.

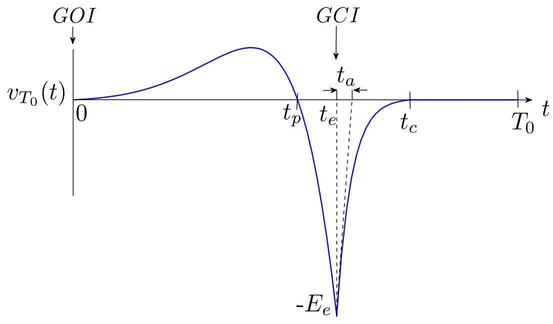

Fig. 1. A sample glottal flow derivative waveform of the LF model. The locations of glottal opening instant (GOI) and glottal closing instant (GCI) are also illustrated.

3. Joint source-filter optimization

Let us denote the observed speech samples by $x[n]$, $0 \leqslant n \leqslant N-1$. Following Eq. (1), the observed samples can be modeled as follows:

$$\begin{aligned} x[n] &= -\sum_{p=1}^{P} a_p x[n-p] + g[n], \quad P \leqslant n \leqslant N-1, \\ \text{where} \quad g[n] &= v[n] + w[n] \\ &= \sum_{k=0}^{K} \alpha_k v_{(\eta_{k+1}-\eta_k)T_s}[n-\eta_k] + w[n]. \end{aligned} \tag{4}$$

α_k is the amplitude of the glottal flow derivative waveform in the kth cycle and $\eta_k \in \mathcal{Z}$, $0 \leqslant k \leqslant K$ are the glottal opening instants (GOIs). K GOIs appear during the observed signal segment. It is assumed that (Childers, 2000) $w[n] \sim \mathcal{N}(0, \sigma^2)$, $\forall n$, i.e. zero-mean white Gaussian noise with variance σ^2. The glottal signal-to-noise ratio is $\mathrm{SNR}_g \triangleq 10\log_{10}\left(\frac{\frac{1}{N}\sum_n v^2[n]}{\sigma^2}\right)$. Note that according to the model in Eq. (4), η_k can take only integer values; hence, any real valued GOIs cannot be modeled exactly using η_k. We assume that this modeling error has an insignificant effect on the accuracy of the glottal parameter estimates for large F_s.

The effect of shimmer and jitter is reflected in the values of α_k and η_k. When there is no shimmer, $\alpha_k = \text{constant}\ \forall k$. Similarly, when there is no jitter, $\eta_{k+1} - \eta_k = \text{constant}\ \forall k$.

Given the model of the observed speech signal samples as in Eq. (4), the goal of the joint source-filter optimization is to find t_p, t_e, t_a, t_c, $\{\alpha_k\}$, $\{\eta_k\}$, and $\{a_p\}$ such that the cost function $J \triangleq E[w^2[n]] \approx \frac{1}{N-P+1}\sum_{n=P}^{N-1} w^2[n]$ is minimized. The cost function can be rewritten as follows:

$$\begin{aligned} J &= \frac{1}{N-P+1}\sum_{n=P}^{N-1} w^2[n] \\ &\propto \sum_{n=P}^{N-1}\left(x[n] + \sum_{p=1}^{P} a_p x[n-p] - v[n]\right)^2 \quad \text{[using Eq. (4)]} \\ &= \sum_{n=P}^{\eta_1-1}\left(x[n] + \sum_{p=1}^{P} a_p x[n-p] - \alpha_0 v_{(\eta_1-\eta_0)T_s}[n-\eta_0]\right)^2 \\ &\quad + \sum_{k=1}^{K-1}\sum_{n=\eta_k}^{\eta_{k+1}-1}\left(x[n] + \sum_{p=1}^{P} a_p x[n-p] - \alpha_k v_{(\eta_{k+1}-\eta_k)T_s}[n-\eta_k]\right)^2 \\ &\quad + \sum_{n=\eta_K}^{N-1}\left(x[n] + \sum_{p=1}^{P} a_p x[n-p] - \alpha_K v_{(\eta_{K+1}-\eta_K)T_s}[n-\eta_K]\right)^2, \end{aligned} \tag{5}$$

where the summation over n has been split into $K+1$ summations, one for each period of $v[n]$.

3.1. No jitter case

Let us consider, for simplicity, $\eta_{k+1} - \eta_k = \text{constant} = N_0$ (no jitter), $\forall k$, i.e. $\eta_k = \eta_0 + kN_0$. We will discuss later how we will handle the case $\eta_{k+1}-\eta_k \neq$ constant. Let us also assume that N_0 is known or estimated from the observed signal and P is known a priori. Then the optimization problem becomes

$$\left\{t_p^\star, t_e^\star, t_a^\star, t_c^\star, \{\alpha_k^\star\}_{k=0}^{K}, \eta_0^\star, \left\{a_p^\star\right\}_{p=1}^{P}\right\} = \arg \min_{\substack{t_p,t_e,t_a,t_c,\\ \{\alpha_k\},\eta_0,\{a_p\}}} J,$$

where

$$\begin{aligned} J = &\sum_{n=P}^{\eta_0+N_0-1}\left(x[n]+\sum_{p=1}^{P}a_p x[n-p]-\alpha_0 v_{N_0T_s}[n-\eta_0]\right)^2 \\ &+\sum_{k=1}^{K-1}\sum_{n=\eta_0+kN_0}^{\eta_0+(k+1)N_0-1}\left(x[n]+\sum_{p=1}^{P}a_p x[n-p]-\alpha_k v_{N_0T_s}[n-\eta_0-kN_0]\right)^2 \\ &+\sum_{n=\eta_0+KN_0}^{N-1}\left(x[n]+\sum_{p=1}^{P}a_p x[n-p]-\alpha_K v_{N_0T_s}[n-\eta_0-KN_0]\right)^2. \end{aligned} \tag{6}$$

Note that for a set of given values of t_p, t_e, t_a, t_c, and η_0, J is a quadratic function of $\{\alpha_k\}$, $\{a_p\}$ and, thus, $\{\alpha_k^\star\}$, $\left\{a_p^\star\right\}$ can be directly obtained by setting $\frac{\partial J}{\partial a_r}=0,\ r=1,\ldots,P$ and $\frac{\partial J}{\partial \alpha_l}=0,\ l=0,\ldots,K$ and solving a set of $P+K+1$ linear equations. $\frac{\partial J}{\partial a_r}=0,\ r=1,\ldots,P$ results in the following equations:

$$\begin{aligned} &\sum_{p=1}^{P}a_p\sum_{n=P}^{N-1}x[n-p]x[n-r]-\left[\alpha_0\sum_{n=P}^{\eta_0+N_0=1}v_{N_0T_s}[n-\eta_0]x[n-r]\right. \\ &\quad+\sum_{k=1}^{K-1}\alpha_k\sum_{n=\eta_0+kN_0}^{\eta_0+(k+1)N_0-1}v_{N_0T_s}[n-\eta_0-kN_0]x[n-r] \\ &\quad\left.+\alpha_k\sum_{n=\eta_0+KN_0}^{N-1}v_{N_0T_s}[n-\eta_0-KN_0]x[n-r]\right] \\ &=-\sum_{n=P}^{N-1}x[n]x[n-r]. \end{aligned} \tag{7}$$

