Modeling and joint estimation of glottal source and vocal tract filter by state-space methods[☆]

Gabriel A. Alzamendi[a,b,c,*], Gastón Schlotthauer[a,b,c]

[a] *Lab. de Señales y Dinámicas no Lineales, Fac. de Ingeniería, Universidad Nacional de Entre Ríos, Argentina*

[b] *Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina*

[c] *Centro de Investigaciones y Transferencia de Entre Ríos (CITER), Argentina*

Abstract

Accurate estimation of the glottal source from a voiced sound is a difficult blind separation problem in speech signal processing. In this work, state-space methods are investigated to enhance the joint estimation of the glottal source and the vocal tract information. The aim of this paper is twofold. First, a stochastic glottal source is proposed, based on deterministic Liljencrants-Fant model and ruled by a stochastic difference equation. Such a representation allows to accurately capture any perturbation occurring at glottal level in real voices. A state-space voice model is formulated considering the stochastic glottal source. Then, combining this voice model and the state-space framework, an inverse filtering method is developed that allows to jointly estimate both glottal source and vocal tract filter. The performance of this method is studied by means of experiments with voices synthesized by applying both the source-filter theory and a physical based voice model. The method is also test using human voice signals. The results demonstrate that accurate estimates of the glottal source and the vocal tract filter can be obtained over several scenarios. Moreover, the method is shown to be robust with respect to different phonation types.

Keywords: Stochastic glottal source, state-space voice model, glottal inverse filtering, joint source-filter estimation

1. Introduction

Glottal inverse filtering consists of the decomposition of a speech waveform into glottal source and vocal tract components [5, 33]. It has become a challenging task in digital speech signal processing since inverse filtering involves a difficult blind separation problem where neither the glottal source nor the vocal tract are known. This non-invasive method has proved to be useful for various purposes, including voice production research, speech coding and analysis,

[☆] Approved manuscript for publication. Caution, it may differ from the final published version.

[*] Corresponding author. Tel: +54-0343-4975100 ext 122. Postal address: Facultad de Ingeniería - UNER, C.C. 47 - Suc. 3 - 3100 Paraná, Entre Ríos - Argentina.

Email addresses: galzamendi@bioingenieria.edu.ar (Gabriel A. Alzamendi), gschlotthauer@conicet.gov.ar (Gastón Schlotthauer) natural speech synthesis, expressive or emotional speech processing and speaker recognition/verification. In biomedical science in particular, inverse filtering has demonstrated to be potentially helpful in applications such as voice disorder detection/diagnose, occupational voice care, pathological voice restoration and clinical depression assessment, among others. A thorough review of inverse filtering and its applications should include [5, 16, 27, 44, 45] and references therein.

Different inverse filtering methods have been developed in accordance with the source-filter theory. Most of them involve the calculation of the vocal tract filter (VTF) and the estimation of the glottal source by deconvolving the speech signal in order to cancel the vocal tract effects. In earlier approaches, tuning of VTF was performed manually by experts. Later, the arrival of *Linear Prediction* and its related methods has given rise to automatic estimation of VTF [1, 3, 18, 32]. An automatic method widely applied in the practice is the so-called *Iterative Adaptive Inverse Filtering* (IAIF) [4, 6]. On the other hand, *joint source-filter optimization* methods have been developed recently, where voice decomposition is achieved by solving the inverse problem of voice production [9, 10, 21, 22, 38]. In the context of inverse problems, a proper model formulation is crucial to guarantee feasible and accurate solutions. Therefore, a flexible voice generation model is mandatory for voice decomposition.

In joint source-filter optimization, both the vocal tract and the glottal source should be explicitly modeled. Although the vocal tract is usually modeled by means of autoregressive filters, time-varying alternatives have recently received more attention because they guarantee a more flexible representation of the vocal tract dynamics. In the glottal source, the harmonic (quasi-periodic) component can be described applying a deterministic model or a combination of predefined basis functions [2, 5, 10]. Deterministic models of glottal source are extensively described in speech literature (e.g., KLGLOTT88, R, LF, FL, R++, EE1 and EE2) [14, 16, 20]. Nevertheless, these glottal models possess two main limitations: *(i)* due to their deterministic formulation, they do not represent the non-modeled features or the perturbations occurring at glottal level in real voices [15, 39], and *(ii)* capturing the harmonic component from a (inverse filtered) glottal waveform generally requires a least-square fitting of non-linear analytical functions [16, 19]. In order to tackle these limitations, we introduce a stochastic linear differential equation for the accurate and flexible representation of the glottal source.

State-space methods allow for the model-guided processing of non-stationary stochastic signals. Their most important characteristics are the following [11, 17]: *(i)* model formulation is straightforward, *(ii)* meaningful statistics (also called estimates) of unobserved processes can be computed analytically, *(iii)* uncertainties and errors are considered in the formulation of state-space models, and *(iv)* algorithms are available for computing the optimal values of model parameters. Given that speech signals are characterized by a non-stationary and stochastic behavior, state-space framework would become specially suitable for joint source-filter optimization methods.

The goal of the present contribution is to investigate the application of state-space methods to the stochastic modeling of voice production and to the joint source-filter optimization. Unlike earlier contributions (e.g., [9, 21, 22, 38]), we have assumed that the glottal source is a non-stationary stochastic phenomenon taking place during phonation. Then, we benefit from this hypothesis in order to improve the accuracy in the estimation of the glottal source and the vocal tract filter. In particular, the aim of this paper is twofold. Firstly, we introduce a time-varying stochastic difference equation for modeling the glottal source. Secondly, we propose a joint source-filter optimization method based on a Gaussian state-space voice model. This method is investigated by means of experiments with voices synthesized by applying both the source-filter theory and a physical based voice model. Finally, for illustrative purposes, we apply this method to a real voice signal.

This paper is structured as follows: In Sec. 2, the stochastic glottal source model and the Gaussian state-space voice model are developed. In Sec. 3, state-space methods are introduced and optimal estimation of voice model parameters is described. In Sec. 4, the voice material and the experimental setup utilized in this work are described. In Sec. 5, the results achieved are exposed and analyzed. Finally, in Sec. 6 the conclusions are presented.

2. Glottal source and voice models

In this paper, only voiced sounds are considered. According to the source-filter theory, in its simplest form, voice production can be described as $\mathsf{S}(z) = \mathsf{V}_g(z)\,\mathsf{G}(z)$, where $\mathsf{S}(z)$ and $\mathsf{V}_g(z)$ are the z-transforms of speech signal s and glottal source v_g, respectively, and $\mathsf{G}(z)$ is VTF transfer functions [12, 37]. Hereafter, v_g represents the derivative of the glottal flow U_g (a.k.a. glottal volume velocity) [2, 14, 16]. In this section we formulate a stochastic model of glottal source v_g, and then we apply it for developing a Gaussian state-space model of voice production.

2.1. Stochastic glottal source (SGS) model

The LF function, proposed by Liljencrants and Fant in [20], is one of the most popular parametric representations of the glottal source v_g. It provides a good fit to waveforms commonly encountered in applications involving glottal inverse filtering [5, 14, 16]. According to it, a glottal source pulse is analytically modeled in time-domain as follows:

$$v_g^{\mathrm{LF}}[n] = \begin{cases} E_0\, e^{\alpha\, n}\, \sin(\omega_g\, n), & 0 \le n \le N_e, \\ \frac{-E_e}{\epsilon\, N_a}\left(e^{-\epsilon\,(n-N_e)} - e^{-\epsilon\,(N_c-N_e)}\right), & N_e < n \le N_c, \\ 0, & N_c < n < N_0, \end{cases} \tag{1}$$

where $\{E_0,\, \alpha,\, \omega_g,\, \epsilon\}$ and $\{E_e, N_p,\, N_e,\, N_a,\, N_0\}$ are called the direct synthesis and the timing parameters, respectively [20]. Here, N_0 is the fundamental period and $f_0 = f_s/N_0$ is the fundamental frequency, with f_s the sampling frequency. The two set of parameters are related by the constrains:

$$\begin{cases} \sum_{n=0}^{N_0-1} v_g^{\mathrm{LF}}[n] = 0, \\ \omega_g = \frac{\pi}{N_p}, \\ \epsilon\, N_a = 1 - e^{-\epsilon\,(N_c-N_e)}, \\ E_e = -E_0\, e^{\alpha\, N_e}\, \sin(\omega_g\, N_e). \end{cases} \tag{2}$$

As an example, one cycle of v_g^{LF} is shown in Fig. 1. In particular, E_e is the absolute value of the minimum located at $n = N_e$ (see the dashed vertical line

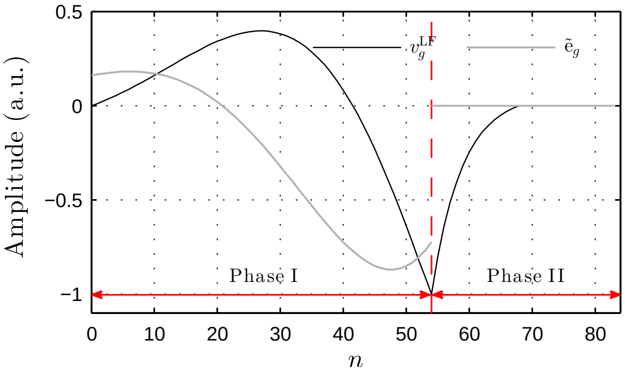

Figure 1: LF glottal source, v_g^{LF}, and the corresponding auxiliary input signal, $\tilde{\mathrm{e}}_g$. Moreover, Phase I and Phase II are indicated by the double arrows.

in the figure). The starting point of every pulse constitutes the *glottal opening instant*. Furthermore, the location of the minimum point in every cycle, where the maximum excitation occurs, is the *glottal closure instant*. Hereafter, Phase I (Phase II) refers to the timespan from a glottal opening (closure) instant to the next closure (opening) instant. In the example in Fig. 1, Phase I and Phase II are also shown.

