Analysis of iterative adaptive and quasi closed phase inverse filtering techniques on OPENGLOT synthetic vowels

Marc Freixes, Joan Claudi Socoró and Francesc Alías

GTM – Grup de recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Sant Joan de La Salle, 42, 08022 Barcelona, Spain
{marc.freixes,joanclaudi.socoro,francesc.alias}@salle.url.edu,

1. Introduction

Three-dimensional (3D) source-filter models have overcome the limitations of their one-dimensional counterparts for an accurate generation of voice through articulatory speech synthesis [1, 2]. To date, these 3D-based numerical simulations have been able to generate vowels [3], diphthongs [4, 5], vowel-consonant-vowel sequences, including fricatives [6, 7]. Despite some preliminary attempts to produce expressive voice, this research line is still in its infancy. For instance, in [8], a very preliminary attempt to generate, using a 3D-based finite element method (FEM), happy and aggressive [a] vowels was introduced by only modifying the spectral tilt of a Liljencrants-Fant (LF) glottal flow model [9] input to MRI-based vocal tract (VT) geometries. That work was later extended in [10], confirming the relevance of properly simulating higher order modes for resembling expressiveness, particularly for tense phonations and/or high fundamental frequencies. Moreover, by modifying the VT characteristics, effects such as the so-called singing formant can be also simulated using FEM [11]. Therefore, the proper modelling and subsequent tuning of both the VT and the glottal source (GS) are key to produce expressive speech, considering their different weight depending on the target speaking style [12, 13].

In order to obtain the GS and the VT estimates from speech, different Glottal Inverse Filtering (GIF) techniques have been introduced in the literature. The main GIF approaches can be grouped into three main categories [14]: mixed-phase decomposition, iterative and adaptive inverse filtering, and closed-phase inverse filtering. Complex Cepstrum Decomposition [15] and Zeros of the Z-Transform [16] are two GIF techniques based on mixed-phase decomposition. The former bases on the cepstral rationale for source-filter decomposition on the frequency domain. The latter distinguishes the VT from the GS components, differentiating the location of zeros on the unit circle, being the one within the circle from VT while the ones outside it from GS as they contain the maximum-phase component of speech (i.e. glottal open phase) [14]. Nevertheless, these straightforward approaches have been overcome by the following techniques, at the cost of higher complexity in their tuning.

On the other hand, the Iterative Adaptive Inverse Filtering (IAIF) algorithm [17] is designed to obtain the GS and VT transfer functions based on linear prediction through a two-step iterative process that entails an initial gross and a subsequent refined estimation. Several variants of this method can be found in the literature, considering a pitch-synchronous analysis plus high-pass filtering [18], replacing the initial step of IAIF by an Iterative Optimal Pre-emphasis (IOP-IAIF) approach [19], or by limiting the filter order of both initial gross and refined stages, an approach denoted as Glottal Flow Model (GFM)-IAIF [20]. Finally, within the third group, it is worth mentioning the Quasi Closed Phase (QCP) analysis, which focuses on the estimation of Closure Glottal Instant (CGIs) of the glottal flow to decouple GS from the VT response [21, 22]. In [23], a QCP variant was introduced by including spectral tilt compensation as a post-processing step to minimise the presence of residual spectral cues from the glottal source on the vocal tract estimation (hereafter denoted as ST-QCP).

The accuracy of GIF methods is typically evaluated using different error measures, such as [14, 24, 25, 20]: Root Mean Square (RMS), Normalized Amplitude Quotient (NAQ), Quasi-Open Quotient (QQO), Harmonic Richness Factor (HRF), H/H2, spectral distortion, parabolic spectrum parameter, and spectral tilt, among others. These studies are generally conducted on clean/noise real or synthetic speech generated by the authors themselves (see e.g., [14, 24, 25, 20]). Hence, comparing the results is difficult as they are highly dependent on the specific conducted experiments. In this regard, the recently developed OPENGLOT [26] may become instrumental as it has been specifically conceived to evaluate GIF methods.

In this work, we conduct a thorough comparison of state-of-the-art IAIF- and QCP-based methods on the OPENGLOT repository I, evaluating their performance through several well-known error measures for different phonation types, fundamental frequencies and vowels.
The paper is organised as follows. Section 2 presents the evaluated GIF methods in a nutshell. Next, Sections 3 and 4 detail the conducted experiments and the obtained results. Finally, the conclusions and future work are presented in Section 5.

2. Glottal inverse filtering methods

This section briefly describes the main elements of the state-of-the-art GIF methods considered in this work.

