2010-01-01

Exploiting Glottal Formant Parameters for Glottal Inverse Filtering and Parameterization

Alan O'Cinneide
Technological University Dublin, alan.ocinneide@tudublin.ie

David Dorran
Technological University Dublin, david.dorran@tudublin.ie

Mikel Gainza
Technological University Dublin, Mikel.Gainza@tudublin.ie

See next page for additional authors

Exploiting Glottal Formant Parameters for Glottal Inverse Filtering and Parameterization

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

Audio Research Group, Dublin Institute of Technology
Kevin Street, Dublin 8, Republic of Ireland
alan.ocinneide@dit.ie, david.dorran@dit.ie, mikel.gainza@dit.ie, eugene.coyle@dit.ie

Abstract

It is crucial for many methods of inverse filtering that the time domain information of the glottal source waveform is known, e.g. the location of the instant of glottal closure. It is often the case that this information is unknown and/or cannot be determined due to e.g. recording conditions which can corrupt the phase spectrum. In these scenarios, alternative strategies are required. This paper describes a method which, given the parameters of the glottal formant of the signal frame, can accurately parameterize the glottal shape source and vocal filter for a broad range of voice quality types and which is robust to the corruption of the phase spectrum.

Index Terms: glottal inverse filtering, frequency domain, glottal models, glottal formant

1. Introduction

Glottal inverse filtering is an operation used to determine the 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 , the synthesis of natural speech and timbral modifications of voice quality.

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 common assumption is that the glottal source has a specific time domain shape. For example, glottal closed phase inverse filtering [2] assumes that the source signal exhibits intervals of null flow between successive glottal pulses. These time regions correspond to the periods when the glottis is closed.

By making assumptions about the time domain shape of the glottal pulse signal during an entire pitch period, researchers have incorporated time domain models of the glottal signal into inverse filtering techniques, e.g. [3]. This step can be seen as an extension of closed phase glottal inverse filtering by also including the portions of the glottal cycle during which the vocal folds are open. Additionally to the optimal vocal tract filter parameters, these methods also simultaneously determine the optimal parameters of the incorporated glottal model, and hence are sometimes referred to as joint estimation techniques.

Glottal inverse filtering methods based upon time domain assumptions like those described require specific information regarding the timing of the glottal pulses. This information is typically supplied in the form of the instant of maximum excitation of the glottal pulse - sometimes referred to as the speech epoch, or the instant of glottal closure. This instant can be obtained by the analysis of an electroglottograph signal recorded simultaneously with the speech signal [4], or via algorithms to detect the epochs from the speech waveform directly, e.g. [5].

Other inverse filtering methods utilize frequency domain assumptions, thereby avoiding the requirement for time domain information. One example of a frequency-domain glottal inverse filtering method is the Iterative Adaptive Inverse Filtering (IAIF) method [6]. IAIF assumes that the glottal source can be viewed as a low-order all-pole filter, the parameters of which are estimated in an iterative procedure directly from the speech signal. This all-pole filter representation of the source is then inversely applied to the analysis frame, canceling the contribution of the glottal source before the estimation of the vocal tract.

However, the adaptive all-pole glottal model constructed in the IAIF approach bears little direct resemblance to the prevalent glottal models, e.g. the KLGLOTT88 model [7] or the Liljencrants-Fant (LF) model [8]. In an effort to maintain the advantages of frequency domain inverse filtering while constraining the source to a particular shape, [9] used the magnitude frequency domain transformation of the KLGLOTT88 model in a exhaustive search approach to determine the optimal vocal tract filter and source parameters.

As is argued in [3], the LF model is a more versatile glottal flow model than the KLGLOTT88 model, reputed to offer more complete coverage over the range of glottal waveforms. This paper outlines a joint estimation technique that is robust to phase corruption of the signal and accurately models the glottal derivative source signal using the LF model. This approach retains the flexibility of frequency domain approaches, while utilizing a preferred source model.

The paper is arranged as follows: the following section gives the necessary background of the acoustic theory of speech production, and the models of its various components. The frequency domain features of the LF model are also described. The third section discusses the frequency domain method for separating the source and filter components. Experiments comparing the method to two other frequency domain based inverse filtering methods are described in the fourth section. The fifth section discusses the results yielded by these experiments. Conclusions are drawn in the final section, which also outlines some directions for future research.