Similarly,

$$\begin{aligned} \frac{\partial J}{\partial \alpha_0}=0 \Rightarrow &\sum_{p=1}^{P}a_p\sum_{n=P}^{\eta_0+N_0-1}x[n-p]v_{N_0T_s}[n-\eta_0] \\ &-\alpha_0\sum_{n=P}^{\eta_0+N_0-1}v^2_{N_0T_s}[n-\eta_0]=-\sum_{n=P}^{\eta_0+N_0-1}x[n]v_{N_0T_s}[n-\eta_0], \end{aligned} \tag{8}$$

$$\begin{aligned} &\frac{\partial J}{\partial \alpha_l}=0, \quad l=1,\ldots,K-1 \\ &\quad\Rightarrow\sum_{p=1}^{P}a_p\sum_{n=\eta_0+lN_0}^{\eta_0+(l+1)N_0-1}x[n-p]v_{N_0T_s}[n-\eta_0-lN_0] \\ &\qquad-\alpha_l\sum_{n=\eta_0+lN_0}^{\eta_0+(l+1)N_0-1}v^2_{N_0T_s}[n-\eta_0-lN_0] \\ &\quad=-\sum_{n=\eta_0+lN_0}^{\eta_0+(l+1)N_0-1}x[n]v_{N_0T_s}[n-\eta_0-lN_0] \end{aligned} \tag{9}$$

and

$$\begin{aligned} \frac{\partial J}{\partial \alpha_K}=0 \Rightarrow &\sum_{p=1}^{P}a_p\sum_{n=\eta_0+KN_0}^{N-1}x[n-p]v_{N_0T_s}[n-\eta_0-KN_0] \\ &-\alpha_K\sum_{n=\eta_0+KN_0}^{N-1}v^2_{N_0T_s}[n-\eta_0-KN_0] \\ &=-\sum_{n=\eta_0+KN_0}^{N-1}x[n]v_{N_0T_s}[n-\eta_0-KN_0]. \end{aligned} \tag{10}$$

Combining the set of linear equations of Eqs. (7)–(10), we can write them in matrix vector form as follows:

$$\boldsymbol{Ra}=\boldsymbol{p}, \tag{11}$$

where

$$\boldsymbol{R}=\begin{pmatrix}\boldsymbol{R}_1 & -\boldsymbol{R}_2\\ -\boldsymbol{R}_2^{\mathrm{T}} & \boldsymbol{R}_3\end{pmatrix},$$

where

$$\boldsymbol{R}_1=\begin{pmatrix}C_{xx}(1,1) & \ldots & C_{xx}(P,1)\\ \vdots & \ldots & \vdots\\ C_{xx}(1,P) & \ldots & C_{xx}(P,P)\end{pmatrix},$$

$$\boldsymbol{R}_2=\begin{pmatrix}C^0_{vx}(\eta_0,1) & C_{vx}(\eta_0+N_0,1) & \ldots & C_{vx}(\eta_0+(K-1)N_0,1) & C^{K+1}_{vx}(\eta_0+KN_0,1)\\ \vdots & \vdots & \ldots & \vdots & \vdots\\ C^0_{vx}(\eta_0,P) & C_{vx}(\eta_0+N_0,P) & \ldots & C_{vx}(\eta_0+(K-1)N_0,P) & C^{K+1}_{vx}(\eta_0+KN_0,P)\end{pmatrix},$$

$$\boldsymbol{R}_3=\begin{pmatrix}E^0_v & 0 & \ldots & 0 & 0\\ 0 & E_v & \ldots & 0 & 0\\ \vdots & \vdots & \ldots & \vdots & \vdots\\ 0 & 0 & \ldots & E_v & 0\\ 0 & 0 & \ldots & 0 & E^{k+1}_v\end{pmatrix},$$

$$\boldsymbol{a}=\begin{pmatrix}a_1\\ \vdots\\ a_p\\ \alpha_0\\ \alpha_1\\ \vdots\\ \alpha_{K-1}\\ \alpha_K\end{pmatrix} \quad\text{and}\quad \boldsymbol{p}=\begin{pmatrix}-C_{xx}(0,1)\\ \vdots\\ -C_{xx}(0,P)\\ C^0_{vx}(\eta_0,0)\\ C_{vx}(\eta_0+N_0,0)\\ \vdots\\ C_{vx}(\eta_0+(K-1)N_0,0)\\ C^{K+1}_{vx}(\eta_0+KN_0,0)\end{pmatrix}.$$

C_{xx}, C_{vx}, C^0_{vx}, C^{K+1}_{vx}, E^0_v, E_v, and E^{K+1}_v are defined as follows:

$$\left\{\begin{array}{l} C_{xx}(p,r) = \sum_{n=P}^{N-1} x[n-p]x[n-r], \\ C_{vx}(p,r) = \sum_{n=P}^{P+N_0-1} v_{N_0T_s}[n-p]x[n-r], \\ C^{0}_{vx}(\eta_0,r) = \sum_{n=P}^{\eta_0+N_0-1} v_{N_0T_s}[n-\eta_0]x[n-r], \\ C^{K+1}_{vx}(\eta_0+KN_0,r) = \sum_{n=\eta_0+KN_0}^{N-1} v_{N_0T_s}[n-\eta_0-KN_0]x[n-r], \\ E^{0}_{v} = \sum_{n=P}^{\eta_0+N_0-1} v^2_{N_0T_s}[n-\eta_0], \\ E_v = \sum_{n=\eta_0+lN_0}^{\eta_0+(l+1)N_0-1} v^2_{N_0T_s}[n-\eta_0-lN_0], \\ E^{K+1}_{v} = \sum_{n=\eta_0+KN_0}^{N-1} v^2_{N_0T_s}[n-\eta_0-KN_0]. \end{array}\right.$$

Note that $\boldsymbol{R}_1$ is identical to the matrix obtained in autocovariance analysis of linear prediction (Rabiner and Schafer, 2010). $\boldsymbol{R}$ is a symmetric matrix and can be interpreted as the covariance matrix of the signal and the glottal flow derivative and thus similar to the auto-covariance analysis of linear prediction (Rabiner and Schafer, 2010), the glottal parameter vector $\boldsymbol{a}$ can be obtained by $\boldsymbol{a} = \boldsymbol{R}^{-1}\boldsymbol{p}$ under the assumption of positive definiteness of $\boldsymbol{R}$.

