Glottal Inverse Filtering with Automatic Filter Order Selection

Alan Ó Cinnéide, David Dorran, Mikel Gainza and Eugene Coyle

Audio Research Group
Dublin Institute of Technology, Kevin Street, Dublin

E-mail: alan.ocinneide@dit.ie, david.dorran@dit.ie, mikel.gainza@dit.ie, eugene.coyle@dit.ie

Abstract — **Joint estimation glottal inverse filtering methods determine the optimal parameters for both a derivative glottal flow model and a vocal tract filter. However, occasionally the best fit model and filter might not correspond to the most realistic glottal source - a fact that is particularly evident when the inverse filtered pulse is compared with the glottal source estimates from adjacent frames. This problem can often be associated with the selection of improper filter order. Adjusting the filter order can often be used to achieve a more satisfactory decomposition of the speech signal, but selecting the most appropriate order remains a problem. This process can be automated by implementing a linear prediction analysis, and smoother, more consistent glottal source derivative signals can be obtained. This paper discusses the problem and presents this possible solution, with algorithmic results on real and synthetic speech.**

Keywords — **glottal inverse filtering, filter order selection**

I Introduction

Glottal inverse filtering is an operation used to determine the glottal excitation source of voiced speech. Based upon the linear source filter theory of speech production [1], these methods estimate the vocal tract filter, the inverse of which is then used to filter the speech analysis frame to yield the derivative glottal flow signal. The revealed waveform is of interest to speech researchers and engineers for a variety of applications including: speaker identification, speech synthesis and voice modification.

Glottal inverse filtering is a blind deconvolution problem, i.e. in order to construct the vocal tract inverse filter, it is necessary to make assumptions about the characteristics of the glottal source. One such assumption is that a glottal pulse has a specific shape during a pitch period. To this end, researchers have incorporated time domain models of the glottal signal into inverse filtering techniques, e.g. [2] [3]. These methods simultaneously determine the optimal parameters of the incorporated glottal model and the vocal tract filter, usually by minimizing the energy of the time domain residual signal. Hence these methods are sometimes referred to as joint estimation techniques. The filter order is usually not optimized by the algorithm and chosen according to assumptions about the vocal tract dimensions.

While it is generally possible achieve realistic separations of the derivative glottal flow and the vocal tract filter following this guideline, occasionally the deconvolution results yield glottal pulses that are substantially different from adjacent analysis frames. Figure 1 illustrates a typical scenario. Despite the minimum error criterion of the deconvolution of each speech pulse, the global evolution of the glottal flow parameters is markedly discontinuous. While pulse-to-pulse deviations are expected within running speech, normal non-pathological speech is presumed to be stationary over 20ms time frames and to change smoothly with time. Sudden discontinuities in the behavior of the source or filter are thus unlikely from a physical perspective.

Inspection of the vocal tract filters associated with these discontinuities often reveals the pres-

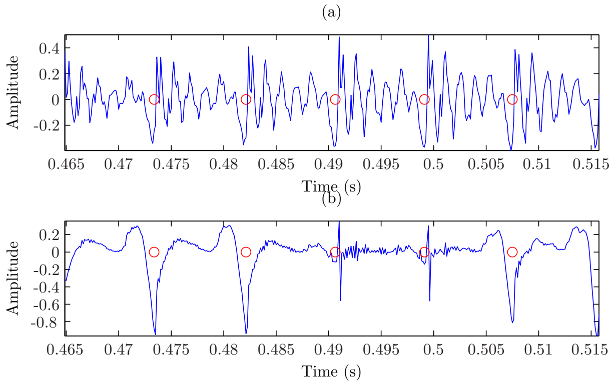

Fig. 1: This figure illustrates the difference in glottal signals that can be obtained using fixed filter order analysis. Panel (a) illustrates a section of sensibly stationary voiced speech. However, following glottal inverse filtering, the source estimate (b) reveals that the third and fourth glottal pulses are considerably different from their immediate neighbors. The instants of glottal closure are shown as red circles.

ence of very low frequency complex conjugate pole pairs, unobserved in adjacent frames. Re-analysis of the speech frame using a reduced filter order can produce a smoother glottal flow estimate, more consistent with its neighboring pulses. However, if too many poles are removed, the analysis will be left unable to capture the total vocal tract resonance. Selecting amongst deconvolution attempts is problematic.

In order to quantify the relative success of joint estimation deconvolutions for a speech pulse for a range of filter orders, this paper proposes a simple method based on linear prediction. Using this test, more consistent glottal source and vocal tract filter separations can be obtained.

