# Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering

Paavo Alku[a] and Carlo Magi[b]
*Department of Signal Processing and Acoustics, Helsinki University of Technology, P.O. Box 3000, Fi-02015 TKK, Finland*

Santeri Yrttiaho
*Department of Signal Processing and Acoustics and Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, P.O. Box 3000, Fi-02015 TKK, Finland*

Tom Bäckström
*Department of Signal Processing and Acoustics, Helsinki University of Technology, P.O. Box 3000, Fi-02015 TKK, Finland*

Brad Story
*Speech Acoustics Laboratory, University of Arizona, Tuscon, Arizona 85721*

(Received 23 November 2007; revised 5 February 2009; accepted 8 February 2009)

Closed phase (CP) covariance analysis is a widely used glottal inverse filtering method based on the estimation of the vocal tract during the glottal CP. Since the length of the CP is typically short, the vocal tract computation with linear prediction (LP) is vulnerable to the covariance frame position. The present study proposes modification of the CP algorithm based on two issues. First, and most importantly, the computation of the vocal tract model is changed from the one used in the conventional LP into a form where a constraint is imposed on the dc gain of the inverse filter in the filter optimization. With this constraint, LP analysis is more prone to give vocal tract models that are justified by the source-filter theory; that is, they show complex conjugate roots in the formant regions rather than unrealistic resonances at low frequencies. Second, the new CP method utilizes a minimum phase inverse filter. The method was evaluated using synthetic vowels produced by physical modeling and natural speech. The results show that the algorithm improves the performance of the CP-type inverse filtering and its robustness with respect to the covariance frame position. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3095801]

PACS number(s): 43.70.Gr, 43.70.Jt [CHS] Pages: 3289–3305

## I. INTRODUCTION

All areas of speech science and technology rely, in one form or another, on understanding how speech is produced by the human voice production system. In the area of voice production research, glottal inverse filtering (IF) refers to methodologies that aim to estimate the source of voiced speech, the glottal volume velocity waveform. The basis for these techniques is provided by the classical source-filter theory, according to which the production of a voiced speech signal can be interpreted as a cascade of three separate processes: the excitation, that is, the glottal volume velocity waveform, the vocal tract filter, and the lip radiation effect (Fant, 1970). In order to compute the first of these processes, IF methodologies estimate the second and third processes typically in forms of linear, time-invariant digital systems and then cancel their contribution from the speech signal by filtering it through the inverse models of the vocal tract and lip radiation effect. Since the lip radiation effect can be estimated at low frequencies as a time-derivative of the flow (Flanagan, 1972), which is easily modeled digitally by a fixed first order finite impulse response (FIR) filter, the key problem in IF methods is the estimation of the vocal tract.

Among the main methodologies used to analyze human voice production, IF belongs to the category of acoustical methods. As alternatives to the acoustical methods, it is possible to investigate voice production with visual inspection of the vocal fold vibrations or with electrical (e.g., Lecluse *et al.*, 1975) or electromagnetic methods (Titze *et al.*, 2000). Visual analysis of the vibrating vocal folds is widely used especially in clinical investigation of voice production. Several techniques, such as video stroboscopy (e.g., Hirano, 1981), digital high-speed stroboscopy (e.g., Eysholdt *et al.*, 1996), and kymography (Švec and Schutte, 1996), have been developed, and many of them are currently used in daily practices in voice clinics. Acquiring visual information about voice production, however, always calls for invasive measurements in which the vocal folds are examined either with a solid endoscope inserted in the mouth or with a flexible fiberscope inserted in the nasal cavity. In contrast to these techniques, a benefit of glottal IF is that the analysis can be computed from the acoustic signal in a truly non-invasive manner. This feature is essential especially in such research areas in which vocal function needs to be investigated under as natural circumstances as possible, for instance, in under-

[a] Electronic mail: paavo.alku@tkk.fi
[b] Deceased in February 2008.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

standing the role of the glottal source in the expression of vocal emotions (Cummings and Clements, 1995; Gobl and Ní Chasaide, 2003; Airas and Alku, 2006) or in studying occupational voice production (Vilkman, 2004; Lehto *et al.*, 2008). In addition to its non-invasive nature, glottal IF provides other favorable features. IF results in a temporal signal, the glottal volume velocity waveform, which is an estimate of a real acoustical waveform of the human voice production process. Due to its direct relationship to the acoustical production of speech, estimates of glottal excitations computed by IF can be modeled with their artificial counterparts to synthesize human voice in speech technology applications (Klatt and Klatt, 1990; Carlson *et al.*, 1991; Childers and Hu, 1994).

Since the introduction of the idea of IF by Miller (1959), many different IF methods have been developed. The methods can be categorized, for example, based on the input signal, which can be either the speech pressure waveform recorded in the free field outside the lips (e.g., Wong *et al.*, 1979; Alku, 1992) or the oral volume velocity captured by a specially designed pneumotachograph mask, also known as the Rothenberg mask (e.g., Rothenberg, 1973; Hertegård *et al.*, 1992). In addition, methods developed to do IF differ depending on whether they need user adjustments in defining the settings of the vocal tract resonances (e.g., Price, 1989; Sundberg *et al.*, 2005) or whether the analysis is completely automatic (e.g., Veeneman and BeMent, 1985). From the methodological point of view, the techniques developed can be categorized based on how the effect of the glottal source is taken into account in the estimation of the vocal tract in the underlying IF method. From this perspective, there are, firstly, methods (e.g., Alku, 1992) that are based on the gross estimation of the glottal contribution during both the closed and open phase of the glottal pulse using all-pole modeling. By canceling the glottal contribution from the speech signal, a model for the vocal tract is computed with linear prediction (LP) (Rabiner and Schafer, 1978) although other spectral envelope fitting techniques such as those based on the penalized likelihood approach (Campedel-Oudot *et al.*, 2001) or cepstrum analysis (Shiga and King, 2004) could, in principle, be used as well. Secondly, the use of a joint optimization of the glottal flow and vocal tract is possible based on synthetic, pre-defined models of the glottal flow (e.g., Milenkovic, 1986; Kasuya *et al.*, 1999; Fröhlich *et al.*, 2001; Fu and Murphy, 2006). Thirdly, it is possible to estimate the glottal flow using closed phase (CP) covariance analysis (Strube, 1974; Wong *et al.*, 1979). This is based on the assumption that there is no contribution from the glottal source to the vocal tract during the CP of the vocal fold vibration cycle. After identification of the CP, covariance analysis is used to compute a parametric all-pole model of the vocal tract using LP.

CP covariance analysis is among the most widely used glottal IF techniques. Since the original presentation of the method by Strube (1974), the CP method has been used as a means to estimate the glottal flow, for instance, in the analysis of the phonation type (Childers and Ahn, 1995), prosodic features of connected speech (Strik and Boves, 1992), vocal emotions (Cummings and Clements, 1995), source-tract interaction (Childers and Wong, 1994), singing (Arroabarren and Carlosena, 2004), and speaker identification (Plumpe *et al.*, 1999). In addition to these various applications, CP analysis has been a target of methodological development. The major focus of this methodological work has been the method of accurately determining the location of the covariance frame, the extraction of the CP of the glottal cycle. In order to determine this important time span from a speech waveform, an approach based on a series of sliding covariance analyses is typically used. In other words, the analysis frame is sequentially moved one sample at a time through the speech signal and the results of each covariance analysis are analyzed in order to determine the CP. Strube (1974) used this approach and identified the glottal closure as an instant when the frame was in a position which yielded the maximum determinant of the covariance matrix. Wong *et al.* (1979) instead defined the CP as the interval when the normalized squared prediction error was minimum, and this technique has been used by several authors since, although sometimes with slight modifications (e.g., Cummings and Clements, 1995). Plumpe *et al.* (1999), however, argued that the use of the prediction error energy in defining the frame position of the covariance analysis might be problematic for sounds which involve gradual closing or opening of the vocal folds. As a remedy, they proposed an idea in which sliding covariance analyses are computed and formant frequency modulations between the open and CP of the glottal cycle are used as a means to define the optimal frame position. Akande and Murphy (2005) suggested a new technique, adaptive estimation of the vocal tract transfer function. In their method, the estimation of the vocal tract is improved by first removing the influence of the glottal source by filtering the speech signal with a dynamic, multi-pole high-pass filter instead of the traditional single-pole pre-emphasis. The covariance analysis is then computed in an adaptive loop where the optimal filter order and frame position are searched for by using phase information of the filter candidates.

All the different CP methods referred to above are based on the identification of the glottal CP from a single source of information provided by the speech pressure waveform. Therefore, they typically involve an epoch detection block in which instants of glottal closure and opening are extracted based on algorithms such as DYPSA (Naylor *et al.*, 2007). Alternatively, if electroglottography (EGG) is available, it is possible to use two information channels so that the position and duration of the CP is estimated from EGG, and then the speech waveform is inverse filtered. This so-called two-channel analysis has been shown to yield reliable results in IF due to improved positioning of the covariance frame (Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986). In this technique, the CP analysis is typically computed by estimating the CP of the glottal cycle as the time interval between the minimum and maximum peaks of the first time-derivative of the EGG waveform (Childers and Ahn, 1995). It is important to notice that even though there have been many modifications to CP analysis since the work by Strube (1974), all the methods developed are based on the same principle in the mathematical modeling of the vocal

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

tract, namely, the use of conventional LP with the covariance criterion described in Rabiner and Schafer (1978).

Even though different variants of CP covariance analysis have been shown to yield successful estimates of the glottal flow by using simple synthesized vowels, this IF methodology has certain shortcomings. Several previous studies have in particular indicated that glottal flow estimates computed by the CP analysis vary greatly depending on the position of the covariance frame (e.g., Larar *et al.*, 1985; Veeneman and BeMent, 1985; Yegnanarayana and Veldhuis, 1998; Riegelsberger and Krishnamurthy, 1993). Given the fundamental assumption of the method, that is the computation of the vocal tract model during an excitation-free time span, this undesirable feature of the CP analysis is understandable. The true length of the glottal CP is typically short, which implies that the amount of data used to define the parametric model of the vocal tract with the covariance analysis is sparse. If the position of this kind of a short data frame is misaligned, the resulting linear predictive filter typically fails to model the vocal tract resonances, which might result in severe distortion of the glottal flow estimates. This problem is particularly severe in voices of high fundamental frequency (F0) because they are produced by using very short lengths in the glottal CP. In order to cope with this problem, previous CP methods typically exploit techniques to improve the extraction of the covariance frame position. In the present work, however, a different approach is suggested based on setting a mathematical constraint in the computation of the inverse model of the vocal tract with LP. The constraint imposes a predefined value for the direct current (dc) gain of the inverse filter as a part of the optimization of the filter coefficients. This results in vocal tract filters whose transfer functions, in comparison to those defined by the conventional covariance analysis, are less prone to include poles in positions in the $z$-domain that are difficult to interpret from the point of view of the classical source-filter theory of vowel production (e.g., on the positive real axis). This new dc-constrained vocal tract model is combined in the present study with an additional procedure, checking of the minimum phase property of the inverse filter, to yield a new CP algorithm.

In the following, typical artifacts caused by the CP analysis are first described using representative examples computed from natural vowels. These examples are then used to motivate the proposed new method to compute LP in vocal tract modeling of the CP analysis. The new method is then tested with both synthetic vowels produced by physical modeling of the human voice production mechanism and with natural speech of both female and male subjects.

## II. METHODS

### A. Sources of distortion in the conventional CP analysis

In this section, two major sources of error in the conventional CP analysis are described with the help of examples. The word “conventional” refers here to the CP analysis in which the vocal tract is modeled with a $p$th order all-pole filter computed by the basic form of the covariance analysis described by Rabiner and Schafer (1978), and the lip radiation effect is modeled with a fixed first order FIR filter. All the analyses described were computed using the sampling frequency of 8 kHz and the order of the vocal tract filter set to $p=12$. The length of the covariance frame was 30 samples (3.75 ms). The instant of glottal closure was extracted, when needed, as the instant of the negative peak of the EGG derivative.