2. Background

2.1. Acoustic Theory of Speech Production

The acoustic theory of speech production [1] views speech $S(z)$ as the convolution 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 un-branched 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 [10]. 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:

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

2.2. The LF Model of the Glottal Flow Derivative

The LF model represents the general flow shape of the glottal flow derivative over one glottal cycle and whose shape can be uniquely described with four parameters. The mathematical formula describing the LF model is a piece-wise function, the evolution of which can be seen in Fig. 1.

The first segment is an exponentially increasing sine function of frequency ω_g and bandwidth α, scaled by the number E_0. This portion of the waveform characterizes the glottal flow derivative from the instant of glottal opening t_o, through the time axis at t_p, to the maximum negative extreme E_e at instant t_e. At this point the second segment of the LF model, often referred to as the return phase, begins. This portion models the glottal closure as a modified exponential function which returns to zero at a rate determined by the steepness of the slope of the tangent to the function at t_e. The distance of this tangent's time axis intercept from t_e is called T_a, and is referred to as the effective duration of the return phase. The parameter ϵ is the decay constant of the exponential. The total number of samples in the pulse is the pitch period, referred to as T_0.

The timing parameters of the model are calculated relative to the glottal opening instant, i.e. $T_p = t_p - t_o$, $T_e = t_e - t_o$, etc. Below are the mathematical equations describing time domain LF model shape using these parameters:

$$u_{LF}(n) = \begin{cases} E_0 e^{\alpha n} \sin \omega_g n & \text{for } 0 \leq n < T_e \\ \frac{-E_e}{\epsilon T_a}(e^{-\epsilon(n-T_e)} - e^{-\epsilon(T_c-T_e)}) & \text{for } T_e \leq n \leq T_c \\ 0 & \text{for } T_c \leq n < T_0 \end{cases} \tag{3}$$

As noted in [8], it is convenient to set $T_c = T_0$. Under this formulation, both the return and closed phase can be very closely approximated as the truncated impulse response of a single positive real pole IIR filter. This arrangement is a suitable approximation for many voice types, and relates the filter pole position μ_{ret} to the T_a parameter by the equation $T_a = \frac{-1}{\ln \mu_{ret}}$ [11].

The LF model parameters are related to one another via certain interdependencies which must be respected in order to generate a valid LF model pulse. Amongst these requirements is the principle of area balance, the requirement that the pulse exhibit no baseline drift over the course of its evolution. This can be represented mathematically as:

$$\int_0^{T_0} u_{LF}(n)dn = 0 \tag{4}$$

Among the different parameter sets that can be used to generate an LF model are the shape parameters. These parameters are the open quotient of the pulse, its asymmetry coefficient, and its return phase coefficient (denoted as O_q, α_m and Q_a, respectively) which can be expressed by the following equations:

$$O_q = \frac{T_e}{T_0} \quad \alpha_m = \frac{T_p}{T_e} \quad Q_a = \frac{T_a}{(1-O_q)T_0} \tag{5}$$

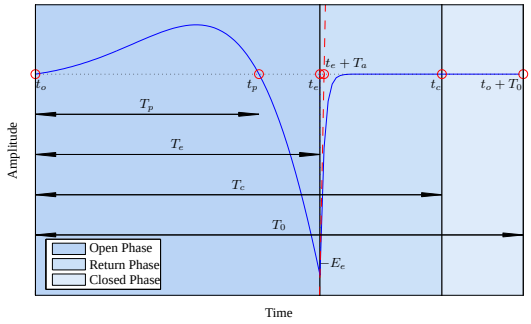

Figure 1: *An LF model of derivative glottal flow, with timing parameters* $(T_o, T_e, T_a, T_c, T_p)$ *and amplitude parameter* E_e. *Also marked are the different phases of the glottal cycle and the tangent at* $(T_e, -E_e)$ *which defines* T_a.

3. A Method of Glottal Inverse Filtering and Source-Filter Parameterization

Within this section, a new method of estimating the glottal source shape and filter parameters of a voiced speech signal is described. The complete algorithm can be described in three basic stages which are, as follows: (1) the estimation of the fundamental frequency of voicing and the glottal formant parameters of the analysis frame, (2) the generation of a discrete set of candidate LF model shapes from the glottal formant parameters and the implications of the LF model principle of area balance, and (3) the application of the inverse magnitude spectrum of these candidates to the signal frame, which then undergoes an autoregressive analysis. The result of this analysis are the all-pole parameters of the vocal tract filter and the shape parameters of the LF model source.

