Quasi closed phase analysis for glottal inverse filtering

Manu Airaksinen1, Brad Story2, Paavo Alku1

1Department of Signal Processing and Acoustics, Aalto University, Finland
2Speech Acoustics Laboratory, University of Arizona, USA

manu.airaksinen@aalto.fi

Abstract

This study presents a new glottal inverse filtering (GIF) technique based on the closed phase analysis over multiple fundamental periods. The proposed Quasi-Closed Phase Analysis (QCP) method utilizes Weighted Linear Prediction (WLP) with a specific Attenuated Main Excitation (AME) weighting function that attenuates the contribution of the glottal source in the linear prediction model optimization. This enables the use of the autocorrelation criterion in linear prediction in comparison to the conventional covariance criterion used in the closed phase analysis. The proposed method was compared to previously developed methods by using a synthetic vowel database created with a physical modeling approach. The obtained objective measures show that the proposed method improves the GIF performance for both low- and high-pitched vowels.

Index Terms: speech analysis, glottal source estimation, glottal inverse filtering, linear prediction

1. Introduction

The glottal volume velocity waveform, or in shorter terms the glottal flow, serves as the main source of excitation for voiced speech. The estimation of the glottal flow is of particular interest, because voiced segments constitute the most prevalent category in most languages [1]. The estimation of the glottal flow can be computed with glottal inverse filtering (GIF), a technique which is based on the use of properly adjusted antiresonances that cancel the effects of the vocal tract and lip radiation from a recorded microphone signal. GIF benefits from the fact that the analysis is non-intrusive, it can be performed solely based on the given speech sound pressure signal, and it can be implemented to perform in a completely automatic manner.

Various GIF methods have been proposed starting from the 1950s (see [2] for a historical review of GIF). In the closed phase (CP) covariance analysis [3], the vocal tract filter is estimated by computing LP (linear prediction) analysis with the covariance criterion over speech samples that are located in the closed phase of the glottal cycle. The Iterative Adaptive Inverse Filtering (IAIF) method [4] utilizes an iterative scheme based on a priori knowledge of the overall shape of the glottal flow and vocal tract envelopes. The mixed-phase decomposition based Complex Cepstrum Decomposition (CCD) [5], also known as the Zeros of the Z-Transform (ZZT) method [6], separates the glottal source and the vocal tract based on decomposing speech into minimum-phase and mixed-phase components. The Mean Squared Phase (MSP) method estimates a best fitting Liljencrants-Fant (LF) model pulse to the signal to obtain the glottal flow estimate [7]. Several previous studies have indicated that the accuracy of GIF methods is, in general, sufficient in estimating the glottal flow from low-pitched male voices [5, 3]. However, when the fundamental frequency (F0)

of speech increases the performance of GIF methods generally deteriorates.

A new glottal inverse filtering algorithm is presented in this study. The method, Quasi-Closed Phase Analysis (QCP), is based on the principles of the closed phase analysis, that is, the estimation of the vocal tract during the closed phase of glottal flow. Differently from the CP method, however, the proposed technique does not compute the vocal tract filter with the covariance method from few samples located in the closed phase. Instead, QCP takes advantage of weighted linear prediction (WLP) in order to utilize all the data samples of the analysis frame but puts more emphasis on the samples which are located in the closed phase. The proposed method is tested by comparing its accuracy in estimating the glottal flow from high-pitched speech. Evaluation is conducted by utilizing test vowels synthesized by a physical modeling approach. The evaluation involves comparing the proposed method with three known techniques: CP, IAIF, and CCD.

This paper is organized as follows: Section 2 describes the WLP analysis which is used in the proposed GIF method in modeling of the vocal tract. Section 3 deals with the new GIF technique, the QCP method. Sections 4 and 5 present the experiments and their results, respectively, and the conclusions are presented in Section 6.

2. Weighted Linear Prediction

Weighted linear prediction (WLP) was first proposed by Ma et al. in 1993 [8]. It is one of several modified LP methods that have been proposed to provide all-pole models that are less affected by F0 and its harmonics [9, 10, 11]. WLP was chosen in the present study as a method to model the vocal tract because it enables, in a straightforward manner, emphasizing the contribution of certain data samples in the computation of linear prediction.