From the v_g^{LF} definition, Eq. (1), we develop a linear time-varying stochastic difference equation for modeling the glottal source. First row can be written as:

$$\begin{aligned} v_g^{\mathrm{LF}}[n] &= E_0\, e^{\alpha\, n}\, \sin\left(\omega_g\, n\right) \\ &= E_0\, e^{\alpha}\, e^{\alpha\,(n-1)}\, \sin\left(\omega_g\left[(n-1)+1\right]\right) \\ &= e^{\alpha}\cos\left(\omega_g\right)\left[E_0\, e^{\alpha\,(n-1)}\sin\left(\omega_g\,(n-1)\right)\right] \\ &\quad + e^{\alpha}\sin\left(\omega_g\right)\left[E_0\, e^{\alpha\,(n-1)}\cos\left(\omega_g\,(n-1)\right)\right]. \end{aligned} \tag{3}$$

Similarly, the second row in (1) can be expressed as:

$$\begin{aligned} v_g^{\mathrm{LF}}[n] &= -\frac{E_e}{\epsilon N_a}\left(e^{-\epsilon(n-N_e)} - e^{-\epsilon(N_c-N_e)}\right) \\ &= -\frac{E_e\, e^{-\epsilon}}{\epsilon\, N_a}\left(e^{-\epsilon(n-1-N_e)} - e^{-\epsilon\,(N_c-1-N_e)}\right) \\ &\approx e^{-\epsilon}\left[-\frac{E_e}{\epsilon\, N_a}\left(e^{-\epsilon(n-1-N_e)} - e^{-\epsilon(N_c-N_e)}\right)\right]. \end{aligned} \tag{4}$$

In the last expression, it is assumed that $N_c - N_e \gg 1$ and therefore $e^{-\epsilon\,(N_c-N_e)} \approx e^{-\epsilon\,(N_c-1-N_e)} \approx 0$.

Combining the two previous results and assuming that the glottal source behaves as a stochastic process, we formulate the stochastic glottal source (SGS) model:

$$v_g^{\mathrm{SGS}}[n+1] = \begin{cases} A_g\, v_g^{\mathrm{SGS}}[n] + B_g\, \tilde{\mathrm{e}}_g[n] + \zeta[n], & 0 \le n \le N_e, \\ C_g\, v_g^{\mathrm{SGS}}[n] + \zeta[n], & N_e < n \le N_0, \end{cases} \tag{5}$$

where $A_g = e^{\alpha}\, \cos\left(\omega_g\right)$, $B_g = e^{\alpha}\, \sin\left(\omega_g\right)$ and $C_g = e^{-\epsilon}$. It is also considered the auxiliary input signal:

$$\tilde{\mathrm{e}}_g[n] = \begin{cases} E_0\, e^{\alpha\, n} \cos\big(\omega_g\, n\big), & 0 \leq n < N_e, \\ 0, & N_e \leq n < N_0, \end{cases} \tag{6}$$

and the Gaussian process $\zeta[n] \sim \mathcal{N}(0, \sigma_\zeta^2)$. Notice that the definition (5) requires only two expressions because, assuming $v_g^{\mathrm{SGS}}[n] \approx 0$ for $N_c < n < N_0$ and $\sigma_\zeta^2 \to 0$, the SGS model is in accordance with the third row in (1). It can also be observed that SGS model does not incorporate the constrains (2). Thus, in order to yield a physically significant glottal pulse, in Sec. 3.4 a penalization is considered on the estimation of SGS model parameters.

The SGS model, defined in (5), possesses four important benefits: (*i*) it is defined in terms of a linear time-varying stochastic difference equation, assuming $\tilde{\mathrm{e}}_g[n]$ is known; (*ii*) it can be studied under the state-space framework; (*iii*) the glottal source waveform is determined by parameters A_g, B_g and C_g, along with the glottal opening and closure instants; and (*iv*) any error or misspecification in the formulation is captured by ζ. Notice that the LF direct synthesis parameters, with the exception of E_0, and the parameters A_g, B_g and C_g are related by:

$$\begin{aligned} \alpha &= \frac{1}{2}\,\ln\big(A_g^2 + B_g^2\big), \\ \omega_g &= \arctan\left(\frac{B_g}{A_g}\right), \\ \epsilon &= -\ln\big(C_g\big). \end{aligned} \tag{7}$$

2.2. State-space voice (SSV) model

Let $s[n]$ for $n = 1, 2, \ldots, N$ be a voiced sound signal, with N the number of samples. Assuming that the opening and closure instants for every glottal pulse in s are known in advance, then Phase I and Phase II are also determined. Denote by $\mathcal{I}_N = \{1, 2, \ldots, N\}$ the time index set corresponding to the voice data $\mathcal{S}_N = \{s[1], s[2], \ldots, s[N]\}$. Then, $\mathcal{I}_{\mathrm{I}}$ and $\mathcal{I}_{\mathrm{II}}$ constitute the time index sets corresponding to Phase I and Phase II, respectively, satisfying the following conditions: $\mathcal{I}_{\mathrm{I}} \cup \mathcal{I}_{\mathrm{II}} = \mathcal{I}_N$, $\mathcal{I}_{\mathrm{I}} \cap \mathcal{I}_{\mathrm{II}} = \emptyset$ and $N = \#\mathcal{I}_{\mathrm{I}} + \#\mathcal{I}_{\mathrm{II}}$, with # the cardinality of a set.

In accordance with the source-filter theory, it is assumed that the voiced sounds are produced by the modulation of the glottal source with the VTF. Here, vocal tract behavior is represented using a time-varying autoregressive filter with exogenous input. Then, the voice signal $s[n]$ can be described in time-domain as follows:

$$s[n] = -\sum_{l=1}^{\rho} a_l[n]\, s[n-l] + G_g\, v_g[n] + \mathrm{v}[n], \tag{8}$$

where ρ is the model order, $a_l[n]$ for $l = 1, 2, \ldots, \rho$ are the time-varying filter coefficients, G_g is a gain term and $\mathrm{v}[n] \sim \mathcal{N}(0, \sigma_{\mathrm{v}}^2)$. In (8), minus sign is introduced only for algebraic convenience.

It is known that, for voiced sounds, the formants and their bandwidths remain approximately constant or show small variations [12, 37, 43]. Thus, it is assumed that VTF coefficients can be modeled as ρ stochastic time series [24, 34]:

$$a_l[n+1] = a_l[n] + \xi_l[n], \qquad l = 1, 2, \ldots, \rho, \tag{9}$$

where ξ_l are jointly Gaussian random processes.

Considering Eq. (8) and taking into account previous hypotheses, next we introduce the linear Gaussian state-space voice (SSV) model. In order to proceed, it is assumed that the voice production can be represented by the (latent) state vector $\mathbf{x}[n] \in \mathbb{R}^p$, with $p = \rho + 1$, given by:

$$\begin{aligned} \mathbf{x}[n] &= \begin{pmatrix} x_{(1)}[n] & x_{(2)}[n] & \ldots & x_{(p-1)}[n] & x_{(p)}[n] \end{pmatrix}^T \\ &= \begin{pmatrix} a_1[n] & a_2[n] & \ldots & a_\rho[n] & v_g[n] \end{pmatrix}^T. \end{aligned} \tag{10}$$

Although similar ideas have been suggested by other authors [21, 22, 38], SSV model differs from those approaches in two aspects: the glottal source is stochastically modeled using the SGS model and, moreover, it is an element of the latent state vector $\mathbf{x}$.

At this point, we can introduce the state and observation equations describing voiced sound signals. For $n \in \mathcal{I}_\mathrm{I}$, the state transition is driven by:

$$\mathbf{x}[n+1] = \mathbf{A}_\mathrm{I}\,\mathbf{x}[n] + \mathbf{B}_\mathrm{I}\,\tilde{\mathrm{e}}_g[n] + \mathbf{w}[n], \tag{11}$$

where $\mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, and matrix $\mathbf{A}_\mathrm{I} \in \mathbb{R}^{p\times p}$ and vector $\mathbf{B}_\mathrm{I} \in \mathbb{R}^p$ are:

$$\mathbf{A}_\mathrm{I} = \begin{pmatrix} \mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & A_g \end{pmatrix} \quad \text{and} \quad \mathbf{B}_\mathrm{I} = \begin{pmatrix} \mathbf{0} \\ B_g \end{pmatrix}. \tag{12}$$

Furthermore, the covariance $\mathbf{Q} \in \mathbb{R}^{q\times q}$ is the positive-definite symmetric matrix, given by:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_\xi & \mathbf{0} \\ \mathbf{0} & \sigma_\zeta^2 \end{pmatrix}, \tag{13}$$

where $\mathbf{Q}_\xi \in \mathbb{R}^{\rho\times\rho}$ is the joint covariance of stochastic processes ξ_l, Eq. (9), and σ_ζ^2 is the variance of ζ, Eq. (5). Thus, matrix $\mathbf{Q}_\xi$ models the stochastic correlation between the VTF filter coefficients.