For the purpose of validating the second and third stages of this joint estimation algorithm, the glottal formant parameters and the local fundamental frequency of speech f_0 are taken as given. The first subsection will discuss the determination of these parameters, while the following subsections will elaborate on the remaining algorithmic steps.

3.1. Estimating the Fundamental Frequency and Parameters of the Glottal Formant

As is often required with many glottal inverse filtering algorithms, an estimate of the signal's fundamental frequency f_0 is necessary. This value facilitates the pitch-synchronous analysis of the signal, and can be estimated via e.g. [13].

The method described here also requires an estimate of the glottal formant parameters. The glottal formant is a band of increased spectral energy in the region of the voiced speech fundamental frequency. The term glottal formant is a misnomer as there is no resonance effect in the same manner as is with the vocal tract: rather, the frequency and degree of the spectral energy boost is related to both the shape and duration of the signal's opening phase. The parameters controlling these aspects of the glottal formant are given in the open phase equation of the LF model: its center frequency is ω_g its bandwidth is α.

There exist methods for determining the glottal formant, e.g. [14]. This method relies on the phase information of the

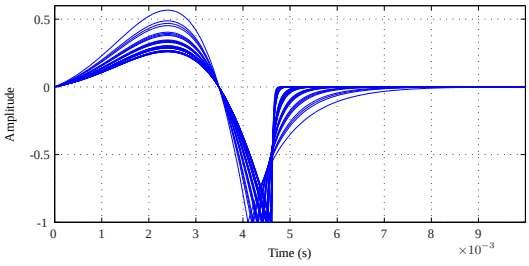

Figure 2: *A set of valid LF model pulses generated using the described method for fixed α and ω_g parameters.*

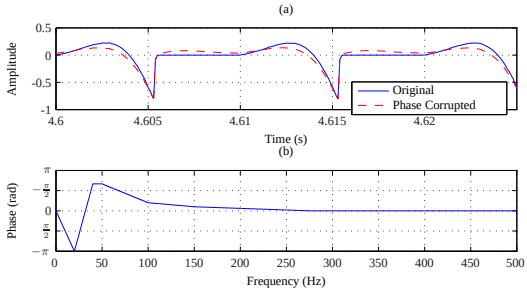

Figure 3: *The above figure shows (a) a comparison of the original and phase corrupted LF model pulses, and (b) the low frequency detail of the phase response used for phase corruption, adapted from* [18].

signal, and is thus unsuitable for analysis for phase-corrupted signals. A first attempt for the estimation of this signal in the frequency domain has been described in [15], which provides accurate results in a limited scenarios. Further work is required to make this technique more robust.

3.2. Determining the LF Model Shape Candidates

Given an estimate of the glottal formant parameters and assuming that the return and closed phases can be modeled together as the truncated impulse response of a single positive real pole IIR filter, a unique LF pulse can then be generated dependent solely on the position of the filter's pole. This section describes how a subset of these shapes can be generated by varying the pole position and solving the area balance equation 4.

First, the formulation of the normalized LF pulse shape $\hat{u}_{LF}$ is determined by setting the E_e parameter to 1. As the two segments of the LF model are continuous at T_e, the value of E_0 can then be expressed as:

$$E_0 = \frac{-1}{e^{\alpha T_e} \sin \omega_g T_e} \tag{6}$$

The area underneath this normalized open phase can be shown to be:

$$\int_0^{T_e} \hat{u}_{LF}(n)dn = \frac{-(e^{\alpha T_e}(\alpha \sin \omega_g T_e - \omega_g \cos \omega_g T_e) + \omega_g)}{(e^{\alpha T_e}(\alpha^2 + \omega_g^2) \sin \omega_g T_e)} \tag{7}$$

Following from the assertion that the LF model return phase can be described as a first-order low-pass IIR filter with pole position amplitude μ_{ret} [11], the area under the normalized return phase of the LF model can be shown to be closely approximated by the following equation:

$$\int_{T_e}^{T_0} \hat{u}_{LF}(n)dn = \int \mu_{ret}^n dn \approx \frac{1}{\ln \mu_{ret}} \tag{8}$$