This paper is organized as follows: first, both the source filter theory of speech production and the glottal inverse filtering method used for this work are briefly described. The usual rules for selecting the filter order of the vocal tract filter are also outlined. Following this, the residual signals that are encountered when a deconvolution is unrealistic are described in more detail. A test to discriminate between the inverse filtered results, reliant on autocorrelation method linear prediction, is then proposed. An experiment demonstrating the efficacy of the algorithm in obtaining more consistent inverse filtering results is described and discussed with reference to real speech examples in the third section. Finally, the work is concluded and possibilities for future work laid out.

II Background

a) Acoustic Theory of Speech Production

The acoustic theory of speech production [1] views speech $S(z)$ as the multiplication of glottal flow signal $G(z)$ with a vocal tract filter $V(z)$ which is then radiated at the lips $L(z)$. In the Z-domain, the process can be represented as follows:

$$S(z) = G(z)V(z)L(z) \tag{1}$$

If the vocal tract can be modeled as an unbranched concatenated set of lossless tubes and wave propagation through the tubes is planar, it can be shown that an all-pole model can be used to represent it [4]. Poles generally occur in complex conjugate pairs which describe a region of resonant energy present in the spectrum called a formant.

As lip radiation $L(z)$ is usually modeled as a differentiating filter and the relationship between the speech chain components assumed linear, it is often combined with the glottal flow $G(z)$ to form the derivative glottal flow $G'(z)$. This reduces the number of elements in the speech production process to two:

$$S(z) = G'(z)V(z) \tag{2}$$

b) Joint Estimation of the Vocal Tract and Glottal Derivative Signal

The method of glottal inverse filtering utilized in the research presented here is the Convex Optimization method of Lu [2]. This approach incorporates the KLGLOTT88 model [5] of the glottal derivative excitation signal into the formulation of the inverse filtering problem. Given the pitch period and instant of glottal closure, the solution of the resulting linear system, efficiently found using quadratic programming, jointly estimates the optimal parameters of both the vocal tract all-pole filter and the glottal model for a range of open quotient values.

For an isolated analysis, the open quotient yielding the minimum least squares error is usually chosen as the solution to the system; for a segment of voiced speech, the five best-fitting candidates of each frame are retained and a dynamic programming algorithm is used to select the sequence of parameters yielding the minimum overall error.

c) Vocal Tract Filter Order Selection

When applying joint estimation techniques to speech signals, the order of the vocal tract filter p must be chosen before analysis. This parameter is determined by the dimensions of the vocal tract and has its foundations in the physical model of the tract as a series of concatenated lossless acoustic tubes [4]. For the average male vocal tract $17cm$ in length, this value is calculated as a pair of poles for every kilohertz of signal bandwidth, i.e.:

$$p = \lfloor \frac{f_s}{1000} + 0.5 \rfloor \tag{3}$$

where f_s is the sampling frequency. Due to their smaller sizes, the value of p is less for the speech of women and children. However, this number is merely a guideline: it is usual to supply a small number of extra poles to this number for analysis flexibility [6].

III Filter Order Selection for Joint Estimation Algorithms

a) Residual Signals of Joint Estimation Glottal Inverse Filtering Algorithms

While it is given that the residual signal following a joint estimation analysis resembles a glottal model in some optimal sense, the waveforms that result from the scenarios where the filter order is over- or under-modeled tend to be contain more high frequency energy than the ideal deconvolution. For instance: in the case where the filter order is insufficient, some of the vocal tract resonance must remain within the excitation signal. For optimization algorithms that minimize the energy of a residual signal, this often means that low amplitude, high frequency formants are improperly modeled. An overestimation of p may include some of the low frequency characteristics of the glottal source derivative within the vocal tract estimate. In each case, the time domain "smoothness" of the resulting waveform is perturbed.

It seems reasonable to assume that more realistic glottal waveform shapes are likely to be relatively smooth in the time domain, exhibiting a higher proportion of low frequency energy and less formant ripple. In general, smoothness is not a reliable indicator of a successful deconvolution as a certain degree of high frequency energy is expected to be found in the glottal signal, but because the results of a joint estimation procedure are already constrained to be a certain time-domain shape, a measure of the relative distribution of energy of the spectrum may be used to indicate residual signals that have been neither over- or under-modeled.

