DC-constrained linear prediction for glottal inverse filtering

Paavo Alku, Carlo Magi, Tom Bäckström

1 Department of Signal Processing and Acoustics, Helsinki University of Technology, Espoo, Finland
paavo.alku@tkk.fi, tom.backstrom@tkk.fi

Abstract
Closed phase covariance (CP) analysis is a glottal inverse filtering method which estimates the vocal tract during the glottal closed phase. Since closed phase durations are typically short, the vocal tract computation with linear prediction is vulnerable to the covariance frame position. This study proposes a modified CP algorithm in which a DC-gain constraint is imposed in optimizing the linear predictive inverse model of the vocal tract. Results show that the algorithm improves the robustness of the CP analysis on the covariance frame position.

Index Terms: voice production, glottal inverse filtering

1. Introduction
Non-invasive estimation of the glottal excitation from speech signals calls for using glottal inverse filtering (IF). The IF 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. From this perspective, there are, firstly, such methods which are based on the gross estimation of the glottal contribution with low-order all-pole modeling which is computed during the closed and open phase of the glottal cycle [1]. Secondly, joint optimization of the glottal flow and vocal tract is possible with synthetic flow models, e.g., [2], [3]. Thirdly, it is possible to estimate the glottal flow with the closed phase (CP) covariance analysis [4], [5]. This implies assuming that there is no contribution from the source to the tract during the closed phase of the glottal cycle. By first extracting this time-span and then by using the samples of the identified closed phase, a parametric all-pole model for the vocal tract is computed with linear prediction (LP) by using the covariance analysis [6].

Since its presentation in [4], the CP technique has been a target of methodological development. The major focus of this work has concentrated on how to determine the location of the covariance frame, that is, to extract the closed phase of the glottal cycle. In order to define this important time span, an approach based on conducting a series of sliding covariance analyses is typically used. Strube [4] used this approach and identified the glottal closure as an instant when the frame was in a position yielding the maximum determinant of the covariance matrix. Wong et al. [5] computed the minimum of the normalized squared prediction error in determining the closed phase. Plumpe et al. [7] proposed an idea in which sliding covariance analyses are computed and formant frequency modulations between the glottal open and closed phase are used as a means to define the optimal frame position. Akande and Murphy [8] suggested a technique in which the estimation of the vocal tract is improved by first removing the effect of the glottal contribution by filtering speech with a dynamic, multi-pole high-pass filter. 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. Finally, if electromyography (EMG) is available, it is possible to use the so-called two channel analysis in which the position and duration of the closed phase is extracted from EMG [9].

Several previous studies have indicated that glottal flow estimates computed by the CP analysis vary greatly depending on the position of the covariance frame, e.g., [10], [11]. Given the fundamental assumption of the method, that is, the computation of the vocal tract model during an excitation-free time span, this feature of the CP analysis is understandable. The true length of the glottal closed phase 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 LP filter typically results in vocal tract models which have roots, both real and complex, at low frequencies. These kinds of false root locations, in turn, result in distortion of the glottal flow estimates which is seen as unnatural peaks at the instant of glottal closure, the so-called “jags” [5], as depicted by an example in Figure 1.

Figure 1: An example of a glottal flow estimated by the conventional CP method. Incorrect roots of the inverse filter at low frequencies distort the waveform at closure instants.

Previous CP methods have tried to tackle the problem caused by false root locations of vocal tract models by exploiting techniques that aim to improve the extraction of the covariance frame position. In this study, 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 pre-defined value for the DC (direct current) 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 comprise poles in such positions in the z-domain (e.g. on the positive real axis) which are difficult to interpret from the point of view of the classical source-tract theory of vowel production.
2. Method

The conventional CP analysis involves modeling the vocal tract with an all-pole filter defined according to the classical LP based on the covariance criterion [6]. Optimization is based on the MSE (mean square error) criterion, which adjusts the filter coefficients mathematically so that the resulting all-pole spectrum matches accurately the high-energetic formant regions of the speech spectrum. However, it is worth emphasizing that the conventional covariance analysis does not pose any additional information to be used 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 unrealistic from the point of view of Fant’s source-tract theory of vowel production and its underlying theory of tube acoustics [12]. This kind of mismatch between the all-pole filter optimized by the conventional covariance analysis and the vocal tract model structure assumed in the source tract theory 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 [9], [10], [11].

The computation of the conventional covariance analysis, however, can be modified by using the concept of constrained linear prediction. The key idea is to impose such restrictions on the LP polynomials prior to the optimization that can be justified, for instance, by the source-tract theory of voice production. Intuitively, this means that instead of allowing the LP 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 the present study, these restrictions are imposed simply on the DC gain of the linear predictive inverse filter, whose coefficients are optimized based on mathematical derivations described below.