To estimate t_p, t_e, t_a, t_c, η_0, we perform a combinatorial search assuming that t_p, t_e, t_a, t_c and $-\eta_0 \in \{nT_s;\ 0 \leqslant n \leqslant (N_0-1)\}$. We find that such discrete approximation leads to small error if the best (closest) discrete time index is selected for the corresponding original continuous parameter for most of the voice types. Considering all possible values of t_p, t_e, t_a, t_c and $-\eta_0$, the cardinality of the search space becomes N_0^5; however, many of these combinations are infeasible due to constraints on the values of timing parameters such as $t_p < t_e < t_e + t_a < t_c$. Also, for example, typical values of (t_c/T_0) for various voice types are not less than 35%; the ratio (t_p/t_c) lies between 0.5 and 0.7 for various voice types (Childers, 2000). Considering all these constraints, we decide to choose the following ranges of the timing parameters – $0 \leqslant \eta_0 \leqslant N_0 - 1$, $0.25 \leqslant \frac{t_c}{N_0 T_s} \leqslant 1$, $0.45 \leqslant \frac{t_p}{t_c} \leqslant 0.75$, $t_p < t_e < t_c$, $0 \leqslant t_a \leqslant t_c - t_e$. Our choice of various timing parameters ensures that their typical values are well within the considered ranges for optimization.

3.2. Jitter case

Based on the discussion above, it is clear that, under the assumption of no jitter, the proposed optimization method can be applied to the observed signal of any duration; no pitch-synchronous analysis is required. In the presence of jitter, the objective function of Eq. (5) can be directly minimized to obtain the estimate of the GOIs $\{\eta_k\}$ in addition to t_p, t_e, t_a, t_c, $\{\alpha_k\}$, and $\{a_p\}$. Instead of searching GOIs $\{\eta_k\}$ directly, one approach is to search over the variation of the pitch period (δN_k) from some average pitch period. As $\eta_{k+1} - \eta_k \neq$ constant, $\forall k$, let us assume that the pitch period (N_0) estimated from the given signal segment is close to the average of $K+1$ pitch periods in the signal segment, i.e. each of the $K+1$ pitch periods in the observed speech segment is within the interval $[N_0(1-Q)\ N_0(1+Q)]$, where N_0Q is a small fraction of N_0. Then, to estimate the GOIs $\{\eta_k\}$, we can search over the deviation $(-\lfloor N_0Q \rfloor \leqslant \delta N_k \leqslant \lceil N_0Q \rceil)$[3] of the kth pitch period from the average pitch period N_0 instead of searching over $\{\eta_k\}$ directly. Choosing $Q = 0.2$ covers the range of jitter for various voice types (Childers, 2000). Even with such alternative optimization variables, the search complexity become $O(N_0^{K+6}Q^{K+1})$ when we need to optimize Eq. (5) in the presence of both shimmer and jitter. For example, for a signal segment of 160 samples (i.e., 20 ms for $F_s = 8$ kHz) and $N_0 = 40$, $K \approx 4$ and therefore, $N_0^{K+6}Q^{K+1} \approx 3.4 \times 10^{15}$. One plausible approach to circumvent this large search complexity issue is to decrease the duration of the speech signal segment so that K is less. It is also clear that the effect of jitter on the optimization problem reduces as we reduce the signal segment duration for analysis. To remove the effect of jitter, ideally, the signal segment between two consecutive GOIs should be used because it is completely free from jitter; but the detection of GOIs is less accurate compared to GCI detection (Drugman and Dutoit, 2009), which can lead to poor estimates of glottal parameters. Hence, for consistency across various voice types and vowels, we restrict the duration of the signal segment to be the duration between two consecutive GCIs for analysis. The pitch period is estimated by taking the difference between corresponding GCIs. Hence, in the presence of jitter, a pre-processing stage for GCI detection is required before the observed signal can be fed into the proposed optimization; note that this pre-processing is not required when we consider only the effect of shimmer. Also note that by restricting the duration of the signal segment under analysis, only the effect of jitter is minimized; the effect of shimmer is handled inside the optimization cost function.

4. Experiments and results

We evaluated the proposed optimization method on both synthetic and real speech. The sampling frequency of the signals considered was $F_s = 8$ kHz. Five different vowels were used in our experiments- /ʌ/, /a/, /æ/, /U/ and /I/. In the case of synthetic speech, the true values of the source and the filter parameters are known, and hence, the estimated parameters can be compared against them and, thus, the effectiveness of the proposed optimization can be judged. In the case of real speech, the true values of the parameters are not known; hence, we have used open quotient (OQ) values based on concurrently obtained electroglottograph (EGG) signals for evaluation.

[3] $\lfloor x \rfloor$ is the largest integer smaller than x and $\lceil x \rceil$ is the smaller integer greater than x.

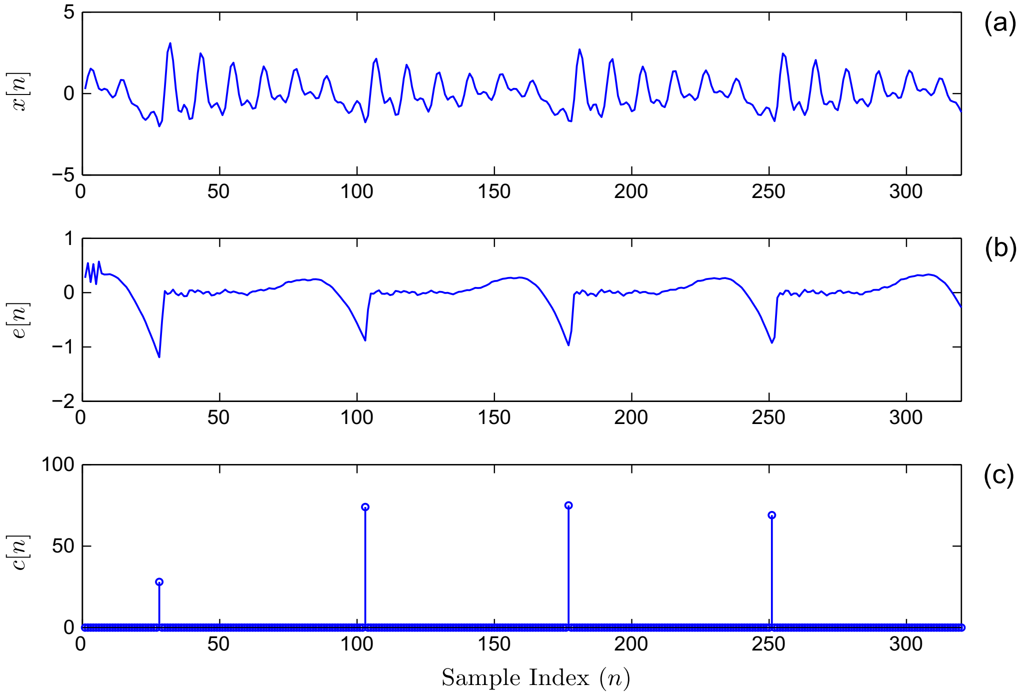

Fig. 2. GCI detection from the residual signal: (a) a sample vowel segment, (b) corresponding LP residual signal $e[n]$, and (c) the count sequence $c[n]$.

