applied sciences

IMPACT FACTOR 2.7

CITESCORE 4.5

Article

Evaluation of Glottal Inverse Filtering Techniques on OPENGLOT Synthetic Male and Female Vowels

Marc Freixes, Luis Joglar-Ongay, Joan Claudi Socoró and Francesc Alías-Pujol

Special Issue

IberSPEECH 2022: Speech and Language Technologies for Iberian Languages

Edited by

Prof. Dr. Francesc Alías, Dr. José Luis Pérez Córdoba, Dr. Zoraida Callejas Carrión and Prof. Dr. António Joaquim da Silva Teixeira

MDPI

https://doi.org/10.3390/app13158775

applied sciences

MDPI

Article

Evaluation of Glottal Inverse Filtering Techniques on OPENGLOT Synthetic Male and Female Vowels †

Marc Freixes *,‡, Luis Joglar-Ongay ‡, Joan Claudi Socoró and Francesc Alías-Pujol

Human-Environment Research (HER), La Salle—Universitat Ramon Llull, Sant Joan de la Salle, 42, 08022 Barcelona, Spain; luis.joglar@salle.url.edu (L.J.-O.); joanclaudi.socoro@salle.url.edu (J.C.S.); francesc.alias@salle.url.edu (F.A.-P.)

* Correspondence: marc.freixes@salle.url.edu

† This paper is an extended version of our paper published in the conference IberSPEECH2022.

‡ These authors contributed equally to this work.

Abstract: Current articulatory-based three-dimensional source–filter models, which allow the production of vowels and diphtongs, still present very limited expressiveness. Glottal inverse filtering (GIF) techniques can become instrumental to identify specific characteristics of both the glottal source signal and the vocal tract transfer function to resemble expressive speech. Several GIF methods have been proposed in the literature; however, their comparison becomes difficult due to the lack of common and exhaustive experimental settings. In this work, first, a two-phase analysis methodology for the comparison of GIF techniques based on a reference dataset is introduced. Next, state-of-the-art GIF techniques based on iterative adaptive inverse filtering (IAIF) and quasi closed phase (QCP) approaches are thoroughly evaluated on OPENGLOT, an open database specifically designed to evaluate GIF, computing well-established GIF error measures after extending male vowels with their female counterparts. The results show that GIF methods obtain better results on male vowels. The QCP-based techniques significantly outperform IAIF-based methods for almost all error metrics and scenarios and are, at the same time, more stable across sex, phonation type, F0, and vowels. The IAIF variants improve the original technique for most error metrics on male vowels, while QCP with spectral tilt compensation achieves a lower spectral tilt error for male vowels than the original QCP.

Keywords: performance evaluation; glottal inverse filtering; glottal source; phonation types; speech analysis; OPENGLOT

Citation: Freixes, M.; Joglar-Ongay, L.; Socoró, J.C.; Alías-Pujol, F. Evaluation of Glottal Inverse Filtering Techniques on OPENGLOT Synthetic Male and Female Vowels. *Appl. Sci.* **2023**, *13*, 8775. https://doi.org/10.3390/app13158775

Academic Editors: Douglas O'Shaughnessy and Javier Hernando

Received: 23 June 2023
Revised: 18 July 2023
Accepted: 25 July 2023
Published: 29 July 2023

1. Introduction

Voice generation based on articulatory speech synthesis has been significantly improved by considering three-dimensional (3D) source–filter models, surpassing the limitations of their one-dimensional counterparts [1,2]. These advanced models have demonstrated their ability to generate various speech utterances, including vowels [3], diphthongs [4,5], and vowel–consonant–vowel sequences incorporating fricatives [6,7]. Despite these accomplishments, the exploration of expressive voice synthesis using these 3D-based numerical simulations is still in its early stages due to its great complexity. In fact, expressiveness in speech, which implies the communication of non-linguistic information about emotions, speaking styles, and mood, among other things, can be conveyed through prosodic (including the modification of the fundamental frequency, F0, duration and energy of phonemes) and voice quality modelling (embracing factors ranging from formant tuning due to vocal tract variations to the modification of the parameters of the glottal source, such as phonation or aspiration noises) (see [8] for a comprehensive review). A preliminary endeavour in this direction was presented by Freixes et al. in [9], where the authors introduced a method that modified the spectral tilt of the glottal flow signal produced using the Liljencrants–Fant (LF) model's [10] input to vocal tract geometries

obtained from magnetic resonance imaging [4], aiming to generate happy and aggressive /a/ vowels. Building upon this initial work, a following study emphasised the importance of accurately simulating higher-order modes to achieve expressiveness [11], especially for tense phonations and high fundamental frequencies. Additionally, the manipulation of the vocal tract characteristics using simulations based on finite element methods (FEM) has enabled the production of effects such as the singing formant in 3D-based articulatory voice generation [12]. Therefore, from these works, it can be concluded that, for the production of expressive speech, a proper model and adjustment of the vocal tract response and the glottal source signal is of paramount importance, as is considering their varying relevance depending on the target speaking style [13,14].

The estimation of the glottal source (GS) and the vocal tract (VT) transfer function from speech signals therefore becomes instrumental. Glottal inverse filtering (GIF) methods have been widely used for this purpose, although alternative approaches based on a joint estimation process have also been proposed [15,16]. Female voices are known to be especially difficult to analyse compared to males [17], especially due to the higher mean fundamental frequency of the GS as well as the smaller larynx and slightly shorter VT [18], which entail formant–frequency differences between both sexes [19], among other effects.

The GIF techniques can be broadly categorized into three main groups, as outlined by Drugman et al. [17]: mixed-phase decomposition, iterative and adaptive inverse filtering, and closed-phase inverse filtering. Among the GIF techniques based on mixed-phase decomposition, complex cepstrum decomposition (CCD) [20] and zeros of the z-transform [21] are notable approaches. CCD relies on the cepstral decomposition of the source–filter model; that is, it entails working on the so-called quefrency domain. On the other hand, zeros of the z-transform distinguishes the VT and GS components of speech by analyzing the location of zeros on the unit circle. Zeros inside the unit circle correspond to the VT resonances, while those outside the circle represent the GS response, which contains the maximum-phase component of speech, i.e., the glottal open phase [17]. However, despite the effectiveness of these straightforward approaches, more advanced techniques have emerged, albeit at the cost of increased complexity in their tuning and implementation.

The iterative adaptive inverse filtering (IAIF) algorithm proposed by Alku et al. [22] offers an alternative approach to estimate the GS and VT transfer functions based on classic linear prediction coding (LPC). This algorithm employs a two-step iterative process that involves an initial gross estimation followed by a refined estimation of GS and VT transfer functions. A subsequent work by Alku [23] included high-pass filtering (HPF) and pitch-synchronous analysis, showing potential improvements, though the results were only analysed qualitatively. More recently, several variants of the IAIF algorithm have been developed and documented in the literature. One such variant, known as iterative optimal pre-emphasis (IOP-IAIF), replaces the initial gross step of IAIF with an iterative pre-emphasis approach [24]. Additionally, a modified version called the glottal flow model (GFM)-IAIF restricts the GS filter order in both the initial gross estimation and the refined stages [25].

Within the third category of GIF techniques, we can find those approaches supported by the estimation of the glottal closing and/or opening instants. Closed-phase covariance analysis uses both estimations to derive an all-pole VT transfer function from speech samples within the closed-phase time region, to this way estimate the glottal flow signal [26]. However, as Wong et al. state in their seminal paper [26], this technique relies on estimations of glottal closure instants (GCIs) and glottal opening instants, being the latter less reliable than the former. Focusing only on the estimation of GCIs, thus simplifying the process, the quasi-closed phase method (QCP) [27] performs VT estimation based on a weighting function that, while uses the whole set of signal samples, emphasizes on the speech samples within the closed-phase region [27,28]. As a potential method to optimize the QCP method's performance, Seshadri et al. in [29] appended a spectral tilt compensation module as a post-processing step to minimize the residual spectral cues from the glottal source (hereafter denoted as ST-QCP).

The performance evaluation of GIF methods typically involves considering diverse error measures; the most commonly employed are [17,25,30,31] the root-mean-square (RMS), normalized amplitude quotient (NAQ), quasi-open quotient (QOQ), harmonic richness factor (HRF), H1H2 (relationship between the first and second harmonics), spectral distortion, parabolic spectrum parameter, and spectral tilt, among others. These measures serve as quantitative indicators of the accuracy of the estimated glottal source signal after applying a specific GIF technique.

Despite considering common error metrics, these evaluations are typically conducted using either real or synthetic speech data generated by the authors themselves (see, e.g., [17,25,30,31]). Therefore, the reliable analysis of the results reported in the literature becomes quite challenging, making the direct comparison between GIF techniques, at least, intricate, or even unfeasible due to the high dependency of the specific experimental settings considered in each study. Table A1 presents the main characteristics of the speech databases used for GIF evaluation in previous literature, listed in terms of pitch range, vowels, type of dataset (synthetic or real speech), sample frequency, and phonation type. To our knowledge, some of the listed databases are not publicly available, which limits the ability to draw comparisons with other methods with exactly the same conditions. To tackle this problem, the recent publication of the OPENGLOT [32] dataset becomes instrumental, especially since it has been specifically designed by researchers from this research field to assess different GIF methods on the same common reference environment. The main pillars of OPENGLOT are twofold: first, it provides a representative variety of test signals for GIF evaluation and, second, it is an open dataset; thus, it allows the evaluation and comparison of any GIF method with respect to previously developed benchmark techniques.

In this paper, first, a two-stage analysis methodology for the comparison of glottal inverse filtering techniques based on a common reference dataset is introduced. Second, current state-of-the-art GIF techniques based on IAIF and QCP approaches, which have been partially compared in the literature, are exhaustively evaluated following this analysis methodology on OPENGLOT Repository I [32] by computing various well-established GIF error measures. As the original repository only contains male formant frequencies for producing vowels with different phonation types and fundamental frequencies, the experiments are extended by including the female counterparts from the original reference study [19]. The paper is organised as follows. Section 2 presents the analysis methodology and the evaluated GIF methods in a nutshell. Next, Sections 3 and 4 detail the conducted experiments and the obtained results, which are discussed in Section 5. Finally, the conclusions and future work are presented in Section 6.