First, the sensitivity of the glottal flow estimate about the position of the covariance frame is demonstrated. Figure 1 shows three glottal flow estimates, which were inverse-filtered from the same token of a male subject uttering the vowel [a] by using a minor change in the position of the covariance frame position: the beginning of the covariance frame in Figs. 1(b) and 1(c) was moved earlier in the signal by two and four samples, respectively, in comparison to the beginning of the covariance frame used in Fig. 1(a). The inverse filters obtained are shown in the $z$-domain in the left panels of Fig. 2, and the amplitude spectra of the corresponding vocal tract filters are depicted in the right panels of the same figure. The example indicates how a minor change in the position of the covariance frame has resulted in a substantial change in the estimated glottal flows. It is worth noticing that the covariance analyses illustrated in Figs. 2(a) and 2(b) have resulted in two inverse filters both of which have one root on the positive real axis in the $z$-domain. In Fig. 2(b), the position of this root is slightly closer to the unit circle than in Fig. 2(a). The CP analysis shown in Fig. 2(c) has, in turn, resulted in an inverse filter with a complex conjugate pair of roots at low frequencies. The effect of an inverse filter root which is located on the positive real axis approaches that of a first order differentiator [i.e., $H(z)=1-z^{-1}$] when the root approaches the unit circle, and a similar effect is also produced by a complex conjugate pair of roots at low frequencies. Consequently, the resulting glottal flow estimate, as shown in Figs. 1(b) and 1(c), becomes similar to a time-derivative of the flow candidate given by an inverse filter with no such roots or when these roots are located in a more neutral position close to the origin of the $z$-plane. This severe distortion of the glottal flow estimate caused by the occurrence of inverse filter roots, both real and complex conjugate pairs, at low frequencies is greatest at time instants when the flow changes most rapidly, that is, near glottal closure. As shown in Figs. 1(b) and 1(c), this distortion[1] is typically seen as sharp negative peaks, called “jags” by Wong *et al.* (1979), of the glottal flow pulses at the instants of closure.

The undesirable distortion of the glottal flow estimates by the occurrence of jags implies that the corresponding all-pole vocal tract model has roots on the positive real axis or at low frequencies, and, consequently, its amplitude spectrum shows boosting of low frequencies. This effect is clearly shown in the example by comparing the right panel of Fig. 2(a) to the corresponding panels in Figs. 2(b) and 2(c). It is worth emphasizing that the source-filter theory of voice production by Fant (1970) assumes that poles of the vocal tract for non-nasalized voiced sounds occur as complex conjugate pairs and the low-frequency emphasis of the vowel spectrum results from the glottal source. Therefore, it can be argued

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

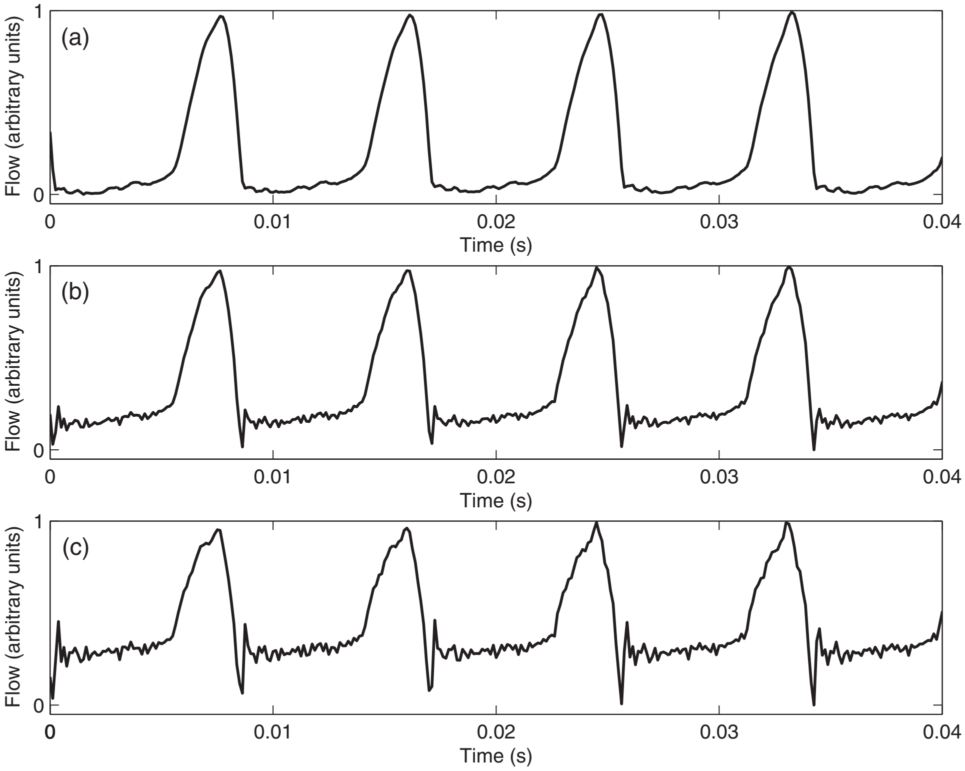

FIG. 1. Glottal flows estimated by IF the vowel [a] uttered by a male speaker by varying the position of the covariance frame in the CP analysis. The covariance frame was placed in the beginning of the CP using the differentiated EGG in panel (a), and its position was moved earlier by two samples in panel (b) and by four samples in panel (c).

that among the three vocal tract models computed by the CP analysis, the one depicted in Fig. 2(a) is the most plausible to represent an amplitude spectrum of an all-pole vocal tract of a vowel sound.

Quality of glottal flows computed by the CP analysis can be made less dependent on the position of the covariance frame by removing the roots of the vocal tract model located on the real axis (Wong *et al.*, 1979; Childers and Ahn, 1995). This is typically done by first solving the roots of the vocal tract model given by LP and then by removing those roots that are located on the positive real axis while preserving the roots on the negative real axis. This procedure was used for the example described in Figs. 1 and 2, and the results are shown in the time domain in Fig. 3 and in the frequency domain in Fig. 4. It can be seen that this standard procedure indeed decreased the distortion caused by the jags, as shown in Fig. 3(b). It is, however, worth noticing that this procedure is blind to complex roots located at low frequencies, which

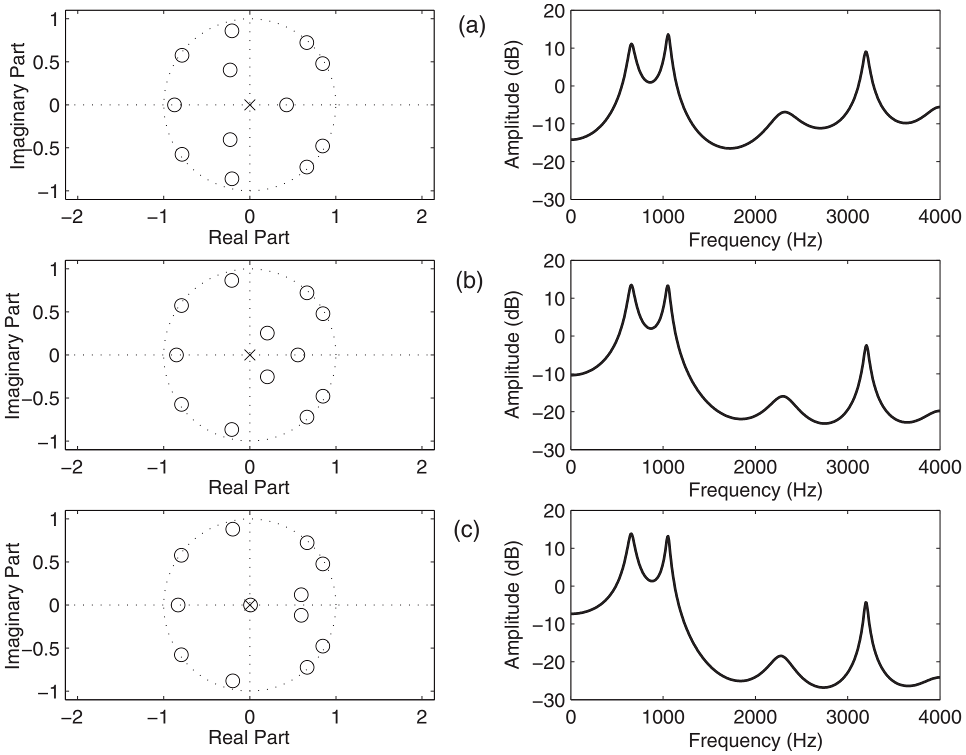

FIG. 2. Transfer functions of inverse filters in the *z*-domain (left panels) and the corresponding amplitude spectra of the all-pole vocal tract models (right panels) used in the CP analyses shown in Fig. 1.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

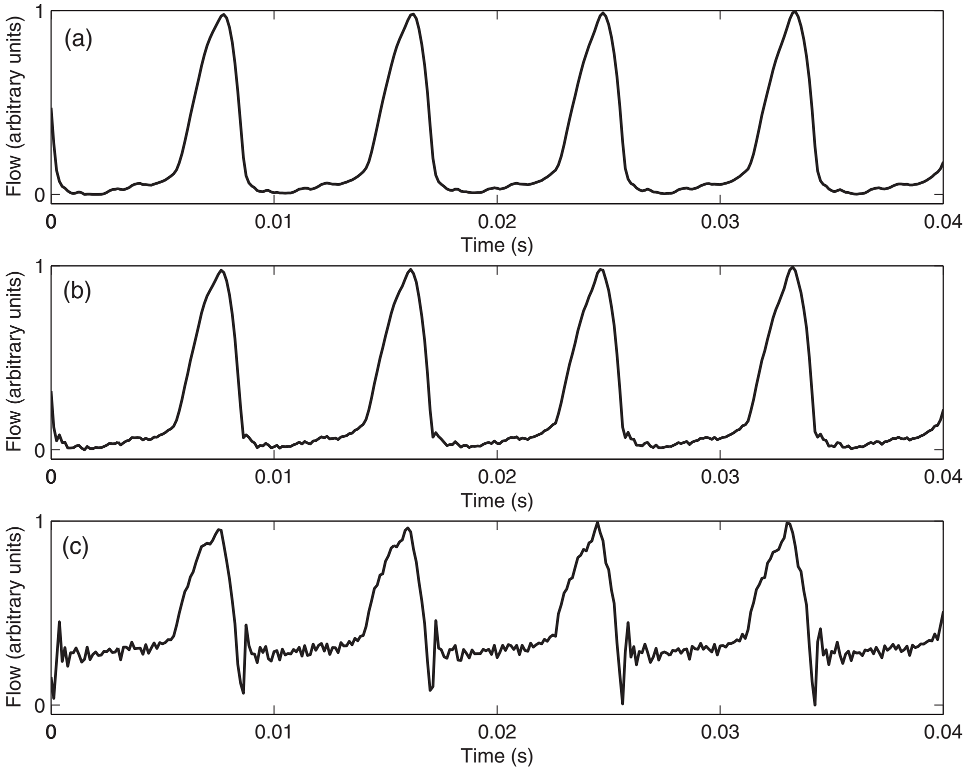

FIG. 3. Glottal flows estimated by IF the same [a] vowel used in Fig. 1. Roots located on the positive real axis were removed before IF. The covariance frame was placed in the beginning of the CP with the help of the differentiated EGG in panel (a), and its position was moved earlier by two samples in panel (b) and by four samples in panel (c).

cause distortion, described in Figs. 1(c) and 3(c), that might be even more severe than that resulting from the roots on the positive real axis.