2.1. Model optimization

WLP uses a similar prediction model as conventional LP:

\[ s_n = \sum_{i=1}^{p} a_i s_{n-i} + e_n, \]  

(1)

where \( s_n \) denotes the \( n \)th sample of a speech wave, \( e_n \) is the \( n \)th sample of an excitation (residual) wave, \( a_i \) is the \( i \)th predictor coefficient, and \( p \) is the LP order. WLP differs from conventional LP in the sense that it imposes temporal weighting on the square of the residual:

\[ E = \sum_{n=n_1}^{n_2} \left( s_n - \sum_{i=1}^{p} a_i s_{n-i} \right)^2 W_n, \]  

(2)
where $E$ is the total residual energy, $n_1 = 1$ and $n_2 = N + p$ for the autocorrelation case, $N$ is the frame length (in samples), and $W_n$ is a weighting function. It is important to note that $W_n$ is not an ordinary time-domain windowing function (e.g. Hamming, Hann etc.) that is widely used, for example, in the computation of autocorrelation terms in LP. In stead, $W_n$ represents a temporal weighting function which have the contribution of selected squared residual samples can be down-graded in the optimization of the filter coefficients. The parameters $a_i$ are obtained by setting the partial derivatives of $E$ with respect to $a_i$ to zero, resulting in the following normal equations:

$$\sum_{k=1}^{p} a_k \sum_{n=n_1}^{n_2} W_n s_{n-k} s_{n-i} = \sum_{n=n_1}^{n_2} W_n s_n s_{n-i},$$

which can be expressed in vector form as

$$\left(\sum_{n=n_1}^{n_2} W_n \cdot s_n s_n^T\right) a = \sum_{n=n_1}^{n_2} W_n \cdot s_n s_n,$$  

where $a = [a_1, a_2, \ldots, a_{n_2}]^T$ and $s_n = [s_{n-1}, s_{n-2}, \ldots, s_{n-p}]^T$. Differently from conventional LP, WLP does not guarantee filter stability even though the autocorrelation criterion is used.

### 2.2. Weighting function

In [8], WLP was computed by using the short-term energy (STE) waveform as the weighting function:

$$W_{n,STE} = \sum_{t=0}^{M-1} s_{n-1-i}^2,$$

where $M$ denotes the length of the energy window. The STE function is easy to compute and it has been shown to roughly emphasize the contribution of speech samples that are located in the closed phase of the glottal cycle [11]. The STE function has been used in a GIF application utilizing stabilised weighted linear prediction (SWLP) [11] in [12]. An alternative weighting waveform, Attenuated Main Excitation (AME), was, however, recently proposed in [13]. The AME weighting function, shown in Fig. 1, is based on forming a temporal waveform which down-grades the contribution of speech samples that are located in the vicinity of the main excitation of the vocal tract near the instant of glottal closure (GCI). In [13], the AME function was used in WLP to estimate formants from high-pitched speech. It was shown that the AME weighting reduced the effect of the glottal source in the computation of WLP models. Consequently, more accurate WLP-based formant estimates were obtained especially for high-pitched speech.

The computation of the AME function is illustrated in Fig. 1 and described as follows: For each fundamental cycle, $W_n$ is equal to 1.0 in all time instants except in the vicinity of GCI where $W_n$ becomes attenuated and equals a small positive constant, denoted by $d$ ($d \approx 0, d > 0$). The value of $d$ used in this paper is $10^{-5}$, which is small enough to ignore the open phase and large enough to keep numerical stability. $W_n$ changes smoothly between $d$ and 1.0 by following a linear ramp whose length can be fixed to a constant of $N_{ramp}$ samples or alternatively be set according to the pitch-proportional Ramp Quotient (RQ) which is defined as $RQ = N_{ramp}/T$. The position and length of the attenuated section are defined by two pitch proportional quantities, $PQ$ (Position Quotient) and $DQ$ (Duration Quotient). $PQ$, defined as $PQ = N_{pq}/T$, indicates the relative starting position of the non-attenuated section of $W_n$ from the previous GCI within the glottal cycle. $DQ$, defined as $DQ = N_{dq}/T$, is a measure for the relative length of the non-attenuated section of $W_n$ inside a glottal cycle.

### 3. Proposed method

The flow diagram of the proposed QCP method is shown in Fig. 2. The algorithm consists of four main computational blocks. First, GCIs are detected from the speech signal to be later used in building the AME function. GCIs can be extracted, for example, by using electroglottography [14] or specific GCI detection algorithms such as [15] or [16]. This study uses accurate GCI information obtained from the synthesis parameters. Second, the AME function is created as explained in Section 2.2 by using the estimated GCIs, from which the fundamental period $T$ is also obtained. Third, the WLP-based vocal tract model, denoted by $V(z)$ in Fig. 2, is defined according to Eq. 4 by using the AME function as $W_n$. On average, the AME function attenuates the effect of (quasi) open phase samples in the optimization of the vocal tract filter coefficients, thus performing a quasi closed phase analysis over multiple fundamental periods of the speech signal $s[n]$ in an analysis frame. Fourth, an estimate for the glottal flow derivative, denoted by $g[n]$ in Fig. 2, is obtained by inverse filtering the input speech signal with $V(z)$.