2. Analysis Methodology and GIF Methods

This section is devoted to describing the analysis methodology followed to compare different GIF methods in a common dataset. Moreover, the state-of-the-art GIF techniques considered in this work are briefly introduced, highlighting the parameters that are tuned for the speech signal's GS and VT decomposition.

2.1. Analysis Methodology

The designed analysis methodology, as depicted in Figure 1, can be divided into two main phases: parameter tuning and error computation. The same speech database is considered as input for both steps. The procedure is repeated for each utterance in the database. The output of the parameter tuning stage is the best parameter configuration for the considered GIF technique given an error metric, whereas a set of typical glottal error measures are obtained as results of the second stage. In both cases, the estimated glottal flow signal, obtained from the GIF method, is evaluated with respect to the glottal flow ground truth, which is included in the dataset's speech audio files. Notice that the methodology includes a block devoted to computing GCIs from the speech signal, which is necessary for those GIF methods that rely on this information as an anchor point. Finally, the glottal error measures, subsequently used for conducting statistical analyses, are computed at the pulse level (based on GCIs) to compare the different GIF methods.

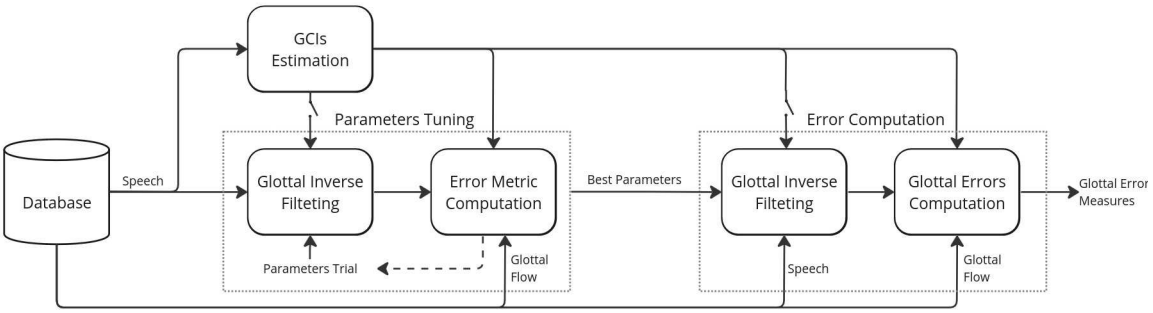


Figure 1. Block diagram of the analysis methodology followed to compare the performance of glottal inverse filtering methods on a common reference speech dataset.

2.2. IAIF-Based Approaches

2.2.1. IAIF

IAIF is a well-known technique to estimate the glottal source in speech signals. It is based on LPC without the need for the detection of the glottal closure phase (e.g., GCI and/or glottal opening instants, from which the percentage of voiced speech periods where the glottis is closed can be derived) [22]. This method is particularly effective in low-signal-to-noise-ratio environments compared to other similar techniques [17]. IAIF involves a two-step iterative process, where the initial estimation of both the GS and VT all-pole transfer functions is followed by a refinement stage. In each iteration, the effect of GS is firstly cancelled by inversely filtering the original speech with a low-order pre-emphasis LPC filter. From its output, the coarse estimation of VT is obtained with a higher-order LPC analysis. The glottal flow estimation derived from the inverse filtering of the original speech signal with this first VT coarse estimation is used as input in the second iteration to improve the accuracy of the low-GS-order LPC model. The glottal source order (henceforth denoted as N_g), fixed at 1 during the first iteration, can be increased during the refinement stage. The vocal tract order (hereafter denoted as N_v) is typically kept constant for both iterations [25,31]. A simple pre-emphasis FIR filter with a leaky integration coefficient is used to model lip radiation (hereafter lip radiation coefficient is denoted as d).

Finally, we also consider including a HPF with a 30 Hz cut-off frequency to remove undesirable fluctuations in the estimated glottal source [23].

2.2.2. IOP-IAIF

The IOP-IAIF variant was proposed by Mokhtari and Ando [24] to improve the cancellation of the spectral tilt of the glottal source remaining from the initial estimation of the vocal tract. It is based on replacing the initial estimation of glottal source with an iterative procedure [24]. This process defines an unconstrained pre-emphasis filter order devoted to minimizing the one-delay correlation of the speech signal at each iteration. The number of iterations applied to a given speech frame is that which obtains a coefficient value lower than a given threshold (typically 0.001). In this IAIF-based approach, in addition to this initial estimation of GS based on an iterative optimization, the other processes related to the coarse estimation of VT as well as the second iteration that obtains a refinement of both GS and VT remain exactly the same as the original IAIF approach. Hence, apart from the VT order and the lip radiation coefficient d, the glottal source is also modelled at the refinement stage using an all-pole filter of a certain order ($N_g \geq 1$).

2.2.3. GFM-IAIF

The GFM-IAIF variant was designed by Perrotin and McLoughlin to reduce the original IAIF algorithm complexity and improve the vocal effort estimation of real speech [25]. As in the IOP-IAIF technique, the GFM-IAIF technique builds on the original IAIF approach, in which the two iterations (a coarse iteration followed by a refinement iteration) both include low-order GS and high-order VT LPC-based modeling. It is based on limiting the

glottal source's linear prediction filter order in both the initial and refinement estimation steps to a value of $N_g = 3$. Therefore, in this case, only two parameters need to be adjusted: N_v and d.

2.3. QCP-Based Approaches

2.3.1. QCP

QCP was proposed by Airaksinen et al. [27] as a new glottal inverse filtering technique that improves on the results from previous closed-phase analysis algorithms [26] using the weighted linear prediction (WLP) technique proposed by Ma et al. [33], thus allowing for more accurate LPC-based models of speech. This was made possible through the definition of an attenuated main excitation (AME) weighting function that emphasizes the closed-phase time regions. Within the closed-phase region of the glottis, the voiced speech signal is solely influenced by VT, which allows fair estimations to be obtained. QCP takes advantage of this fact while enabling the processing of unconstrained length speech signals of sustained vocalizations. In constrast to IAIF-based techniques, in the QCP approach, only one high-order VT estimation is performed, and the glottal flow signal estimation is directly obtained while inverse filtering the original speech signal after cancelling the lip radiation effect.

In contrast to IAIF-based methods, the AME weighting function of the QCP technique needs the computation of GCIs, which, together with the following three parameters, define its specific shape: the position quotient (PQ), which represents the relative starting position of the non-attenuated section; the duration quotient (DQ), which represents the relative length of the non-attenuated section; and the ramp quotient (RQ), which defines the relative duration of the transition ramp that connects attenuated and non-attenuated sections of each voiced speech period. Then, QCP can be tuned to a specific speech signal with these three control parameters, the vocal tract all-pole filter order N_v, and the lip radiation coefficient d.

2.3.2. ST-QCP

The vocal tract filter obtained with QCP can still present a residual spectral tilt. In order to improve this aspect, a compensation procedure was proposed by Seshadri et al. [29] using a first-order linear prediction filter to transfer the residual spectral tilt from the vocal tract to the glottal flow's estimated signal. In this paper, the spectral tilt compensation is performed as follows. First, the original VT frequency transfer function is computed, and from this, the least squares 1-order LPC fitting based on [34] is performed; finally, the QCP-based estimation of the glottal flow signal is filtered, with the obtained filter presenting the compensated glottal flow estimation.

3. Experiments

The analysis carried out on the speech signals was based on a constant frame rate with a 50 ms Hanning window and a 50% overlap across the five considered GIF methods. To ensure accurate results, only the stationary part of the vowels was considered, excluding the initial and final 5% of the audio files. Furthermore, to ensure the continuity of the signal frame-to-frame, a processing method that preserved the initial and final memory state for every filter over the applied methods was applied. On the other hand, the estimation of GCIs, which play a key role in both QCP-based approaches and the calculation of the error metrics (detailed in Section 3.3), was obtained using a speech event detection technique called *SEDREAMS*, which is based on the residual excitation and mean-based signal [35].

The code used for the experiments can be found at this public repository: https://github.com/SpeechSalleBcn/inverse-filtering-evaluation (first version published on 23 June 2023).

3.1. OPENGLOT Dataset

OPENGLOT is composed of several speech datasets and was conceived to become a coherent and common environment for the evaluation of GIF methods. It was developed and released in 2019 by Alku et al. [32]. In this work, the performance of the GIF methods

was evaluated using the first repository of OPENGLOT as a common reference. This dataset consists of a collection of synthetic vowels obtained by filtering glottal flow signals generated with an LF model [10]. The audio files, sampled at 8 kHz, include both the synthetic vowels and the glottal flow signals that constitute the ground truth.

Specifically, 56 glottal flow signals were generated as excitation to cover four different phonation types (from lax to tense: whispery, breathy, normal, and creaky) and an F0 range between 100 and 360 Hz with steps of 20 Hz. On the other hand, the VTs were modelled as digital 8th-order all-pole filters defined with 4 formants. Six different vowels were considered (*a*, *e*, *i*, *o*, *u*, *æ*) using the male formant frequencies extracted from Gold and Rabiner [36], which, in turn, refer to the previous study conducted by Peterson and Barney [19]. For a more comprehensive assessment, the corresponding female vowels have been also generated considering the female formant frequencies also reported in [19]. As a result, the experiments have been conducted on a total of 672 vowels. The code used for the generation of the OPENGLOT Repository I vowels is available at the main OPENGLOT website (http://research.spa.aalto.fi/projects/openglot, accessed on 5 May 2023).

3.2. GIF Parameter Tuning

Each of the evaluated GIF methods considers several parameters that should be fine-tuned to obtain the optimal results. During the first phase of the analysis methodology, grid-search optimisation was conducted for each method and speech signal in the dataset, driven by the median absolute waveform error (MAE-Wave) as the optimization error metric (see Figure 1). This metric is computed by averaging the root-mean-square (RMS) error between the normalised estimated glottal flow and the ground truth across all the speech signal periods [30]. This normalisation is twofold, consisting, first, of a time alignment by peak-picking the autocorrelation against the ground-truth glottal flow. Secondly, it is normalised pulse-by-pulse using the GCIs as time marks by applying a scale and the DC offset normalization factors obtained by minimising the total squared error between the ground truth and the estimated glottal flow.

Table A2 presents the GIF methods and the parameters explored in the previous literature. The parameter ranges used in the present work were chosen accordingly. Furthermore, none of the reviewed studies included an analysis of all the IAIF and QCP variants.