Thus, the area balance equation can be re-expressed in terms of ω_g, α, T_e and μ_{ret}:

$$\int_0^{T_0} \hat{u}_{LF}(n)dn = \int_0^{T_e} \hat{u}_{LF}(n)dn + \int_{T_e}^{T_0} \hat{u}_{LF}(n)dn = 0 \tag{9}$$

$$\Rightarrow \frac{-(e^{\alpha T_e}(\alpha \sin \omega_g T_e - \omega_g \cos \omega_g T_e) + \omega_g)}{(e^{\alpha T_e}(\alpha^2 + \omega_g^2) \sin \omega_g T_e)} + \frac{1}{\ln \mu_{ret}} = 0 \tag{10}$$

If the values of ω_g and α are given, a limited subset of LF model shapes can be determined by sampling μ_{ret} for a range of discrete values, e.g. $\mu_{ret} = 0 : 0.01 : 0.9$.

The shape parameters of the pulse can be determined using the estimate of the pitch period $T_0 = \frac{f_s}{f_0}$ and the formulae given in equation 5. As these parameters are subject to certain limits, those sets of parameters where one of its members falls outside these limits are removed from the final set of shape candidates. One such subset can be seen in Fig. 2.

3.3. Determining the Vocal Tract Filter

Once a set of normalized LF model candidates have been determined, each LF model candidate is inverse applied to the spectrum of the Hann-windowed speech frame. As the ideal glottal derivative pulse shape would theoretically leave only the vocal tract resonances and these resonances are assumed to be from an all-pole filter, the quality of the de-convolution can be measured from the unmodelable residue that remains following an autoregressive analysis. In this work, these spectra were analyzed using discrete all-pole analysis [16], which determines the optimal all-pole filter based upon the minimum Itakura-Saito distance measure. The discrete μ_{ret} parameter that yields the minimum of this value is then refined using a simplex optimization method [17] to obtain the final shape parameter estimates.

4. Experiments and Discussion

In order to validate the algorithm, two experiments were performed: one to assess the performance of the algorithm using the usual speech model described by equation 2, another to determine the performance during a more realistic, interactive speech model. At a sampling frequency of 10kHz, two 5-second LF model source signals were generated corresponding to a male and female voice ($f_0 = 100$ and 180Hz respectively). Modulated Gaussian noise was added to the source signal that is filtered using the interactive speech model for increased naturalness. The parameters of the LF model shape and the signal-to-noise ratio of the pulses to the added aspiration noise were taken from [12], and selected to represent a smoothly changing voice quality from pressed to modal to breathy.

Using the gender-appropriate all-pole vocal tract transfer function of the vowel /æ/, two types of speech signals were generated. One adhered exactly to the linear speech model; the other simulates the interaction effects known to be present in natural speech by implementing a different filter during the open phase of the signal, where the center frequency and bandwidth of the first formant are both increased. As it is desired to ultimately apply this technique to phase corrupted signals, both synthetic segments are phase-distorted by applying an all-pass filter which mimics the response of a PC sound card, see Fig. 3.

From each of these signals, a total of 1000 frames $(2T_0 + 1)$ samples in length were extracted at different points. Each frame

underwent the above-described analysis. This validation assumes that the glottal formant has been estimated by some method and thus its parameters are simply passed to the algorithm. The glottal shape parameters deduced by the algorithm are assessed by determining their deviations from the true values. The results of the algorithm are analyzed by the means μ and standard deviations σ of the deviations of O_q, α_m and Q_a from their true values. As an additional method of expressing the similarity between the original and determined normalized LF model pulse shapes, the correlation function ρ of the pulses is also determined, the global mean of which is calculated. These results are shown in Tables 1 and 2.

f_0	ΔO_q		$\Delta\alpha_m$		ΔQ_a		ρ
	μ	σ	μ	σ	μ	σ	μ
100	-0.0003	0.0037	-0.0000	0.0047	0.0011	0.0105	0.9980
180	0.0007	0.0043	-0.0014	0.0062	-0.0013	0.0115	0.9952

Table 1: *The deviation of the glottal shape parameters from their true values estimated for the phase corrupted synthetic speech signal following a linear speech model.*

f_0	ΔO_q		$\Delta\alpha_m$		ΔQ_a		ρ
	μ	σ	μ	σ	μ	σ	μ
100	-0.0048	0.0066	0.0062	0.0077	0.0127	0.0192	0.9902
180	-0.0037	0.0076	0.0045	0.0095	0.0102	0.0210	0.9871