In order to optimize the AME parameters, a large set of test vowels were created by the Lijencrants-Fant (LF) pulseform [17] as excitation. 625 different LF pulses were created by using values reported in [18]. $F_0$ varied from 75 Hz to 405 Hz with a 10 Hz increment. The vocal tract was adjusted as in [19] by simulating three different vowels ([a], [e], [i]). In total, the procedure yielded 63750 test vowels.

The optimal combination of the AME parameters was determined by varying the values of $PQ$, $DQ$, and $N_{ramp}$ or $RQ$, and by searching for the minimum of an overall error measure. This overall error measure was a combination of several parametrization schemes that have been used in GIF. The involved error measures were the average absolute relative errors of NAQ [20], QOQ [21], HRF [22], the average absolute error of H1H2 [23], and the mean squared error. The analysis resulted in the following optimal AME parameter values: $PQ = 0.05$, $DQ = 0.5$, $d = 0.15$ and $N_{ramp} = 10$.
$DQ = 0.7$, and a fixed $N_{ramp} = 7$. Additional tuning to the AME parameters was performed by testing the hypothesis that the optimal $DQ$ parameters vary depending on $F_0$. With $PQ$ and $N_{ramp}$ set fixed, the optimal $DQ$ parameters were determined for each $F_0$ bin of the test set. It was found that the optimal $DQ$ parameter indeed clearly shifted depending on $F_0$, and this trend was roughly approximated to the final pitch-adaptive version of the AME function. The assigned AME function values for the baseline and final AME weighting functions are presented in Table 1.

4. Experiments

The accuracy of the proposed QCP method was compared to a set of selected GIF techniques. The comparison was conducted using test vowels produced by a physical modeling approach. The test vowels were inverse filtered with the selected GIF methods and the obtained glottal flow estimates were expressed using glottal flow parameterization methods that could be automatically computed from the obtained source waveforms. The used parameters were NAQ, QQQ, H1H2, and HRF. Finally, the GIF methods were compared by computing the absolute relative error between the parameter value that was extracted from the source signal produced by the physical model and the one estimated by GIF.

Section 4.1 introduces the selected GIF methods and their configurations, and Section 4.2 describes the synthetic vowel test set that was obtained via physical modeling of human speech production.

4.1. GIF methods to be compared and their configurations

The following three prominent GIF techniques were selected for the comparison: Closed Phase Covariance method (CP) [3], Iterative Adaptive Inverse Filtering (IAIF) [4], and the Complex Cepstrum Decomposition (CCD) [5]. As the fourth benchmark, the WLP analysis with STE weighting function was used. All the GIF analyses were computed with the sampling rate of 8 kHz, the vocal tract order $N = 7$, and a first-order pre-emphasis with its zero at $z = 0.99$. The length of the analysis frame was 30 ms. For each inverse filter, $z$-domain root locations were solved and the roots outside the unit circle were replaced with their minimumphase reciprocals. The values of GCI and $T$ were obtained directly from the glottal flow signals generated by the physical models (see Section 4.2).

The CP method was implemented using the traditional covariance LP analysis by using the two pitch-period analysis for vowels with $F_0 \geq 200$ Hz [24]. The IAIF method was implemented as a pitch-asynchronous version with prediction orders $p = 10$ and $g = 4$. The CCD method was obtained from the GLOAT toolbox [25], which is a straight implementation of the method as described in [5]. As in [8], the STE function (Eq. 5) was computed with $M = 12$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$PQ$</th>
<th>$DQ$</th>
<th>$N_{ramp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.05</td>
<td>0.7</td>
<td>7</td>
</tr>
<tr>
<td>Pitch-adaptive</td>
<td>0.05</td>
<td>0.95</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.55</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.7</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1: The obtained parameters values for the AME weighting function.

Figure 3: Average error measures with their 95% confidence intervals of the glottal source parameters for the tested GIF methods. The “Low” group consists of samples with $F_0 \leq 200$ Hz, and the “High” group consists of samples with $F_0 > 200$ Hz.

4.2. Test vowels synthesized via physical modeling

Accuracy assessment of a GIF method calls for using synthetic speech with known glottal flow signals. Therefore, previous investigations have exclusively utilized voices generated by source-filter modeling. This kind of evaluation, however, might not be truly objective because the test material and the methods to be assessed are based on similar models. Therefore, the present study takes advantage of a different approach, the physical modelling of the speech production mechanism, in generation of the test utterances with known glottal flow signals.