2.1. IAIF-based approaches

2.1.1. Original IAIF approach

IAIF is a GIF technique based on linear prediction [17], which follows a two-step iterative process, being more effective than similar techniques in very low signal-to-noise ratio environments [14]. The initial gross estimation of both glottal source and vocal tract all-pole transfer functions are optimized through a second refinement stage. While in the first iteration the pre-emphasis order is set to 1, during the refinement stage it can be increased to \( N_g \geq 1 \). Usually vocal tract order \( N_v \) value is fixed for both iterations [25, 20].

Lip radiation is modelled as a simple preemphasis FIR filter with a leaky integration coefficient \( d \) that can be inverted to obtain the glottal flow from its derivative when inverse vocal tract has been applied to the speech signal [20, 25].

In [18], a pitch-synchronous IAIF variant was proposed. Besides repeating the IAIF analysis of the speech signal pitch-synchronously using the pitch marks obtained from the initial IAIF output, it also includes a high-pass linear phase IIF filter (HPF) with cutoff frequency of 30 Hz to remove undesirable fluctuations of the estimated glottal source.

2.1.2. IOP-IAIF

IOP was proposed in [19] to replace IAIF’s initial step to increase the ability to remove the spectral tilt due to glottal source bias in the first gross estimation of the vocal tract. Instead of modelling the spectral tilt with a simple first-order all-pole filter, an iterative procedure is implemented to cancel the one-delay correlation of the speech signal by monitoring the obtained linear prediction coefficient. The iterative process finishes when this coefficient becomes lower than a given threshold (typically 0.001). This leads to an unconstrained first stage preemphasis equivalent filter order.

2.1.3. GFM-IAIF

In [20], the GFM-based variant of the iterative adaptive inverse filtering technique was proposed. In this approach, basically both the iterative pre-emphasis of IOP-IAIF and the second refined glottal source filter are limited to order 3 in the aim of reducing the algorithm complexity and improving the vocal effort estimation of real speech. Also, the initial HPF is not applied here. Therefore, GFM-IAIF does fix \( N_g = 3 \), while \( N_v \) and \( d \) are the parameters to be selected.

2.2. QCP-based approaches

2.2.1. Original QCP approach

QCP analysis is an inverse filtering technique that builds on the principles of closed phase covariance analysis [27], where the estimation of the vocal tract uses all the samples of the analysed speech frame instead of only those where the glottis is closed [21]. An Attenuated Main Excitation (AME) weighting function is defined to emphasize the speech samples where closed phase occurs, allowing for an accurate estimation of the vocal tract by means of weighted linear prediction. Differently from the previous IAIF-based described techniques, the estimation of Closure Glottal Instants (CGLs) is mandatory for the AME weighting function definition.

Apart from the vocal tract all-pole filter order \( N_v \) and the lip radiation coefficient \( d \), also three parameters related to the AME function must be defined: Position Quotient (PQ), the relative starting position of the non-attenuated section; Duration Quotient (DQ), relative length of the non-attenuated section; and Ramp Quotient (RQ), that encodes the relative duration of the transition ramp that connects attenuated with non-attenuated sections of each speech period.

2.2.2. ST-QCP approach

According to [23], a post-processing step can be added to the original QCP technique so as to compensate the remaining spectral tilt of the vocal tract filter prior to inverse filtering. To that end, the residual spectral tilt of the VT filter resulting from QCP algorithm is estimated using a first-order linear prediction filter and subsequently removed from the estimated glottal flow.

2.3. Parameters’ tuning

As aforementioned, the implementation of the IAIF- and QCP-based GIF methods entails the adjustment of different parameters, which are obtained through some kind of optimization process [20, 25]. The averaged root mean square error between the reference glottal flow signal and a normalized version of its estimated counterpart is considered in the optimization process. When this measure is averaged across all signal periods, then the median absolute waveform error (MAE-Wave) is obtained [24]. Nevertheless, it is to note that before computing the error measure, a normalization of the estimated glottal is typically performed through time-alignment besides both absolute scale and DC adjustment (e.g., see [25]).

3. Experiments

The speech signals are analysed at a constant frame rate using 50 ms Hanning windows with 50% of overlap for the five considered GIF methods. In addition, frame-based processing ensuring continuity of signal from frame-to-frame is applied maintaining initial and final memory conditions for every filter within the applied methods. GCI, which are used for both QCP-based approaches but also for the computation of MAE-Wave error and the performance error measures detailed in Section 3.3, have been estimated using the speech event detection based on the residual excitation and mean-based signal technique described in [28].