In addition to the distortion caused by the occurrence of inverse filter real and complex roots at low frequencies as described above, the estimation of the glottal flow with the CP analysis might be affected by another issue. Namely, the computation of the linear predictive analysis with the covariance analysis might yield an inverse filter that is not minimum phase; that is, the filter has roots outside the unit circle in the $z$-domain. Although this property of the covariance analysis is well-known in the theory of LP (Rabiner and Schafer, 1978), it is, unfortunately, typically ignored in most glottal IF studies (exceptions are Akande and Murphy, 2005; Bozkurt *et al.*, 2005; Bäckström and Alku, 2006). A possible explanation of why the occurrence of non-minimum phase filters gets so little attention in glottal wave analysis is the fact that IF is always computed via FIR filtering. Hence,

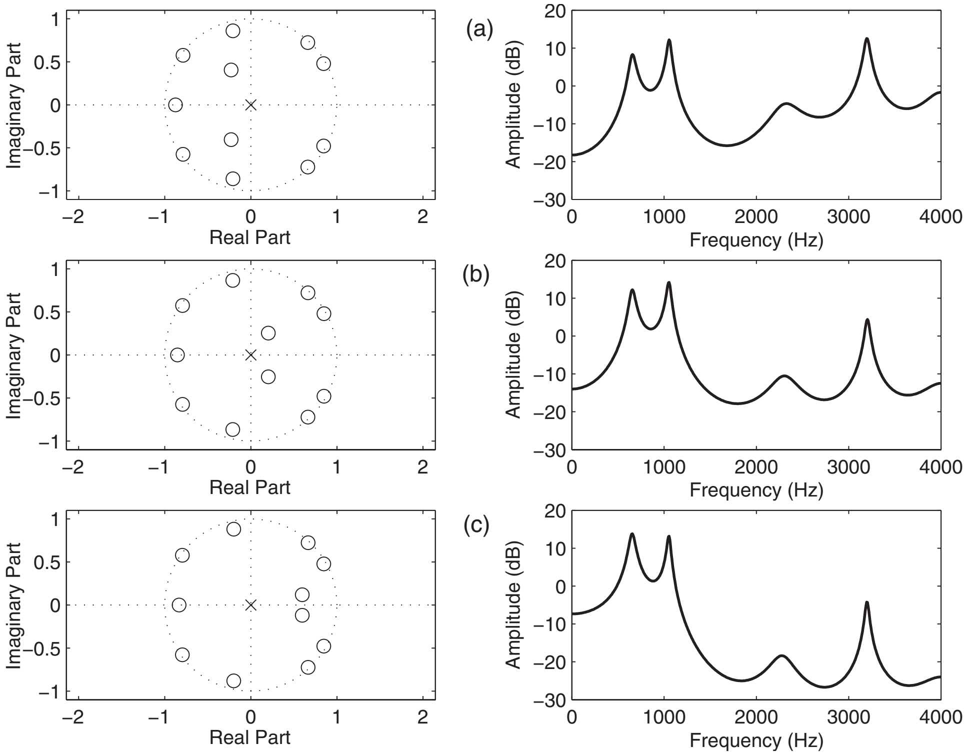

FIG. 4. Transfer functions of inverse filters in the $z$-domain (left panels) and the corresponding amplitude spectra of the all-pole vocal tract models (right panels) used in the CP analyses shown in Fig. 3.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

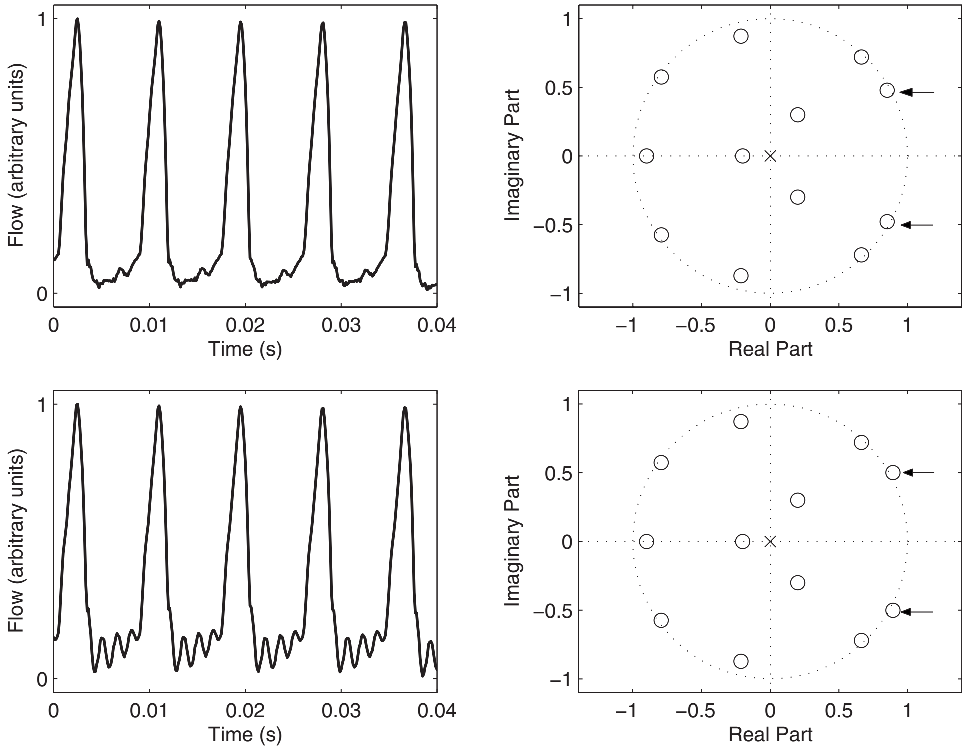

FIG. 5. Glottal flows estimated by the CP analysis (left panels) and inverse filter transfer functions in the $z$-domain (right panels) in the case of (a) minimum phase and (b) non-minimum phase IF. Radii of all roots in minimum phase filtering are less than unity. In non-minimum phase filtering, the complex conjugate root pair indicated by arrows in panel (a) is replaced by its mirror image pair outside the unit circle. The root radius of the indicated complex conjugate pair is 0.98 in panel (a) and 1.02 in panel (b).

non-minimum phase filters do not cause stability problems, which, of course, would be the case if non-minimum phase filters were used in all-pole synthesis, such as in speech coding or synthesis. Even though stability problems are not met in glottal in IF, the use of non-minimum phase inverse filters does cause other kinds of artifacts, as demonstrated below.

According to the source-filter theory of speech production, the glottal flow is filtered by a physiological filter, the vocal tract, which is regarded as a stable all-pole system for vowels and liquids. In the $z$-domain, this kind of system must have all its poles inside the unit circle (Oppenheim and Schafer, 1989). An optimal inverse filter cancels the effects of the vocal tract by matching each pole inside the unit circle with a zero of a FIR filter. However, it is well-known in the theory of digital signal processing that zeros of a FIR filter can be replaced by their mirror image partners; that is, a zero at $z=z_1$ is replaced by $z=1/z_1^*$, without changing the shape of the amplitude spectrum of the filter (Oppenheim and Schafer, 1989). In other words, an inverse filter that is minimum phase can be replaced with a non-minimum phase FIR by replacing any of its roots with a corresponding mirror image outside the unit circle without changing the shape of inverse filter's amplitude response. Therefore, from the point of view of canceling the amplitude response of the all-pole vocal tract, there are several inverse filters, of which one is minimum phase and others are non-minimum phase, that can be considered equal. These candidates are, however, different in terms of their phase characteristics, and canceling the effects of an all-pole vocal tract with a non-minimum phase inverse filter produces phase distortion, which might severely affect the shape of the glottal flow estimate. This distortion is especially strong in cases where zeros in the inverse filter located in the vicinity of the lowest two formants are moved from inside the unit circle to the outside. Figure 5 shows an example of this effect. In Fig. 5(a), a glottal flow estimated with a minimum phase inverse filter is shown in the left panel, and the $z$-plane representation of the corresponding inverse filter is shown on the right. This inverse filter was deliberately modified by replacing one complex conjugate root pair located inside the unit circle by its corresponding mirror image pair located outside the circle. The root pair selected corresponds to the inverse model of the first formant and is represented in the $z$-plane graph of Fig. 5(a) by the complex conjugate pair having the lowest angular frequency (indicated by arrows). Even though the modification caused only a minor change in the root radius (original radius: 0.98, modified radius: 1.02), the change from the minimum phase structure into the non-minimum phase form is manifested as increased ripple during the CP of the glottal cycle, as shown in the left panel of Fig. 5(b).

## B. The improved CP analysis

A new approach is proposed in the present study to compute IF with the CP analysis. The proposed technique aims to reduce the effects of the two major artifacts, occurrence of low-frequency roots of the inverse filter and occurrence of inverse filter roots outside the unit circle, described in the previous section. The main part of the method, to be described in Sec. II B 1, is represented by a new mathematical algorithm to define a linear predictive inverse filter. The novel way to compute vocal tract inverse filters is then combined, as described in Sec. II B 2, with an additional processing stage to yield the new glottal IF algorithm.

### *1. Computation of the vocal tract inverse filter with constrained linear prediction*

The conventional CP analysis involves modeling the vocal tract with an all-pole filter defined according to the clas-

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

sical LP based on the covariance criterion (Rabiner and Schafer, 1978). The filter coefficients of $p$th order inverse filter are searched for by using a straightforward optimization where the energy of the prediction error is minimized over the covariance frame. In principle, this kind of optimization based on the mean square error (MSE) criterion treats all the frequencies equally, and the filter coefficients are mathematically adjusted so that the resulting all-pole spectrum accurately matches the high-energy formant regions of the speech spectrum (Makhoul, 1975). However, it is worth emphasizing that the conventional covariance analysis does not use any additional information in the optimization process, for example, to bias the location of roots of the resulting all-pole filter. This inherent feature of the conventional covariance analysis implies that roots of the resulting all-pole model of the vocal tract might be located in such a position in the $z$-domain (e.g., on the positive real axis) that is correct from the point of view of MSE-based optimization but unrealistic from the point of view of the source-filter theory of vowel production and its underlying theory of tube modeling of the vocal tract acoustics. In his fundamental work, Fant (1970) related vocal tract shapes derived from x-rays to the acoustic theory of different tube shapes and developed the source-filter theory of speech production. According to this theory, the transfer function of voiced speech, defined as the ratio of the Laplace transforms of the sound pressure at the lips to the glottal volume velocity, includes only complex poles in the $s$-domain. According to the discrete time version of this theory (e.g., Markel and Gray, 1976), the $z$-domain transfer function of the vocal tract is expressed for vowel sounds as an all-pole filter of order $2K$, which models $K$ formants as a cascade of $K$ second order blocks, each representing an individual resonance of a certain center frequency and bandwidth. In other words, there might be a mismatch in root locations of vocal tract filters between those optimized by the conventional covariance analysis and those assumed both in the source-filter theory and its underlying acoustical theory of tube shapes. It is likely that this mismatch becomes prevalent especially in cases when the covariance frame consists of a small number of data samples. Hence, the phenomenon discussed is related to the sensitivity of the CP analysis about the position of the covariance frame, a drawback discussed in several previous studies (e.g., Larar *et al.*, 1985; Veeneman and BeMent, 1985; Yegnanarayana and Veldhuis, 1998; Riegelsberger and Krishnamurthy, 1993).

