2010-01-01

Towards a Method to Determine the Glottal Formant Parameters of Voiced Speech without Time-Domain Reference

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

Eugene Coyle
Technological University Dublin, Eugene.Coyle@tudublin.ie

Towards a Method to Determine the Glottal Formant Parameters of Voiced Speech without Time-Domain References

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 — **This paper presents an approach to estimate the glottal formant parameters of the voicing source in the frequency-domain. The method is based on a simplified pole-zero interpretation of the prevalent Liljencrants-Fant (LF) model of glottal flow, and gives approximations for a broad range of pulses shapes. An advantage of the method is that, unlike other methods, it does not rely on time-domain references.**

Keywords — **glottal source parameterisation, LF model, glottal formant**

I Introduction

Parameterising the voicing source of speech has many applications in speech including: speaker identification [1], the synthesis of natural speech [2] and timbral modifications of voice quality [3].

One frequency-domain aspect of the voicing source is a region of increased spectral energy referred to as the glottal formant. A prevalent method for estimating the glottal formant parameters is to fit a model to the inverse filtered speech signal in the time-domain [4]. Following the convolution of the speech pulse with the vocal tract inverse filter, the various landmarks representing important time points of the glottal source are directly estimated from the time-domain waveform. These initial approximations are then refined by using an optimisation algorithm to minimise a suitable error criterion, e.g. time-domain least squares.

A second method to determine the glottal formant parameters exploits the phase properties of the glottal source to separate the maximum-phase glottal formant from the minimum-phase characteristics of the vocal tract resonances and glottal pulse return phase [5]. By appropriately positioning the analysis frame over the instant of glottal closure and calculating the zeros of the Z-transform, the open phase of the voicing source can be reconstructed from a subset of the calculated roots. The glottal formant parameters can then be determined by an analysis of this signal.

A pre-requisite of both of these methods is knowledge of the evolution of the glottal waveform in the time-domain. These instants can be obtained by the analysis of an electroglottograph signal recorded simultaneously with the speech signal [6], or via various algorithms to detect the epochs directly from the speech waveform, e.g. [7]. However, in the case of signals that have been phase corrupted, by e.g. non-linear phase recording equipment, these instants may not be recoverable without a phase-correction filter, the design of such a filter is not always feasible [8].

This paper develops a strategy for determining the glottal formant characteristics from a glottal pulse signal without reference to its time evolution. The outline of this paper is as follows: the following section gives the necessary background on the acoustic theory of speech production and a detailed description of a prevalent model of a pulse of the derivative glottal flow, the LF model. Here the glottal formant parameters are described along with its time-domain correlates. The third section shows how the LF model can be expressed as a high-order pole-zero filter, and develops an

approximation. Experiments validating this approach are given in the fourth section. The fifth section discusses the results yielded by these experiments. Conclusions are drawn in the final section, which also outlines the direction of future research.

II Background

a) *Acoustic Theory of Speech Production*

The acoustic theory of speech production [9] views speech as the convolution of glottal flow signal with a vocal tract filter which is then radiated at the lips. In the Z-domain, the process can be represented as follows:

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

where $S(z)$ represents the speech waveform, $G(z)$ the glottal flow, $V(z)$ the vocal tract filter, and $L(z)$ represents lip radiation.

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 $G'(z)$.

b) *The LF Model of the Glottal Flow Derivative*

The LF model [10] 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, consisting of two segments, the evolution of which can be seen in Fig. 1. The first segment is an exponentially increasing sine function, characterizing the glottal flow derivative from the instant of glottal opening t_o, through the time axis at t_p, to the instant of maximum negative extreme at 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 total number of samples in the pulse is the pitch period, T_0. The value is related to the fundamental frequency f_0 and the sampling frequency f_s by the expression:

$$T_0 = \frac{f_s}{f_0} \tag{2}$$

In order to correctly place the pulse in time, the timing instants are calculated to be relative to the instant of glottal opening, i.e. $T_o = 0$, $T_p = t_p - t_o$, $T_e = t_e - t_o$ and $T_c = t_c - t_o$. Below are the mathematical equations describing time-domain LF model shape using these parameters:

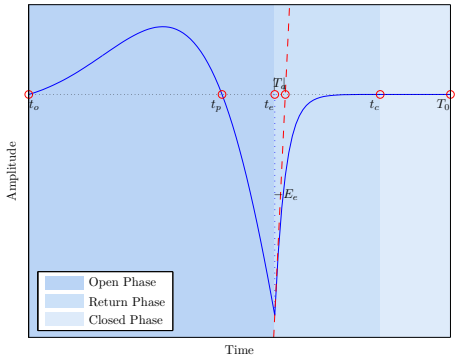

Fig. 1: An LF model of derivative glottal flow, with timing parameters $(t_o, t_e, t_a, t_c, t_p)$ and amplitude parameter E_e. Also marked are the different phases of the glottal cycle and the tangent at $(t_e, -E_e)$ which defines T_a.

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

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

$$O_q = \frac{T_e}{T_0} \quad \alpha_m = \frac{T_p}{T_e} \quad Q_a = \frac{T_a}{(1 - O_q)T_0} \tag{4}$$

These quantities are convenient because they bear a more meaningful relationship with the glottal behaviour than the timing parameters individually. For example, the open quotient is related to the time during which the glottal cycle is in its open phase during the pulse, and therefore an indication of the breathiness of the voice [11]. Another advantage is the normalised limits of these ratios.

c) *Frequency-domain Features of the LF Model*

In their work on the frequency spectra of the glottal models, Doval and d'Alessandro [12] describe the main frequency-domain features of the glottal source models. The salient spectral features of these models are the previously mentioned glottal formant and the spectral tilt. These features are referred to in Fig. 2.