Thus, given a speech frame inverse filtered by a joint estimation algorithm, a measurement of the spectral balance of the frame can indicate the relative quality of deconvolution. An existing method to test the distribution of the energy within a signal's spectrum is first-order linear predictive analysis.

b) First Order Linear Predictive Spectral Analysis

Performing a first order linear predictive autocorrelation analysis of a signal frame fits the magnitude response of a single pole filter to the signal's spectrum. The position of this pole in the Z-domain gives a rough approximation to the overall spectral envelope - the closer the pole is to the unit circle, the higher the ratio that low frequency energy is to the upper end of the spectrum, see Figure 2. Smoother time domain waveforms will exhibit poles in these positions, while noisier signal frames will exhibit poles closer towards the origin, or indeed within the negative portion of the Z-plane.

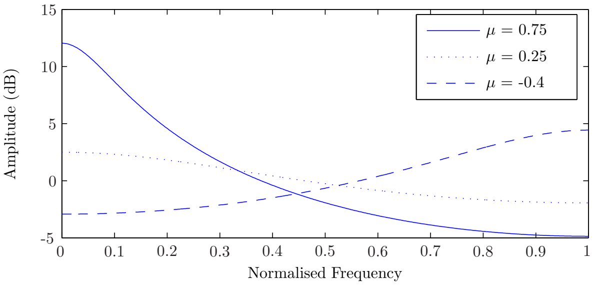

Fig. 2: This figure shows how the spectral balance of a signal can be described by a single real pole a distance of μ from the origin in the Z-plane.

An alternative viewpoint is to recognize that the linear prediction coefficient μ of a signal $s[n]$ is the ratio between the samples of the autocorrelation sequence $r[n]$ corresponding to lags 1 and 0 [4]. The samples of noisy signals tend to be less correlated than signals dominated by low frequency energy. This can be illustrated by the following equation:

$$\mu = \frac{r(1)}{r(0)} \tag{4}$$

This signal parameter has previously been used in voiced/unvoiced decision algorithms [7] and more recently in a method to automatic glottal inverse filtering without glottal closing instant information [8].

c) Method Summary

The algorithmic steps of this method may be summarized as follows: for each period, given an estimate of the pitch period and glottal closing instant, the vocal tract filter is estimated by the a joint estimation inverse filtering algorithm for a range of filter orders. Each filter is then used to inverse filter the original speech segment, and each of the resulting residual signals undergoes first order autocorrelation method linear predictive analysis. The residual signal corresponding to the root of the filter which is found to be closest to the unit circle is selected as the best representative of the derivative glottal flow. An example of the process can be seen in Figure 3.

IV Validation of Algorithm

The technique proposed in this paper promises to automate the selection of an appropriate filter order that should give more consistent glottal flow pulses minimizing the need for dynamic programming and also coping with the requirement

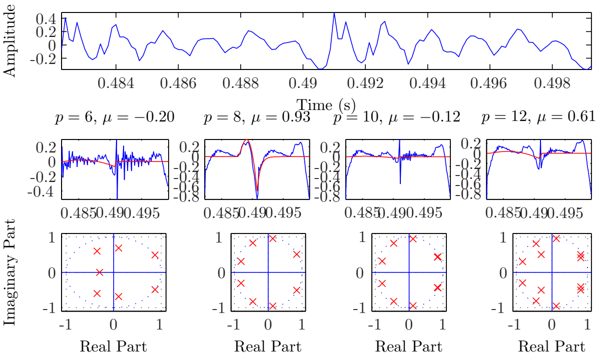

Fig. 3: This figure illustrates the variety of different residual pulses obtained after inverse filtering the top speech frame with different filter orders. The most likely glottal pulse is obtained using filter order $p = 8$ and distinguishes itself from the other possibilities by exhibiting the highest value of μ $(= 0.93)$.

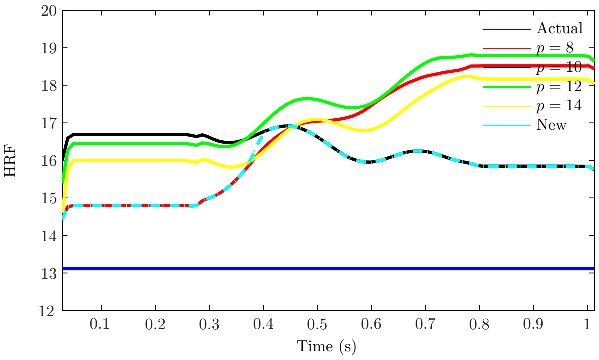

Fig. 4: This figure graphs the evolving HRF parameter estimated from derivative glottal flow signals corresponding to various filter orders. The filter orders associated with each filter order are given in the legend - the curve associated with the method outlined in this work is referred to as 'New'.

of changing filter order. In order to demonstrate this, two experiments are devised using synthetic and real speech signals.