Based on the concept of *constrained* LP, the computation of the conventional covariance analysis, however, can be modified in order to reduce the distortion that originates from such vocal tract model roots that are located in unrealistic positions in the $z$-domain. The key idea is to impose such restrictions on the linear predictive polynomials *prior* to the optimization that can be justified by the source-filter theory of voice production. Intuitively, this means that instead of allowing the linear predictive model to locate its roots freely in the $z$-domain based solely on the MSE criterion, the optimization is given certain restrictions in the predictor structure, which then result in more realistic root locations. In order to implement restrictions that end up in equations which can be solved in closed form, one has to first find a method to express the constraint in a form of a concise mathematical equation and then use the selected equation in the minimization problem. One such convenient constraint can be expressed with the help of the dc gain of the linear predictive inverse filter. The rationales to apply this quantity are as follows. First, the dc gain of a digital FIR filter can be expressed in a very compressed and mathematically straightforward manner as a linear sum of the predictor coefficients [see Eq. (4) below]. Consequently, the optimization of the constrained linear predictive filter is mathematically straightforward, ending up with a matrix equation [see Eq. (9)] that can be solved noniteratively in a similar manner as the corresponding normal equations of the conventional LP. Second, it is known from the classical source-filter theory of voice production that the vocal tract transfer function of non-nasalized sounds approaches unity at zero frequency provided that the losses through vibration of the cavity walls are small (Fant, 1970, pp. 42–44). In conventional LP, the dc gain of the inverse filter is not constrained, and, consequently, it is possible that the amplitude response of the vocal tract model computed by the covariance analysis shows excessive boost at zero frequency. If the covariance frame is short and placed incorrectly, it might even happen that the amplitude response of the obtained vocal tract model shows larger gain at zero frequency than at formants, which violates the assumptions of the source-filter theory and its underlying acoustical theory of tube shapes. Hence, by imposing a pre-defined constraint on the dc gain of the linear predictive inverse filter, one might expect to get such linear predictive vocal tract models whose amplitude response shows better correspondence with Fant's source-filter theory; that is, the transfer function indicates peaks at formant frequencies, while the gain at zero frequency is clearly smaller and approaches unity. It must be emphasized, however, that even though the proposed idea to assign the dc gain of the inverse filter into a pre-defined value is undoubtedly mathematically straightforward, this technique does not involve imposing explicit constraints on the root positions *per se* prior to the optimization. In other words, the exact $z$-domain root locations of the vocal tract model are still determined by the MSE-type optimization, yet the likelihood for these roots to become located in such positions that they create an excessive boost at low frequency is less than in the case of the conventional LP. Mathematical derivations to optimize the proposed idea of the dc-constrained LP will be described below.

In the conventional LP, the error signal, known as the residual, can be expressed in matrix form as follows:

$$e_n = x_n + \sum_{k=1}^{p} a_k x_{n-k} = \sum_{k=0}^{p} a_k x_{n-k} = \mathbf{a}^T \mathbf{x}_n, \qquad (1)$$

where $\mathbf{a}=[a_0, \ldots, a_p]^T$, with $a_0=1$, and the signal vector is $\mathbf{x}_n=[x_n \ldots x_{n-p}]^T$. The coefficient vector $\mathbf{a}$ is optimized according to the MSE criterion by searching for such parameters that minimize the square of the residual. In the covariance method, this minimization of the residual energy is computed over a finite time span (Rabiner and Schafer,

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

1978). By denoting this time span with $0 \leqslant n \leqslant N-1$, the prediction error energy $E(\mathbf{a})$ can be written as

$$E(\mathbf{a}) = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} \mathbf{a}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{a} = \mathbf{a}^T \left[ \sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^T \right] \mathbf{a} = \mathbf{a}^T \boldsymbol{\Phi} \mathbf{a}, \tag{2}$$

where matrix $\boldsymbol{\Phi}$ is the covariance matrix defined from speech samples as

$$\boldsymbol{\Phi} = \sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^T \in R^{(p+1)\times(p+1)}. \tag{3}$$

It is worth noticing that the computation of matrix $\boldsymbol{\Phi}$ requires speech samples located inside the energy minimization frame, that is, $x_n$, where $0 \leqslant n \leqslant N-1$, plus $p$ samples occurring before this frame, that is, $x_n$, where $-p \leqslant n < 0$. The optimal filter coefficients can be computed easily by minimizing the prediction error energy $E(\mathbf{a})$ with respect to the coefficient vector $\mathbf{a}$. This yields $\mathbf{a} = \sigma^2 \boldsymbol{\Phi}^{-1} \mathbf{u}$, where $\boldsymbol{\sigma}^2 = (\mathbf{u}^T \boldsymbol{\Phi}^{-1} \mathbf{u})^{-1}$ is the residual energy given by the optimized predictor and $\mathbf{u} = [10 \cdots 0]^T$.

The conventional LP can be modified by imposing constraints on the minimization problem presented above. A mathematically straightforward way to define one such constraint is to set a certain pre-defined value for the frequency response of the linear predictive inverse filter at zero frequency. By denoting the transfer function of a $p$th order constrained inverse filter $C(z)$, the following equation can be written:

$$C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j0}) = C(1) = \sum_{k=0}^{p} c_k = l_{\mathrm{dc}}, \tag{4}$$

where $c_k$, $0 \leqslant k \leqslant p$, are the filter coefficients of the constrained inverse filter and $l_{\mathrm{dc}}$ is a pre-defined real value for the gain of the filter at dc. Using matrix notation, the dc-constrained minimization problem can now be formulated as follows: minimize $\mathbf{c}^T \boldsymbol{\Phi} \mathbf{c}$ subject to $\boldsymbol{\Gamma}^T \mathbf{c} = \mathbf{b}$, where $\mathbf{c} = [c_0 \cdots c_p]^T$ is the filter coefficient vector with $c_0 = 1$, $\mathbf{b} = [1 l_{\mathrm{dc}}]^T$, and $\boldsymbol{\Gamma}$ is a $(p+1) \times 2$ constraint matrix defined as

$$\Gamma = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & 1 \end{bmatrix}. \tag{5}$$

The covariance matrix defined in Eq. (3) is positive definite. Therefore, the quadratic function to be minimized in the dc-constrained problem is convex. Thus, in order to solve the minimization problem, the Lagrange multiplier method (Bazaraa *et al.*, 1993) can be used. This procedure begins with the definition of a new objective function,

$$\eta(\mathbf{c},\mathbf{g}) = \mathbf{c}^T \boldsymbol{\Phi} \mathbf{c} - 2\mathbf{g}^T (\boldsymbol{\Gamma}^T \mathbf{c} - \mathbf{b}), \tag{6}$$

where $\mathbf{g} = [g_1 g_2]^T > \mathbf{0}$ is the Lagrange multiplier vector. The objective function of Eq. (6) can be minimized by setting its derivative with respect to vector $\mathbf{c}$ to zero. By taking into account that matrix $\boldsymbol{\Phi}$ is symmetric (i.e., $\boldsymbol{\Phi} = \boldsymbol{\Phi}^T$), this results in the following equation:

$$\begin{aligned} \nabla_c \eta(\mathbf{c},\mathbf{g}) &= \mathbf{c}^T(\boldsymbol{\Phi}^T + \boldsymbol{\Phi}) - 2\mathbf{g}^T \boldsymbol{\Gamma}^T = 2\mathbf{c}^T \boldsymbol{\Phi} - 2\mathbf{g}^T \boldsymbol{\Gamma}^T \\ &= 2(\boldsymbol{\Phi}\mathbf{c} - \boldsymbol{\Gamma}\mathbf{g}) = 0. \end{aligned} \tag{7}$$

By combining Eq. (7) with the equation of the constraint (i.e., $\boldsymbol{\Gamma}^T \mathbf{c} - \mathbf{b} = 0$), vector $\mathbf{c}$ can be solved from the group of equations

$$\begin{aligned} &\boldsymbol{\Phi}\mathbf{c} - \boldsymbol{\Gamma}\mathbf{g} = 0, \\ &\boldsymbol{\Gamma}^T \mathbf{c} - \mathbf{b} = 0, \end{aligned} \tag{8}$$

which yields the optimal coefficients of the constrained inverse filter:

$$\mathbf{c} = \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma})^{-1} \mathbf{b}. \tag{9}$$

In summary, the optimal dc-constrained inverse filter, a FIR filter of order $p$ given in Eq. (4) is obtained by solving for the vector $\mathbf{c}$ according to Eq. (9), in which the covariance matrix $\boldsymbol{\Phi}$ is defined by Eq. (3) from the speech signal $x_n$, matrix $\boldsymbol{\Gamma}$ is defined by Eq. (5), and matrix $\mathbf{b} = [1 l_{\mathrm{dc}}]^T$, where $l_{\mathrm{dc}}$ is the desired inverse filter gain at dc.

### *2. Checking the minimum phase property*

In order to eliminate the occurrence of non-minimum phase filters, the roots of the inverse filter are solved, and if the filter is not minimum phase, those roots that are located outside the unit circle are replaced by their mirror image partners inside the circle. In principle, it is possible that the constrained LP computed according to Eq. (9) yields an inverse filter that has roots on the positive real axis. Due to the use of the dc constraint, the risk for this to happen is, however, clearly smaller than in the case of the conventional covariance analysis. Because the roots of $C(z)$ are solved for in order to eliminate the occurrence of non-minimum phase filters, it is trivial also to check simultaneously whether there are any roots on the positive real axis inside the unit circle. If so, these roots are simply removed, in a procedure similar to that used in the conventional CP analysis (Wong *et al.*, 1979).

### *3. Summary of the new algorithm*

In summary, the new glottal IF algorithm can be presented by combining the procedures described in Secs. II B 1 and II B 2. The estimation of the glottal flow with this new CP-based IF algorithm consists of the following stages.

(1) Prior to the analysis, the speech pressure waveform is filtered through a linear-phase high-pass FIR with its cut-off frequency adjusted to 70 Hz. The purpose of this filter is to remove annoying low-frequency components picked up by the microphone during the recordings of the speech signals. The output of this stage, the high-pass filtered speech sound, is denoted by $S_{\mathrm{hp}}(n)$ below.

(2) The position of the covariance frame is computed using any of the previously developed methods based on, for

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

example, the maximum determinant of the covariance matrix (Wong *et al.*, 1979) or the EGG (Krishnamurthy and Childers, 1986).

(3) Vocal tract transfer function $C(z)$ is computed according to Eq. (9) by defining the elements of the covariance matrix in Eq. (3) from $S_{\mathrm{hp}}(n)$ by using the covariance frame defined in stage (2).

(4) Roots of $C(z)$ defined in stage (3) are solved. Those roots of $C(z)$ that are located outside the unit circle are replaced by their corresponding mirror image partner inside the unit circle. Any real roots located on the positive real axis are removed.

(5) Finally, the estimate of the glottal volume velocity waveform is obtained by filtering $S_{\mathrm{hp}}(n)$ through $C(z)$ defined in stage (4) and by canceling the lip radiation effect with a first order infinite impulse response filter, with its pole close to the unit circle (e.g., at $z=0.99$).

The algorithm runs in a frame-based manner, and the adjustable parameters are recommended to be set to values typically used in CP analysis: frame length: 50 ms; order of the vocal tract model: 12 (with sampling frequency of 8 kHz); the length of the covariance frame: 30 samples (a value that equals the order of the vocal tract model multiplied by 2.5). In the experiments conducted in the present study, the parameter $l_{\mathrm{dc}}$ used in the computation of the dc-constrained vocal tract inverse filters was adjusted so that the amplitude response of the vocal tract filter at dc was always equal to unity.[2]

## III. MATERIALS AND EXPERIMENTS

In order to evaluate the performance of the new CP analysis technique, experiments were conducted using both natural and synthetic speech. The purpose of these experiments was to investigate whether the new modified covariance analysis based on the concept of constrained LP, when supplemented with the minimum phase requirement of the inverse filter, would make IF with the CP analysis less vulnerable to the position of the covariance frame.