For $n \in \mathcal{I}_\mathrm{II}$, the state transition is driven by:

$$\mathbf{x}[n+1] = \mathbf{A}_\mathrm{II}\,\mathbf{x}[n] + \mathbf{B}_\mathrm{II}\,\tilde{\mathrm{e}}_g[n] + \mathbf{w}[n], \tag{14}$$

where $\mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, and matrix $\mathbf{A}_\mathrm{II} \in \mathbb{R}^{p\times p}$ and vector $\mathbf{B}_\mathrm{II} \in \mathbb{R}^p$ are:

$$\mathbf{A}_\mathrm{II} = \begin{pmatrix} \mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & C_g \end{pmatrix} \quad \text{and} \quad \mathbf{B}_\mathrm{II} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}. \tag{15}$$

Matrix $\mathbf{Q}$ fulfills the structure and the requirements described above. It is important to remember that $\tilde{\mathrm{e}}_g[n] = 0$ for $n \in \mathcal{I}_\mathrm{II}$, according to (6).

Considering Eq. (8) for voice generation, along with Eqs. (11) and (14), the observation equation is obtained. For $n \in \mathcal{I}_N$, it is formulated as:

$$s[n] = \mathbf{H}[n]\,\mathbf{x}[n] + \mathrm{v}[n], \tag{16}$$

with row vector $\mathbf{H}[n] \in \mathbb{R}^p$:

$$\mathbf{H}[n] = \begin{pmatrix} -s[n-1] & -s[n-2] & \cdots & -s[n-\rho] & G_g \end{pmatrix}. \tag{17}$$

In summary, the SSV model is constituted by Eqs. (11), (14) and (16). Due to its stochastic formulation, this model is able to represent the effects of non-modeled phenomena or the perturbations occurring during phonation.

3. State-space methods

In this section, state-space methods applied in this work are briefly described. Taking into account the SSV model introduced in Sec. 2, only methods for linear Gaussian state-space models are here considered.

3.1. State filtering

State filtering, also called Kalman filtering, is an iterative forward procedure for conveniently recovering latent states, conditional on past and present data. It consists in computing the filtered states $\hat{\mathbf{x}}[n|n] = \mathcal{E}\left\{\mathbf{x}[n] \,|\, \mathcal{S}_n\right\}$, and its covariance matrix $\mathbf{P}[n|n] = \mathcal{E}\left\{\mathbf{x}[n]\,\mathbf{x}[n]^T \,|\, \mathcal{S}_n\right\}$, for each $n = 1, 2, \ldots, N$ [11]. This is an optimal procedure, in the sense of minimizing the mean square estimation error [8]. Due to its simplicity and robustness, it becomes a useful method in real-time applications. In this work, the so-called *contemporaneous Kalman filter* is applied (see e.g., [11, 17]). For initialization, it is assumed $\mathbf{x}[0] \sim \mathcal{N}(\hat{\mathbf{x}}_0, \hat{\mathbf{P}}_0)$, where $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{P}}_0$ are known.

3.2. State smoothing

State smoothing considers all information available in the data and, therefore, generates more accurate estimations of latent states [8]. It involves the computation of the smoothed state $\hat{\mathbf{x}}[n|N] = \mathcal{E}\left\{\mathbf{x}[n] \,|\, \mathcal{S}_N\right\}$, and its covariance matrix $\mathbf{P}[n|N] = \mathcal{E}\left\{\mathbf{x}[n]\,\mathbf{x}[n]^T \,|\, \mathcal{S}_N\right\}$, for each $n = 1, 2, \ldots, N$. In this work, the so-called *fixed interval Kalman smoothing* is used, consisting on a two-steps procedure. In a first step, the forward Kalman filter is applied to the data, and next, in a second step, the backward Kalman smoother improves the estimations. This is a non-causal method making use of future information to improve the estimations, and is suitable for processing stored signals or in delay-tolerant real-time applications. Please refer to [17, 26], for further information about state smoothing.

3.3. Different smoothed estimates

Other smoothed estimates can be obtained from a given time series conditional on a state-space model. They are relevant in practice because carry information about the phenomenon under investigation, are useful for goodness-of-fit evaluation, and play a critical role in the computation of model parameters [13, 25, 30]. Here, smoothed estimates considered in this article are briefly introduced.

First, initial state smoothing is explained. It consists in computing the smoothed initial state $\hat{\mathbf{x}}[0|N] = \mathcal{E}\left\{\mathbf{x}[0] \,\middle|\, \mathcal{S}_N\right\}$ and its covariance matrix $\hat{\mathbf{P}}[0|N]$ conditional on the full data and the state-space model. These parameters are computed by means of the Kalman smoother [8, 13].

The disturbance smoothing involves the estimation of the smoothed state disturbances $\hat{\mathbf{w}}[n-1|N] = \mathcal{E}\left\{\mathbf{w}[n-1]\middle|\, \mathcal{S}_N\right\}$ and the observation disturbances $\hat{\mathrm{v}}[n|N] = \mathcal{E}\left\{\mathrm{v}[n]\middle|\, \mathcal{S}_N\right\}$, along with their covariances $\hat{\mathbf{P}}_{\mathrm{w}}[n-1|N]$ and $\hat{P}_{\mathrm{v}}[n|N]$, for $n = 1, 2, \ldots, N$, respectively. This method also requires a two-steps procedure, where the forward Kalman filter is applied, followed by the backward disturbance smoothing recursion. For further information regarding disturbance smoothing, see [17, 28].

Taking into account the smoothed states, including the initial state, the smoothed state autocorrelation matrix is defined for $n = 0, 1, \ldots, N$:

$$\hat{\mathbf{C}}[n|N] = \hat{\mathbf{P}}[n|N] + \hat{\mathbf{x}}[n|N]\,\hat{\mathbf{x}}[n|N]^T. \tag{18}$$

Similarly, smoothed disturbance autocorrelation matrices $\hat{C}_{\mathrm{v}}[n|N]$ and $\hat{\mathbf{C}}_{\mathrm{w}}[n|N]$ of observation v and state $\mathbf{w}$ disturbances, respectively, can also be computed.

Finally, smoothed one-step ahead state cross-correlation matrix is considered, consisting on computing $\hat{\mathbf{C}}_{n-1,n}[n|N] = \mathcal{E}\left\{\mathbf{x}[n-1]\,\mathbf{x}[n]^T \,|\, \mathcal{S}_N\right\}$, for $n = 1, 2, \ldots, N$. Notice that $\hat{\mathbf{C}}_{n,n-1}[1|N] = \hat{\mathbf{C}}_{n-1,n}[1|N]^T$. For more information on it and others smoothed correlation estimates relevant to time series analysis, refer to [17, Sec. 4.5].

3.4. Estimation of SSV model parameters

The proposed SSV model depends on a set of parameters $\mathbf{\Theta}$ which in general are unknown and, therefore, must be calculated from the voice data $\mathcal{S}_N$. In particular, the extensional definition of this set is $\mathbf{\Theta} = \{\sigma_{\mathrm{v}}^2, \mathring{\mathbf{Q}}, \hat{\mathbf{x}}_0, \hat{\mathbf{P}}_0, A_g, B_g, C_g, G_g\}$, assuming that $\mathbf{Q} = \sigma_{\mathrm{v}}^2\,\mathring{\mathbf{Q}}$. In the SSV model, glottal source amplitude depends on the product $G_g\,E_0$. Fixing E_0, then G_g can be accordingly computed. Hereafter, $E_0 = 0.001$.

There are several techniques based on state-space methods for computing $\mathbf{\Theta}$, relying on the formulation and the solution of an optimization problem. Here, the iterative *Expectation-Maximization* (EM) algorithm is applied for solving the penalized optimization problem formulated as [13, 29]:

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}\in\mathcal{D}}{\arg\max}\;\; \mathcal{E}\left\{\ln\, \mathcal{L}(\mathbf{\Theta}|\mathcal{S}_N)\right\} - \lambda\,\Phi(\mathbf{\Theta}), \tag{19}$$

where $\mathcal{D}$ is the domain of definition of $\mathbf{\Theta}$, $\mathcal{L}$ is the *log-likelihood function*, Φ is a non-negative penalization function and λ is a penalization factor.

Different penalization functions can be chosen, depending on which aspect of the optimization procedure must be controlled. Here, Φ is given by:

$$\Phi(\mathbf{\Theta}) = \frac{1}{2}\left[\left(A_g - \tilde{A}_g\right)^2 + \left(B_g - \tilde{B}_g\right)^2 + \left(C_g - \tilde{C}_g\right)^2\right], \tag{20}$$

where $\tilde{A}_g$, $\tilde{B}_g$ and $\tilde{C}_g$ are known. As can be inferred, the effect of this penalization is to stabilize the estimation of SGS model parameters, see Eq. (5). This is mandatory because, as stated in Sec. 2.1, SGS model approximates the non-linear LF model and, as a consequence, unreliable parameters may be obtained giving rise to unsuitable glottal sources and turning the optimization procedure unstable. If $\tilde{A}_g$, $\tilde{B}_g$ and $\tilde{C}_g$ ensure an acceptable glottal waveform, then the penalization term favors physically significant solutions.

In order to solve the optimization problem (19) by using the EM method, rules for iteratively computing the parameters in $\mathbf{\Theta}$ were required. For computing σ_{v}^2 and $\mathring{\mathbf{Q}}$ the equations introduced in [17, 30] were considered. Parameters $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{P}}_0$ were obtained by means of initial state smoothing (Sec. 3.3).