We assume that the observed signal suffers from both shimmer and jitter, in general. Thus, as discussed in Section 3, detection of GCI is a necessary step before the optimization is deployed. We first discuss our approach for GCI detection.

4.1. GCI detection

For GCI detection, we use the linear-prediction (LP) residual signal $e[n]$ obtained from the observed speech signal, $x[n]$, $0 \leqslant n \leqslant N-1$. At first, $x[n]$ is pre-emphasized and from the pre-emphasized signal, LP coefficients are calculated using standard autocorrelation analysis (Markel and Gray, 1976). The estimated LP filter is then used to perform inverse filtering of $x[n]$ and, thus, $e[n]$ is obtained. As an illustrative example, $x[n]$ and the corresponding $e[n]$ are shown in Fig. 2(a) and (b) respectively. An estimate of the pitch period $(\widehat{N}_0)$ is obtained from $e[n]$ using an approach outlined in (Shimamura et al., 2001).

We design a count sequence $c[n]$ of length N, which is initially set to zero. For every segment of $e[n]$ of length $\widehat{N}_0$, let the maximum of the $|e[n]|$, $k \leqslant n \leqslant k+\widehat{N}_0-1$ occurs[4] at m_k; then, $c[m_k]$ is incremented by one. This is performed for $1 \leqslant k \leqslant N-\widehat{N}_0+1$. This yields the final count signal $c[n]$ illustrated in Fig. 2(c). Note that the count is high only at the places where GCI occurs. The locations of the top $\left\lfloor \frac{N}{\widehat{N}_0} \right\rfloor$ values from $c[n]$ are selected and declared to be the estimates of GCI for the observed signal segment.

After the GCIs are detected, a GCI c_k is randomly chosen and the signal segment between the randomly chosen GCI (c_k) and the next GCI (c_{k+1}) is used for optimizing the source and filter parameters. $c_{k+1}-c_k$ is used as the prior estimate of N_0 in the optimization process.

4.2. Results for synthetic speech

We synthetically generated five vowels (/ʌ/, /a/, /æ/, /U/, and /I/) using the source-filter model of speech production for different voice types (Childers, 2000) namely, Modal, Vocal Fry, Breathy, Falsetto, Harsh. By varying the filter parameters (corresponding to different vowels) and source parameters (corresponding to different voice types), we would like to evaluate the performance of the proposed source-filter optimization across various source and filter types. The formant frequencies and the bandwidths of /ʌ/, /a/, /æ/, /U/, and /I/ are taken from (Flanagan, 1965) which are listed in Table 1. The glottal flow derivative parameters for different voice types were simulated following (Childers, 2000, Table A7.3, p. 315).

For a specific vowel and voice type combination, 100 different realizations of vowel sounds were generated by randomly choosing t_p, t_e, t_a, t_c within a range of 2% around the ideal values given in (Childers, 2000) (also listed in Table 2). The SNR of the glottal derivative waveform SNR_g is simulated using values given in (Childers, 2000). The amplitude of the glottal flow derivative in each pitch

[4] The absolute value of $e[n]$ is taken before selecting the maximum value because the signal polarity might be different for different speech signals (Saratxaga et al., 2009).

Table 1
Formant frequencies and bandwidths (in Hz) to synthetically generate five vowels /ʌ/, /a/, /æ/, /U/, and /I/.

Vowels	Formants (F_i) and bandwidths (B_i)					
	F_1	B_1	F_2	B_2	F_3	B_3
/ʌ/	700	56	1400	52	2500	105
/a/	800	60	1200	50	2600	105
/æ/	720	65	1800	90	2500	155
/U/	520	50	1200	60	2300	90
/I/	470	42	2100	70	2800	143

Table 2
Various LF parameter values to synthesize glottal waveforms.

Voice	t_p (%)	t_e (%)	t_a (%)	t_c (%)	Jitter (%)	SNR$_g$ (dB)
Modal	41.0	55.0	0.9	58.0	2.0	40.0
Vocal Fry	48.0	59.0	2.7	72.0	10.0	20.0
Breathy	46.0	66.0	2.7	77.0	5.0	20.0
Falsetto	50.0	80.0	8.0	100.0	2.0	50.0
Harsh	25.0	30.0	1.0	50.0	10.0	10.0

cycle is chosen randomly to be between 1 and 1.5; thus, the shimmer in the synthetic vowel is simulated. Note that the amplitudes are randomly determined and hence need not be smoothly varying from one pitch cycle to the next. Similarly the jitter is simulated by varying the pitch period randomly from one cycle to the next according to jitter values given in (Childers, 2000); thus, the pitch period from one cycle to the next need not be assumed to be smoothly changing either. It is important to note that the synthetic vowels with shimmer and jitter are completely artificial; the purpose of these experiments is to examine the performance of the proposed optimization under hypothetically extreme shimmer and jitter conditions. LF parameter and noise model parameter values used in synthesizing the glottal source signal are listed in Table 2. The average fundamental frequency for all simulations was chosen to be 108 Hz. 108 Hz corresponds to a typical male voice.

For each realization of the synthesized vowels (0.8 s long), a 20 ms segment was randomly selected for estimating source and filter parameters using the proposed optimization approach. At first this 20 ms segment is passed through the GCI detection module and only the signal segment between two consecutive GCIs is deployed for the optimization. Fig. 3 illustrates the original glottal flow derivative and the glottal flow derivative computed using