Table 2: *The deviation of the glottal shape parameters estimated for the pitch corrupted synthetic speech signal where the parameters of the vocal tract filter change during the open phase of the glottal pulse.*

Studies by Henrich et al [19] suggest that the just noticeable difference (JND) for the open quotient is linearly related to its initial value by the equation $JND_{O_q} \simeq \beta O_q$, where β is 0.14 for untrained listeners and 0.1 for trained listeners. They also report that the smallest JND_{α_m} for trained listeners is 0.022. In all the experimental cases, the results of the described algorithm indicate that these values are below the perceptual threshold of human listeners. While the deviations of these parameters increase slightly in the case of the more realistic, interactive speech simulation, given robust glottal formant parameters the method can generally determine a perceptually equivalent parameter set.

The mean of the correlation coefficient gives an indication of the similarity of the source waveforms. In both cases, the analysis of the lower pitch signal determines a parameter set that is closer to the original, while the more realistic scenario presents a slightly worse performance. This was expected due to the unmodeled elements of the signal.

5. Conclusions and Future Work

This paper describes a method of inverse filtering that exploits prior information of the glottal formant parameters that does not require the determination of specific timing parameters. Using an estimate of these parameters for a frame of voiced speech, a discrete set of the LF model candidate source shapes can be determined, which are then individually applied to the signal frame. The parameter set yielding the lowest energy is refined using an optimization procedure to determine the best fit glottal shape parameters.

The method was demonstrated to successfully obtain good parameterizations of the LF model source signal. The degradation of the results was noted in the more realistic speech scenario involving a modulation first formant during the open phase of the source flow. Additionally, results were also worse in the case of higher fundamental frequency. However, despite this decrease in accuracy in these circumstances, it can be concluded that this technique is suitable for inverse filtering and source parameterization as the deviations from the ideal are within perceptual bounds.

The determination of the glottal formant parameters is critical to this method: future research will concentrate on the robust estimation of these parameters from the speech frame. In order to obtain a full description of the source, the above algorithm should be expanded to estimate the source amplitude E_e. Another concern is the continuity of the parameters during running speech; the parameter trajectories across the speech frames of an utterance could be smoothed by e.g. the Viterbi algorithm.

6. References

[1] G. Fant, *Acoustic theory of speech production*, Walter de Gruyter, 1970.

[2] D. Wong et al., "Least squares glottal inverse filtering from the acoustic speech," *IEEE T. Acoust. Speech*, 1979, pp. 350-355.

[3] H. Lu, "Toward a High Quality Singing Synthesizer with Vocal Texture Control," Ph.D. thesis, Stanford Uni., 2002.

[4] A. Krishnamurthy and D. Childers, "Two-channel speech analysis," *IEEE T. Acoust. Speech*, vol. 34, 1986, pp. 730-743.

[5] P.A. Naylor et al., "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE T. Audio Speech*, vol. 15, 2007, pp. 34-43.

[6] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, 1992, pp. 109-117.

[7] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, 1990, pp. 820-857.

[8] G. Fant et al., "A four-parameter model of glottal flow," *STL-QPSR*, 1985.

[9] I. Arroabarren and A. Carlosena, "Glottal spectrum based inverse filtering," *EUROSPEECH*, 2003, pp. 57-60.

[10] J. Markel and A.J. Gray, Linear Prediction of Speech, Springer, 1982.

[11] A. Ó Cinnéide et al., "On the Appearance of a Real Root at 0Hz in the Results of Glottal Closed Phase Linear Prediction," *EUSIPCO 2010*.

[12] D.G. Childers, Speech Processing and Synthesis Toolboxes, Wiley, 1999.

[13] S. Ahmadi and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE T. Speech Audi. Proces.*, vol. 7, 1999, pp. 333-338.

[14] T. Drugman et al., "Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase," *EUSIPCO 2008*.

[15] A. Ó Cinnéide et al., "Towards a Method to Determine the Glottal Formant Parameters of Voiced Speech without Time domain References," *ISSC 2010*.

[16] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE T. Signal Proces.*, vol. 39, 1991.

[17] J.A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, 1965, pp. 308-313.

[18] O. Akande, "Speech analysis techniques for glottal source and noise estimation in voice signals," Ph.D. thesis, Uni. Limerick, 2004.

[19] N. Henrich et al., "Just noticeable differences of open quotient and asymmetry coefficient in singing voice," *J. Voice*, vol. 17, 2003, pp. 481-494.