Next, the rules for computing glottal source parameters A_g, B_g, C_g and G_g are described. The proofs can be found in the supplementary material complementing this article. In first place, parameters A_g and B_g were obtained as follows:

$$\begin{aligned} \hat{A}_g^{\mathrm{opt}} &= \frac{\gamma_1\,\theta_{22} - \gamma_2\,\theta_{12}}{\theta_{11}\,\theta_{22} - \theta_{12}\,\theta_{21}}, \\ \hat{B}_g^{\mathrm{opt}} &= \frac{\gamma_2\,\theta_{11} - \gamma_1\,\theta_{21}}{\theta_{11}\,\theta_{22} - \theta_{12}\,\theta_{21}}. \end{aligned} \tag{21}$$

The variables involved in the previous equations are given by:

$$\begin{aligned} \theta_{11} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{\mathrm{I}}_{n-1})_{(p,p)} + \lambda\,\sigma_{\mathrm{v}}^2, \\ \theta_{12} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{\mathrm{I}}_{\tilde{\mathrm{e}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(p)}, \\ \theta_{21} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{\mathrm{I}}_{\tilde{\mathrm{e}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(p)}, \\ \theta_{22} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,\tilde{\mathrm{e}}^{\mathrm{I}}_{n-1} + \lambda\,\sigma_{\mathrm{v}}^2, \\ \gamma_1 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{\mathrm{I}}_{n,n-1})_{(p,p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{\mathrm{I}}_{n-1})_{(i,p)} + \lambda\,\sigma_{\mathrm{v}}^2\,\tilde{A}_g, \\ \gamma_2 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{\mathrm{I}}_{\tilde{\mathrm{e}}_{n-1}\,\hat{\mathrm{x}}_{n}})_{(p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{\mathrm{I}}_{\tilde{\mathrm{e}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(i)} + \lambda\,\sigma_{\mathrm{v}}^2\,\tilde{B}_g, \end{aligned} \tag{22}$$

with $(\cdot)_{(i)}$ the i-th element of a vector and, similarly, $(\cdot)_{(i,j)}$ the i,j-th entry in a matrix. These results were obtained assuming that both $\sigma_{\mathrm{v}}^2 > 0$ and $\mathring{\mathbf{Q}}^{-1}$ are known. In practice, the most recent values computed during the EM method are chosen.

Previous equations depend on the parameters:

$$\begin{aligned} \tilde{\mathbf{C}}^{\mathrm{I}}_{n,n-1} &= \sum_{n\in\mathcal{I}_{\mathrm{I}}} \hat{\mathbf{C}}_{n,n-1}[n|N], \\ \tilde{\mathbf{C}}^{\mathrm{I}}_{n-1} &= \sum_{n\in\mathcal{I}_{\mathrm{I}}} \hat{\mathbf{C}}[n-1|N], \\ \tilde{\mathbf{C}}^{\mathrm{I}}_{\tilde{\mathrm{e}}_{n-1}\,\hat{\mathrm{x}}_{n-1}} &= \sum_{n\in\mathcal{I}_{\mathrm{I}}} \tilde{\mathrm{e}}_g[n-1]\,\hat{\mathbf{x}}[n-1|N] \\ \tilde{\mathbf{C}}^{\mathrm{I}}_{\tilde{\mathrm{e}}_{n-1}\,\hat{\mathrm{x}}_{n}} &= \sum_{n\in\mathcal{I}_{\mathrm{I}}} \tilde{\mathrm{e}}_g[n-1]\,\hat{\mathbf{x}}[n|N], \\ \tilde{\mathrm{e}}^{\mathrm{I}}_{n-1} &= \sum_{n\in\mathcal{I}_{\mathrm{I}}} \tilde{\mathrm{e}}_g[n-1]^2. \end{aligned} \tag{23}$$

In the case of C_g, the following rule was obtained:

$$\hat{C}_g^{\mathrm{opt}} = \frac{\gamma_3}{\theta_3}, \tag{24}$$

where

$$\begin{aligned}\theta_3 &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{\mathrm{II}}_{n-1})_{(p,p)} + \lambda\,\sigma_{\mathrm{v}}^2,\\ \gamma_3 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{\mathrm{II}}_{n,n-1})_{(p,p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{\mathrm{II}}_{n-1})_{(i,p)} + \lambda\,\sigma_{\mathrm{v}}^2\,\tilde{C}_g.\end{aligned} \tag{25}$$

As in the previous case, it is assumed that both $\sigma_{\mathrm{v}}^2 > 0$ and $\mathring{\mathbf{Q}}^{-1}$ are known. Furthermore, the remaining parameters are:

$$\begin{aligned}\tilde{\mathbf{C}}^{\mathrm{II}}_{n,n-1} &= \sum_{n\in\mathcal{I}_{\mathrm{II}}} \hat{\mathbf{C}}_{n,n-1}[n|N],\\ \tilde{\mathbf{C}}^{\mathrm{II}}_{n-1} &= \sum_{n\in\mathcal{I}_{\mathrm{II}}} \hat{\mathbf{C}}[n-1|N].\end{aligned} \tag{26}$$

Finally, the gain G_g is calculated by:

$$\hat{G}_g^{\mathrm{opt}} = \frac{\mu}{\eta}, \tag{27}$$

where

$$\begin{aligned}\mu &= \sum_{n\in\mathcal{I}_N}\left[s[n]\,(\hat{\mathbf{x}}[n|N])_{(p)} + \sum_{i=1}^{\rho} s[n-i]\,(\hat{\mathbf{C}}[n|N])_{(i,p)}\right],\\ \eta &= \sum_{n\in\mathcal{I}_N}(\hat{\mathbf{C}}[n|N])_{(p,p)}.\end{aligned} \tag{28}$$

Flowchart in Fig. 2 summarizes the implemented optimization procedure considering all the expressions described above. To start with, voice signal $s[n]$ with $n \in \mathcal{I}_N$, model order ρ, glottal opening and closure instants, and parameters $\tilde{A}_g$, $\tilde{B}_g$ and $\tilde{C}_g$ are mandatory. First of all, $\mathcal{I}_{\mathrm{I}}$ and $\mathcal{I}_{\mathrm{II}}$ are established from the glottal instants. SSV model is implemented as follows: $\sigma_{\mathrm{v}}^2 = \mathrm{Var}\,\{s\}$, $G_g = 1$, $A_g = \tilde{A}_g$, $B_g = \tilde{B}_g$ and $C_g = \tilde{C}_g$. Here, initial values of $\mathring{\mathbf{Q}}$ and $\hat{\mathbf{P}}_0$ were heuristically defined in order to produce satisfactory results. Taking into account definition (10), initial state $\hat{\mathbf{x}}_0$ is set from the glottal source and the vocal tract filter estimated by using IAIF method (see Sec. 4.3). Next, a coarse estimate of σ_{v}^2 is obtained through the EM method, keeping fixed the remaining parameters in $\mathbf{\Theta}$, and the iteration index is initialized as $j = 0$.

Afterward the previous initial steps, the iterative optimization is started. State-space methods are applied in order to generate the smoothed estimates and, then, SSV model parameters are computed. These constitute the E and M steps in the EM method, respectively. Next, the cost function is evaluated and the convergence is analyzed. If the increment between previous and current cost values is lower than a given threshold, it is assumed that optimal parameters are obtained. Otherwise, optimization method does not converge, parameters $\tilde{A}_g$, $\tilde{B}_g$ and $\tilde{C}_g$ are upgraded, and the process is repeated. Depending on the accuracy required, a threshold in the range $(10^{-8},\, 10^{-3})$ has been considered. Parameters $\tilde{A}_g$, $\tilde{B}_g$ and $\tilde{C}_g$ are upgraded by using the most recent estimates of SGS parameters. Alternatively, penalization factor λ could be gradually decreased during the process. On one hand, this allows to force a smooth convergence of glottal source parameters and, on the other hand, to modify the effect of $\Phi(\mathbf{\Theta})$ on the optimization procedure. An initial λ in the range $(10^4,\, 10^8)$ has been considered.

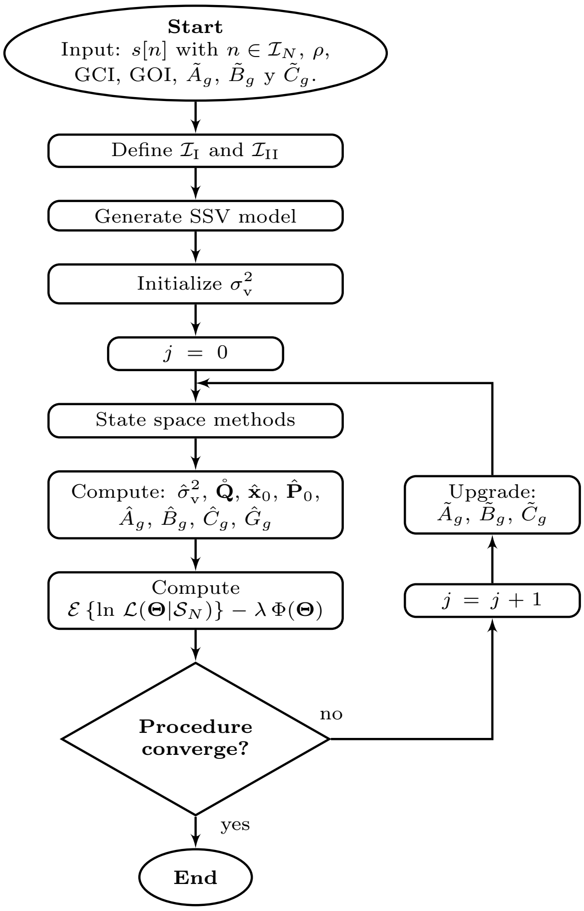
Start
Input: $s[n]$ with $n \in \mathcal{I}_N$, ρ,
GCI, GOI, $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$.
Define $\mathcal{I}_{\mathrm{I}}$ and $\mathcal{I}_{\mathrm{II}}$
Generate SSV model
Initialize σ_{v}^2
$j = 0$
State space methods
Compute: $\hat{\sigma}_{\mathrm{v}}^2$, $\mathring{\mathbf{Q}}$, $\hat{\mathbf{x}}_0$, $\hat{\mathbf{P}}_0$,
$\hat{A}_g$, $\hat{B}_g$, $\hat{C}_g$, $\hat{G}_g$
Compute
$\mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta}|\mathcal{S}_N)\right\} - \lambda\,\Phi(\boldsymbol{\Theta})$
Procedure
converge?
no
yes
End
Upgrade:
$\tilde{A}_g$, $\tilde{B}_g$, $\tilde{C}_g$
$j = j + 1$

