Biomedical Signal Processing and Control xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

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

ARTICLE INFO

Article history:
Received 7 June 2016
Received in revised form
30 December 2016
Accepted 30 December 2016
Available online xxx

Keywords:
Stochastic glottal source
State-space voice model
Glottal inverse filtering
Joint source-filter estimation

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.

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, 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

* Corresponding author at: Facultad de Ingeniería – UNER, C.C. 47 – Suc. 3, 3100 Paraná, Entre Ríos, Argentina.
E-mail addresses: galzamendi@bioingenieria.edu.ar (G.A. Alzamendi), gschlotthauer@conicet.gov.ar (G. Schlotthauer).

http://dx.doi.org/10.1016/j.bspc.2016.12.022
1746-8094/

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 Section 2, the stochastic glottal source model and the Gaussian state-space voice model are developed. In Section 3, state-space methods are introduced and optimal estimation of voice model parameters is described. In Section 4, the voice material and the experimental setup utilized in this work are described. In Section 5, the results achieved are exposed and analyzed. Finally, in Section 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 $S(z) = V_g(z)\,G(z)$, where $S(z)$ and $V_g(z)$ are the z-transforms of speech signal s and glottal source v_g, respectively, and $G(z)$ 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:

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

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:

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

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 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:

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

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

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

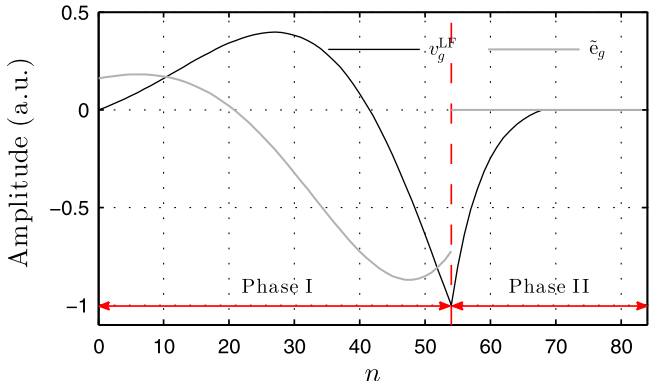

Fig. 1. LF glottal source, v_g^{LF}, and the corresponding auxiliary input signal, $\tilde{e}_g$. Moreover, Phase I and Phase II are indicated by the double arrows.

Please cite this article in press as: G.A. Alzamendi, G. Schlotthauer, Modeling and joint estimation of glottal source and vocal tract filter by state-space methods, Biomed. Signal Process. Control (2016), http://dx.doi.org/10.1016/j.bspc.2016.12.022