### A. Speech and EGG recordings

Simultaneous speech pressure waveform and EGG signals were recorded from 13 subjects (six females). The ages of the subjects varied between 29 and 43 (mean of 32), and none of them had experienced voice disorders. The speaking task was to produce the vowel [a] five times by using sustained phonation. Vowel [a] was used because it has a high first formant (F1).[3] Subjects were allowed to use the fundamental frequency of their own choice, but they were encouraged not to use a pitch that is noticeably higher than in their normal speech. The duration of each phonation was at least 1 s. The production was done by two types of phonation: normal and pressed. These two phonation types were selected because they are more likely to involve a CP in the vocal fold vibration, which would not be the case in, for example, breathy phonation (Alku and Vilkman, 1996). This, in turn, implies that the basic assumption of the CP analysis, that is, the existence of a distinct CP within the glottal cycle, should be valid. Consequently, using these two modes, one would expect to be able to demonstrate effectively the dependency of the CP analysis on the position of the covariance frame. The recordings were perceptually monitored by an experienced phonetician who trained the subjects to create the two registers properly. Phonations were repeated until the phonation type was satisfactory.

Speech pressure waves were captured by a condenser microphone (Brüel & Kjær 4188) that was attached to a sound level meter (Brüel & Kjær Mediator 2238) serving also as a microphone amplifier, and the EGG was recorded simultaneously (Glottal Enterprise MC2-1). The mouth-to-microphone distance was 40 cm. In order to avoid inconsistency in the synchronization of speech and EGG, the microphone distance was carefully monitored in the recordings, and its value was checked prior to each phonation. Speech and EGG waveforms were digitized using a (DAT) digital audio tape recorder (Sony DTC-690) by adopting the sampling rate of 48 kHz and the resolution of 16 bits.

The speech and EGG signals were digitally transferred from the DAT tape into a computer. Before conducting the IF analysis, the sampling frequency of both signals was downsampled to 8 kHz. The propagation delay of the acoustic signal from the glottis to the microphone was estimated by using the vocal tract length of 15 and 17 cm for females and males, respectively, the mouth-to-microphone distance of 40 cm, and the speed of sound value of 350 m/s. These values yielded the propagation delay of 1.57 and 1.63 ms for female and male speakers, respectively. The fundamental frequency of each vowel sound was computed by searching for the peak of the autocorrelation function from the differentiated EGG signal. For female speakers, the mean F0 was 195 Hz (min: 178 Hz, max: 211 Hz) and 199 Hz (min: 182 Hz, max: 216 Hz) in normal and pressed phonation, respectively. For males, the mean F0 was 104 Hz (min: 90 Hz, max: 119 Hz) and 114 Hz (min: 95 Hz, max: 148 Hz) in normal and pressed phonation, respectively.

### B. Synthetic vowels

A fundamental problem present both in developing new IF algorithms and in comparing existing methods is the fact that assessing the performance of an IF technique is complicated. When IF is used to estimate the glottal flow of natural speech, it is actually never possible to assess in detail how closely the obtained waveform corresponds to the true glottal flow generated by the vibrating vocal folds. It is, however, possible to assess the accuracy of IF by using synthetic speech that has been created using artificial glottal waveform. This kind of evaluation, however, is not truly objective because speech synthesis and IF analysis are typically based on similar models of the human voice production apparatus, for example, the traditional linear source-filter model (Fant, 1970).

In the current study, a different strategy was used in order to evaluate the performance of different CP analysis methods in the estimation of the glottal flow. The idea is to use *physical modeling* of the vocal folds and the vocal tract in order to simulate time-varying waveforms of the glottal

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

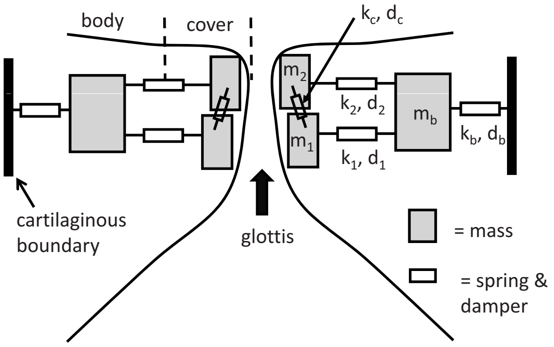

FIG. 6. Schematic diagram of the lumped-element vocal fold model. The cover-body structure of each vocal fold is represented by three masses that are coupled to each other by spring and damping elements. Bilateral symmetry was assumed for all simulations.

flow and radiated acoustic pressure. By using the simulated pressure waveform as an input to an IF method, it is possible to determine how closely the obtained estimate of the voice source matches the simulated glottal flow. This approach is different from using synthetic speech excited by an artificial form of the glottal excitation because the glottal flow waveform results from the interaction of the self-sustained oscillation of the vocal folds with subglottal and supraglottal pressures, as would occur during real speech production. Hence, the glottal flow waveform generated by this model is expected to provide a more stringent and realistic test of the IF method than would be permitted by a parametric flow waveform model where no source-tract interaction is incorporated.[4]

The sound pressure and glottal flow waveforms used to test the new IF technique were generated with a computational model of the vocal folds and acoustic wave propagation. Specifically, self-sustained vocal fold vibration was simulated with three masses coupled to one another through stiffness and damping elements (Story and Titze, 1995). A schematic diagram of the model is shown in Fig. 6, where the arrangement of the masses was designed to emulate the body-cover structure of the vocal folds (Hirano, 1974). The input parameters consisted of lung pressure, prephonatory glottal half-width (adduction), resting vocal fold length and thickness, and normalized activation levels of the cricothyroid (CT) and thyroarytenoid (TA) muscles. These values were transformed to mechanical parameters of the model, such as mass, stiffness, and damping, according to the "rules" proposed by Titze and Story (2002). The vocal fold model was coupled to the pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations as specified by Titze (2002), thus allowing for self-sustained oscillation. Bilateral symmetry was assumed for all simulations such that identical vibrations occur within both the left and right folds. Nine different fundamental frequency values (105, 115, 130, 145, 205, 210, 230, 255, and 310 Hz), which roughly approximate the ranges typical of adult male and female speech (e.g., Hollien *et al.*, 1971; Hollien and Shipp, 1972; Stoicheff, 1981), were generated by modifying the resting vocal fold length and activation levels of the CT and TA muscles; all other input parameters were held constant. The input parameters for all nine cases are shown in Table I. Those cases with the resting length ($L_o$) equal to 1.6 cm were intended to be representative of the male F0 range, whereas those with $L_o$=0.9 cm were intended to be in the female F0 range.

Acoustic wave propagation in both the trachea and vocal tract was computed in time synchrony with the vocal fold model. This was performed with a wave-reflection approach (e.g., Strube, 1982; Liljencrants, 1985) where the area functions of the vocal tract and trachea were discretized into short cylindrical sections or tubelets. Reflection and transmission coefficients were calculated at the junctions of consecutive tubelets, at each time sample. From these, pressure and volume velocity were then computed to propagate the acoustic waves through the system. The glottal flow was determined by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis as specified by Titze (2002). At the lip termination, the forward and backward traveling pressure wave components were subjected to a radiation load modeled as a resistance in parallel with an inductance (Flanagan, 1972), intended to approximate a piston in an infinite plane baffle. The output pressure is assumed to be representative of the pressure radiated at the lips. To the extent that the piston-in-a-baffle reasonably approximates the radiation load, the calculated output pressure can also be assumed to be representative of the pressure that would be transduced by a microphone in a non-reflective environment. The specific implementation of the vocal tract

TABLE I. Input parameters for the vocal fold model used to generate the nine different fundamental frequencies. Notation is identical to that used in Titze and Story (2002). The $a_{\mathrm{CT}}$ and $a_{\mathrm{TA}}$ are normalized activation levels (can range from 0 to 1) of the CT and TA muscles, respectively. $L_o$ and $T_o$ are the resting length and thickness of the vocal folds, respectively. $\xi_{01}$ and $\xi_{02}$ are the prephonatory glottal half-widths at the inferior and superior edges of vocal folds, respectively, and $P_L$ is the respiratory pressure applied at the entrance of the trachea (see Fig. 7). The value of $P_L$ shown in the table is equivalent to a pressure of 8 cm $H_2O$.

| | Fundamental frequency (Hz) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter value | 105 | 115 | 130 | 145 | 205 | 210 | 230 | 255 | 310 |
| $a_{\mathrm{CT}}$ | 0.1 | 0.4 | 0.1 | 0.4 | 0.2 | 0.3 | 0.3 | 0.4 | 0.7 |
| $a_{\mathrm{TA}}$ | 0.1 | 0.1 | 0.4 | 0.4 | 0.2 | 0.2 | 0.3 | 0.4 | 0.5 |
| $L_o$ (cm) | 1.6 | 1.6 | 1.6 | 1.6 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| $T_o$ (cm) | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 |
| $\xi_{01}$ (cm) | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| $\xi_{02}$ (cm) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| $P_L$ (dyn/cm$^2$) | 7840 | 7840 | 7840 | 7840 | 7840 | 7840 | 7840 | 7840 | 7840 |

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

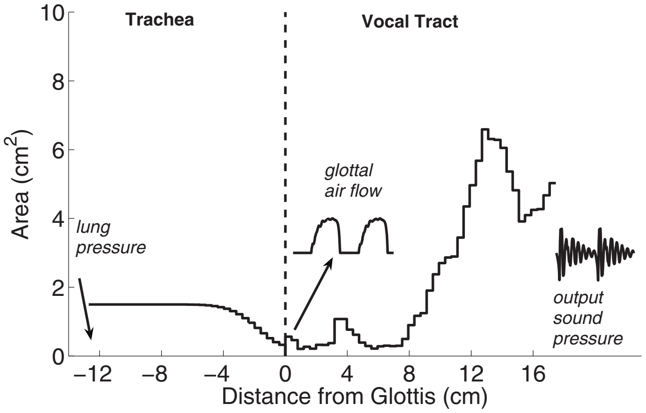

FIG. 7. Area function representation of the trachea and vocal tract used to simulate the male [a] vowel. The vocal fold model of Fig. 6 would be located at the 0 cm point indicated by the dashed vertical line. Examples of the glottal flow and output pressure waveforms are shown near the locations at which they would be generated.

model used for this study was presented in Story (1995) and included energy losses due to viscosity, yielding walls, heat conduction, as well as radiation at the lips.

In the model, a specific vocal tract shape is represented as an area function. For this study, glottal flow and output pressure waveforms were generated based on the area function for the [a] vowel reported by Story *et al.* (1996). For simulations of this vowel with the four lowest fundamental frequencies (105, 115, 130, and 145 Hz), the vocal tract length was set to 17.46 cm. For the five higher F0 speech simulations, exactly the same [a] vowel area function was used, but the length was non-uniformly scaled to 14.28 cm with scaling factors based on those reported by Fitch and Giedd (1999). The purpose of the shortened tract length was to provide an approximation of a possible female-like vocal tract to coincide with the higher F0 simulations. Although a measured female area function could have been used (e.g., Story, 2005), scaling the length of the male [a] vowel was done so that all cases resulted from fairly simple modifications of the same basic model.

A conceptualization of the complete model is given in Fig. 7, where the vocal fold model is shown to be located between the trachea and the vocal tract. The vocal tract is shown configured with the shape and length of the adult male [a] vowel, and the trachea is a uniform tube with a cross-sectional area of 1.5 cm$^2$ but tapered to 0.3 cm$^2$ near the glottis. An example glottal flow waveform is indicated near the middle of the figure. Note that the ripples in the waveform are largely due to interaction of the flow with the formant oscillations in the vocal tract. The coupling of the trachea to the vocal tract (via glottal area), however, will slightly alter the overall resonant structure of the system and, hence, will also contribute to glottal waveform shape. The sound pressure waveform radiated at the lips is also shown at the lip end of the area function and, as mentioned previously, can be considered analogous to a microphone signal recorded for a speaker.