Table 1. Parameters tuned for each GIF method evaluated. Vocal tract order is referred to as N_v, glottal source order as N_v, and lip radiation as d, HPF represents a high-pass filter flag, DQ, PQ and RQ denote duration, position and ramps quotient, whilst ST stands for the spectral tilt compensation binary flag. The ranges are written as initial value:increment:final value.

	N_v	N_g	d	HPF	DQ	PQ	RQ	ST
IAIF	6:1:14	3:1:6	0.8:0.01:0.99	0/1				
IOP-IAIF	6:1:14	3:1:6	0.8:0.01:0.99					
GFM-IAIF	6:1:14	3	0.8:0.01:0.99	0/1				
QCP	6:1:14	3:1:6	0.8:0.01:0.99		0.4:0.05:1	0:0.025:0.2	0:0.05:0.2	0
ST-QCP	6:1:14	3:1:6	0.8:0.01:0.99		0.4:0.05:1	0:0.025:0.2	0:0.05:0.2	1

As mentioned in Section 2, there are two common parameters for all methods that must be defined: the vocal tract order N_v and the lip radiation coefficient d. Following Perrotin and McLoughlin's study [25], the N_v coefficient ranged from $[Fs/1000] - 2 = 6$ to $[Fs/1000] + 6 = 14$ with increments of 1, which is sufficient for modeling vowel vocal tract resonances. Moreover, the d coefficient varied from 0.8 to 0.99, with increments of 0.01. Furthermore, IAIF, IOP-IAIF, and GFM-IAIF also had the glottal source order (N_g) parameter. For the first two methods, a range from 3 to 6 with increments of 1 was considered, whilst in the case of the GFM-IAIF method, it was fixed at 3 according to Perrotin and McLoughlin definition [25]. In addition, for IAIF and IOP-IAIF, the HPF activation was also considered as an optimisation parameter. Lastly, the parameters of the QCP, based on Airaksinen et al. [28], ranged as follows: DQ from 0.4 to 1 with a step of

0.05, PQ from 0 to 0.2 with a step of 0.025, and RQ from 0 to 0.2 with a step of 0.05. Table 1 summarizes all the parameter ranges and values for each GIF method.

3.3. Error Measures

For the assessment of the compared GIF techniques, we used a set of well-known temporal and frequency-domain performance measures computed at the pulse level. A total of 28.193 (14.097 for female and for 14.096 male, respectively) pulses were considered. These measures allowed us to evaluate the estimated signals of glottal flow in comparison with their respective ground-truth counterparts from a global perspective but also for each subset of speech features (e.g., phonation types, fundamental frequencies, or vowel type).

First, the relative RMS distance (in %) between normalized and ground-truth glottal signals was obtained for every considered period of the signal according to the CGIs [30]. Moreover, the following four error measures related to voice quality were also calculated for both the estimated and ground-truth glottal signals: (i) average normalized amplitude quotient (NAQ, in %), defined as the relative duration of the glottal closing phase, which has been used in the estimation of voice quality variation along the breathy-to-pressed continuum [31,37]; (ii) H1H2 (in dB), a measure widely used to characterize voice quality [17] and defined as the difference between the amplitudes of the first and second harmonics of the glottal flow spectrum; (iii) the harmonic richness factor (HRF, in dB), which is defined as the ratio of the sum of the amplitudes at the harmonics in the glottal waveform to the amplitude of the component at the fundamental frequency and which has shown good correlation with the phonation types [31,38]; and (iv) spectral tilt (ST, in dB/decade), which describes the spectral slope of the GS using a simple linear regression of the harmonic amplitudes below 5 kHz [25] and has also been found to correlate well with phonation type [38] as well as vocal effort [39].

Following other similar studies, such as [25,31], the NAQ, H1H2, HRF and ST measures computed at the pulse level for both the glottal flow output of a given GIF technique and the corresponding reference signal were used to obtain a certain error measure; this was, specifically, the absolute error (in dB) for H1H2, HRF, and ST, while it was the absolute relative error (in %) for NAQ. The obtained results were analysed globally and also per vowel, phonation type, and F_0 value. In all cases, the results for both the male and female vocal tracts were also compared in order to analyse their influence on the set of selected GIF methods.

4. Results

This section describes the results obtained from the conducted experiments to evaluate and compare the performance of the considered IAIF- and QCP-based GIF methods. Five measure errors were used for this evaluation: RMS, NAQ, H1H2, HRF, and ST error. Figures 2–5 show the distribution in boxplots of these measures for the evaluation of the five GIF approaches, separated by sex and analysing the results by method globally, as well as grouped by F0, phonation type, and vowel, respectively. Figure 3 presents the results by F0, grouping them into three F0 intervals [28] composed of 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, while Figure A1, in turn, presents the results for the complete sweep of F0 values from 100 to 360 Hz.

The comparison between the obtained results for every pair of GIF methods along each of the following analyses was supported with the Wilcoxon signed-rank paired test [40] with the Holm–Bonferroni correction, and the outcomes of all these tests are included in Appendix C.

4.1. Global Results

Figure 2 depicts the global boxplot distributions of the evaluation error metrics for the speech signals of the dataset for each GIF method divided in two columns by sex. Moreover, their median values are included in Tables 2 and 3 to aid in their quantitative comparison.

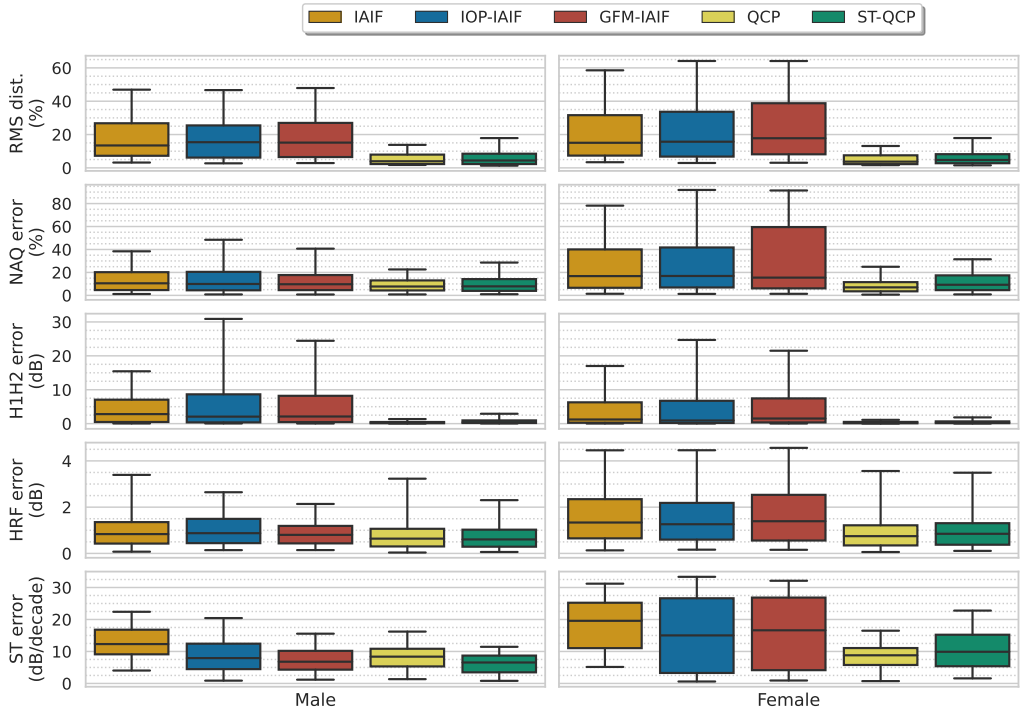


Figure 2. Distributions and median of error measures of RMS, NAQ, H1H2, HRF, and ST between the estimated glottal flow and the ground-truth signals for each GIF method evaluated as boxplots with whiskers from 5% up to 95%. Columns shows results for the male and female subsets, respectively.

Table 2. Median of the 5 error measures between the estimated glottal flow and the ground-truth (original glottal flow) signals for male vowels, computed at the pulse level for the 14.096 considered pulses. The best result per error metric is highlighted in bold.

	IAIF	IOP-IAIF	GFM-IAIF	QCP	ST-QCP
RMS distance (%)	13.45	15.41	15.16	**3.97**	4.44
NAQ error (%)	10.53	9.94	9.71	**7.74**	7.87
H1H2 error (dB)	2.81	2.07	2.1	**0.27**	0.41
HRF error (dB)	0.84	0.87	0.8	0.64	**0.6**
ST error (dB/decade)	12.31	7.93	6.79	8.38	**6.55**

Table 3. Median of the 5 error measures between the estimated glottal flow and the ground-truth (original glottal flow) signals for female vowels, computed at pulse level for the 14.097 considered pulses. The best result per per error metric is highlighted in bold.

	IAIF	IOP-IAIF	GFM-IAIF	QCP	ST-QCP
RMS distance (%)	15.06	15.7	17.79	**3.74**	4.74
NAQ error (%)	16.8	16.91	15.48	**6.99**	9.25
H1H2 error (dB)	1.22	0.93	1.51	**0.25**	0.34
HRF error (dB)	1.34	1.26	1.39	**0.75**	0.85
ST error (dB/decade)	19.58	15.02	16.62	**8.8**	9.9

As a general trend, for both sexes, it can be observed that QCP-based approaches obtain lower error metrics than the IAIF-based techniques, with 98 out of the 100 pair-to-pair comparisons being statistically significant, as can be seen in Table A3. When it comes to comparing the performance of the five GIF approaches with respect to sex, it can be observed that the error metrics obtained in the female speech corpus are, in general

terms, despite H1H2, higher than those obtained in the male speech data. Moreover, the QCP-based approaches present quite a more stable response in terms of sex variation than IAIF-based counterparts. Moreover, it is worthwhile to note that QCP presents the most stable performance across sexes (with the lowest median RMS, NAQ, and H1H2 for female speech, while ST-QCT presents the lowest HFR and ST-QCP for the male data), while GFM-IAIF is more sensitive, as shown by the increases in the HRF and ST error medians as well as the wider distributions in the female dataset.