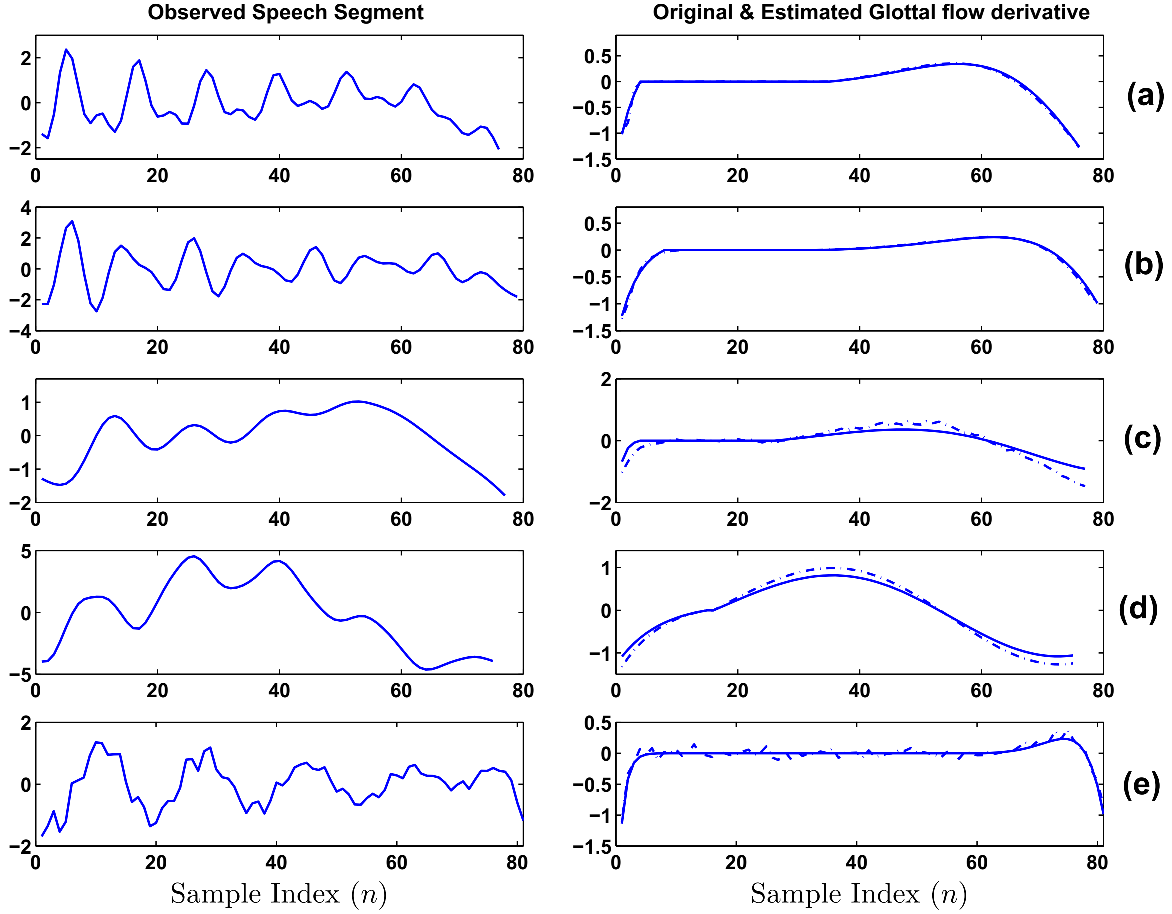

Fig. 3. Illustration of the glottal flow derivative computed using estimated glottal parameters for different vowels and voice types: (a) Modal /ʌ/, (b) Vocal Fry /a/, (c) Breathy /æ/, (d) Falsetto /U/ and (e) Harsh /I/. The original glottal flow derivatives are plotted in dash-dotted line and the estimated ones are plotted in solid lines.

the estimated glottal parameters for five different vowels and voice types combinations.

The left column in Fig. 3 plots the signal segment used for optimization; the column on the right shows the glottal flow derivative obtained using estimated glottal parameters overlaid on the original glottal flow derivative signal. From the chosen signal segments in Fig. 3, it appears that the estimated glottal flow derivative in the case of Breathy /æ/ and Falsetto /U/ are worse compared to the ones in the case of /ʌ/, /a/, and /I/. In general, however, the estimated glottal flow derivatives follow the original ones for all the cases presented in Fig. 3.

For a more systematic and comprehensive evaluation, following (Fu et al., 2006), we report the absolute percentage error of the estimated parameters over 100 realizations for each vowel and voice type combination. Thus we report the results for the estimated t_p, t_e, t_a, t_c and the estimated amplitudes α_0 and α_1 of the glottal flow derivative in two cycles appearing in the observed signal segment between two GCIs. For each parameter θ and its estimate $\hat{\theta}$, the absolute percentage error δ_r is computed as follows:

$$\delta_r = \frac{|\hat{\theta} - \theta|}{\theta} \times 100\%. \tag{12}$$

The mean and standard deviations of δ_r over 100 realizations are reported for different voice types in Tables 3–7.

The last row in each table demonstrates the least possible error because of finite grid search for the timing parameters. We call this *Model error*. For computing the *Model error*, the discrete-time points (integer multiples of T_s) closest to the timing parameters are chosen as an estimate of the respective parameters. There is no such bound for error in the case of α_0 and α_1. It is clear from Tables 3–7 that the *Model error* is close to zero for all of the timing parameters except for t_a in the case of Modal and Harsh voice types. This is because the t_a for these voice types are smaller than the resolution (T_s) of the search grid for the timing parameters. Low *Model error* demonstrates that such a finite grid search method can yield estimated timing parameters very close to the true ones under most of the voice types.

From Tables 3–7, it appears that the estimation error is higher for t_a than other parameters. This is mainly because t_a is of the order of the resolution T_s of our search grid. Also, it is clear that the estimation errors for the Harsh voice type are higher, in general, compared to those for other voice types. This is because SNR_g is assumed to be 10 for Harsh voice types, while SNR_g is 40, 20, 20, and 50 for Modal, Vocal Fry, Breathy, and Falsetto voice types, respectively (Childers, 2000). The jitter values (%) are 2, 10, 5, 2, 10 for Modal, Vocal Fry, Breathy, Falsetto, and Harsh voice types, respectively. The error of t_c for different voice types shows a trend similar to their relative jitter values. For example, δ_r for t_c is greater in the case of Vocal Fry and Harsh voice types compared to other voice types.

We have also performed experiments with an average fundamental frequency of 201Hz, corresponding to a female voice. On average, absolute percentage errors in the estimates of the timing parameters, α_0 and α_1 in the case of the female voice, remain similar to those of male voice. For example, the absolute percentage errors for various parameters for the Modal voice type are, on average, absolute 2–3% more in the case of higher pitch while the absolute percentage errors for the Harsh voice type are, on average, absolute 4–5% less in the case of higher pitch.

Table 3
Absolute percentage error of the estimated glottal parameters for Modal speech.

Vowel	Mean (SD) δ_r (Modal)					
	t_p	t_e	t_a	t_c	α_0	α_1
/ʌ/	3.85 (2.65)	3.33 (2.26)	82.08 (61.51)	2.28 (1.48)	5.99 (2.57)	6.21 (3.92)
/a/	3.64 (2.56)	3.15 (2.13)	81.95 (61.36)	2.28 (1.42)	5.46 (3.01)	6.07 (4.66)
/æ/	4.66 (2.50)	3.78 (1.72)	68.28 (61.01)	20.82 (23.93)	58.12 (42.29)	14.09 (7.33)
/U/	3.80 (2.27)	3.27 (1.86)	86.34 (64.22)	2.08 (1.40)	5.19 (2.96)	4.59 (3.59)
/I/	3.80 (2.32)	3.20 (1.87)	77.33 (57.81)	2.48 (4.30)	6.48 (4.72)	6.68 (5.51)
Model error	1.02 (0.47)	0.61 (0.40)	50.20 (0.83)	0.51 (0.32)	–	–

Table 4
Absolute percentage error of the estimated glottal parameters for Vocal Fry speech.