Figure 2: Flowchart summarizing the optimization procedure for the computation of unknown parameters $\boldsymbol{\Theta}$ in SSV model. As a result, optimal values of σ_{v}^2, $\mathring{\mathbf{Q}}$, $\hat{\mathbf{x}}_0$, $\hat{\mathbf{P}}_0$, A_g, B_g, C_g and G_g are obtained.

4. Experiments

In this section, the voice material and the reference methods used in this article are briefly described.

4.1. LF glottal source based voice signals

Voice material was generated by synthesizing sustained vowels /a/ using the source-filter theory. The signals were obtained processing periodic LF glottal functions, described by Eqs. (1) and (2), by an autoregressive VTF with the first four formants $\{800, 1200, 2600, 3200\}$ Hz and the bandwidths $\{60, 50, 105, 110\}$ Hz. LF glottal pulses were generated considering random time parameters, as proposed in [22].

Different scenarios were investigated in the experiments, modifying the signal-to-noise ratio (SNR), the glottal source-to-aspiration noise ratio (GNR), and the fundamental frequency (f_0). Only one of these parameters was modified at a time, starting from initial values SNR=60 dB, GNR=60 dB, and f_0=108 Hz. For each setting, 100 signals with a duration of 0.5 s were synthesized considering a sampling frequency of $f_s = 10$ kHz. Notice that, for these signals, both the glottal source waveform and the vocal tract power-spectrum were known in advance.

4.2. Physical model based voice signals

The proposed SSV model and the previously described voice material are both based on the source-filter theory. As a consequence, the above described signals may produce biased results. Thus, voice material obtained by a physical model of voice production was also involved in testing the proposed methods. This idea has been previously applied to the examination of inverse filtering methods [1, 7, 9, 23].

In physical models, vocal fold dynamics and sound wave propagation in the vocal tract are described by analogy with physical and acoustical elements. The length and shape of the vocal tract is described by a series of acoustic tubes with variable cross-section areas. Moreover, the acoustical coupling between the time-varying pressures below and above the glottis, the glottal air flow and the driving forces acting on the vocal folds are also explicitly modeled. More information regarding physical models of voice production could be found in [40, 41].

In order to test the proposed method in more challenging scenarios, the voice material proposed in [38] was used. It consists of signals corresponding to sustained vowels /a/ and /i/ representative of an adult male speaker for three different qualities: pressed, modal and breathy [38]. For each case, one example was available with a constant fundamental frequency of $f_0 = 105$ Hz and a duration of 0.7 s. The sampling frequency was $f_s = 44.1$ kHz. Also, the glottal flow and the first four formants, calculated from the vocal tract assuming the glottal end is closed, were informed. First, the signals were low-pass filtered and downsampled to $f_s = 10$ kHz. Then, they were divided into 12 non-overlapped portions, each of them considering 5 glottal cycles (approximately 50 ms long).

4.3. Experimental setup

The voice signals were individually processed. Parameters $\tilde{A}_g$, $\tilde{B}_g$ and $\tilde{C}_g$ were firstly obtained, by fitting the LF model to the inverse filtered glottal source using IAIF method (see below). Then, optimal values of SSV parameters $\mathbf{\Theta}$ were calculated, applying the method described in Sec. 3.4. As a result, a SSV model fitted to the voice data was obtained. Next, state-space methods were applied in order to compute the smoothed states $\hat{\mathbf{x}}[n|N]$. According to Eq. (10), estimates of coefficients $\hat{a}_l[n]$, for $l = 1, 2, \ldots, \rho$, and the glottal source $\hat{v}_g[n]$ were therefore obtained. Finally, the glottal flow $\hat{U}_g[n]$ was computed by the integration of the glottal source. This procedure constitutes the state-space based voice inverse filtering (SSIF) method.

For comparison purposes, the glottal source and the vocal tract spectral information were also estimated applying two others widely used methods:

- *Iterative Adaptive Inverse Filtering (IAIF):* This method was developed for the automatic decomposition of voiced sounds. It consists on iteratively compute AR models of the glottal source and the vocal tract from a voice signal. In the first place, a glottal source (order 1) AR model is estimated, the source effect is canceled, and a vocal tract (order $\mathtt{p}$) AR model is then computed. Next, model estimates are refined by repeating previous procedures. In this case, a higher-order ($\mathtt{q} > 1$) AR model for representing the glottal source is considered. Generally, IAIF method is applied pitch-synchronously considering windowed segments of voice signal centered according to the glottal closure instants. For a throughout description, see [4, 6] and the references therein. In this work, the implementation provided in the *TKK Aparat* toolbox was applied [2], freely available at `http://sourceforge.net/projects/aparat`.

- *Linear Prediction (LP):* This is one of the most widely used methods for the estimation of vocal tract filter. It computes the vocal tract (order $\mathtt{p}$) AR model that minimizes the auto-regressive prediction error for a given windowed voice signal. In the standard formulation of LP, glottal source is poorly modeled. Therefore, this method is not suitable for glottal source estimation by inverse filtering. For a detailed description of LP and its limitations, see [12, 18, 32, 37]. Here, the classical autocorrelation LP method was considered.

5. Results

5.1. LF glottal source based voice signals

For this voice material, best results were obtained with $\rho = \mathtt{p} = 8$ and $\mathtt{q} = 2$. In Fig. 3, examples obtained through the proposed SSIF method are shown, for a voice signal synthesized with SNR=60 dB, GNR=30 dB, f_0=108 Hz. At the top, a 30 ms-length signal waveform is displayed. In the second row, the LF glottal source (gray line) and the estimates provided by SSIF (thick line) and IAIF (thin line) methods are displayed. In the third row, the estimated glottal flows along with the theoretical (LF) signal are presented. It can be appreciated that both methods suitably estimate the glottal information, in comparison with the original signals. Notice also that SSIF yields the less fluctuating glottal estimates. The latter is more evident in the closed phase. At the bottom, the VTF power spectrum (gray line) and the mean power spectra estimated by SSIF (thick line) and IAIF (thin line) methods are presented. In the case of SSIF, the instantaneous power spectra were first constructed [34] and, then, the mean spectrum was obtained. It can be appreciated that, in this example, SSIF captures the spectral behavior more accurately than IAIF, particularly in the region of the third and fourth formants. It is important to point out that in Fig. 3 all SSIF and IAIF estimates are shifted down and up, respectively, in order to improve the visualization.

Estimated information was objectively assessed for analyzing the performance of the proposed method. On one hand, glottal source estimation was evaluated by computing the relative root-mean-square estimation error in percentage:

$$e_{v_g} = 100 \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{v}_g[n] - v_g[n]}{E_e} \right)^2} \quad \%, \tag{29}$$

where $\hat{v}_g$ and v_g are the estimated and the theoretical glottal sources, respectively, and E_e is taken as a reference value. In Tab. 1, errors e_{v_g} for SSIF and

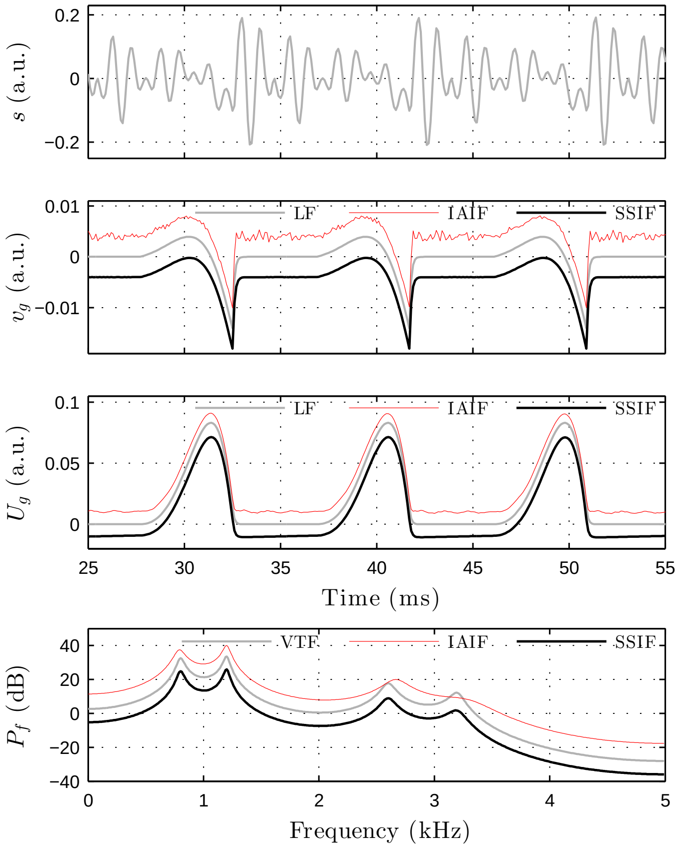

Figure 3: Inverse filtering of a sustained vowel /a/ synthesized applying the LF glottal source (SNR=60 dB, GNR=30 dB, f_0=108 Hz). *Top:* 30 ms of voice waveform. *Second row:* estimation of the glottal source v_g by using SSIF and IAIF methods, in comparison with the LF glottal function considered in the synthesis. *Third row:* glottal flows U_g obtained from the source estimates, along with the original (LF) glottal flow. *Bottom:* power spectra P_f computed by SSIF and IAIF methods, along with the spectral response of the VTF considered in the synthesis. In all cases SSIF and IAIF estimates are shifted down and up, respectively, for a better visualization.