3.1. OPENGLOT dataset

OPENGLOT is composed of four repositories [26]: three of which contain synthetic speech signals obtained throughout different experiment setups together with their ground truth glottal flow inputs, and one more that contains real speech plus electroglossotogram signals.

In this work, the first repository of OPENGLOT has been used as it contains synthetic vowels sampled at \( F_s = 8 \) kHz generated with LF excitation [9], considering a digital 8th-order all-pole filter model of the vocal tract (defined with 4 for-
Figure 1: Distances between the original and the estimated glottal flow signals. First column depicts overall results, while columns from 2 to 4 contain the results detached by phonation type, F0 and vowel, respectively. The differences between GIF methods are statistically significant with $p < 0.01$ for all cases except those marked with * ($0.01 \leq p < 0.05$) and ns (not significant, $p \geq 0.05$).

ments). Concretely, the simulated vocal tracts correspond to six different vowels $\{a, e, i, o, u, ae\}$, while glottal flow signals vary taking into account 14 fundamental frequencies ($F_0$ values from 100 to 360 Hz, with step of 20 Hz) and four phonation types (from lax to tense: whispery, breathy, normal and creaky) changing the control parameters of the LF model. As a result, the dataset used comprises 2016 synthetic vowels obtained from 336 reference glottal flow signals.

3.2. GIF parameters tuning

As aforementioned, the parameters of both GIF methods families have to be adjusted. To this aim, a grid-search optimization strategy has been applied for the IAIF-based approaches, which comprises not more than 4 parameters, being two of them discrete. On the other hand, for tuning the parameters of QCP variants, a surrogate-based optimization has been considered as it has been proved to accelerate the tuning process as the number of parameters to tune is significantly higher than in IAIF-based approaches (i.e., three continuous parameters need to be tuned for the AME function) [29], making their tuning through grid-search almost unfeasible.

Following [20], we have defined several parameter ranges and steps for the grid-search strategy. The vocal tract order $N_v$ has been varied from $\lfloor F_s/1000 \rfloor - 2 = 6$ to $\lfloor F_s/1000 \rfloor + 6 = 14$ with increments of 1, which are enough for the modelling of the speech vowels resonances. Moreover, the lip radiation coefficient $d$ has been varied from 0.8 to 0.99, with increments of 0.01. Finally, the glottal source order $N_g$ has been varied from 3 to 6 with increments of 1 for IAIF and IOP-IAIF, while it has been kept to order 3 for GFM-IAIF according to its definition. The HPF binary activation has been considered as another optimization parameter for both IAIF and IOP-IAIF techniques. On the other hand, the surrogate-based optimization has been implemented by considering the following set of boundaries per parameter, based on [22]: $N_v \in [6, 14]$, $d \in [0.8, 0.99]$, $DQ \in [0.4, 1]$, $PQ \in [0, 0.2]$ and $RQ \in [0, 0.2]$. The estimated glottal flow considered in the optimization process is time-aligned and subsequently scaled. The time offset applied is obtained through peak-picking of the autocorrelation between the estimated and the ground-truth glottal flow. The scaling is applied pulse-by-pulse using the GCIs as time boundary references, considering absolute scale and DC offset normalization factors obtained through the minimization of the total squared error between the ground-truth and the three-factor normalized glottal flow estimation.

3.3. Error measures

Several well-known time- and frequency-domain performance measures are computed at pulse level to evaluate the quality of the estimated glottal flow signals with respect to a given ground-truth, specifically: i) RMS distance, from which MAE-Wave error driving optimal parameter selection has been computed; ii) NAQ, a time-domain based dimensionless parameter that measures the relative duration of the glottal closing
phase that has proved useful predict the voice quality variation along the breathy-to-pressed continuum [30, 31]; iii) H1H2 (in dB), which is the difference between the amplitudes of the first and second harmonics of the glottal flow spectrum, a measure widely used to characterize voice quality [14]; iv) HRF (in %) that compares the amplitudes of the harmonics located above the first harmonic with the amplitude of the first harmonic, showing a good correlation with the phonation types [25, 31]; v) Spectral Tilt (ST in dB/decade) that represents the spectral slope of the GS computed from the linear regression of the harmonic amplitudes below 5 kHz [20], and it has been found to correlate with voice quality aspects of speech such as its phonation type [31] and its vocal effort [32].