In particular, for the male corpus, QCP presents the lowest RMS, NAQ, and H1H2 errors, while ST-QCP outperforms the original version of the technique in terms of the HRF and ST error metrics. Moreover, it is worth mentioning that, for this sex, the IOP and GFM variants of IAIF outperform IAIF in terms of the NAQ, H1H2, and ST error, with the latter also being lower than that obtained by QCP. In the female speech corpus, QCP-based approaches surpass IAIF-based approaches for all error metrics, with the QCP method presenting the lowest values of all methods, even outperforming ST-QCP in the ST error metric. For this sex, both the IOP and GFM variants improve the original IAIF in terms of the ST error, as does H1H2 plus HRF for IOP and NAQ for GFM, respectively.

4.2. F0

Regarding the performance of the inverse filtering techniques for the F0 variable, the general tendency is for RMS, NAQ, and H1H2 for both male and female subsets to present larger errors for higher F0s, with a slight decrease in the higher frequencies for the case of male data for QCP-based methods and IAIF-IOP. This result can be observed from Figure 3 (for a detailed analysis, the reader is referred to the F0 sweep analysis included in Appendix B). This difference between lower and higher frequencies is emphasised in the female case. It can be observed that IAIF-based methods are more sensitive to F0 changes, while QCP approaches are more stable and obtain better results across the whole F0 range, especially for RMS and H1H2 error metrics.

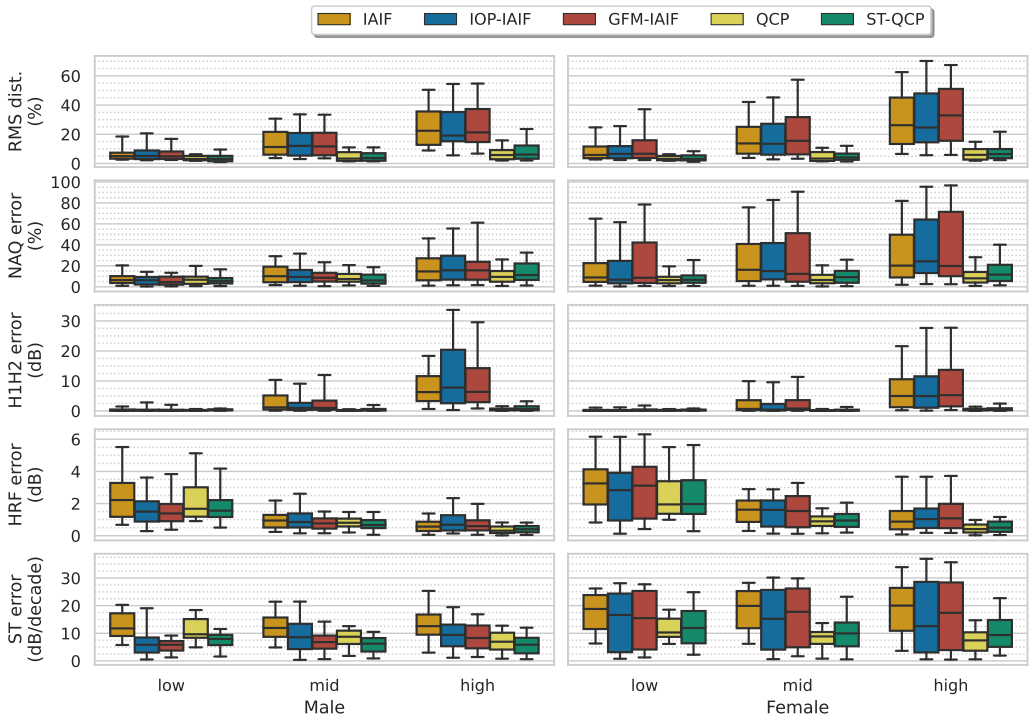


Figure 3. Distributions and median of error measures of RMS, NAQ, H1H2, HRF, and ST between the estimated glottal flow and the ground-truth signals for each GIF method evaluated as boxplots with whiskers from 5% to 95%, grouped by F0 range. Columns shows results for the male and female subsets, respectively.

On the contrary, the evaluation through HRF presents better results for high frequencies for both male and female speech data. For the ST error, IOP-IAIF and GFM-IAIF are the best performing methods for lower frequencies in the male subset, whilst the ST-QCP method presents lower values in the male high F0 range. In the case of the female subset, the QCP-based approaches are better, obtaining the best results with the original version of the algorithm.

In this case, as shown in Table A4, of the 150 pair-to-pair Wilcoxon statistical analyses conducted, 94% in the case of male data and 98.6% in the case of female data are significant.

4.3. Phonation Type

Regarding the evaluation of the behaviour of the GIF methods according to the different phonation types, a general decrease in the RMS and NAQ error metrics can be observed in Figure 4 when the voice is less tense (e.g., whispery). On the other hand, H1H2 increases for whispery and IAIF-based methods for both sexes. For these three errors, QCP and ST-QCP achieve the best results for both the male and female subsets. It is worth mentioning that H1H2 presents a very stable response across phonation types for both sexes. A general observation in the case of the female corpus is the wider spread of the distributions for IAIF-based methods, especially for tense phonations compared to the male distributions, except for H1H2. In terms of the HRF and ST errors, the IAIF methods show larger errors for relaxed phonation types, whereas IOP-IAIF presents larger error values for tense phonation types while being the best method next to ST-QCP for whispery for male speech data. Furthermore, for male data, GFM-IAIF and ST-QCP show more stable behaviour across the phonation type range. On the other hand, for the female case, the IOP-IAIF results do not improve for lax phonations, and QCP-based approaches are clearly better than IAIF-based methods. Finally, the Wilcoxon analysis shows, as presented in Table A5, statistically significant differences compared pair-to-pair in 95.5% of the cases for the male subset and 98.5% of the cases for the female subset. The breathy phonation type has more non-significant differences, with four cases for male data and one for female data.

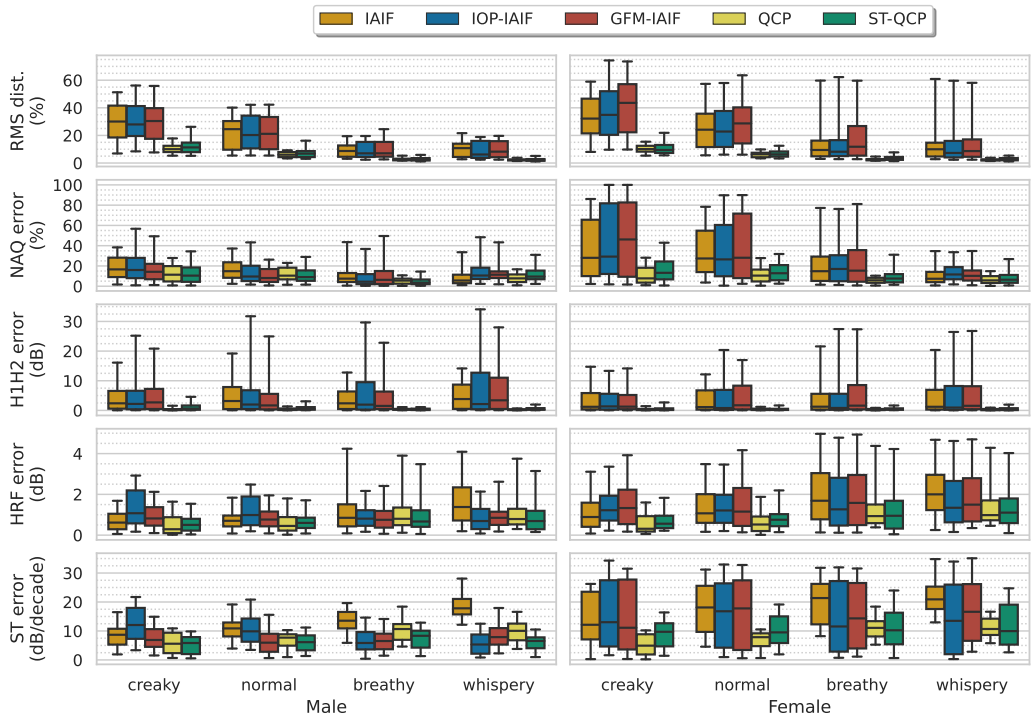


Figure 4. Distributions and median of error measures of RMS, NAQ, H1H2, HRF, and ST between the estimated glottal flow and the ground-truth signals for each GIF method evaluated as boxplots with whiskers from 5% to 95%, grouped by phonation type. Columns shows results for the male and female subsets, respectively.

4.4. Vowels

Figure 5 depicts the results grouped per vowel. It can be seen how, in terms of RMS, the best results are obtained by QCP-based approaches for both sexes, with these being the most stable across all vowels, as shown in Figure A1. IAIF and its variants behave considerably worse, especially for /i/ and /u/ for male speech and /i/, /e/, and /æ/ for female speech. When looking at H1H2 outcomes, again, QCP and ST-QCP behaviours are the best and are also very stable across vowels for both sexes. The NAQ errors show lower error values for IAIF-based methods and are, in most cases, comparable to those obtained by QCP, except for /i/ for both male and female and /e/ and /æ/ for female speech. For the HRF measure, the results of IAIF-based methods are again closer to those obtained by QCP-based techniques. These results are quite stable with some exceptions, such as IOP-IAIF for /u/ and /o/ for the male case and the three IAIF methods for /i/, /e/, and /æ/ for female data. Finally, it is interesting to notice how the ST error shows a very different result for male and female speech. In the first case, all the methods behave in a very stable manner across the vowels, with ST-QCP having the best response, closely followed by IOP-IAIF with the best result for /e/ and GFM-IAIF for /a/ and /u/, while IAIF obtains the worst ST values. On the other hand, QCP presents the best results for vowels with lower formant frequencies, and the IAIF variants demonstrate the worst results. For higher-formant-frequency vowels, on the contrary, IOP-IAIF and GFM-IAIF are the methods with lower errors.