In summary, the model is a simplified but physically-motivated representation of a speaker in which glottal airflow and output pressure waveforms result from self-sustained oscillation of the vocal folds and their interaction with propagating pressure waves within the trachea and vocal tract. The model generates both the signal on which IF is typically performed (microphone signal) and the signal that it seeks to determine (glottal flow), thus providing a reasonably realistic test case for IF algorithms.

## C. Experiments

Four representative examples of glottal flow pulse forms computed by the proposed CP algorithm are shown in Fig. 8.

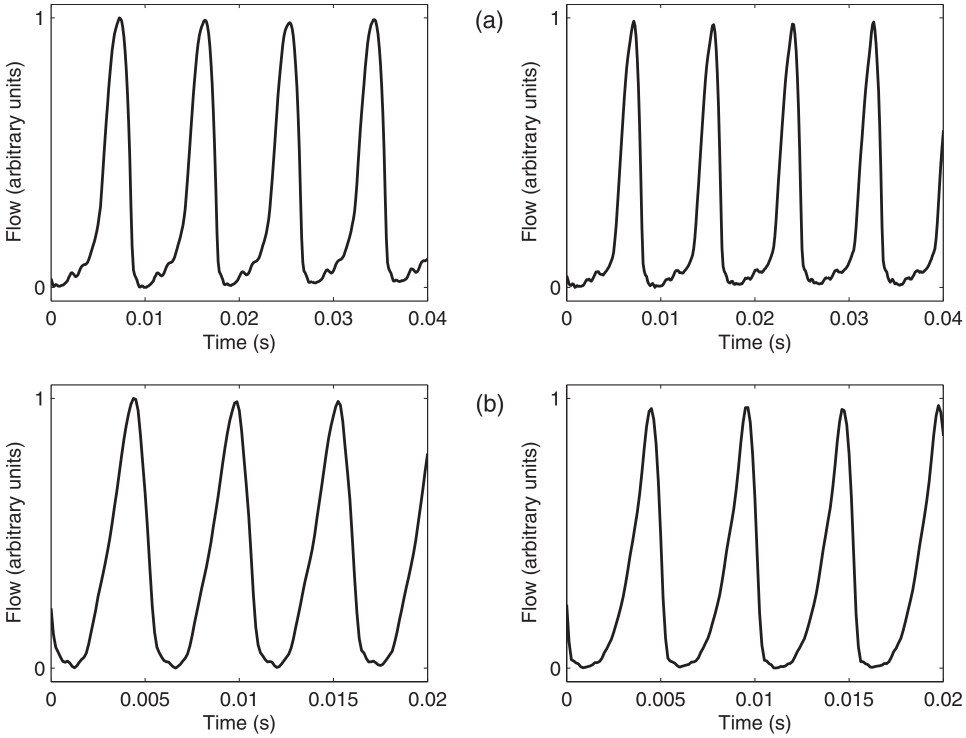

FIG. 8. Examples of glottal flows estimated by the proposed CP$_{\text{con}}$ method. IF was computed from [a] vowels produced by a male (panels a) and a female (panels b) speaker using normal (left panels) and pressed (right panels) phonation.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

The examples shown in Figs. 8(a) and 8(b) were computed from a male and female speaker, respectively, by using both normal and pressed phonations of the vowel [a]. All these estimates of the glottal excitation were computed by using parameter values given at the end of Sec. II B 3. The beginning of the covariance frame was adjusted to a time instant three samples after the negative peak of the EGG derivative. It can be seen in Fig. 8 that none of the estimated glottal pulse forms show abrupt high amplitude peaks at the end of the closing phase, indicating that inverse filter roots are most likely located correctly in the formant region rather than in unrealistic positions at low frequency. CP can be identified rather easily from all the examples shown. However, the waveforms estimated from utterances spoken by the male speaker show a small ripple component. This ripple might be due to incomplete canceling of some of the higher formants by the inverse filter. Alternatively, this component might be explained by the existence of nonlinear coupling between the source and the tract, which cannot be taken into account in CP analysis because it is based on linear modeling of the voice production system.

The performance of the proposed CP analysis algorithm was tested by conducting two major experiments, one of which used synthetic vowels and the other natural speech. Both experiments involved estimating the glottal flow with three CP analysis types. The first one, denoted by $\mathrm{CP_{bas}}$ in the rest of the paper, is represented by the basic CP analysis in which the vocal tract model computed by the covariance analysis is used as such in IF. The second one, denoted by $\mathrm{CP_{rem}}$, is the most widely used form of the CP analysis in which the roots of the inverse filter polynomial computed by the covariance analysis are solved, and those located on the positive real axis are removed before IF. The third type, denoted by $\mathrm{CP_{con}}$, is the proposed method based on the constrained LP described in Sec. II B.

In both experiments, the robustness of each CP analysis to the position of the covariance frame was evaluated by varying the beginning of the frame position near its optimal value, $n_{\mathrm{opt}}$, the instant of glottal closure. For synthetic vowels, $n_{\mathrm{opt}}$ was first adjusted by using the derivative of the flow pulse generated by the physical vocal fold model. In this procedure, the optimal beginning of the covariance frame was set to the time instant after the negative peak of the flow derivative when the waveform returns to the zero level. For each synthetic vowel, the beginning of the covariance frame was then varied in 11 steps by defining the start index as $n = n_{\mathrm{opt}} + i$, where $i = -5$ to $+5$. (In other words, the optimal frame position corresponds to index value $i = 0$.) For natural vowels, the position of the covariance frame was varied by first extracting the glottal closure as the time instant when the EGG derivative reached a negative peak within a glottal cycle. Again, 11 frame positions were analyzed around this instant of glottal closure.

For synthetic sounds, there is no variation between periods, and, therefore, only a single cycle was analyzed. The total number of CP analyses conducted for synthetic speech was 297 (3 CP methods × 9 F0 values × 11 frame positions per cycle). For natural vowels, the analysis was repeated for six consecutive glottal cycles. Hence, the total number of CP analyses conducted for natural speech was 5148 (3 CP methods × 2 phonation types × 13 speakers × 11 frame positions per cycle × 6 cycles). The estimated glottal flows were parametrized using two frequency-domain measures. The first of these, H1H2, is defined as the difference in decibel between the amplitudes of the fundamental and the second harmonic of the source spectrum (Titze and Sundberg, 1992). The second parameter, the harmonic richness factor (HRF), is defined from the spectrum of the glottal flow as the difference in decibel between the sum of the harmonic amplitudes above the fundamental and the amplitude of the fundamental (Childers and Lee, 1991). (Notice the difference in the computation of the spectral ratio between the two parameters: if only the second harmonic is included in HRF, then its value becomes equal to H1H2 multiplied by −1.) These parameters were selected for two reasons. First, both of them can be computed automatically without any user adjustments. In CP analysis with a varying frame position, this is highly justified because the glottal flow waveforms, especially those computed with $\mathrm{CP_{bas}}$, are sometimes so severely distorted that their reliable parametrization with, for example, time-based glottal flow quotients is not possible. Second, both H1H2 and HRF are known to reflect the spectral decay of the glottal excitation: a slowly decaying source spectrum is reflected by a small H1H2 and a large HRF value. Hence, if the glottal flow estimate is severely distorted by artifacts seen as jags in the closing phase, as shown in Figs. 1(b), 1(c), and 3(c), one is expected to get a decreased H1H2 value and an increased HRF value because the spectrum of the distorted glottal flow approaches that of the impulse train, that is, a flat spectral envelope. Since HRF takes into account a larger number of spectral harmonics, one can argue that its value reflects more reliable changes in the glottal flow than H1H2. Therefore, HRF alone might represent a sufficient spectral parameter to be used from the point of view of the present study. H1H2 is, however, a more widely used parameter in voice production studies, which justifies its selection as an additional voice source parameter in the present investigation.

## IV. RESULTS

### A. Experiment 1: Synthetic vowels

Robustness of the different CP analyses to the covariance frame position is demonstrated for the synthetic vowels by the data given in Table II. H1H2 and HRF values were first computed in each covariance frame position with each of the three CP techniques. For both H1H2 and HRF, the difference between the parameters extracted from the original flow and the estimated flow was computed. The data in Table II show the absolute value of this difference computed as an average pooled over 11 frame positions. The obtained results indicate that the error in both H1H2 and HRF due to the variation of the CP frame position is smallest for all vowels with F0 less than 310 Hz when IF is computed with the proposed new method. The average value of H1H2, when pooled over all vowels with F0 less than 310 Hz, equaled to 2.6, 0.9, and 0.5 dB for $\mathrm{CP_{bas}}$, $\mathrm{CP_{rem}}$, and $\mathrm{CP_{con}}$, respectively. For HRF, the average value equaled to 7.8, 3.8, and 2.4 dB for $\mathrm{CP_{bas}}$, $\mathrm{CP_{rem}}$, and $\mathrm{CP_{con}}$, respectively. For the synthetic

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

TABLE II. Effect of the covariance frame position on H1H2 and HRF using vowels synthesized by physical modeling. Absolute value of the difference (in dB) was computed between parameter values extracted from the original flows and from the glottal flows estimated by IF. Inverse filtering was computed by three CP algorithms: $CP_{bas}$, $CP_{rem}$, and $CP_{con}$. Data were averaged over 11 different frame positions starting around the instant of glottal closure.

| F0 (Hz) | Diff in H1H2 (dB) | | | Diff in HRF (dB) | | |
|---|---|---|---|---|---|---|
| | $CP_{bas}$ | $CP_{rem}$ | $CP_{con}$ | $CP_{bas}$ | $CP_{rem}$ | $CP_{con}$ |
| 105 | 1.36 | 0.06 | 0.03 | 5.08 | 2.30 | 1.87 |
| 115 | 2.93 | 0.14 | 0.08 | 9.15 | 2.37 | 1.75 |
| 130 | 1.81 | 0.13 | 0.06 | 6.14 | 2.23 | 1.63 |
| 145 | 3.42 | 0.10 | 0.07 | 10.36 | 1.74 | 1.59 |
| 205 | 2.98 | 1.66 | 0.83 | 8.80 | 6.13 | 2.72 |
| 210 | 2.67 | 1.40 | 0.82 | 8.06 | 5.91 | 2.91 |
| 230 | 3.17 | 1.28 | 0.90 | 8.31 | 4.00 | 3.21 |
| 255 | 2.40 | 2.13 | 1.24 | 6.35 | 5.44 | 3.40 |
| 310 | 0.69 | 0.69 | 3.38 | 6.24 | 6.24 | 5.02 |

vowel with the largest F0 value, the best result was also given by $CP_{con}$ when the parametrization was performed with HRF. However, H1H2 indicated a surprisingly small error for this high-pitch vowel when IF was conducted with $CP_{bas}$ and $CP_{rem}$. The waveforms, however, were greatly distorted, but the levels of the fundamental and the second harmonic, that is, those sole spectral components used in the computation of H1H2, were only marginally affected. It is, though, worth emphasizing that the length of the glottal CP for this high-pitch vowel with $F0=310$ Hz is only ten samples (1.25 ms). This implies that the underlying assumption underlying all the three assessed IF techniques, that is, the existence of sufficiently long CP, is greatly violated. Hence, the surprisingly small value of H1H2 difference for this signal is explained mainly by the shortcomings of the simple spectral parameter rather than by the successful voice source estimation. In summary, the experiments conducted with synthetic vowels indicate that the proposed CP algorithm was the least vulnerable to the covariance frame position among the three techniques when voices of different F0 were compared.

## B. Experiment 2: Natural vowels

The standard deviations (std) and means of the H1H2 and HRF values extracted from the glottal flows computed from natural vowels of varying covariance frame positions were compared with repeated measures analyses of variance (ANOVAs). The data were analyzed with sex × method × phonation ANOVAs where "sex" included male and female sexes, factor "method" included three different CP algorithms, $CP_{bas}$, $CP_{rem}$, and $CP_{con}$, and factor "phonation" included phonation types normal and pressed. H1H2 and HRF data were analyzed with separate ANOVAs, and Newman–Keuls tests were used as a means of *post hoc* analysis for pairwise differences in the data. The standard deviations and mean values of H1H2 and HRF obtained from the 66 window positions (11 frame positions of 6 cycles) are shown in Fig. 9. The main and interaction effects of the corresponding ANOVA results are given in Table III.