Vowel	Mean (SD) δ_r (Vocal Fry)					
	t_p	t_e	t_a	t_c	α_0	α_1
/ʌ/	3.82 (3.51)	3.12 (2.65)	32.52 (37.58)	14.21 (17.86)	10.68 (6.30)	10.70 (7.40)
/a/	4.44 (3.36)	3.75 (2.85)	25.14 (27.66)	15.95 (18.64)	9.46 (5.61)	8.46 (5.94)
/æ/	3.59 (2.48)	3.52 (2.57)	25.87 (26.73)	10.91 (14.04)	28.50 (19.59)	15.99 (11.11)
/U/	5.54 (4.18)	4.66 (3.23)	36.31 (39.19)	15.20 (17.90)	10.59 (7.00)	7.48 (4.67)
/I/	4.61 (3.42)	4.31 (2.83)	32.85 (27.61)	12.61 (17.18)	16.06 (8.56)	15.81 (8.56)
Model error	0.89 (0.29)	0.62 (0.34)	0.52 (0.30)	0.48 (0.26)	–	–

Table 5
Absolute percentage error of the estimated glottal parameters for Breathy speech.

Vowel	Mean (SD) δ_r (Breathy)					
	t_p	t_e	t_a	t_c	α_0	α_1
/ʌ/	2.69 (2.05)	2.67 (1.90)	38.71 (52.50)	7.48 (9.82)	6.54 (4.98)	5.98 (4.30)
/a/	3.44 (2.64)	3.66 (2.99)	37.26 (56.30)	9.40 (10.97)	7.80 (7.68)	7.93 (7.60)
/æ/	2.32 (1.87)	2.00 (1.41)	43.33 (37.90)	4.17 (4.25)	30.39 (21.50)	17.80 (14.53)
/U/	3.59 (2.18)	3.47 (2.19)	34.31 (54.45)	8.72 (10.32)	8.49 (8.35)	7.99 (6.43)
/I/	2.85 (2.08)	2.70 (1.77)	28.30 (36.65)	6.75 (8.63)	7.25 (6.29)	8.25 (6.72)
Model error	0.51 (0.30)	0.49 (0.31)	0.51 (0.30)	0.48 (0.27)	–	–

Table 6
Absolute percentage error of the estimated glottal parameters for Falsetto speech.

Vowel	Mean (SD) δ_r (Falsetto)					
	t_p	t_e	t_a	t_c	α_0	α_1
/ʌ/	2.16 (1.22)	1.37 (1.30)	15.43 (14.71)	1.27 (1.43)	35.72 (19.52)	35.89 (19.64)
/a/	2.28 (1.07)	1.86 (1.41)	14.04 (16.46)	1.28 (0.96)	34.73 (18.23)	34.74 (18.17)
/æ/	2.39 (1.40)	1.24 (1.04)	27.54 (20.49)	2.83 (2.77)	37.23 (21.00)	37.87 (20.68)
/U/	2.16 (1.19)	1.47 (1.00)	13.37 (14.99)	1.42 (1.54)	30.89 (19.52)	30.71 (19.06)
/I/	2.43 (1.43)	1.92 (1.70)	17.04 (19.63)	1.95 (2.67)	53.71 (22.38)	53.87 (21.89)
Model error	0.52 (0.29)	0.43 (0.23)	1.34 (0.58)	0.35 (0.19)	–	–

Table 7
Absolute percentage error of the estimated glottal parameters for Harsh speech.

Vowel	Mean (SD) δ_r (Harsh)					
	t_p	t_e	t_a	t_c	α_0	α_1
/ʌ/	37.45 (40.19)	30.55 (33.29)	69.33 (61.87)	20.94 (29.09)	16.56 (10.32)	12.84 (7.68)
/a/	31.26 (31.05)	24.85 (25.40)	78.75 (80.95)	14.84 (24.08)	16.31 (11.71)	12.13 (6.54)
/æ/	29.49 (31.05)	23.03 (25.15)	192.15 (146.66)	13.60 (20.10)	21.70 (17.70)	11.69 (8.65)
/U/	34.43 (37.50)	27.56 (31.05)	75.72 (88.74)	14.41 (23.96)	23.52 (24.67)	12.60 (6.91)
/I/	40.35 (46.89)	31.99 (38.42)	69.04 (72.78)	20.63 (31.50)	15.72 (8.80)	14.16 (7.74)
Model error	2.20 (0.29)	0.93 (0.56)	35.12 (0.77)	0.51 (0.30)	–	–

4.3. Results for natural speech

We have used sustained vowel samples spoken by male and female speakers from the CD available with Childer's book (Childers, 2000). Fig. 4 illustrates the estimates of the glottal flow derivatives when a 30 ms long speech segment is used for optimization corresponding to five vowels /ʌ/, /a/, /æ/, /U/ and /I/ spoken by randomly chosen five male and five female speakers. The speech segment is taken from a sustained portion of the vowel. A 30 ms segment is chosen only for visual illustration. The glottal flow derivatives in Fig. 4 are estimated assuming there is no jitter in the speech segments, i.e. Eq. (6) is solved directly using 30 ms long speech segment and no GCI is detected. This visually demonstrates that when there is no jitter or the effect of jitter is minimal, the optimization problem of Eq. (6) can be directly solved in a non pitch-synchronous fashion.

For a quantitative evaluation, we randomly select three male and female speakers. For each speaker, five vowel utterances corresponding to /ʌ/, /a/, /æ/, /U/, and /I/ are used for the experiments. Since the actual glottal flow derivative parameters for natural speech are not known, we follow an evaluation strategy similar to that outlined in (Fu et al., 2006; Frohlich et al., 2001). Open quotients (OQ_{EGG}) estimated from the EGG signal, also available in the CD (Childers, 2000), are used as reference for our evaluation. The OQ_{EGG} is estimated following (Krishnamurthy et al., 1986), in which GOI and GCI are detected from the peaks of the differentiated EGG signal and OQ_{EGG} is obtained by taking the ratio of the duration between GCI and GOI and the duration between two consecutive GOIs.

For each speaker and vowel combination, a similar analysis is performed as for the synthetic vowel case – a randomly chosen 30 ms speech segment is first passed through the GCI detection module and only the signal segment between two GCIs is used for optimization. Since the order of the filter is not known in the case of real speech, *P*