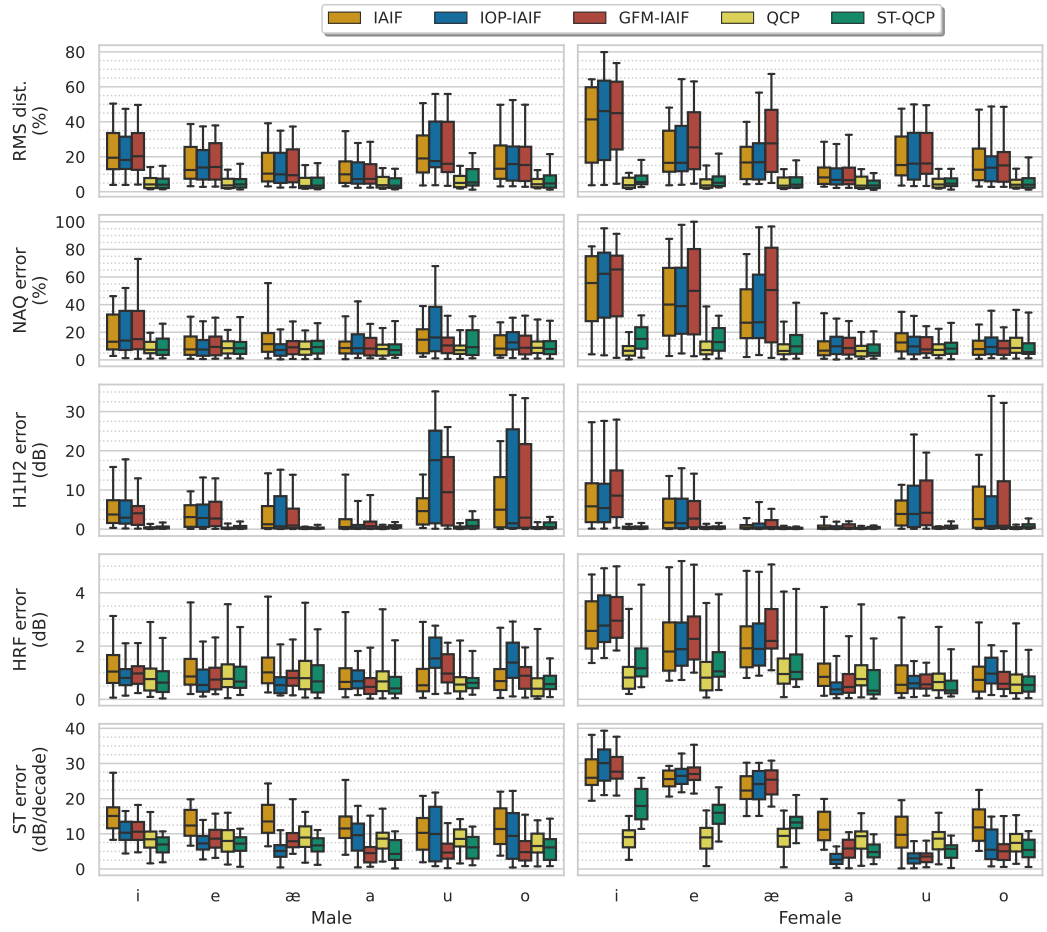
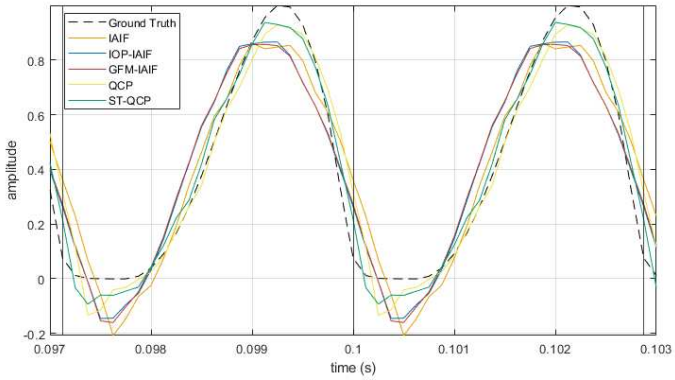


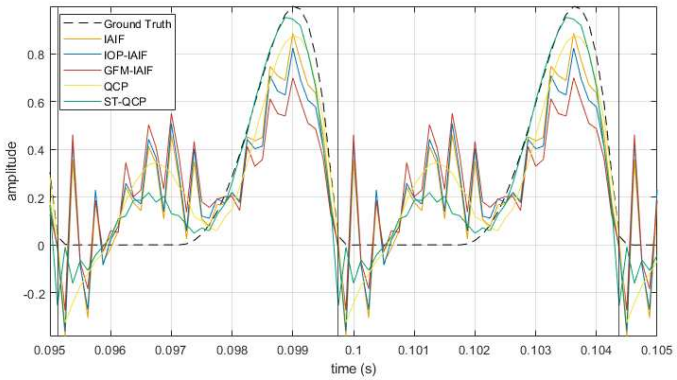
Figure 5. Distributions and median of error measures of RMS, NAQ, H1H2, HRF, and ST between the estimated glottal flow and the ground-truth signals for each GIF method evaluated shown as boxplots with whiskers from 5% to 95% grouped by vowel. Columns show results for the male and female subsets, respectively.

In terms of statistics, this case had the lowest statistically significant differences for the pair-to-pair Wilcoxon analysis and was significantly different for 92% out of the 300 pairs for the male subset and 96% in the case of the female subset. It is also apparent, as shown in Table A6, that the NAQ measure has the most non-significant number of pairs, with 11 out of 60 for male data and 4 for female data.

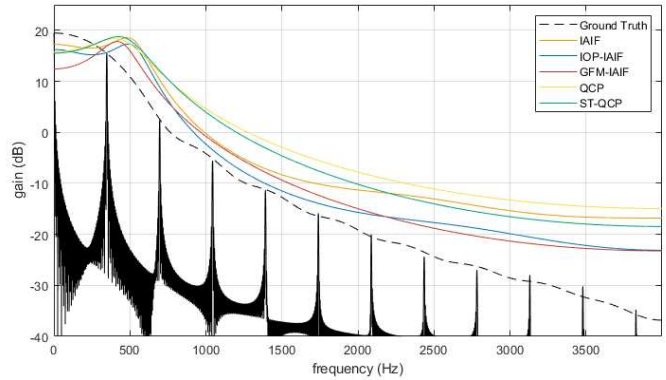
Figure 6 shows two examples of both normalised glottal flow output signals and spectra of ground-truth and GS estimations obtained for the optimised versions of each GIF method. The two chosen examples correspond to an /i/ vowel for male and female data, respectively, because there is a clear difference between GIF results within these two subsets, as can be appreciated in Figure 5. In particular, for each subset (i.e., /i/ for the male subset and /i/ for the female subset), the selected example is the one that obtained a lower sum of distances between its MAE-Wave value and the corresponding MAE-Wave mean value of each of the five GIF methods. The /i/ of the male example corresponds to a signal with breathy phonation and an F0 value of 360 Hz, while the /i/ of the female example has a normal phonation and an F0 value of 220 Hz. As can be seen in Figure 6a,b, the two selected examples show clearly worse results for the female estimated signal in comparison to the male counterpart for all GIF methods, which is also visible in the /i/ region of Figure 5. Specifically, the female example obtains mean relative increments with regard to the male example of 47.7% for MAE-Wave, 521.7% for NAQ, 88.2% for H1H2, 67.8% for HRF, and 167.1% for ST. In addition, mean relative increments of each of the GIF methods across the five error measures are 120.4% for IAIF, 111% for IOP-IAIF, 130.4% for GFM-IAIF, 65.2% for QCP, and 465.5% for ST-QCP.



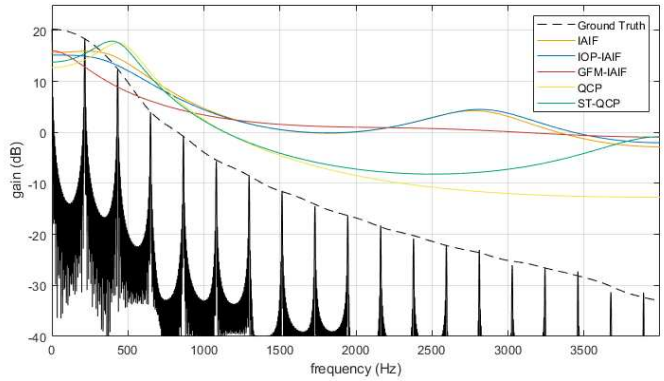
(**a**) Glottal flow signal comparison for male /i/, breathy at 360 Hz.



(**b**) Glottal flow signal comparison for a female /i/, normal at 220 Hz.



(**c**) Glottal flow spectra comparison for male /i/, breathy at 360 Hz



(**d**) Glottal flow spectra comparison for female /i/, normal at 220 Hz

Figure 6. Ground-truth and normalised glottal flow signal outputs and spectra of the five GIF methods for a male /i/ with breathy phonation and F0 of 360 Hz on the left column and a female /i/ with normal phonation and F0 of 220 Hz on the right column.

As discussed before, ST-QCP and QCP clearly outperform the three IAIF-based approaches. In Figure 6a these differences become apparent, with ST-QCP being the method that attains a greater degree of similarity to the ground-truth signal. In Figure 6b these differences are more salient because the normalised glottal flow output signals of IAIF-based approaches present more noisy behaviour than their QCP-based counterparts. When it comes to the spectral representations of the glottal flow estimations, Figure 6c shows estimations that denote fewer differences in spectral tilt with regard to the ground-truth signal, while in Figure 6d, these differences are visibly higher, specially for the three IAIF variants.

5. Discussion

Some of the general trends observed from the results of this study can be linked with other similar studies found in the literature. For example, the better performance of QCP compared to the original IAIF approach in terms of NAQ, H1H2, and HRF was also found in the work of Airaksinen et al. in [27,28]. However, it is worth mentioning that the error measures obtained in their study and the ones presented in this paper are not comparable. Airaksinen et al. used a synthetic vowel database constructed via the physical modeling of human speech production with different characteristics (e.g., the F0 range), and they fixed the GIF parameters, such as the vocal tract order, glottal source order, and lip radiation coefficient to fixed values for the set of analysed speech signals. Additionally, in the study by Chien et al. [30], the authors found that methods based on weighted linear prediction, such as QCP approaches (denoted as sparse linear prediction and weighted linear prediction) obtained better performance with regard to IAIF using physically modelled synthetic sustained vowels with different ranges of physical parameter variations (e.g., F0, subglottal pressures, and vowel types). Furthermore, in the present work, it has been observed that there is a clear relationship between specific error measures and F0, as well as phonation type. In regard to F0, for higher F0 values, higher RMS and H1H2 errors are also obtained, as in the work of Perrotin and McLoughlin [25], but the set of IAIF-based approaches obtain different patterns than our study (e.g., our results for the GFM-IAIF approach do not present with more stable behavior, nor is the technique with better performance across the entire set of signals).