In the conventional LP, the residual signal can be expressed in matrix from as follows

\[ e_n = x_n + \sum_{k=0}^{p} a_k x_{n-k} = \sum_{k=0}^{p} a_k x_{n-k} = a^T x_n \]  
\[ (1) \]

where \( a = [a_0 \ldots a_p]^T \) with \( a_0 = 1 \) and the signal vector is \( x_n = [x_n \ldots x_{n-p}]^T \). The coefficient vector \( a \) 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 [6]. By denoting this time span with \( 0 \leq n \leq N-1 \), the prediction error energy \( E(a) \) can be written as

\[ E(a) = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} a^T x_n x_n^T a = a^T \sum_{n=0}^{N-1} x_n x_n^T a^T = a^T \Phi a \]  
\[ (2) \]

where matrix \( \Phi \) is the covariance matrix defined from speech samples as

\[ \Phi = \sum_{n=0}^{N-1} x_n x_n^T \in \mathbb{R}^{(p+1)(p+1)} \]  
\[ (3) \]

The optimal coefficients can be computed by minimizing the prediction error energy \( E(a) \) with respect to \( a \). This yields

\[ a = \Phi^{-1} \Sigma^{-1} u \]  
\[ (4) \]

where \( \Sigma = (u^T \Phi^{-1} u)^{-1} \) is the residual energy given by the optimized predictor and \( u = [1 \ 0 \ldots 0]^T \).

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

\[ C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j0}) = C(1) = \sum_{k=0}^{p} c_k = l_{DC} \]  
\[ (5) \]

where \( c_k, \ 0 \leq k \leq p \), are the filter coefficients of the constrained inverse filter and \( l_{DC} \) is a predefined real value for the gain of the filter at DC. Using matrix notation, the DC-constrained minimization problem can now be formulated as follows

\[ \text{minimize } c^T \Phi c, \]  
\[ (6) \]

subject to \( \Gamma c = b \), where \( c = [c_0 \ldots c_p]^T \) is the filter coefficient vector with \( c_0 = 1 \), \( b = [1 \ l_{DC}] \), and \( \Gamma = a \times 1 \) 2 constraint matrix defined as

\[ \Gamma = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T \]  
\[ (7) \]

The covariance matrix defined in Eq. 3 is positive definite. Therefore, the quadratic function to be minimized is convex, and the Lagrange multiplier method [13] can be used in order to solve the minimization problem in Eq. 6. This procedure begins with the definition of a new objective function

\[ \eta(c,g) = c^T \Phi c - 2g^T (\Gamma^T c - b) \]  
\[ (8) \]

where \( g = [g_1 \ g_2]^T \) 3 > 0 is the Lagrange multiplier vector. It is well known that \( c \) minimizes objective function \( \eta \) if and only if it satisfies the linear equation

\[ \nabla_2 \eta(c,g) = 2(\Phi c - \Gamma g) = 0, \]  
\[ (9) \]

where \( \nabla_2 \) is the gradient operator with respect to vector \( c \). By solving the vector \( c \) from the following group of equations,

\[ \begin{bmatrix} \Phi & -\Gamma \end{bmatrix} g = 0 \]  
\[ (10) \]

the following equation is finally obtained for the optimal coefficients of the constrained inverse filter:

\[ c = \Phi^{-1} \Gamma^T (\Phi^{-1} \Gamma^{-1})^{-1} b. \]  
\[ (11) \]

In summary, the optimal DC constrained inverse filter, a FIR filter of order \( p \), is obtained by solving for the vector \( c \) according to Eq. 11, in which the covariance matrix \( \Phi \) is defined by Eq. 3 from the speech signal \( x_n \) and matrices \( b \) and \( \Gamma \) are defined by Eqs. 6 and 7. The transfer function of the FIR determined this way, to be used in the following in glottal inverse filtering, can be written by using the solved coefficient vector \( c \)

\[ C(z) = \sum_{k=0}^{p} c_k z^{-k}. \]  
\[ (12) \]

3. Materials and experiments

In order to evaluate the performance of the new CP analysis technique, experiments were conducted using natural speech and EGG. Our purpose was to investigate whether the new method makes inverse filtering less vulnerable to the position of the covariance frame. The experiments are described by first depicting the procedures used in data collection.
3.1. Speech and EGG recordings

Simultaneous speech pressure waveform and EGG signals were recorded from four female and four male subjects. The subjects produced the vowel /a/ using normal sustained phonation. Speech pressure waves were captured by a condenser microphone (Brüel & Kjær 4188) and the EGG was recorded simultaneously (Glottal Enterprise MC2-1). The mouth-to-microphone distance was 40 cm and its value was checked prior to each phonation. The propagation delay of the acoustic signal from the glottis to the microphone was estimated by using the vocal tract length of 15 cm and 17 cm for females and males, respectively, and the speed of sound value of 350 m/s. After recordings, the sampling frequency of the signals was down-sampled to 8 kHz.