The harmonic richness factor (HRF) [9] of the inverse filtered signal was chosen to the objective frequency domain measure to characterize the derivative glottal flow. This value is calculated as the difference in decibels between the sum of the harmonic amplitudes above the fundamental and the amplitude of the fundamental. It is calculated as:

$$HRF = 20\log_{10}\frac{\sum_{i\geq 2}^{N} H_i}{H_1} \tag{5}$$

where H_i is the magnitude of the i^{th} harmonic in the frequency domain and N is the number of the highest harmonic taken into account. In this work, N is calculated as $N = \lfloor\frac{f_s}{2f_0}\rfloor$, i.e. all harmonics available in the signal bandwidth. High HRF values indicate the presence of energy in the upper harmonics of the spectrum of the signal relative to the fundamental, and often correspond to more impulsive type signals. Conversely, low HRF values characterizes less energy in the upper harmonics and are indicative of a more sinusoidal signal.

This parameter was used because they can be automatically calculated from the signal and does not require any user interaction. It has previously been used to test the robustness of the DC-constrained closed phase inverse filtering technique to the position of the analysis interval [10] and was found to approximately characterize various voice types [9].

a) Experiments

In order to validate the selection criterion for signals with changing vocal tract filters, the following experiment was performed on a synthetic speech segment. A diphthong-type signal was created by convolving a train of LF model [11] pulses combined with additive Gaussian noise with a set of gradually changing vocal tract filters of different orders. The transition segment was generated by interpolating line spectral frequencies of two different filters. As the signal progress from a static eight order filter, it will gradually become more appropriate to change the filter order for successful deconvolutions.

The HRF values were determined according to equation 5 using a signal frame three times the size of the local pitch period, centered over each glottal closing instant. The evolution of this parameter for each of the inverse filtered signals are plotted in Figure 4.

An experiment with real speech segments was performed to demonstrate the ability of the algorithm to obtain more consistent glottal pulse derivative waveforms than fixed filter order methods. For test material, five speech waveforms for each of the speakers from the freely available CMU-Arctic database [?] were obtained. The signals were down-sampled to 8kHz and estimates of the instants of glottal closure were determined by a method similar to the one described in [12]. The fundamental frequency f_0 and voiced segments of the signal were determined by the SWIPE' algorithm [13].

The voiced segments were analyzed using the aforementioned method of inverse filtering using various fixed filter orders, in addition to the new method. The HRF of each of the signals was calculated as before. The standard deviations σ of this parameter for each of the speakers and filter orders are contained in Table 1.

Speaker	Filter order				
Male	8	10	12	14	New
bdl	4.355	5.331	5.342	5.177	4.368
jmk	6.157	8.549	7.670	7.646	4.491
awb	3.626	4.255	4.899	5.289	4.289
rms	4.208	6.690	6.377	6.052	3.632
ksp	6.364	7.720	6.965	7.352	4.102
Female	6	8	10	12	New
slt	6.579	7.570	8.284	7.930	6.401
clb	8.981	10.445	10.243	10.261	9.024

Table 1: The standard deviations of the HRF parameters for each filter order analysis estimated for the six speakers from inverse filtered segments of their speech.

b) Results and Discussion

Figure 4 shows the behavior of various fixed filter analyses to reproduce the characteristics of diphthong type signal that would require a change in the filter order. As clearly be seen, the HRF of the actual source signal is relatively static, and is approximately $6dB$. On the other hand, the high fixed order filter analyses (12 and 14) fluctuate quite rapidly over the duration of the entire signal. The analysis of order 8 initially follows the original source signal closely until the transition segment begins and gradually more of the filter requires an addition pole pair to model all the formants. The evidence is reversed for the 10 order filter: during the 8 pole section of the signal, the extra poles models a portion of the glottal signal in such a way that the signal becomes quite pulse like, with a relatively high HRF. Once a 10 pole filter becomes appropriate, the HRF of the signal drops toward the actual HRF value. Only the method outlined in this paper produces a signal whose HRF values consistently follows that of the original.

Though this initial experiment is limited in experimental data, it validates the basic concept of filter order selection. The error witnessed in these signals can be attributed to a number of factors. First, it must be noted that the inability of the KLGLOTT88 model to model the variation possible of the more complex LF model. Additionally, the additive noise into the signal undoubtedly corrupts the estimates to some extent.

The second experiment analyzed real speech signals in a similar fashion. As the actual glottal source in these cases is unknown, the experiment measured the consistency of the glottal signals by measuring their HRF values for many different sets of speech pulses. The work of Plumpe et al. [14] used glottal features for speaker identification purposes, therefore it is reasonable to assume that successful glottal inverse filtering procedures should reveal a limited variety of glottal source parameters. This fact will be particularly evident if using a carefully recorded database, such as the CMU-Arctic database.