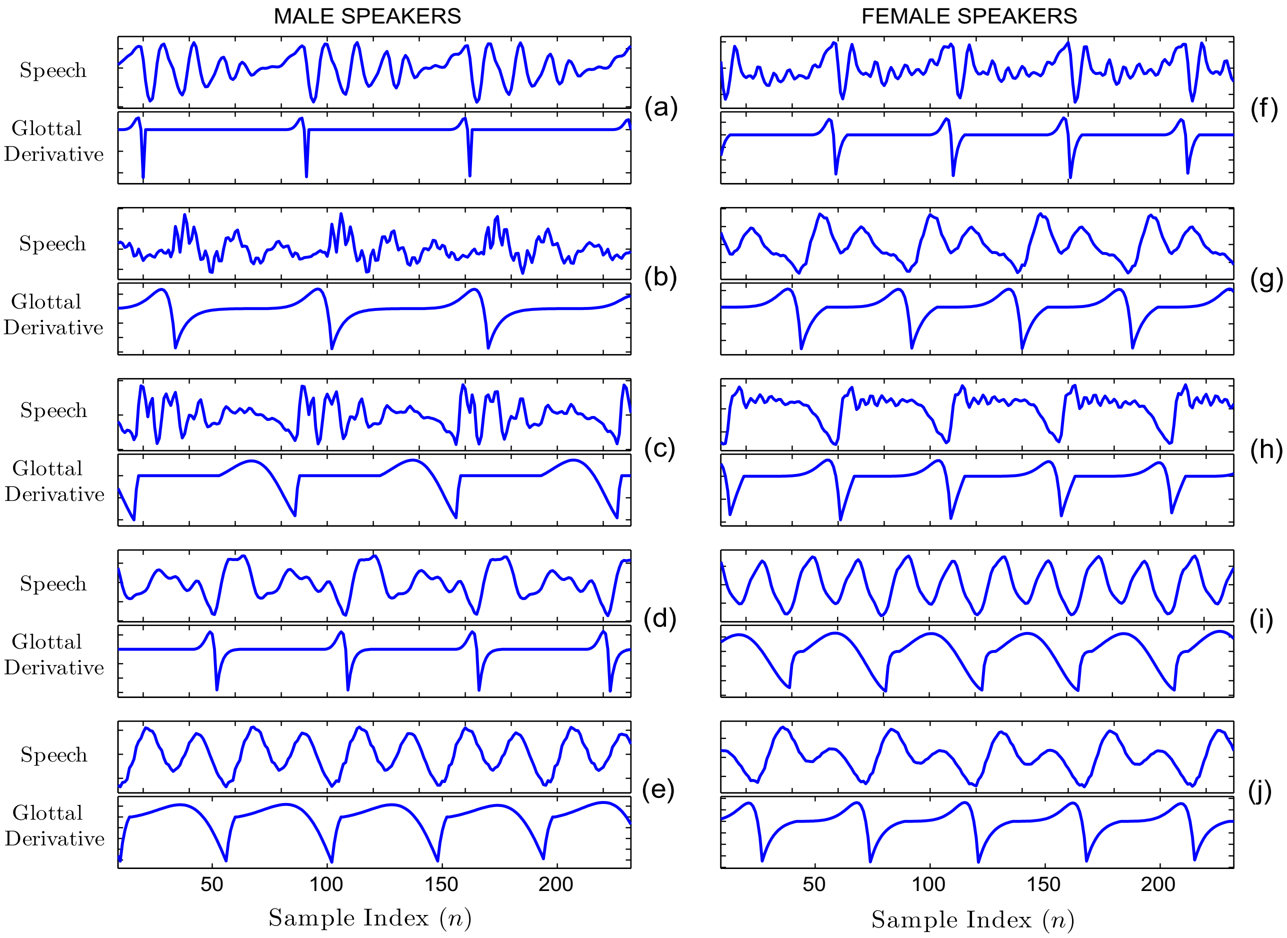

Fig. 4. The natural speech segments and the corresponding estimated glottal flow derivatives for five vowels /ʌ/, /a/, /æ/, /U/, and /I/ spoken by five male speakers (a)–(e) and five female speakers (f)–(j).

is kept fixed at 8 in the optimization for all vowels, assuming four formants appear in the range of 0–4 kHz and that each formant is modeled by 2 filter coefficients. After the glottal parameters are estimated, $(t_e^{\star}/\widehat{N}_0 T_s)$ is used as the estimate of the open quotient (OQ). Table 8 shows the mean and standard deviation of the absolute percentage error of the estimated OQ for the three chosen female speakers. For comparison, we also provide the absolute percentage error of the estimated OQ using iterative adaptive inverse filtering (IAIF) proposed by Alku (1992). IAIF is performed using APARAT software (Airas, 2008).

Similar results in the case of three male speakers are shown in Table 9. From Tables 8 and 9, it can be seen that the mean δ_r obtained by the proposed optimization is smaller for most of the cases compared to those obtained by APARAT.

We compute the correlation between jitter values of an utterance and the δ_r of the estimated OQ to analyze if there is any trend in the quality of estimate for different jitter values. Jitters are computed from the GCI estimated from the observed speech segment used for optimization. Average jitter values for each vowel across all speakers (separately

Table 8
Absolute percentage error of the estimated open quotient by the proposed optimization and APARAT (Airas, 2008) for real speech by three female speakers (F03, F14, F26).

Vowel	Mean (SD) δ_r (real speech)					
	Proposed optimization			APARAT		
	F03	F14	F26	F03	F14	F26
/ʌ/	29.62 (10.88)	17.49 (15.35)	22.73 (21.77)	57.40 (76.54)	111.32 (125.93)	58.60 (69.46)
/a/	26.52 (3.95)	23.65 (27.46)	19.33 (10.40)	13.79 (4.38)	76.38 (4.79)	28.15 (3.45)
/æ/	20.12 (12.09)	8.80 (6.62)	11.82 (12.34)	24.97 (8.43)	57.90 (9.16)	27.51 (9.39)
/U/	10.47 (2.77)	1.18 (1.39)	15.28 (3.11)	7.66 (2.69)	50.01 (6.30)	22.49 (9.09)
/I/	13.47 (20.28)	4.00 (5.81)	21.03 (10.81)	3.05 (3.32)	55.71 (20.97)	27.73 (6.26)

Table 9
Absolute percentage error of the estimated open quotient by the proposed optimization and APARAT (Airas, 2008) for real speech by three male speakers (M01, M10, M21).

Vowel	Mean (SD) δ_r (real speech)					
	Proposed optimization			APARAT		
	M01	M10	M21	M01	M10	M21
/ʌ/	24.16 (5.20)	31.39 (17.59)	16.56 (4.18)	146.83 (119.80)	42.07 (93.19)	121.99 (119.99)
/a/	24.35 (5.36)	34.14 (14.13)	42.82 (28.94)	60.09 (70.89)	57.21 (4.62)	46.62 (2.55)
/æ/	60.82 (22.81)	27.38 (12.77)	76.50 (3.09)	88.48 (57.46)	48.99 (7.18)	63.28 (27.90)
/U/	33.77 (28.52)	59.54 (31.39)	17.16 (1.98)	47.48 (11.55)	20.04 (16.22)	54.24 (5.23)
/I/	45.45 (0.90)	56.49 (6.60)	18.72 (21.87)	45.45 (1.21)	27.18 (7.17)	55.22 (12.96)

Table 10
Average jitter values across all speakers for each vowels and the correlation between the jitter values (obtained from the estimated GCI) and δ_r obtained by the proposed optimization separately for male and female subjects.