Concerning phonation type, our results mark a clear decrease in performance in terms of the RMS for tenser phonations (e.g., creaky). The same trend was observed for IOP-IAIF and GFM-IAIF by Perrotin and McLoughlin in [25] for low R_d values (a parameter that correlates with phonation type). Furthermore, in terms of HRF and ST errors, these two techniques also respond worse for tenser phonations, while IAIF obtains better performance, as in [25]. However, it must be noticed that the four simulated phonation types along the OPENGLOT Repository I do not represent specific values of this R_d parameter; additionally, in the RMS error computation procedure reported in [25], the mean and scale normalization process were not specified.

As a global trend, it can be noticed that QCP-based methods outperform their IAIF-based counterparts. However, there are still some specific situations where IAIF-based techniques obtain better results, e.g., in terms of HRF and ST for low F0 ranges or NAQ for whisper phonation (in this last case, both the original IAIF and GFM-IAIF obtain the distribution with the lowest values together with QCP). Nevertheless, it is worth mentioning that QCP-based methods use extra information with respect to IAIF-based ones due to the closing phase timings that are provided by the CGIs, which, in turn, are used as input to define the AME function. This fact could partially explain the better performance observed in QCP-based methods, together with the fact that CGIs are also used for glottal error measure computation. However, when working on real speech data, the CGIs may be difficult to extract, thus affecting the performance of the QCP-based approaches negatively. This is a relevant issue that should be studied in future works.

In the present work, an extension of OPENGLOT Repository I was generated to include vowels produced with female formant vocal tracts. By maintaining the same ranges of exploration for the other signal attributes, like F0, vowel, or phonation type, we can

study the influence of the formants in the inverse filtering results. However, it must be appreciated that sex-related dependencies in other signal attributes, such as the F0 vocal range, were not included in the study. As a general trend, our results revealed that the performance of GIF methods decreases with female vocal tract signals with regard to male vocal tracts for all error measures but H1H2. This worsening would probably increase if the F0 vocal range was also adjusted to each type of vocal tract (male vs. female), which is something that might be explored in future works.

The development and study of GIF methods has relied on the use of synthetic speech because it provides a ground truth that facilitates the evaluation and comparison of GIF methods, as well as the tuning of the GIF parameters. Nevertheless, in order to move to the analysis of real speech (i.e., without ground-truth references of glottal flow signals), other evaluation metrics should be considered. Furthermore, a strategy for the selection of optimal GIF parameters should be devised.

Regarding the addition of expressiveness to the numerical generation of voice, previous works have been developed using ST-QCP to analyse the characteristics of GS and VT in aggressive and happy female vowels [9,14]. However, the results obtained in the present study suggest that maybe it would be better to use QCP without the spectral tilt correction when dealing with female speech.

The analysis methodology followed in this study includes the tuning of GIF parameters that was driven by a certain error metric. This error metric was defined as the MAE-Wave, which accounts for differences in the glottal flow time-domain waveform, as in the works by Mokhtari et al. [31] or Perrotin and McLoughlin [25]. However, the error metric can also be defined as a weighted sum of different glottal error measures (e.g., including others based on the frequency domain), similar to what Airaksinen et al. performed to determine the AME function parameters of the QCP method in [28].

6. Conclusions

This work analysed the performance of state-of-the-art IAIF- and QCP-based glottal inverse filtering methods on the OPENGLOT Repository I extended with female vowels. The main conclusions are that QCP-based methods achieve the best global performance according to the considered error metrics in addition to presenting with more stable behaviour across sex, phonation type, F0, and vowels. These techniques present statistically significant lower error metrics than their IAIF counterparts. As a general trend, the results obtained for the female vowels are worse than those obtained for the male ones, except for H1H2.

When looking at results in terms of F0, the study reveals a worsening of performance for high F0 values in several glottal error measures, such as RMS, NAQ and H1H2, with this behaviour being even more prominent in female vowels for IAIF-based approaches. Additionally, there is a general decrease in RMS and NAQ values when moving from tense to lax phonations, and IAIF-based methods obtain poorer performances for specific vowels (e.g., /i/ and /u/ for male data and /i/, /e/ and /æ/ for female data), with the QCP-based counterparts being more stable across all vowels. Moreover, the IOP-IAIF and GFM-IAIF variants outperform the original IAIF technique for most glottal error measures in male vowels. Regarding the spectral tilt error, ST-QCP achieves a lower value for male vowels than the original QCP, while the opposite occurs with female vowels.

Future work will consider the inclusion of vocal tract-based errors as additional error measures for both GIF methods' performance analysis and their tuning. Moreover, other speech datasets that rely on physical modelling and/or real speech could be considered to contrast the obtained results of the considered GIF methods and other similar approaches following the designed analysis methodology. Finally, GIF results could also be analysed accurately to model the specific glottal source and vocal tract patterns in order to provide articulatory-based numerical simulations with the desired expressiveness using three-dimensional realistic geometries.

Supplementary Materials: The results of the optimization process with the computed error measures were saved as csv files that can be downloaded at: https://www.mdpi.com/article/10.3390/app13158775/s1.

Author Contributions: Conceptualization, M.F., J.C.S. and F.A.-P.; methodology, M.F., J.C.S. and F.A.-P.; software, M.F., L. J.-O. and J.C.S.; validation, M.F. and L.J.-O.; formal analysis, M.F.; investigation, M.F., L.J.-O., J.C.S. and F.A.-P.; resources, M.F.; data curation, M.F.; writing—original draft preparation, M.F., L.J.-O., J.C.S. and F.A.-P.; writing—review and editing, M.F., L.J.-O., J.C.S. and F.A.-P.; visualization, M.F., L.J.-O. and J.C.S.; supervision, M.F.; project administration, F.A.-P.; funding acquisition, F.A.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Agencia Estatal de Investigación (AEI) through the FEMVoQ project (PID2020-120441GB-I00/AEI/10.13039/501100011033. The authors also thank the Departament de Recerca i Universitats (Generalitat de Catalunya) for their support under Grant Ref. 2021 SGR 01396.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is based on the OPENGLOT Repository I [32]. The original corpus and the code its generation can be found at: http://research.spa.aalto.fi/projects/openglot/ (accessed on 5 May 2023). The code for the generation of the extended OPENGLOT Repository I, together with the code to replicate the experiments and the visualisation performed here, can be found at: https://github.com/SpeechSalleBcn/inverse-filtering-evaluation (first version published on 23 June 2023). The results of the optimization process with the error measures computed can be found in the form of csv files as supplementary materials.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AME	Attenuated main excitation
DQ	Duration quotient
FEM	Finite element method
GCI	Glottal closure instant
GIF	Glottal inverse filtering
GFM	Glottal flow model
GS	Glottal source
H1H2	difference in amplitude between the first and second harmonics
HRF	Harmonic richness factor
HPF	High-pass filtering
IAIF	Iterative adaptive inverse filtering
IOP	Iterative optimal pre-emphasis
LF	Liljencrants–Fant model
LPC	Linear predictive coding
MAE-Wave	Median absolute waveform error
NAQ	Normalized amplitude quotient
PQ	Position quotient
QCP	Quasi-closed phase
RMS	Root-mean-square-error
RQ	Ramp quotient
ST	Spectral tilt
VT	Vocal tract

Appendix A

Two tables are presented summarizing the ranges of parameters and the characteristics of the datasets used in the papers present in the literature review of this study.

Table A1. Summary of the datasets used for the evaluation of GIF methods in the papers of the literature review, regarding the 5 methods evaluated in the present work. The ranges are written as initial value:increment:final value.

Paper	Pitch	Vowels	Type	Sample Freq.	Phonation	Other
[17]	100:5:240 Hz	14	Synthetic (LF)	16 kHz	Oq: 0.3:0.05:0.9 α_m: 0.55:0.05:0.8	SNR (dB): 10:10:80
[20]	60:20:180 Hz	/a/ /@/ /i/ /y/	Synthetic (LF)		Oq: 0.4:0.05:0.9 α_m: 0.6:0.05:0.9	male
[20]	Flat pitch	/a/	Real		Decreasing Oq	
[22]	Two values for the pitch period for female and male	/a/	Synthetic [36]	8 kHz	Breathy Normal Pressed	8th order All-pole filter Male Female
[23]	F: 133–200 Hz M: 67–100 Hz	/a/ /i/	Synthetic [36]	8 kHz		Male Female
[24]		/a/	Real	44.1 kHz down to 8 kHz	Weak, breathy Breathy Modal Loud, slightly tense Shouted & Tense	
[25]	100:5:240 Hz	10	Synthetic (LF)	16 kHz	Rd: 0.4-2.7	
[25]			Real			
[26]			Real	48 kHz down to 16 kHz	Normal Lombard	11 sentences 2 to 9 s
[28]	75:10:405	/a/ /e/ /i/	Synthetic (LF)	8 kHz	625 different LF pulses	Optimize AME 8th order All-pole filter
[28]	80:10:400	/e/ /o/ /æ/	Synthetic (LF)	8 kHz	4 LF values interlaced with the optim. set	Test set 8th order All-pole filter
[28]	100:50:450	/a/ /i/ /ae/	Physical Model	8 kHz		Test set Male Female 5 year-old
[30]	90:30:210 Hz	/i/ /e/ /ɛ/ /ä/ /o/ /u/	Physical Model Two-mass, triangular-glottis vocal folds and transmission-line vocal tract	48 kHz down to 16 kHz	pressed slightly pressed modal slightly breathy and breathy	VocalTractLab 2.1 {500, 708, 1000, 1414, 2000}Pa 0.6 s
[30]	5 median target fundamental frequencies	Utterances derived from: "Lea und Doreen mögen Bananen."	Physical Model Two-mass, triangular-glottis vocal folds and transmission-line vocal tract	48 kHz down to 16 kHz	5 median voice qualities	VocalTractLab 2.1 125 utterances 5 median pressure levels
[31]	92, 110, 131, 156, 185, 220, 262, 311, 370, 440 Hz	/a/ /æ/ /i/ /ə/ /u/ /o/	Physical Model	4, 8, 12, 16 kHz	11 steps from weak & breathy to strong & pressed	Vocal tract and trachea specified by 44 and 34 cross-sectional areas.