IAIF methods are reported. The mean values and the corresponding standard deviations, in parenthesis, are presented for different levels of SNR, GNR and f_0. Best results are printed in bold font, and statistically different groups according to the *Wilcoxon sum rank test* [31, Sec. 7.5] are indicated. This non-parametric test assumes as null hypothesis that two populations are equal, against the alternative that the latter is not true. It can be appreciated that SSIF yields accurate glottal source estimates, except for very low SNR (high acoustic noise level) where mean error is greater than 8 %. Moreover, notice that SSIF generates the smaller mean errors and the smaller standard deviations compared with IAIF, over all the considered scenarios.

The accuracy in the estimation of VTF spectral information was also assessed. As it was stated before, for the voice material here considered the theoretical VTF power-spectrum P_f^{VT} was known in advance. Therefore, the

Table 1: Glottal source estimation error e_{v_g} (in %) for SSIF and IAIF methods in the LF glottal source based voice material, for different levels of SNR, GNR and f_0. The mean values are presented and the standard deviations are shown in parenthesis. The statistically different results are indicated (* $0.05 \geq p > 0.01$, ** $0.01 \geq p > 0.001$, *** $0.001 \geq p$).

	SNR (dB)			**GNR** (dB)			f_0 (Hz)	
	0-20	25-40	45-60	0-20	25-40	45-60	88-118	188-218
SSIF	**8.12**	**3.64**	**1.40**	**2.87**	**1.05**	**0.97**	**1.04**	**4.98**
	(2.70)	(1.30)	(1.37)	(1.76)	(1.35)	(1.36)	(1.15)	(1.14)
	29.10	13.71	5.52	10.49	2.93	2.75	6.78	23.27
IAIF	(4.52)	(2.86)	(2.46)	(5.85)	(0.40)	(0.37)	(7.31)	(15.48)
	***	***	***	***	***	***	***	***

Table 2: Log-spectral error e_{P_f} (in dB) for SSIF, IAIF and LP methods in the LF glottal source based voice material, for different levels of SNR, GNR and f_0. The mean values are presented and the standard deviations are shown in parenthesis. The statistically different results are indicated (* $0.05 \geq p > 0.01$, ** $0.01 \geq p > 0.001$, *** $0.001 \geq p$).

	SNR (dB)			**GNR** (dB)			f_0 (Hz)	
	0-20	25-40	45-60	0-20	25-40	45-60	88-118	188-218
SSIF	13.49	**9.09**	**2.46**	**2.46**	**0.71**	**0.62**	**0.81**	**2.10**
	(1.18)	(1.45)	(2.12)	(1.40)	(1.07)	(0.91)	(1.01)	(1.71)
	15.25	9.43	4.09	4.07	3.00	2.86	4.49	4.14
IAIF	(1.47)	(1.68)	(1.31)	(1.35)	(0.83)	(0.72)	(1.60)	(1.34)
	***	**	***	***	***	***	***	***
	12.93	9.92	9.66	8.59	9.64	9.65	10.01	9.91
LPC	(1.40)	(0.29)	(0.05)	(1.23)	(0.11)	(0.06)	(0.31)	(0.67)
	***	***	***	***	***	***	***	***

estimates were compared using the log-spectral error, in dB, over L frequency bins $\{f_1, f_2, \ldots, f_L\}$:

$$e_{P_f} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left[10\,\log_{10}\left(\frac{\hat{P}_f[f_l]}{P_f^{VT}[f_l]}\right)\right]^2} \quad \text{dB}, \tag{30}$$

where $\hat{P}_f$ is the estimated power-spectrum. Frequencies in the range 0-5 kHz were considered, with $L = 512$. In Tab. 2, errors e_{P_f} for SSIF, IAIF and LP methods are presented. The mean values and the corresponding standard deviations, in parenthesis, are reported for different levels of SNR, GNR and f_0. Best results are exhibited in bold font. It is also indicated whether the results obtained by SSIF are statistically different from those generated by IAIF or LP, respectively, according to the *Wilcoxon sum rank test*. It can be observed that SSIF generates smaller mean errors and smaller standard deviations than the other methods considered, except for very low SNR where the LP performs better. As expected, the IAIF method outperforms LP, exhibiting smaller mean errors, over almost all the analyzed scenarios.

These results suggest that the stochastic models proposed in Sec. 2, in combination with the state-space methods, give raise to accurate estimates of glottal source and vocal tract spectral information in the voice material here considered. The latter was observed over all the analyzed scenarios, except un-

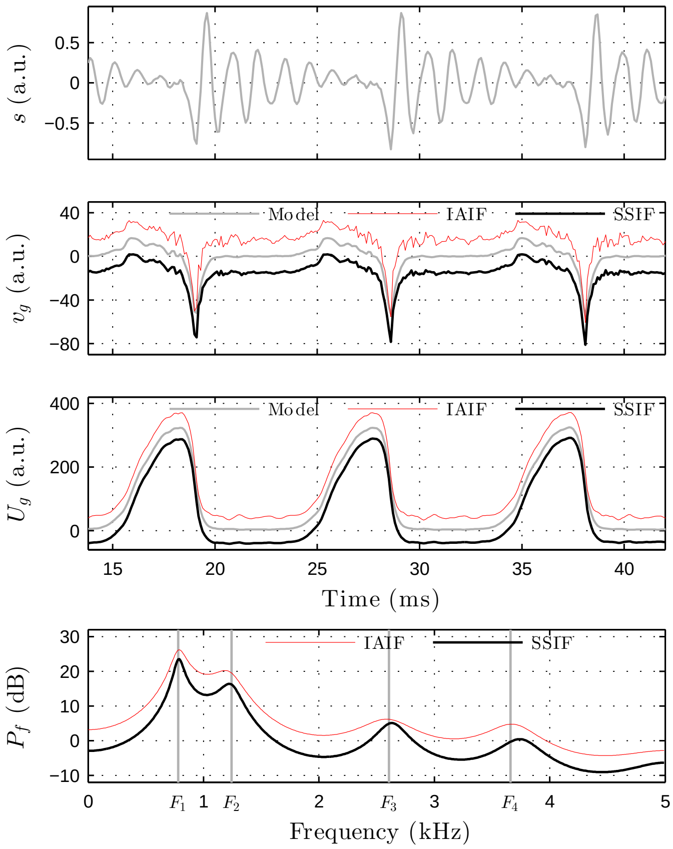

Figure 4: Inverse filtering of a sustained vowel /a/ generated from a physical model of voice production. *Top:* voice signal waveform. *Second row:* estimation of the glottal source v_g by using SSIF and IAIF methods, in comparison with the simulated glottal function. *Third row:* glottal flows U_g obtained from the source estimates, along with the simulated glottal flow. *Bottom:* power spectra P_f computed by SSIF and IAIF methods. Vertical thick lines mark the theoretical location of the first four formants. In all cases SSIF and IAIF estimates are shifted down and up, respectively.

der severe acoustic noise conditions. The proposed SSIF method proved to be robust over a wide range of GNR and f_0. Nevertheless, the estimates deteriorated significantly when acoustic noise was increased. This was also observed for IAIF and LP methods.

5.2. Physical model based voice signals

Next, the results obtained in the physical model based voice material are discussed. In this case, the best results were obtained with $\rho = \mathsf{p} = 10$ and $\mathsf{q} = 4$. Examples of inverse filtering a modal vowel /a/ are shown in Fig. 4. At the top, a 30 ms-length signal waveform is displayed. In the second row, the simulated glottal source (gray line) and the estimates obtained by SSIF (thick line) and IAIF (thin line) methods are presented. Similarly, the simulated glottal flow and the estimates are shown in the third row. It can be observed that the estimates yielded by both methods fit properly the physical (latent) glottal signals. However, it can be seen that SSIF produces smoother signals, showing a less fluctuating behavior in the closed phase, compared with IAIF. At the bottom, the power spectra computed by SSIF (thick line) and IAIF (thin line) methods are presented. Vertical gray lines mark the theoretical location of the first four formants. In this case, both methods suitably capture the vocal tract resonances, except for the fourth formant where it seems that IAIF outperforms SSIF. Once again, SSIF (IAIF) estimates are shifted down (up) in order to improve the visualization.