The standard deviations σ of the HRF parameter for each of the speakers and filter orders are contained in Table 1. Five sentences were analyzed from each speaker, an average of 900 individual pulses in total analyzed. For the majority (4 out of 7) of speakers, there was a clear tendency for the proposed algorithm to produce a narrower range of glottal pulse estimate, with the greatest improvement noticed for speaker *ksp*. Of the speakers where the new method failed to give the most narrow range of estimates, the difference is of the region of 1% of the standard deviation, except for the speaker *awb*, where the difference is 14.2% of the mean.

It is noteworthy that the range of glottal parameters is measured to be markedly wider for those of women. This may be due to the higher pitch and increased breathiness of female voices which have made these voice types difficult to model.

V Conclusions and Future Work

This paper remarks on the phenomenon that can sometimes be observed when attempting to separate and parameterize voiced speech signals using fixed filter order, joint estimation analysis; namely, the scenario where the method produces unexpected results, particularly from a perspective of the continuity of parameters. It was explained that the problem may perhaps originate with the appropriateness of the incorporated glottal model and the misestimation of the signal's filter order. The paper then showed that, by testing a number of different filter orders in tandem and selecting the best output based upon a linear predictive analysis, smoother and more consistent results can be obtained. This was shown by with real speech using a number of well-known frequency domain parameters.

Furthermore, the new selective method of joint estimation glottal inverse filtering avoids the need to specify the filter order and can handle speech signals where the filter order describing the vocal tract might change, as was shown with synthetic speech signals. This flexibility of the analysis makes it an appropriate tool for the robust separation of continuous voiced speech.

Though this work utilizes an implementation of this specific algorithm for glottal inverse filtering, it is believed that the observations and algorithms developed this work will be applicable to other joint estimation glottal inverse filtering methods.

Because essentially the described algorithm chooses between parallel fixed filter order analyses, the system is well-suited for the deconvolution of a speech segment with changing filter order.

Additionally, there are a number of elements of the speech production system that may adversely effect the results of a glottal inverse filtering procedure, e.g. inability to adequately model source-filter interaction, turbulent aspiration noise, limitations of the various models etc. Even with the correct number of poles, it is possible that an algorithm neglecting these element might fail to converge to an acceptable solution.

The linear prediction method used in this work to quantify inverse filtering results was used originally in [8] to refine estimates of the instant of glottal closure. One problem with this approach was that occasionally the method would select a waveform accordingly, yet it was not always certain that the results would be inverse filtering. The inclusion of a glottal model as in this paper rectifies that issue: the time-domain residual signal must adhere to a stricter shape. The ability of this new algorithm to refine estimates of the instant of glottal closure has yet to be explored and is a possible avenue of future work.

References

[1] Gunnar Fant. *Acoustic theory of speech production with calculations based on X-ray ...* Walter de Gruyter, 1970.

[2] Hui-Ling Lu. *Toward a High Quality Singing Synthesizer with Vocal Texture Control.* Ph.d., Stanford University, 2002.

[3] Damien Vincent, Olivier Rosec, and Thierry Chonavel. Estimation of LF glottal source parameters based on an ARX model. In *INTERSPEECH*, volume 4, pages 333–336, Lisbon, Portugal, 2005.

[4] J.D. Markel and A.H. Jr. Gray. *Linear Prediction of Speech.* Springer, 1982.

[5] Dennis H. Klatt and L C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[6] Lawrence R. Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, 1978.

[7] Bishnu S. Atal and Lawrence R. Rabiner. A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classifcation with Applications to Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):201–212, 1976.

[8] Elliot Moore and Mark Clements. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. In *ICASSP 2004*, volume 1. IEEE, 2004.

[9] Donald G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *The Journal of the Acoustical Society of America*, 1991.

[10] Paavo Alku, Carlo Magi, S Yrttiaho, Tom Bäckström, and Brad H. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *The Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

[11] Gunnar Fant, J Liljencrants, and Q Lin. A four-parameter model of glottal flow. *STL-QPSR*, 1985.

[12] Thomas Drugman and Thierry Dutoit. Glottal Closure and Opening Instant Detection from Speech Signals. In *Interspeech 2009*, pages 0–3, 2009.

[13] Arturo Camacho. *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator.* Ph.d., University of Florida, 2007.

[14] Michael David Plumpe. *Modeling of the glottal flow derivative waveform with application to speaker identification.* Masters, Massachusetts Institute of Technology, 1997.