Table A2. Summary of the GIF methods and the ranges of parameters evaluated in the papers of the literature review regarding the 5 methods evaluated in the present work. Vocal tract order is referred to as N_v, glottal source order as N_v, and lip radiation as d. The ranges are written as initial value:increment:final value.

Paper	IAIF	IOP	GFM	QCP
[17]	Aparat default options N_v = 10 N_g = 2 d = 0.99			
[22]	N_v = 8:2:12 N_g = 2 d: Lip radiation effect cancelled by integrating the estimation of the glottal flow derivative.			
[23]	N_v = 10; N_g = 4 d: Lip radiation effect cancelled by integrating the estimation of the glottal flow derivative.			
[24]	N_v = 8:2:18 N_g = 4 d = 0.8:0.01:0.99	N_v = 8:2:18 N_g = 4 d = 0.8:0.01:0.99		
[25]	N_v = 14:2:22 N_g = 3:1:6 d = 0.8:0.01:0.99	N_v = 14:2:22 N_g = 3:1:6 d = 0.8:0.01:0.99	N_v = 14:2:22 N_g = 3 d = 0.8:0.01:0.99	
[28]	N_v = 10 N_g = 4 d = 0.99			N_v = 10 N_g = 4 d = 0.99 DQ = 0.4:0.05:1 PQ = 0:0.025:0.2 RQ = 0:0.05:0.2
[30]	N_v = 20 N_g = 4			
[31]	d = 0.75:0.001:0.999 N_v = 2:2:10 (4 kHz) N_v = 6:2:14 (8 kHz) N_v = 10:2:16 (12 kHz) N_v = 14:2:22 (16 kHz) N_g = 3:1:6	d = 0.75:0.001:0.999 N_v = 2:2:10 (4 kHz) N_v = 6:2:14 (8 kHz) N_v = 10:2:16 (12 kHz) N_v = 14:2:22 (16 kHz) N_g = 3:1:6	d = 0.75:0.001:0.999 N_v = 2:2:10 (4 kHz) N_v = 6:2:14 (8 kHz) N_v = 10:2:16 (12 kHz) N_v = 14:2:22 (16 kHz) N_g = 3	

Appendix B

This appendix contains the Figure A1, which shows the medians of the GIF error measures for each F0 value of the OPENGLOT Repository I range.

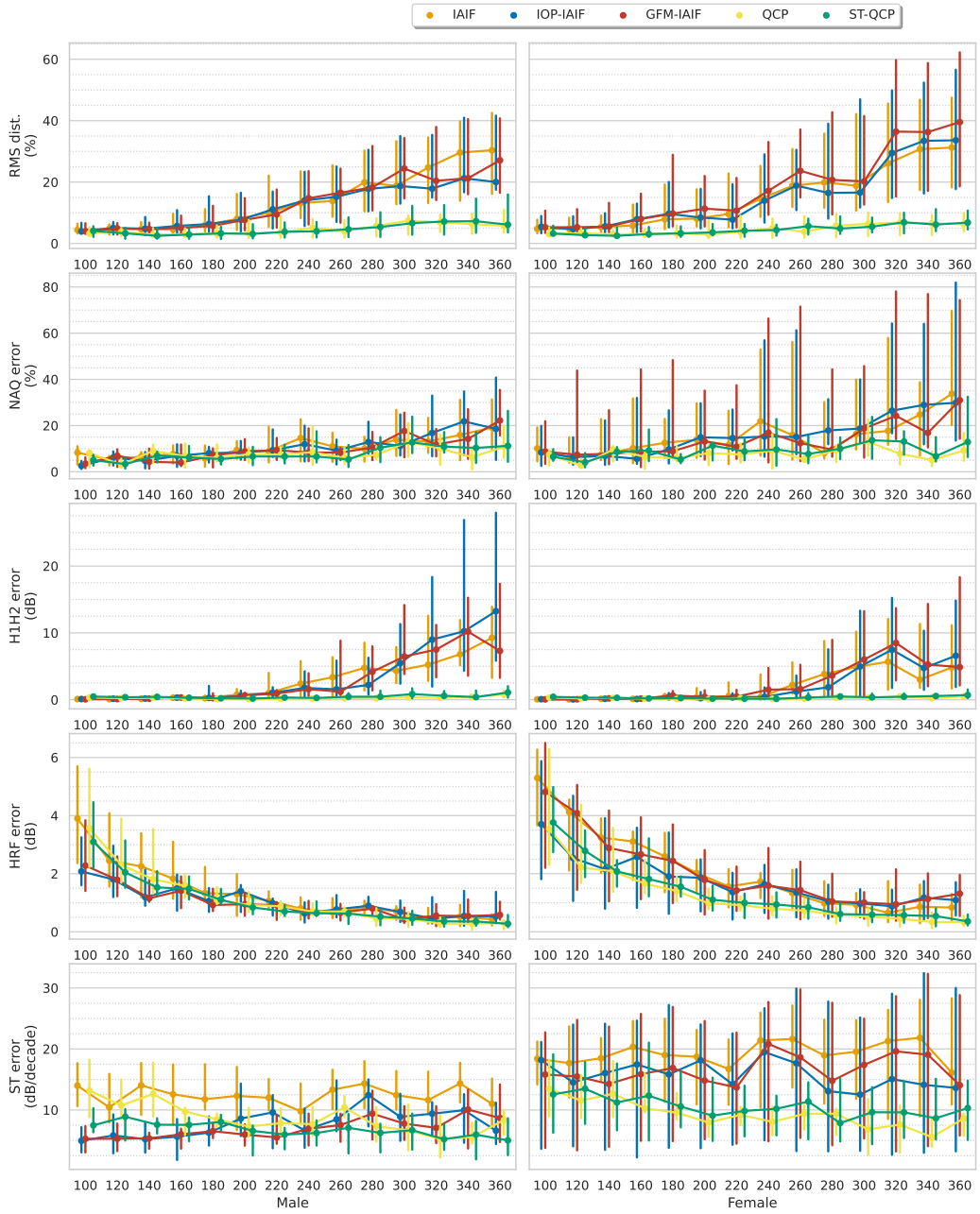


Figure A1. Medians of the measures RMS, NAQ, H1H2, HRF, and ST between the estimated glottal flow and the ground-truth signals for each evaluated GIF method as points, with quartiles 1 and 3 as errors bars for each F0 in the corpus range. Columns shows results for the male and female subsets, respectively.

Appendix C

This appendix includes the results of the statistical analyses conducted to evaluate to what extent the obtained error metrics for each GIF method and evaluated scenario are statistically significant or not, according to the Wilcoxon signed-rank paired test with Holm–Bonferroni correction.

Table A3. Wilcoxon statistical analysis for errors by GIF method, * show statistical significant differences with $p > 0.05$, while n represent non-significant differences.

		Male				Female			
Error	**GIF Method**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**
RMSE	IAIF	n	*	*	*	*	*	*	*
	IOP-IAIF		*	*	*		*	*	*
	GFM-IAIF			*	*			*	*
	QCP				*				*
NAQ	IAIF	*	*	*	*	*	*	*	*
	IOP-IAIF		*	*	*		*	*	*
	GFM-IAIF			*	*			*	*
	QCP				*				*
H1H2	IAIF	*	*	*	*	*	*	*	*
	IOP-IAIF		*	*	*		*	*	*
	GFM-IAIF			*	*			*	*
	QCP				*				*
HRF	IAIF	*	*	*	*	*	*	*	*
	IOP-IAIF		*	*	*		*	*	*
	GFM-IAIF			n	*			*	*
	QCP				*				*
Spectral Tilt	IAIF	*	*	*	*	*	*	*	*
	IOP-IAIF		*	*	*		*	*	*
	GFM-IAIF			*	*			*	*
	QCP				*				*

Table A4. Wilcoxon statistical analysis for errors by F0 range, where *l m h* denote low, mid and high, * show statistical significant differences with $p > 0.05$, and n represents non-significant differences.

		Male				Female			
Error	**GIF Method**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**
	F0 Range	*l m h*	*l m h*	*l m h*	*l m h*	*l m h*	*l m h*	*l m h*	*l m h*
RMSE	IAIF	***	***	***	***	***	***	***	***
	IOP-IAIF		*n*	***	***		***	***	***
	GFM-IAIF			***	***			***	***
	QCP				***				***
NAQ	IAIF	***	**n	n**	***	***	***	***	***
	IOP-IAIF		n*n	***	n**		***	***	***
	GFM-IAIF			***	***			***	***
	QCP				***				***
H1H2	IAIF	n**	***	***	***	n**	***	***	***
	IOP-IAIF		***	***	***		***	***	***
	GFM-IAIF			n**	***			***	***
	QCP				***				***
HRF	IAIF	***	***	***	***	***	*n*	***	***
	IOP-IAIF		***	***	***		***	***	***
	GFM-IAIF			***	***			***	***
	QCP				***				***
Spectral Tilt	IAIF	***	***	***	***	***	***	***	***
	IOP-IAIF		n**	***	***		***	***	***
	GFM-IAIF			***	***			***	***
	QCP				***				***

Table A5. Wilcoxon statistical analysis for errors by phonation type, where *c n b w* denote creaky, normal, breathy and whispery, respectively, * show statistical significant differences with $p > 0.05$, while n represents non-significant differences.

		Male				Female			
Error	**GIF Method**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**
	Vocal Effort	*c n b w*	*c n b w*	*c n b w*	*c n b w*	*c n b w*	*c n b w*	*c n b w*	*c n b w*
RMSE	IAIF	****	****	****	****	****	****	****	****
	IOP-IAIF		n***	****	****		****	****	****
	GFM-IAIF			****	****			****	****
	QCP				**nn				****
NAQ	IAIF	****	**n*	***n	****	****	****	****	****
	IOP-IAIF		****	****	****		****	****	****
	GFM-IAIF			****	****			****	****
	QCP				n***				****
H1H2	IAIF	****	n***	****	****	****	****	****	****
	IOP-IAIF		****	****	****		n***	****	****
	GFM-IAIF			****	****			****	****
	QCP				**n*				****
HRF	IAIF	****	****	****	****	****	**n*	****	****
	IOP-IAIF		****	****	**n*		****	****	****
	GFM-IAIF			****	****			****	****
	QCP				****				***n
Spectral Tilt	IAIF	****	****	****	****	****	****	****	****
	IOP-IAIF		****	****	****		****	****	****
	GFM-IAIF			****	****			****	****
	QCP				****				****

Table A6. Wilcoxon statistical analysis for errors by vowel, where *i e æ a o u* denote the vowel, * show statistically significant differences with $p > 0.05$, while n represent non-significant differences.