3.2. Experiments

The performance of the proposed CP analysis algorithm was compared to the most widely used conventional 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 inverse filtering [5], [14]. The robustness of both CP analyses to the position of the covariance frame was evaluated by varying the frame position near the instant of glottal closure, which was estimated as the time instant when the EGG derivative reached a negative peak within a glottal cycle. One such instant of glottal closure was extracted from each vowel. The covariance frame with the duration of 30 samples (3.75 ms) was then positioned to start from this instant. The position of the covariance frame was varied ten times by incrementing its beginning by one sample. Hence, ten different glottal flow estimates were obtained by both the conventional CP analysis and by the proposed new technique for each vowel. The duration of the glottal flow estimates was 300 samples (37.5 ms). All the analyses were conducted by using $p=12$ as the order of the vocal tract model, and the lip radiation effect was modeled with a first order FIR whose zero was at $z=0.98$. The value of the DC-constraint (Eq. 5) was set to $l_{DC}=-1.0$ (i.e., the DC level of the vocal tract models given by the proposed CP method was always 0 dB). Finally, the estimated glottal flows were parameterized using a frequency domain measure, the harmonic richness factor (HRF) [15], which is well-suited for the present purpose because it can be computed automatically without any user adjustments.

4. Results

Harmonic richness factor is shown as a function of the covariance frame position in Figures 2 and 3 for female and male speakers, respectively. In both figures, the upper graphs depict HRF values computed by the conventional CP analysis and the lower panel shows results obtained by the proposed technique. Standard deviations of the HRF values obtained by changing the frame position are given for each speaker in Table 1.

As shown in Table 1, changing the covariance frame position caused deviations in HRF values which were larger for each speaker when inverse filtering was conducted with the conventional CP method than with the proposed method. The results of male voices showed in general small differences between the glottal flows computed by the two CP techniques. However, for female vowels the variation of the HRF value as a function of the covariance frame position was clearly larger with the conventional CP method. For female voices, the difference between the glottal flow estimates computed by the two techniques was large particularly in such cases when the covariance frame was located more than five samples after the instant of glottal closure extracted from EGG. These major results are demonstrated by representative examples of the glottal flow estimates shown in Figures 4 and 5.

![Figure 2: Harmonic richness factor (HRF) when the covariance frame position is incremented. CP-analysis was computed with (a) the conventional, and (b) the new method. Four female speakers.](image)

![Figure 3: Harmonic richness factor (HRF) when the covariance frame position is incremented. CP-analysis was computed with (a) the conventional, and (b) the new method. Four male speakers.](image)
to reduce this distortion and hence improve the estimation accuracy and its robustness with respect to the CP frame position.

Figure 4: Examples of glottal flows computed by (a) the conventional and (b) the new CP method. The position of the covariance frame was seven samples after the minimum peak of the EGG derivative.

Figure 5: Examples of glottal flows computed by (a) the conventional and (b) the new CP method. The position of the covariance frame was eight samples after the minimum peak of the EGG derivative.

Table 1. Standard deviations (in dB) of HRF values obtained from glottal flows computed by varying the covariance frame position in the conventional CP analysis (CP_con) and in the new method (CP_new). Female speakers: F1-F4, male speakers: M1-M4.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>CP_con</th>
<th>CP_new</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>13.8</td>
<td>1.4</td>
</tr>
<tr>
<td>F2</td>
<td>10.8</td>
<td>1.9</td>
</tr>
<tr>
<td>F3</td>
<td>7.6</td>
<td>3.2</td>
</tr>
<tr>
<td>F4</td>
<td>9.9</td>
<td>1.5</td>
</tr>
<tr>
<td>M1</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>M2</td>
<td>1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>M3</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>M4</td>
<td>3.9</td>
<td>0.4</td>
</tr>
</tbody>
</table>

5. Conclusions

Closed phase (CP) covariance analysis, a widely used glottal inverse filtering method, computes a parametric model of the vocal tract by conducting LP analysis over a frame that is located in the closed phase of the glottal cycle. Since the length of the closed phase is typically short, the resulting all-pole model is highly vulnerable with respect to the extraction of the frame position. The present study proposes a new type of CP analysis in which a constraint is imposed on the DC gain of the linear predictive inverse filter prior to the optimization of the coefficients. With this constraint, the LP analysis is more prone to give vocal tract models that show complex conjugate roots in the vicinity of formant regions rather than unrealistic resonances at low frequencies. Our preliminary results indicate the proposed idea reduced the occurrence of impulse-like “jags” in the estimation of the glottal flow with closed phase analysis. This improvement of the estimation accuracy was largest for female voices.

6. Acknowledgements

This study was supported by the Academy of Finland (project no. 111848) and by the COST action 2103, “Advanced Voice Function Assessment”.

7. References