Vowel	Correlation between jitter and δ_r			
	Female		Male	
	Average jitter (%)	Corr. coef.	Average jitter (%)	Corr. coef.
/ʌ/	9.19	−0.20	2.53	0.026
/a/	1.89	−0.17	1.95	0.00
/æ/	2.74	−0.30	0.95	−0.25
/U/	5.69	−0.22	3.04	−0.35
/I/	3.31	−0.04	1.97	0.13

for male and female) are shown in Table 10. Also the correlation coefficients in those respective cases are reported. If the proposed optimization performed poorly in the case of high jitter values, it would result in high positive correlation coefficients. However, most of the correlation coefficients are small, and hence, the values indicate that there is no significant trend between jitter and δ_r obtained by the proposed optimization.

However, as mentioned in (Fu et al., 2006), there could be measurement errors in EGG. For example, there are various factors that influence the quality of the EGG signal such as electrode placement, skin-electrode resistance, poor conductivity due to fat tissue, and head movement (Baken and Orlikoff, 1999) and hence, the OQ_{EGG} might not reflect the true value of OQ. In that respect, the performance of the proposed optimization should be judged based on results obtained for the synthetic speech.

5. Conclusions

We proposed a joint source-filter optimization for glottal source estimation which is robust to shimmer and jitter. The proposed optimization is based on the source-filter theory of speech production, wherein the LF model is used for the glottal source and an all-pole model is used for the vocal-tract filter; being a joint optimization, it captures the source-filter interaction under the considered cost function for estimating source and vocal-tract filter parameters. A key feature of the proposed optimization is that the variability in the speech due to variation in the source signal, such as due to shimmer and jitter, is considered within the optimization framework. The variation in the amplitude of the glottal flow derivative (shimmer) is estimated by properly designing the optimization cost function; the variation in the pitch period from one pitch cycle to the next (jitter) is handled by selecting a signal segment of appropriate length so that the effect of jitter is minimal.

Experimental evaluation of the proposed optimization method using both synthetic and natural speech for different vowels and voice types indicates its robustness, particularly to shimmer and jitter. For example, the absolute percentage errors in the case of the Modal and Vocal Fry speech are similar for most of the glottal parameters although Modal speech suffers from 2% jitter while Vocal Fry speech suffers from 10% jitter. Similarly, under a maximum shimmer of 50% for synthetic speech, the average absolute percentage error in estimated amplitudes of glottal flow derivative is about 20% for various voice types and vowels.

In the proposed optimization framework, the time-parameters of the LF model are estimated using a combinatorial search over a feasible range of values. One limitation of such a grid search is that it cannot estimate the actual real valued glottal timing parameters exactly. Although the accuracy obtained by such a grid search approach could be useful depending on the application under consideration, a better optimization problem needs to be designed where the optimization variables corresponding to the glottal timing parameters are real valued. However, the trade-off between the complexity of the optimization and the accuracy of the estimates need to be addressed.

Acknowledgements

This research work was supported by the National Science Foundation (NSF) and the US Army.

References

Airas, M., 2008. An environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocology 33, 49–64.

Airas, M., Alku, P., 2006. Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalized amplitude quotient. Phonetica 63, 26–46.

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Comm. 11 (June), 109–118.

Baken, R.J., Orlikoff, R.L., 1999. Clinical Measurement of Speech and Voice, second ed. Singular.

Carre, R., 1981. Vocal source-vocal tract coupling. Effects on the vowel spectrum. In: IVth FASE Symposium, April 1981, Venice .

Childers, D.G., 2000. Speech Processing and Synthesis Toolboxes. Wiley, New York.

Ding, W., Kasuya, H., Adachi, S., 1995. Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. IEICE Trans. Inf. Systems E78-D (6), 738–743.

Drugman, T., Dutoit, T., 2009. Glottal closure and opening instant detection from speech signals. In: Proc. Interspeech, Brighton, UK, September 2009, pp. 2891–2894.

Fant, G., Lin, Q., 1985. A four-parameter model of glottal flow. In: STL-QPSR 4/85, R. Inst. Technol. (KTH), Stockholm, Sweden.

Flanagan, J.L., 1965. Speech Analysis Synthesis and Perception. Academic Press Inc., Publishers, New York.

Frohlich, M., Michaelis, D., Strube, H.W., 2001. SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. J. Acoust. Soc. Amer. 110 (1), 479–488.

Fu, Q., Murphy, P., 2006. Robust glottal source estimation based on joint source-filter model optimization. IEEE Trans. Audio Speech Lang. Process. 14 (2), 492–501.

Hall, M.G., Oppenheim, A.V., Willsky, A.S., 1983. Time-varying parametric modeling of speech. Signal Process. 5 (3), 267–285.

Hess, W., 1983. Pitch Determination of Speech Signals – Algorithms and Devices. Springer-Verlag, Berlin, Germany.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Amer. 87 (February), 820–856.

Krishnamurthy, A.K., 1992. Glottal source estimation using a sum-of-exponentials model. IEEE Trans. Signal Process. 40 (3), 682–686.

Krishnamurthy, A.K., Childers, D.G., 1986. Two-channel speech analysis. IEEE Trans. Acoust. Speech Signal Process. ASSP-34 (4), 730–743.

Markel, J.D., Gray, A.H., 1976. Linear Prediction of Speech. Springer-Verlag, Berlin.

Miller, R.L., 1959. Nature of the vocal cord wave. J. Acoust. Soc. Amer. 31, 667–677.

Moore, E., Clements, M., Peifer, J., Weisser, L., 2003. Investigating the role of glottal features in classifying clinical depression. In: Proc. 25th Annual IEEE Internat. Conf., September 2003, Vol. 3, pp. 2849–2852.

Plumpe, M.D., Quatieri, T.F., Reynolds, D.A., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process. 7 (5), 569–586.

Quatieri, T.F., 2001. Discrete-time Speech Signal Processing: Principles and Practice, first ed. Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L., Schafer, R., 2010. Theory and Applications of Digital Speech Processing. Prentice-Hall.

Rosenberg, A.E., 1971. Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Amer. 49, 583–590.

Saratxaga, I., Erro, D., Hernáez, I., Sainz, I., Navas, E., 2009. Use of harmonic phase information for polarity detection in speech signals. In: Proc. Interspeech, Brighton, UK, pp. 1075–1078.

Shimamura, T., Kobayashi, H., 2001. Weighted autocorrelation for pitch extraction of noisy speech. IEEE Trans. Speech Audio Process. 8 (7), 727–730.

Strik, H., 1998. Automatic parameterization of differentiated glottal flow: comparing methods by means of synthetic flow pulses. J. Acoust. Soc. Amer. 103 (5), 2659–2669.

Veldhuis, R., 1998. A computationally efficient alternative for the Liljencrant–Fant model and its perceptual evaluation. J. Acoust. Soc. Amer. 103 (1), 566–571.

Wong, D.Y., Markel Jr, J.D., Gray, A.H., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Acoust. Speech Signal Proc. ASSP-27 (4), 350–355.

Yoshiyuki, H., 1982. Jitter and shimmer differences among sustained vowel phonations. J. Speech Hearing Res. 25 (March), 12–14.