		Male				Female			
Error	**GIF Method**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**	**IOP-IAIF**	**GFM-IAIF**	**QCP**	**ST-QCP**
	vowel	*i e æ a u o*	*i e æ a u o*	*i e æ a u o*	*i e æ a u o*	*i e æ a u o*	*i e æ a u o*	*i e æ a u o*	*i e æ a u o*
RMSE	IAIF	******	******	******	******	******	******	******	******
	IOP-IAIF		******	******	******		*****n	******	******
	GFM-IAIF			******	******			******	******
	QCP				******				****n*
NAQ	IAIF	******	***n*n	******	******	******	*****n	******	*****n
	IOP-IAIF		******	******	*n****		n*****	******	******
	GFM-IAIF			**n**n	**n**n			******	*****n
	QCP				*nnn*n				******
H1H2	IAIF	******	nn*n**	******	******	******	******	******	******
	IOP-IAIF		n*n**n	******	***n**		*n***n	******	******
	GFM-IAIF			******	******			******	******
	QCP				******				**n***
HRF	IAIF	******	*****n	******	******	******	******	******	******
	IOP-IAIF		******	n**n**	******		******	******	******
	GFM-IAIF			******	******			*****n	***n**
	QCP				******				******
Spectral Tilt	IAIF	****n*	******	******	******	******	******	******	******
	IOP-IAIF		******	******	******		****n*	******	******
	GFM-IAIF			**n***	******			******	******
	QCP				******				******

References

1. Birkholz, P.; Jackel, D.; Kroger, B. Construction And Control Of A Three-Dimensional Vocal Tract Model. In Proceedings of the 2006 IEEE ICASSP Proceedings, Toulouse, France, 14–19 May 2006; Volume 1, p. I. [CrossRef]
2. Blandin, R.; Arnela, M.; Laboissière, R.; Pelorson, X.; Guasch, O.; Hirtum, A.V.; Laval, X. Effects of higher order propagation modes in vocal tract like geometries. *J. Acoust. Soc. Am.* **2015**, *137*, 832–838. [CrossRef]
3. Arnela, M.; Dabbaghchian, S.; Blandin, R.; Guasch, O.; Engwall, O.; Van Hirtum, A.; Pelorson, X. Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds. *J. Acoust. Soc. Am.* **2016**, *140*, 1707–1718. [CrossRef]
4. Arnela, M.; Dabbaghchian, S.; Guasch, O.; Engwall, O. MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2173–2182. [CrossRef]
5. Dabbaghchian, S.; Arnela, M.; Engwall, O.; Guasch, O. Simulation of vowel-vowel utterances using a 3D biomechanical-acoustic model. *Int. J. Numer. Methods Biomed. Eng.* **2021**, *37*, e3407. [CrossRef] [PubMed]

6. Arnela, M.; Guasch, O. Finite element simulation of /asa/ in a three-dimensional vocal tract using a simplified aeroacoustic source model. In Proceedings of the 23rd International Congress on Acoustics (ICA), Aachen, Germany, 9–13 September 2019; pp. 1802–1809.
7. Pont, A.; Guasch, O.; Arnela, M. Finite element generation of sibilants /s/ and /z/ using random distributions of Kirchhoff vortices. *Int. J. Numer. Methods Biomed. Eng.* **2020**, *36*, e3302. [CrossRef] [PubMed]
8. Schröder, M. Expressive speech synthesis: Past, present, and possible futures. In *Affective Information Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 111–126.
9. Freixes, M.; Arnela, M.; Alías, F.; Socoró, J.C. GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]. In Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW), Vienna, Austria, 20–22 September 2019; pp. 132–136. [CrossRef]
10. Fant, G.; Liljencrants, J.; Lin, Q. A four-parameter model of glottal flow. *Speech Transm. Lab. Q. Prog. Status Rep. (STL-QPSR)* **1985**, *26*, 1–13.[CrossRef]
11. Freixes, M.; Arnela, M.; Socoró, J.C.; Alías, F.; Guasch, O. Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels. *Appl. Sci.* **2019**, *9*, 4535. [CrossRef]
12. Arnela, M.; Guasch, O.; Freixes, M. Finite element generation of sung vowels tuning 3D MRI-based vocal tracts. In Proceedings of the 27th International Congress on Sound and Vibration (ICSV27), Graz, Austria, Online, 11–16 July 2021; pp. 1–8.
13. Li, Y.; Li, J.; Akagi, M. Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space. *J. Acoust. Soc. Am.* **2018**, *144*, 908. [CrossRef]
14. Freixes, M.; Socoró, J.C.; Alías, F. Contribution of Vocal Tract and Glottal Source Spectral Cues in the Generation of Acted Happy and Aggressive Spanish Vowels. *Appl. Sci.* **2022**, *12*, 2055. [CrossRef]
15. Li, K.; Unoki, M.; Li, Y.; Dang, J.; Akagi, M. Study on Simultaneous Estimation of Glottal Source and Vocal Tract Parameters by ARMAX-LF Model for Speech Analysis/Synthesis. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 36–43.
16. Schleusing, O.; Kinnunen, T.; Story, B.; Vesin, J.M. Joint Source-Filter Optimization for Accurate Vocal Tract Estimation Using Differential Evolution. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1560–1572. [CrossRef]
17. Drugman, T.; Bozkurt, B.; Dutoit, T. A Comparative Study of Glottal Source Estimation Techniques. *Comput. Speech Lang.* **2012**, *26*, 20–34. [CrossRef]
18. Klatt, D.H.; Klatt, L.C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* **1990**, *87*, 820–857. [CrossRef] [PubMed]
19. Peterson, G.E.; Barney, H.L. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* **1952**, *24*, 175–184. [CrossRef]
20. Drugman, T.; Dutoit, T. Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation. In Proceedings of the INTERSPEECH 2009, 10th Annual Conference of the International Speech, Brighton, UK, 6–10 September 2009; pp. 116–119.
21. Bozkurt, B.; Dutoit, T. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In Proceedings of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis, Geneva, Switzerland, 27–29 August 2003.
22. Alku, P.; Vilkman, E.; Laine, U. Analysis of glottal waveform in different phonation types using the new IAIF-method. In Proceedings of the 12th lnternational Congress Phonetic Sciences, Aix-en-Provence, France, 19–24 August 1991; Volume 4, pp. 362–365.
23. Alku, P. Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Commun.* **1992**, *11*, 109–118. [CrossRef]
24. Mokhtari, P.; Ando, H. Iterative Optimal Preemphasis for Improved Glottal-Flow Estimation by Iterative Adaptive Inverse Filtering. In Proceedings of the INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 24–27 August 2017; pp. 1044–1048. [CrossRef]
25. Perrotin, O.; McLoughlin, I. A Spectral Glottal Flow Model for Source-filter Separation of Speech. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7160–7164. [CrossRef]
26. Wong, D.; Markel, J.; Gray, A. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 350–355. [CrossRef]
27. Airaksinen, M.; Story, B.; Alku, P. Quasi closed phase analysis for glottal inverse filtering. In Proceedings of the INTERSPEECH'2013, Lyon, France, 25–29 August 2013.
28. Airaksinen, M.; Raitio, T.; Story, B.; Alku, P. Quasi Closed Phase Glottal Inverse Filtering Analysis with Weighted Linear Prediction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 596–607. [CrossRef]
29. Seshadri, S.; Juvela, L.; Räsänen, O.; Alku, P. Vocal Effort based Speaking Style Conversion using Vocoder Features and Parallel Learning. *IEEE Access* **2019**, *7*, 17230–17246. [CrossRef]
30. Chien, Y.R.; Mehta, D.D.; Guðnason, J.; Zañartu, M.; Quatieri, T.F. Evaluation of Glottal Inverse Filtering Algorithms Using a Physiologically Based Articulatory Speech Synthesizer. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1718–1730. [CrossRef] [PubMed]

31. Mokhtari, P.; Story, B.; Alku, P.; Ando, H. Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production. *Speech Commun.* **2018**, *104*, 24–38. [CrossRef]
32. Alku, P.; Murtola, T.; Malinen, J.; Kuortti, J.; Story, B.; Airaksinen, M.; Salmi, M.; Vilkman, E.; Geneid, A. OPENGLOT–An Open Environment for the Evaluation of Glottal Inverse Filtering. *Speech Commun.* **2019**, *107*, 38–47. [CrossRef]
33. Ma, C.; Kamp, Y.; Willems, L. Robust signal selection for linear prediction analysis of voiced speech. *Speech Commun.* **1993**, *12*, 69–81. [CrossRef]
34. Levy, E.C. Complex-curve fitting. *IRE Trans. Autom. Control.* **1959**, *AC-4*, 37–43. [CrossRef]
35. Drugman, T.; Dutoit, T. Glottal closure and opening instant detection from speech signals. In Proceedings of the INTERSPEECH 2009, 10th Annual Conference of the International Speech, Brighton, UK, 6–10 September 2009; pp. 2891–2894.
36. Gold, B.; Rabiner, L. Analysis of digital and analog formant synthesizers. *IEEE Trans. Audio Electroacoust.* **1968**, *16*, 81–94. [CrossRef]
37. Alku, P.; Bäckström, T.; Vilkman, E. Normalized amplitude quotient for parametrization of the glottal flow. *J. Acoust. Soc. Am.* **2002**, *112*, 701–710. [CrossRef]
38. Childers, D.G.; Lee, C.K. Vocal quality factors: Analysis, synthesis, and perception. *J. Acoust. Soc. Am.* **1991**, *90*, 2394–2410. [CrossRef] [PubMed]
39. Summers, V.; Pisoni, D.; Bernacki, R.; Pedlow, R.; Stokes, M. Effects of noise on speech production: Acoustic and perceptual analyses. *J. Acoust. Soc. Am.* **1988**, *84*, 917–928. [CrossRef] [PubMed]
40. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]