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).