Following [20], all previous measures have been computed at pulse level, comparing the glottal flow output of a given GIF technique with respect to the corresponding reference signal. Absolute error is computed for H1H2, HRF, and ST, while absolute relative error (in %) is calculated for NAQ. The obtained results have been analysed globally and also per vowel, phonation type and F0 value. Regarding the latter, three frequency ranges have been considered to group the results [22]: low-pitch (i.e., \( F_0 < 190 \) Hz), mid-pitch (i.e. 190 Hz \( \leq F_0 < 280 \) Hz) and high-pitch (i.e. \( F_0 \geq 280 \) Hz) signals.

### 4. Results

Figure 1 depicts the distribution of the five performance measure errors obtained for the analysed GIF methods in the OPENGLOT repository I. The Wilcoxon signed-rank paired test [33] with Holm-Bonferroni correction has been applied to every pair of GIF methods’ distributions in order to determine if the differences between different GIF approaches are statistically significant. It is to note that 94.7% out of the 700 distributions’ comparisons are statistically significant (with \( p < 0.05 \)), being their main differences highlighted in the following paragraphs.

The first column of Figure 1 shows the global distributions for all the signals of the corpus, while the median values are depicted in Table 1. It can be observed that QCP-based approaches obtain lower RMS, NAQ, H1H2 and HRF errors. On the other hand, in terms of ST, the IOP and GFM variants outperform the original IAIF and QCP, while ST-QCP is the most competitive GIF approach.

When looking at results in terms of the different phonation types (see the 2nd column of Figure 1), it can be observed that RMS and NAQ errors decrease for all methods inasmuch as the voice is less tense (e.g., whispery), being QCP-based methods the ones with lower values, while H1H2 does increase for whispered and IAIF-based methods. In terms of HRF and ST, the original IAIF and QCP variants entail larger errors for relaxed phonation, whereas IOP-IAIF presents larger values for tense phonation. Moreover, both GFM-IAIF and ST-QCP achieve more stabilized values across phonation values, being ST-QCP the one with clearly better performance in almost all phonation types, despite in some cases (e.g., in breathy phonation, IOP-IAIF and QCP presents no statistical differences for ST).

Regarding the performance of the inverse filtering techniques for the low, mid and high F0 intervals (see the 3rd column of Figure 1), the general tendency is to obtain higher RMS, NAQ and H1H2 errors for higher F0s, while the opposite happens for HRF. It is remarkable that both QCP variants obtain significantly lower values for high F0 values, especially for RMS and H1H2. Only IOP-IAIF presents a better performance in terms of spectral tilt for low pitched vowels, while QCP-based methods remain the best in mid and high F0 ranges.

Finally, from the results per vowel (see the 4th column of Figure 1), it can be observed that close rounded vowels /o/ and /u/ present higher values of H1H2 for the IAIF-based methods, being also higher for /u/ in terms of RMS and for /o/ in terms of NAQ. Again, the QCP variants present a better behaviour than their IAIF counterparts for almost all vowels and error measures, being overtaken only in certain cases by IOP-IAIF (for /o/ obtains better HRF and for /u/ obtains better NAD, HRF and ST) or GFM-IAIF (for /u/ obtains better ST).

### 5. Conclusions

In this work, a performance comparison between IAIF- and QCP-based state-of-the-art glottal inverse filtering methods has been conducted on a specifically designed reference dataset named OPENGLOT, concretely on its repository I composed of synthetic vowels with different phonation types and F0s. After properly adjusting the parameters of each inverse filtering variant, the results show that, in general terms, QCP-based methods achieve lower glottal flow error measures than the IAIF-based counterparts, for RMS, NAQ and HRF metrics, being almost all differences statistically significant. Regarding the spectral tilt measure, both IAIF variants show significant improvements with respect to original IAIF, except for IOP-IAIF when creaky or close rounded vowels /o/ and /u/ are analysed. It is to note that the spectral tilt compensation applied to the QCP output improves the original method performance with similar (or even better) levels than IAIF-based techniques for relaxed phonation types like breathy and whispery, low-pitched vowels.

Future work will focus, on the one hand, on extending the conducted analyses to real speech to contrast and complement the results obtained on synthetic vowels. On the other hand, besides taking into account the glottal flow estimation, further investigations should also include the vocal tract response estimation in order model them for the generation of reliable expressive speech through 3D-based articulatory simulations.

### 6. Acknowledgements

This research has been partially funded by the Agencia Estatal de Investigación (AEI) through the FEMVoQ project (PID2020-120441GB-I00/AEI/10.13039/501100011033).
References