The performance of the proposed method for the physical model based voice material was also analyzed. Accuracy in the estimation of the glottal source was objectively assessed by computing the estimation error e_{v_g}, defined in Eq. (29). In Tab. 3, errors e_{v_g} for SSIF and IAIF methods are reported. The mean values and the corresponding standard deviations, in parenthesis, are presented for vowels /a/ and /i/, produced with pressed, modal and breathy qualities. Best results are indicated in bold font, and statistically different groups according to the *Wilcoxon sum rank test* are marked. It can be seen that the proposed SSIF accurately estimates the glottal source, except for the breathy vowels /a/ where mean error is greater than 7 %. Furthermore, SSIF outperforms IAIF over all the considered scenarios, yielding the smallest mean errors and the smallest standard deviations. Notice that the mean errors are higher than those reported in Tab. 1, except for a very low SNR.

As stated before, for physical model voice material the first four formants are known in advance. Therefore, the estimation of VTF spectral information was evaluated considering the relative root-mean-square formant estimation error in percentage:

$$e_{F_k} = 100 \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{F}_k[n] - F_k[n]}{F_k[n]} \right)^2} \quad \%, \tag{31}$$

where F_k and $\hat{F}_k$ are the theoretical and the estimated k-th formant, respectively. In Tab. 4, errors e_{F_k} for SSIF, IAIF and LP methods are reported. The mean values and the corresponding standard deviations, in parenthesis, are informed for $\{F_1, F_2, F_3\}$ corresponding to vowels /a/ and /i/ for pressed, modal and breathy qualities. Best results are exhibited in bold font. It is also indicated whether the results obtained by SSIF are statistically different from those generated by IAIF or LP, respectively, according to the *Wilcoxon sum rank test*. It can be seen that the obtained results are considerably diverse. In general, the mean errors produced by SSIF are less than or equal to those obtained by IAIF or LP, and they yield the smallest standard deviations, over most of the considered cases. Moreover, SSIF clearly outperforms the other two methods in the estimation of F_2 in vowels /a/ and F_1 in vowels /i/, respectively.

The last results support the hypothesis that the proposed SSIF, considering the stochastic models formulated in Sec. 2, is able to produce accurate estimates of glottal source and vocal tract information. In addition, the method shows to be robust with respect to the voiced phoneme and the phonation type. As expected, SSIF and the other considered methods yielded the worst results for breathy signals. This could be, in part, explained by the fact that breathy voices are characterized by a shorter closed phase and a lesser glottal excitation energy, compared with modal and pressed phonations. In this scenario, the main hypotheses of the source-filter theory are not fulfilled and the voice

Table 3: Glottal source estimation error e_{v_g} (in %) for SSIF and IAIF methods in the physical model based voice material, corresponding to vowels /a/ and /i/ for voice qualities: pressed, modal and breathy. The mean values are presented and the standard deviations are shown in parenthesis. The statistically different results are indicated (* $0.05 \geq p > 0.01$, ** $0.01 \geq p > 0.001$, *** $0.001 \geq p$).

		Pressed	**Modal**	**Breathy**
/a/	SSIF	**3.91** (0.27)	**3.85** (0.26)	**7.33** (0.61)
	IAIF	6.60 (1.04) ***	7.60 (0.24) ***	12.22 (1.45) ***
/i/	SSIF	**3.89** (0.29)	**4.25** (0.31)	**4.71** (0.32)
	IAIF	8.70 (0.70) ***	6.75 (0.22) ***	17.47 (0.59) ***

Table 4: Relative root-mean-square formant estimation error e_{P_f} (in %) for SSIF, IAIF and LP methods in the physical model based voice material, for $\{F_1, F_2, F_3\}$ corresponding to vowels /a/ and /i/ for voice qualities: pressed, modal and breathy. The mean values are presented and the standard deviations are shown in parenthesis. The statistically different results are indicated (* $0.05 \geq p > 0.01$, ** $0.01 \geq p > 0.001$, *** $0.001 \geq p$).

		Pressed			**Modal**			**Breathy**		
		F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
/a/	SSIF	0.39 (0.17)	**0.50** (0.27)	0.72 (0.07)	**0.36** (0.08)	**0.24** (0.20)	**1.56** (0.34)	**2.44** (0.71)	**3.87** (4.12)	**3.57** (1.23)
	IAIF	**0.31** (0.15)	1.54 (0.38) ***	0.68 (0.13)	0.41 (0.29)	1.64 (0.98) ***	1.84 (1.18)	3.69 (2.80)	11.47 (6.25) ***	4.88 (2.44)
	LP	5.62 (0.16) ***	2.00 (0.11) ***	**0.48** (0.22) **	1.47 (0.54) ***	2.32 (0.88) ***	2.22 (1.61)	3.36 (2.13)	8.27 (4.13) *	4.69 (2.67)
/i/	SSIF	**0.44** (0.13)	0.52 (0.09)	**0.08** (0.06)	**0.49** (0.10)	**0.21** (0.08)	**0.81** (0.30)	**0.57** (0.30)	**0.69** (0.48)	6.82 (1.19)
	IAIF	6.84 (2.47) ***	1.75 (0.30) ***	2.85 (0.39) ***	0.96 (0.10) ***	1.28 (0.69) ***	1.45 (1.02)	2.71 (1.59) ***	2.70 (1.95) *	6.46 (4.08)
	LP	9.96 (0.14) ***	**0.24** (0.16) ***	0.31 (0.15) ***	4.06 (0.55) ***	1.48 (0.69) ***	1.50 (1.09)	4.50 (1.61) ***	2.70 (2.54) *	**5.34** (4.33)

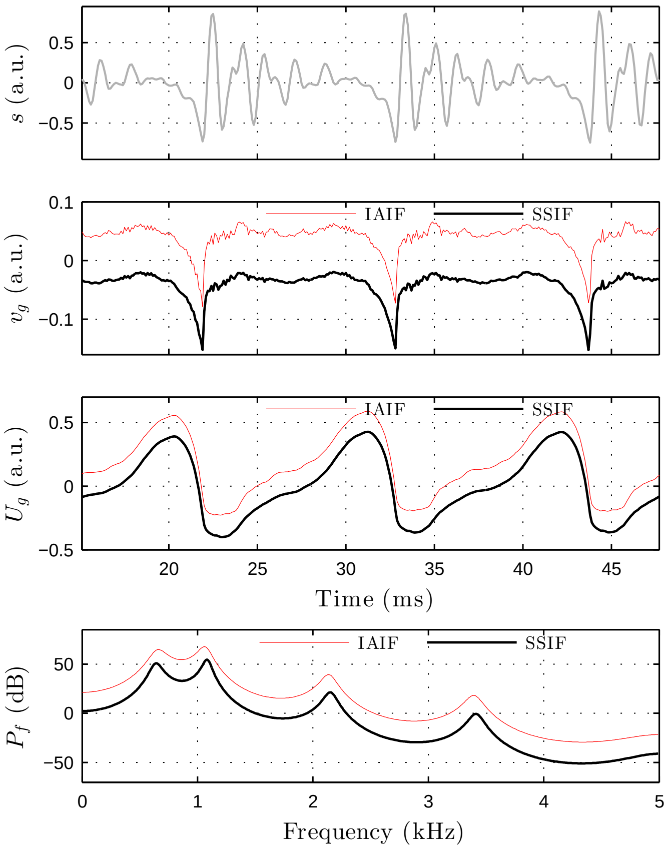

Figure 5: Inverse filtering of a real voice signal corresponding to a low pitch vowel /a/ uttered by an adult male *Top:* voice signal waveform. *Second row:* estimation of the glottal source v_g by using the proposed SSIF and IAIF. *Third row:* glottal flows U_g obtained from the source estimates. *Bottom:* power spectra P_f computed through SSIF and IAIF methods. In all cases SSIF and IAIF estimates are shifted down and up, respectively, for a better visualization.

decomposition fails [19, 38, 44].

5.3. Real voice signals

In this section, the performance of the proposed SSIF for glottal inverse filtering a real voice signal is illustrated. A voice signal selected from the *Saarbruecken Voice Database* [36] is considered, corresponding to a low pitch vowel /a/ uttered by an adult male. In the database, the electroglottographic (EGG) signal was also available. Both voice and EGG signals were recorded simultaneously in a quiet room, and digitized at a 50 kHz sampling frequency. For further information, see [35].

First, the signals were downsampled to a 10 kHz sampling frequency, and the glottal opening and closure instants were estimated from the EGG applying the *SIGMA* algorithm [42]. Then, the optimal parameters for the SSV model were computed, and with them the SSIF was performed. Here, it was considered that $\rho = \mathsf{p} = 10$ and $\mathsf{q} = 4$.

Fig. 5 summarizes the results of using the proposed approach. At the top, a 30 ms-length signal waveform is displayed. In the second row, the glottal source estimates obtained by SSIF (thick line) and IAIF (thin line) methods are presented. It can be appreciated that the estimates closely resemble those signals presented in Fig. 4. Once again, even when the results were similar, the estimates yielded by SSIF were smoother than those obtained by IAIF. Glottal flows resulting from the source estimates are shown in the third row. In this case, these signals are similar to each other, but differ considerably from previous results presented in Figs. 3 and 4. However, these estimates show a good agreement with results of other studies (see e.g., [1, 3, 9, 45]). On the other hand, vocal tract power spectra computed by SSIF (thick line) and IAIF (thin line) methods are plotted at the bottom panel. Notice that both power spectra are similar. Moreover, SSIF seems to produce narrower formant bandwidths showing pronounced formant peaks compared with IAIF.