The standard deviation of both H1H2 and HRF differed significantly between the IF methods. *Post hoc* analyses showed that the standard deviations of H1H2 and HRF were, on the average, smaller when the new CP method

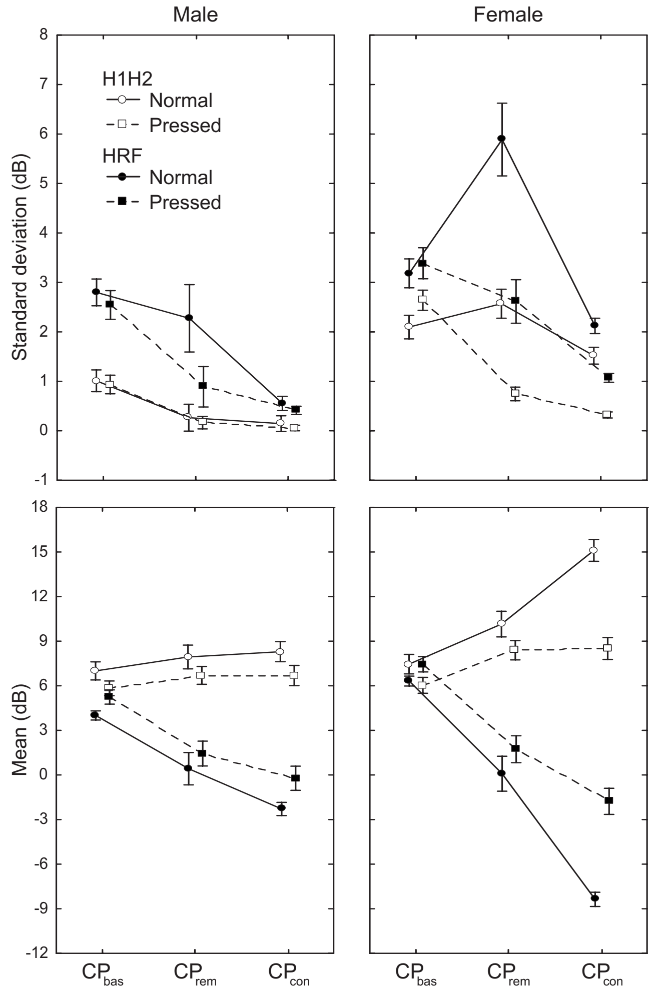

FIG. 9. Standard deviations (top panels) and means (bottom panels) of H1H2 and HRF according to the speaker sex and the type of phonation for CP analyses computed by $CP_{bas}$, $CP_{rem}$, and $CP_{con}$. Error bars represent standard error of the mean.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

TABLE III. ANOVA results for standard deviations and means of H1H2 (upper table) and HRF (lower table). The degrees of freedom (DF), Greenhouse–Geisser epsilons ($\varepsilon$), $F$-values, and the associated probability ($p$) values are shown for each ANOVA effect. Analyses were conducted for utterances produced by 13 speakers using 11 covariance frame positions per glottal cycle and 6 successive periods.

| H1H2 | | Standard deviations | | | Means | | |
|---|---|---|---|---|---|---|---|
| Effects and degrees of freedom (Df1, Df2) | | $\varepsilon$ | $F$ | $p$ | $\varepsilon$ | $F$ | $p$ |
| Sex | 1, 11 | | 83.86 | <0.001 | | 9.38 | <0.05 |
| Method | 2, 22 | 0.73 | 47.22 | <0.001 | 0.69 | 88.85 | <0.001 |
| Method × sex | 2, 22 | 0.73 | 4.12 | <0.05 | 0.69 | 37.70 | <0.001 |
| Phonation | 1, 11 | 1.00 | 10.42 | <0.01 | 1.00 | 20.43 | <0.001 |
| Phonation × sex | 1, 11 | 1.00 | 6.80 | <0.05 | 1.00 | 3.58 | ns |
| Method × phonation | 2, 22 | 0.95 | 16.26 | <0.001 | 0.68 | 23.22 | <0.001 |
| Method × phonation × sex | 2, 22 | 0.95 | 15.57 | <0.001 | 0.68 | 16.81 | <0.001 |

| HRF | | Standard deviations | | | Means | | |
|---|---|---|---|---|---|---|---|
| Effects and degrees of freedom (Df1, Df2) | | $\varepsilon$ | $F$ | $p$ | $\varepsilon$ | $F$ | $p$ |
| Sex | 1, 11 | | 49.33 | <0.001 | | 0.83 | ns |
| Method | 2, 22 | 0.61 | 26.43 | <0.001 | 0.87 | 262.27 | <0.001 |
| Method × sex | 2, 22 | 0.61 | 6.26 | <0.05 | 0.87 | 30.99 | <0.001 |
| Phonation | 1, 11 | 1.00 | 18.12 | <0.01 | 1.00 | 15.22 | <0.01 |
| Phonation × sex | 1, 11 | 1.00 | 2.86 | ns | 1.00 | 2.06 | ns |
| Method × phonation | 2, 22 | 0.77 | 12.83 | <0.001 | 0.67 | 9.29 | <0.01 |
| Method × phonation × sex | 2, 22 | 0.77 | 3.10 | ns | 0.67 | 4.58 | <0.05 |

ns=not significant

(H1H2-std=0.5, HRF-std=1.0) was used than when either $\text{CP}_{\text{bas}}$ (H1H2-std=1.6, HRF-std=3.0) or $\text{CP}_{\text{rem}}$ (H1H2-std =0.9, HRF-std=2.8) was used.

For H1H2, the difference between $\text{CP}_{\text{bas}}$ and $\text{CP}_{\text{rem}}$ was also significant. Additional effects on H1H2 and HRF variability were observed for sex and phonation type. The H1H2 and HRF standard deviations were larger for female (H1H2-std=1.6, HRF-std=3.0) than for male (H1H2-std =0.4, HRF-std=1.6) speakers. Further, the variability of H1H2 and HRF was larger for the normal (H1H2-std=1.2, HRF-std=2.7) than for the pressed (H1H2-std=0.8, HRF-std=1.8) type of phonation. Finally, significant method × sex, method × phonation, and method × phonation × sex interactions were found for both H1H2 and HRF, and a phonation × sex interaction was additionally significant for the H1H2.

The results indicated a statistically significant effect of CP method on the mean H1H2 and HRF values. The mean H1H2 and HRF values increased and decreased, respectively, when the IF algorithm $\text{CP}_{\text{bas}}$ (H1H2=6.6 and HRF=5.7) was changed to $\text{CP}_{\text{rem}}$ (H1H2=8.2 and HRF=0.9) and, then, further to the new $\text{CP}_{\text{con}}$ algorithm (H1H2=9.5 and HRF= −3.0). While HRF mean values were similar for both sexes, the average H1H2 values were larger for female (9.3) than for male (7.1) speakers. Additionally, a smaller mean H1H2 and a larger mean HRF value was observed for the pressed phonation (H1H2=7.0 and HRF=2.3) than for the normal phonation (H1H2=9.2 and HRF=0.1). Finally, significant method × sex, method × phonation, and method × phonation × sex interactions were found for both H1H2 and HRF data.

## V. CONCLUSIONS

CP covariance analysis, a widely used glottal IF method, computes a parametric model of the vocal tract by conducting linear predictive analysis over a frame that is located in the CP of the glottal cycle. Since the length of the CP is typically short, the resulting all-pole model is highly vulnerable with respect to the extraction of the frame position. Even a minor change in the frame position might greatly affect the $z$-domain locations of the roots of the all-pole model given by LP. This undesirable feature of the conventional CP analysis typically results in vocal tract models, which have roots, both real and complex, at low frequencies or roots that are located outside of the unit circle. These kinds of false root locations, in turn, result in distortion of the glottal flow estimates, which is typically seen as unnatural peaks at the instant of glottal closure, the so-called jags, or as increased formant ripple during the CP.

The present study proposed an improved version of the CP analysis based on a combination of two algorithmic issues. First, and most importantly, a constraint is imposed on the dc gain of the inverse filter prior to the optimization of the coefficients. With this constraint, linear predictive analysis is more prone to give vocal tract models that can be justified from the point of view of the source-filter theory of vowel production; that is, they show complex conjugate roots in the vicinity of formant regions rather than unrealistic resonances at low frequencies. Second, the new CP method utilizes an inverse filter that is minimum phase, a property that is not typically used in glottal IF.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

The new glottal IF method, $CP_{con}$, was compared to two CP analysis techniques by using both synthetic vowels produced by physical modeling of the voice production apparatus and natural vowels produced by male and female speakers. In summary, the experiments conducted with synthetic vowels having F0 from 105 to 310 Hz indicate that the proposed CP method gave glottal flow estimates with better robustness to the covariance frame position than the conventional CP methods. The result suggests that the parametric model of the vocal tract computed with the dc-constrained linear predictive analysis is less prone to distortion by the problem typically met in the CP analysis, namely, the involvement of samples outside the CP in the computation of the vocal tract. This problem violates the basic assumption of the CP analysis that the estimation of the vocal tract transfer function is made during the excitation-free time span. It can be argued that this violation is larger for voices of high pitch because they typically show short CPs in the glottal excitation. Violation results in the occurrence of unjustified inverse filter roots at low frequencies, which, in turn, distorts the resulting glottal flow estimates. Based on the results achieved with synthetic speech, the involvement of the dc constraint in the optimization process of the vocal tract model, however, seems to reduce this distortion and hence improve the estimation robustness with respect to the CP frame position. It must be emphasized, though, that if the amount of data samples during the glottal CP becomes extremely small, which was the case in analyzing the vowel with $F0=310$ Hz in the present investigation, distortion of the glottal flow estimates becomes large with all CP techniques.

The experiments conducted with natural speech indicate that the deviation of H1H2 and HRF due to the varying of the covariance frame position inside the glottal cycle was larger for female speech than for male vowels and the deviation was also larger in normal than in pressed phonation. These results are in line with findings reported in previous studies (e.g., Veeneman and BeMent, 1985) as well as with experiments conducted in the present investigation with synthetic speech, indicating that the robustness of the CP analysis with respect to the frame position tends to decrease for shorter CP intervals, as in higher F0 speech or in normal as opposed to pressed phonation. The proposed new CP method, importantly, gave the smallest deviation of H1H2 and HRF, suggesting that the involvement of the dc constraint reduces the sensitivity of the CP analysis to the covariance frame position and that this holds true also for natural vowels. This finding is also supported by the fact that the mean levels of H1H2 and HRF were found to be largest and smallest, respectively, when IF was computed with $CP_{con}$. In other words, the average spectral decay of the glottal flow pulse forms computed by varying the frame position was steeper with $CP_{con}$ than with the other two CP methods. This is explained by the frequency-domain effect produced by distortion represented by impulse-like jags: the larger their contribution, the flatter the spectrum.

In summary, the proposed IF method constitutes a potential means to compute the CP covariance analysis to estimate the glottal flow from speech pressure signals. It reduces distortion caused by one of the major drawbacks of the conventional CP analysis, the sensitivity of the analysis to the position of the covariance frame. The computational load of the new method is only slightly larger than that of the conventional CP method. In addition, the method can be implemented in a manner similar to the conventional one, that is, either based solely on the speech pressure signal or in a two-channel mode where an EGG signal is used to help extract the covariance frame position. Therefore, there are no obstacles in principle for the implementation of the proposed method in environments where the conventional analysis is used. One has to keep in mind, though, that the new method does not change the basic assumptions of the CP analysis, namely, that the voice source and vocal tract are linearly separable, and there is a CP of finite duration during which there is no excitation by the source of the tract.

## ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (Project No. 111848) and by the COST action 2103, "Advanced Voice Function Assessment."

[1] It is worth emphasizing that glottal pulses estimated from natural speech sometimes show fluctuation, typically referred to as "ripple," after the instant of glottal closure. This component might correspond to actual phenomena or it may result from incorrect inverse filter settings. If the pulse waveform is fluctuating after the instant of the glottal closure, it is, though, difficult, if not impossible, to define accurately which part of the fluctuation corresponds to real phenomena and which part results from incorrect IF. If, however, the flow waveform shows an abrupt peak at the end of the closing phase, such as in Fig. 1(c), and if this component is removed by, for example, a minor change in the position of the analysis frame, it is more likely that the component represents an artifact than a real phenomenon.

[2] By using Eq. (4), the gain of the vocal tract filter at dc, denoted by $G_{dc}$, is defined as the absolute value of the inverse of the frequency response of the constrained predictor at $\omega=0$: $G_{dc}=|1/C(e^{j0})|=|1/l_{dc}|$. In principle, the requirement $G_{dc}=1$ can be satisfied by assigning either $l_{dc}=1$ or $l_{dc}=-1$. Although both of these values result in vocal tract filters of equal gain at dc, they end up as different constrained transfer functions. In order to test the difference between the two values of $l_{dc}$, the glottal flows were estimated from the synthetic vowels described in Sec. III B by using H1H2 and HRF parameters described in Sec. III C and by conducting the constrained IF analysis by assigning both $l_{dc}=1$ and $l_{dc}=-1$. The results indicated clearly that the choice $l_{dc}=-1$ yielded glottal flow estimates that were closer to the original flows generated by the physical modeling approach.

[3] In the area of glottal IF, most studies analyze vowels with high first formant such as [a] or [ae]. The reason for this is the fact that the separation of the source and the tract becomes increasingly difficult from a mathematical point of view if the first formant is low. This is due to the fact that the strong harmonics at low frequencies bias the estimation of the first formant in all-pole modeling (El-Jaroudi and Makhoul, 1991). This, in turn, results in severe distortion of the glottal flow estimates.

[4] It should be noted that while synthetic vowels produced by the physical modeling approach mimic real speech production by involving source-tract interaction, this effect is not taken into account in CP analysis, which simply assumes that the source and tract are linearly separable (Strube, 1974; Wong *et al.*, 1979). The proposed dc-constrained LP is a new mathematical method to compute the vocal tract model of CP analysis, but it does not in any way change the underlying assumption of the linear coupling between the source and the tract. Therefore, the use of physically-motivated synthetic speech was justified by a need to have more realistic artificial vowels as test material, not by a goal to analyze how source-tract interaction affects different versions of the CP technique, all of which are based on the linear source-filter theory and are therefore unable to take into account the coupling between the source and the tract.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

Airas, M., and Alku, P. (**2006**). "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient," Phonetica **63**, 26–46.

Akande, O., and Murphy, P. (**2005**). "Estimation of the vocal tract transfer function with application to glottal wave analysis," Speech Commun. **46**, 15–36.

Alku, P. (**1992**). "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," Speech Commun. **11**, 109–118.

Alku, P., and Vilkman, E. (**1996**). "A comparison of glottal voice source quantification parameters in breathy, normal, and pressed phonation of female and male speakers," Folia Phoniatr Logop **48**, 240–254.

Arroabarren, I., and Carlosena, A. (**2004**). "Vibrato in singing voice: The link between source-filter and sinusoidal models," EURASIP J. Appl. Signal Process. **7**, 1007–1020.

Bäckström, T., and Alku, P. (**2006**). "Harmonic all-pole modelling for glottal inverse filtering," in CD Proceedings of the seventh Nordic Signal Processing Symposium, Reykjavik, Iceland.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (**1993**). *Nonlinear Programming: Theory and Algorithms* (Wiley, New York).

Bozkurt, B., Doval, B., D'Alessandro, C., and Dutoit, T. (**2005**). "Zeros of z-transform representation with application to source-filter separation of speech," IEEE Signal Process. Lett. **12**, 344–347.

Campedel-Oudot, M., Cappe, O., and Moulines, E. (**2001**). "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," IEEE Trans. Speech Audio Process. **9**, 469–481.

Carlson, R., Granström, B., and Karlsson, I. (**1991**). "Experiments with voice modelling in speech synthesis," Speech Commun. **10**, 481–489.

Childers, D., and Ahn, C. (**1995**). "Modeling the glottal volume-velocity waveform for three voice types," J. Acoust. Soc. Am. **97**, 505–519.

Childers, D., and Hu, H. (**1994**). "Speech synthesis by glottal excited linear prediction," J. Acoust. Soc. Am. **96**, 2026–2036.

Childers, D., and Lee, C. (**1991**). "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90**, 2394–2410.

Childers, D., and Wong, C.-F. (**1994**). "Measuring and modeling vocal source-tract interaction," IEEE Trans. Biomed. Eng. **41**, 663–671.

Cummings, K. E., and Clements, M. A. (**1995**). "Analysis of the glottal excitation of emotionally styled and stressed speech," J. Acoust. Soc. Am. **98**, 88–98.

El-Jaroudi, A., and Makhoul, J. (**1991**). "Discrete all-pole modeling," IEEE Trans. Signal Process. **39**, 411–423.

Eysholdt, U., Tigges, M., Wittenberg, T., and Pröschel, U. (**1996**). "Direct evaluation of high-speed recordings of vocal fold vibrations," Folia Phoniatr Logop **48**, 163–170.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fitch, T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522.

Flanagan, J. (**1972**). *Speech Analysis, Synthesis and Perception* (Springer, New York).

Fröhlich, M., Michaelis, D., and Strube, H. (**2001**). "SIM—Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," J. Acoust. Soc. Am. **110**, 479–488.

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

Gobl, C., and Ní Chasaide, A. (**2003**). "The role of voice quality in communicating emotion, mood and attitude," Speech Commun. **40**, 189–212.

Hertegård, S., Gauffin, J., and Karlsson, I. (**1992**). "Physiological correlates of the inverse filtered flow waveform," J. Voice **6**, 224–234.

Hirano, M. (**1974**). "Morphological structure of the vocal cord as a vibrator and its variations," Folia Phoniatr Logop **26**, 89–94.

Hirano, M. (**1981**). *Clinical Examination of Voice* (Springer, New York).

Hollien, H., Dew, D., and Philips, P. (**1971**). "Phonational frequency ranges of adults," J. Speech Hear. Res. **14**, 755–760.

Hollien, H., and Shipp, T. (**1972**). "Speaking fundamental frequency and chronologic age in males," J. Speech Hear. Res. **15**, 155–159.

Kasuya, H., Maekawa, K., and Kiritani, S. (**1999**). "Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics," in Proceedings of the International Congress on Phonetic Sciences, San Francisco, CA, pp. 2505–2512.

Klatt, D., and Klatt, L. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Krishnamurthy, A., and Childers, D. (**1986**). "Two-channel speech analysis," IEEE Trans. Acoust., Speech, Signal Process. **34**, 730–743.

Larar, J., Alsaka, Y., and Childers, D. (**1985**). "Variability in closed phase analysis of speech," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, pp. 1089–1092.

Lecluse, F., Brocaar, M., and Verschuure, J. (**1975**). "The electroglottography and its relation to glottal activity," Folia Phoniatr. **17**, 215–224.

Lehto, L., Laaksonen, L., Vilkman, E., and Alku, P. (**2008**). "Changes in objective acoustic measurements and subjective voice complaints in call-center customer-service advisors during one working day," J. Voice **22**, 164–177.

Liljencrants, J. (**1985**). "Speech synthesis with a reflection-type line analog," DS dissertation, Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden.

Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.

Markel, J., and Gray, A., Jr. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, Berlin).

Milenkovic, P. (**1986**). "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. Acoust., Speech, Signal Process. **34**, 28–42.

Miller, R. (**1959**). "Nature of the vocal cord wave," J. Acoust. Soc. Am. **31**, 667–677.

Naylor, P., Kounoudes, A., Gudnason, J., and Brookes, M. (**2007**). "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," IEEE Trans. Audio, Speech, Lang. Process. **15**, 34–43.

Oppenheim, A., and Schafer, R. (**1989**). *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).

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

Price, P. (**1989**). "Male and female voice source characteristics: Inverse filtering results," Speech Commun. **8**, 261–277.

Rabiner, L., and Schafer, R. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).

Riegelsberger, E., and Krishnamurthy, A. (**1993**). "Glottal source estimation: Methods of applying the LF-model to inverse filtering," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, MN, Vol. **2**, pp. 542–545.

Rothenberg, M. (**1973**). "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," J. Acoust. Soc. Am. **53**, 1632–1645.

Shiga, Y., and King, S. (**2004**). "Accurate spectral envelope estimation for articulation-to-speech synthesis," in the CD Proceedings of the Fifth ISCA Speech Synthesis Workshop, Pittsburgh, PA.

Stoicheff, M. L. (**1981**). "Speaking fundamental frequency characteristics of nonsmoking female adults," J. Speech Hear. Res. **24**, 437–441.

Story, B. (**1995**). "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract," Ph.D. dissertation, University of Iowa.

Story, B. (**2005**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**, 3834–3859.

Story, B., and Titze, I. (**1995**). "Voice simulation with a body-cover model of the vocal folds," J. Acoust. Soc. Am. **97**, 1249–1260.

Story, B., Titze, I., and Hoffman, E. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**, 537–554.

Strik, H., and Boves, L. (**1992**). "On the relation between voice source parameters and prosodic features in connected speech," Speech Commun. **11**, 167–174.

Strube, H. (**1974**). "Determination of the instant of glottal closure from the speech wave," J. Acoust. Soc. Am. **56**, 1625–1629.

Strube, H. (**1982**). "Time-varying wave digital filters for modeling analog systems," IEEE Trans. Acoust., Speech, Signal Process. **30**, 864–868.

Sundberg, J., Fahlstedt, E., and Morell, A. (**2005**). "Effects on the glottal voice source of vocal loudness variation in untrained female and male voices," J. Acoust. Soc. Am. **117**, 879–885.

Švec, J., and Schutte, H. (**1996**). "Videokymography: High-speed line scanning of vocal fold vibration," J. Voice **10**, 201–205.

Titze, I. (**2002**). "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model," J. Acoust. Soc. Am. **111**, 367–376.

Titze, I., and Story, B. (**2002**). "Rules for controlling low-dimensional vocal fold models with muscle activities," J. Acoust. Soc. Am. **112**, 1064–1076.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21

Titze, I., Story, B., Burnett, G., Holzrichter, J., Ng, L., and Lea, W. (**2000**). "Comparison between electroglottography and electromagnetic glottography," J. Acoust. Soc. Am. **107**, 581–588.

Titze, I., and Sundberg, J. (**1992**). "Vocal intensity in speakers and singers," J. Acoust. Soc. Am. **91**, 2936–2946.

Veeneman, D., and BeMent, S. (**1985**). "Automatic glottal inverse filtering from speech and electroglottographic signals," IEEE Trans. Acoust., Speech, Signal Process. **33**, 369–377.

Vilkman, E. (**2004**). "Occupational safety and health aspects of voice and speech professions," Folia Phoniatr Logop **56**, 220–253.

Wong, D., Markel, J., and Gray, A., Jr. (**1979**). "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Process. **27**, 350–355.

Yegnanarayana, B., and Veldhuis, N. (**1998**). "Extraction of vocal-tract system characteristics from speech signals," IEEE Trans. Speech Audio Process. **6**, 313–327.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.189.194.69 On: Thu, 18 Dec 2014 03:41:21