A computational model of the speech production system was used to generate vowels representative of an adult male, adult female, and approximately a five-year-old child. The voice source component used consisted of a kinematic representation of the medial surface of the vocal folds [26, 27] for which the $F_0$, surface bulging, adduction, length, and thickness are control parameters. As the vocal fold surfaces are set into vibration the model produces a glottal area signal that is coupled to the acoustic pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations [28]. The resulting glottal flow was determined by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis. The vocal tract shape was specified by an area function representative of an [a], [i], and [æ] vowel. The acoustic wave propagation in the subglottal and supraglottal airspaces was computed with a wave-reflection model that included energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips [29]. Detailed parameter values used for the physical model are presented in [13].

Each of the vowels for the male, the female, and the child were generated with eight $F_0$ values, ranging from 100 Hz to
Table 2: Average errors of the glottal source parameters for the physical model test set. Minimum values for each group are marked with bold font.

<table>
<thead>
<tr>
<th>Method</th>
<th>NAQ (%)</th>
<th>QOQ (%)</th>
<th>H1H2 (dB)</th>
<th>HRF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Low</td>
<td>High</td>
<td>All</td>
</tr>
<tr>
<td>QCP</td>
<td>33</td>
<td>8</td>
<td>48</td>
<td>13</td>
</tr>
<tr>
<td>WLP-STE</td>
<td>63</td>
<td>13</td>
<td>92</td>
<td>28</td>
</tr>
<tr>
<td>CCD</td>
<td>119</td>
<td>81</td>
<td>143</td>
<td>34</td>
</tr>
<tr>
<td>CP</td>
<td>52</td>
<td>15</td>
<td>76</td>
<td>24</td>
</tr>
<tr>
<td>IAIF</td>
<td>73</td>
<td>28</td>
<td>100</td>
<td>29</td>
</tr>
</tbody>
</table>

5. Results

The absolute relative errors of NAQ, QOQ, and HRF as well as the absolute error of H1H2 are given in numerical forms in Table 2. The same results are expressed graphically in Fig. 3. Data in both illustrations are separated into three categories (“Low”, “High”, and “All”) according to F0 and averaged within each category. Data were separated into low-pitched and high-pitched speech with a F0 threshold of 200 Hz.

The results indicate that QCP performs best by a large margin in all cases except for HRF in the low-pitched samples. CP gave the second best overall performance. This was not surprising given the fact that the positioning of the covariance window was able to be done reliably by using the glottal area function of the physical sound synthesis technique. The performance of WLP with STE weighting was substantially lower than that of QCP hence indicating that the choice of \( W_m \) has a large effect when the WLP analysis is used in GIF. Even though STE weighting resulted in clearly lower estimation accuracy than AME weighting, the former method achieved, on average, the third best scores among all compared methods.

The poor performance of the CCD method is surprising, because, based on previous studies [5], it was expected to give good accuracy in estimating the glottal source. A possible explanation is that the mixed-phase model of speech production assumed in CCD does not hold true for the synthetic vowel that were generated by the physical modeling approach. In addition, the CCD method only estimates the opening and closing phases of the glottal excitation, thus excluding the return phase. In some cases this causes significant deterioration in the estimated glottal source parameter values.

Examples of glottal sources estimated from a male and female talker are shown, respectively, in the upper and lower panel of Fig. 4. The figure depicts results obtained by two methods (QCP and IAIF) together with the source waveform generated by the physical model. The QCP method shows waveforms that are very close to those generated by the physical model, whereas the glottal flows computed by IAIF indicate fluctuations during the closed phase for the male speaker, and the form of the flow during the open phase is distorted for both speakers.

450 Hz in 50-Hz increments. Although the full range of these F0s is unlikely to be produced by either the male, female, or the child, conducting the experiment with the entire range was desirable for ease in comparison. Vowel duration was 0.4 seconds and F0 was maintained constant during the utterance. Sampling frequency was finally set to 8 kHz.

6. Conclusions

A new glottal inverse filtering technique, Quasi Closed Phase Analysis (QCP), that performs a closed phase type analysis over a time frame of multiple fundamental periods was introduced in this paper. The QCP method uses a specific weighting function to attenuate the contribution of the (quasi) open phase in the computation of the Weighted Linear Prediction (WLP) coefficients, resulting in a good estimate of the vocal tract transfer function. The method was evaluated with objective measures against well known GIF methods. The objective measures were obtained by inverse filtering a synthetic vowel database constructed via physical modeling of human speech production.

The results indicate that the proposed method outperforms the other four GIF methods in a vast majority of cases, both with low and high pitched vowels. There are two main drawbacks to the proposed method. First, the method requires information about the glottal closure instants (GCIs) to form the appropriate WLP weighting function, which will cause degradation in its performance in real-world situations where accurate GCI data is not available. Second, the proposed method does not guarantee filter stability, which must be assessed in GIF applications where all-pole synthesis is needed.

7. Acknowledgements

The research leading to these results has received funding from the European Communities Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678 and from the Academy of Finland (project no. 256961).
8. References