6. Conclusion

In this work, a stochastic glottal source model was developed, based on the widely accepted Liljencrants-Fant function and in accordance with the state-space theory. Among others benefits, this model allows the description of the glottal source as a non-stationary stochastic phenomenon. As a consequence, this model is suitable to represent the alterations or perturbations normally observed in real voices. As far as the authors know, no other model of the glottal source with this capability has been previously developed.

Additionally, a Gaussian state-space model for voiced sounds production was introduced, combining the classical source-filter theory, the proposed stochastic glottal source model, and the state-space framework. Therefore, a state-space based voice glottal inverse filtering method was implemented combining this voice production model and the state-space methods. The simulations here presented suggest that this approach yielded accurate estimates of the glottal source and the vocal tract information over diverse scenarios. It was also proved that the proposed stochastic glottal source is able to suitably represent different glottal source dynamics. However, a more thorough analysis is required in order to confirm these findings.

An important issue needs to be pointed out: the glottal opening and closure instants are essential components for the implementation of the model here proposed. Therefore, the state-space based voice glottal inverse filtering method highly depends on the quality of these values. Even though several methods have been developed in the last years to deal with this issue, the computation of this information is a very difficult task in practice. Future work will focus on the optimal estimation of the glottal opening and closure instants, based on the proposed voice production state-space model.

7. Acknowledgments

This work was partially supported by PICT-ANPCYT 2012- 2954, PID-UNER 6136, PIO-UNER-CONICET 146-201401- 00014-CO, CITER and CONICET. The authors would like to thank Prof. Brad H. Story, from the Speech Acoustics and Physiology Lab, University of Arizona, Tucson, for kindly providing the physical model based voice material. The authors would like to thank

Dr. Paola Catalfamo Formento from the Universidad Nacional de Entre Ríos and CITER, Argentina, for her helpful advice in the manuscript preparation.

References

[1] M. Airaksinen, T. Raitio, B. H. Story, and P. Alku. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):596–607, 2014. ISSN 2329-9290. doi: 10.1109/TASLP.2013.2294585.

[2] M. Airas. TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008. doi: 10.1080/14015430701855333.

[3] O. O. Akande and P. J. Murphy. Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Communication*, 46(1):15 – 36, 2005. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/j.specom.2005.01.007.

[4] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2):109 – 118, 1992. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/0167-6393(92)90005-R.

[5] P. Alku. Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011. ISSN 0973-7677. doi: 10.1007/s12046-011-0041-5.

[6] P. Alku, B. Story, and M. Airas. Estimation of the voice source from speech pressure signals: evaluation of an inverse filtering technique using physical modelling of voice production. *Folia Phoniatrica et Logopaedica*, 58(2): 102–113, 2006.

[7] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical Society of America*, 134(2), 2013.

[8] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Courier Corporation, 2005. ISBN 978-0-486-13689-9.

[9] H. Auvinen, T. Raitio, M. Airaksinen, S. Siltanen, B. H. Story, and P. Alku. Automatic glottal inverse filtering with the Markov chain Monte Carlo method. *Computer Speech & Language*, 28(5):1139 – 1155, 2014. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2013.09.004.

[10] M. A. Berezina, D. Rudoy, and P. J. Wolfe. Autoregressive modeling of voiced speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5042–5045, 2010. doi: 10.1109/ICASSP.2010.5495058.

[11] J. Candy. *Model-Based Signal Processing*. John Wiley & Sons, New Jersey, USA, 2005. ISBN 9780471732662.

[12] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, New York, 1993.

[13] V. Digalakis, J. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 1(4):431–442, 1993. ISSN 1063-6676. doi: 10.1109/89.242489.

[14] B. Doval, C. d'Alessandro, and N. Henrich. The spectrum of glottal flow models. *Acta Acustica united with Acustica*, 92(6):1026–1046, 2006.

[15] C. Drioli and A. Calanca. Speaker adaptive voice source modeling with applications to speech coding and processing. *Computer Speech & Language*, 28(5):1195 – 1208, 2014. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2014.01.002.

[16] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana. Glottal source processing: From analysis to applications. *Computer Speech & Language*, 28(5):1117 – 1138, 2014. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2014.03.003.

[17] J. Durbin and S. Koopman. *Time Series Analysis by State Space Methods*. Oxford Univ Pr (Sd), New York, USA, 1 edition, 2001. ISBN 0198523548.

[18] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *Signal Processing, IEEE Transactions on*, 39(2):411–423, 1991. ISSN 1053-587X. doi: 10.1109/78.80824.

[19] G. Fant. Some problems in voice source analysis. *Speech Communication*, 13(1):7 – 22, 1993. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/0167-6393(93)90055-P.

[20] G. Fant, J. Liljencrants, and Q. G. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985.

[21] Q. Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(2):492–501, 2006. ISSN 1558-7916. doi: 10.1109/TSA.2005.857807.

[22] P. K. Ghosh and S. S. Narayanan. Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter. *Speech Communication*, 53(1):98 – 109, 2011. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/j.specom.2010.07.004.

[23] J. Guðnason, D. D. Mehta, and T. F. Quatieri. Evaluation of speech inverse filtering techniques using a physiologically based synthesizer. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4245–4249, April 2015. doi: 10.1109/ICASSP.2015.7178771.

[24] M. G. Hall, A. V. Oppenheim, and A. S. Willsky. Time-varying parametric modeling of speech. *Signal Processing*, 5(3):267 – 285, 1983. ISSN 0165-1684. doi: http://dx.doi.org/10.1016/0165-1684(83)90074-9.

[25] A. C. Harvey and S. J. Koopman. Diagnostic checking of unobserved-components time series models. *Journal of Business & Economic Statistics*, 10(4):377–389, 1992. doi: 10.1080/07350015.1992.10509913.

[26] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Courier Corporation, 2007. ISBN 978-0-486-46274-5.

[27] K. Järvinen, A.-M. Laukkanen, and A. Geneid. Voice quality in native and foreign languages investigated by inverse filtering and perceptual analyses. *Journal of Voice*, 2016. ISSN 0892-1997. doi: http://dx.doi.org/10.1016/j.jvoice.2016.05.003. Article in press.

[28] S. J. Koopman. Disturbance smoother for state space models. *Biometrika*, 80(1):117–126, 1993. doi: 10.1093/biomet/80.1.117.

[29] S. J. Koopman. Exact initial kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association*, 92(440):1630–1638, 1997. doi: 10.1080/01621459.1997.10473685.

[30] S. J. Koopman and N. Shephard. Exact Score for Time Series Models in State Space Form. *Biometrika*, 79(4):823–826, 1992. ISSN 0006-3444. doi: 10.2307/2337237.

[31] P. H. Kvam and B. Vidakovic. *Nonparametric Statistics with Applications to Science and Engineering*. John Wiley & Sons, New Jersey, USA, 2007.

[32] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975. ISSN 0018-9219. doi: 10.1109/PROC.1975.9792.

[33] R. L. Miller. Nature of the vocal cord wave. *The Journal of the Acoustical Society of America*, 31(6), 1959.

[34] A. Poulimenos and S. Fassois. Parametric time-domain methods for non-stationary random vibration modelling and analysis - A critical survey and comparison. *Mechanical Systems and Signal Processing*, 20(4):763 – 816, 2006. ISSN 0888-3270. doi: http://dx.doi.org/10.1016/j.ymssp.2005.10.003.

[35] M. Pützer and J. Koreman. A german database of patterns of pathological vocal fold vibration. *Phonus*, 3:143–153, 1997. Institute of Phonetics, University of the Saarland.

[36] M. Pützer and W. J. Barry. Saarbruecken voice database. Institute of Phonetics, University of the Saarland. URL http://www.test.org/doe/. Accessed September 1, 2016.

[37] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, Upper Saddle River, NJ, 1 edition edition, Nov. 2001. ISBN 978-0-13-242942-9.

[38] O. Schleusing, T. Kinnunen, B. Story, and J.-M. Vesin. Joint source-filter optimization for accurate vocal tract estimation using differential evolution. *IEEE Transactions on Audio, Speech, and Language Processing*, 21 (8):1560–1572, 2013. doi: 10.1109/TASL.2013.2255275.

[39] Y.-L. Shue and A. Alwan. A new voice source model based on high-speed imaging and its application to voice source estimation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5134–5137, 2010. doi: 10.1109/ICASSP.2010.5495030.

[40] B. H. Story. *Physiologically-Based Speech Simulation Using an Enhanced Wave-Reflection Model of the Vocal Tract.* PhD thesis, University of Iowa, 1995.

[41] B. H. Story. Phrase-level speech simulation with an airway modulation model of speech production. *Computer Speech & Language*, 27(4):989 – 1010, 2013. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2012.10.005.

[42] M. Thomas and P. Naylor. The SIGMA algorithm: A glottal activity detector for electroglottographic signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1557–1566, 2009. ISSN 1558-7916. doi: 10.1109/TASL.2009.2022430.

[43] I. R. Titze. *Principles of Voice Production.* The National Center for Voice and Speech, Iowa, USA, 2 edition, 2000. ISBN 0-87414-122-2.

[44] J. Walker and P. Murphy. A review of glottal waveform analysis. In Y. Stylianou, M. Faundez-Zanuy, and A. Esposito, editors, *Progress in Nonlinear Speech Processing*, volume 4391 of *Lecture Notes in Computer Science*, pages 1–21. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-71503-0. doi: 10.1007/978-3-540-71505-4_1.

[45] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka. Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1929–1939, 2013. ISSN 1558-7916. doi: 10.1109/TASL.2013.2263138.