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

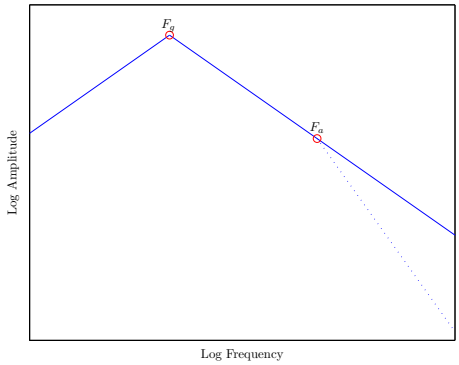

Fig. 2: A stylized LF model spectrum, as in [12]. The lines represent the asymptotic behaviour of the spectrum, while the breakpoints represent the frequencies F_g and F_a respectively. The dotted line represents the change that occurs in the spectrum when an LF pulse exhibits a return phase.

boost is related to both the time-domain duration and shape of the signal's opening phase. Its parameters are given in the open phase equation of the LF model: its center frequency is ω_g while its bandwidth is α. The term F_g is sometimes used to represent the frequency of the glottal formant.

The other aspect of the glottal flow is its decrease in spectral energy with increasing frequency, referred to as the spectral tilt. Contrary to the glottal formant, the spectral tilt is due to the behaviour of the time-domain waveform during its return phase. The most abrupt change, i.e. no return phase, imparts a slope of $-6dB/oct$ on the spectrum; a more gradual return phase introduces a steeper spectral slope. As was illustrated in [13], this behavior is very similar to a first-order low pass IIR filter, and dependent on T_a and the length of the closed phase. The cut-off frequency of this filter is commonly referred to as F_a.

III The LF Model as a Pole-Zero Filter

This section develops an alternative interpretation of the LF model in order to make accessible, through frequency-domain analysis, the parameters describing the glottal formant. For this purpose, a pole-zero interpretation of the LF model is derived. The model is developed such that the parameters of the filter poles correspond to the glottal formant, while the zeros of the filter control the spectral tilt. Once a full model has been described, an approximation to the denominator of this filter is developed. The physical limitations of the parameters of this filter reveal that the variation of this approximation filter is very narrow, and so a general approximation can be derived. The mathematical details follow here.

As previously mentioned, the open phase of the LF model is an exponentially increasing sine wave: this can be fully modelled as the impulse response of an unstable 2^{nd} order IIR filter, truncated at the point T_e. The Z-transform of the open phase then takes the form:

$$H_{op}(z) = \frac{1}{1 - 2e^{\alpha}\cos\omega_g z^{-1} + e^{2\alpha}z^{-2}} \tag{5}$$

Recall that the parameters of this filter ω_g and α are the centre frequency and bandwidth of the previously mentioned glottal formant.

In order to completely represent the LF model as a pole-zero model, it is desirous to construct a filter which will cancel the exponentially expanding sine wave beyond the point T_e to create the glottal closed phase. In the case of glottal waveforms that close abruptly, i.e. immediately after the instant of glottal closure, it can be shown that this cancelling filter can be constructed according to the following transfer function:

$$H_{cp}(z) = 1 + b_{T_e+1}z^{-(T_e+1)} + b_{T_e+2}z^{-(T_e+2)} \tag{6}$$

where

$$b_{T_e+1} = \frac{e^{\alpha T_e}(\sin\omega_g(T_e - 1) - 2\cos\omega_g \sin\omega_g T_e)}{\sin\omega_g} \tag{7}$$

$$b_{T_e+2} = \frac{e^{\alpha(T_e+1)}\sin\omega_g T_e}{\sin\omega_g} \tag{8}$$

The FIR filter H_{cp} can be seen as a truncating/cancelling filter, whose numerous zeros impart a certain degree of spectral tilt and ripple onto the spectrum of the LF model.

In order to extend this representation to the more general case where the LF model pulse does exhibit a return phase, rather than the cancel the expanding sine wave of the open phase, H_{cp} must instead supply it.

The LF model return phase is usually approximated by a low-pass IIR filter of the form [13]:

$$H_{ret}(z) = \frac{1}{1 - \mu_{ret}z^{-1}} \tag{9}$$

The impulse response of the filter $H_{ret}(z)$ can be approximated by an FIR filter of sufficiently high order. Such a behaviour can be incorporated into the filter H_{cp} by augmenting it such that the n^{th} term beyond b_{T_e+2} can be determined from the following equation:

$$b_{T_e+2+n} = \mu_{ret}^n b_{T_e+2} \tag{10}$$

Thus, the generalised H_{cp} filter which cancels the open phase and adds a return phase can be expressed:

$$H_{cp} = 1 + b_{T_e+1}z^{-(T_e+1)} + b_{T_e+2}R(z) \tag{11}$$

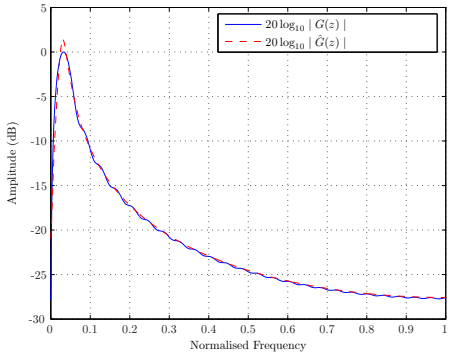

Fig. 3: The log magnitude frequency response comparison of $G(z)$ and $\hat{G}(z)$, adjusted for scale.

where

$$R(z) = z^{-\tau} + \mu_{ret} z^{-(\tau+1)} + \cdots + \mu_{ret}^{n} z^{-(\tau+n)} + \cdots \tag{12}$$

and $\tau = T_e + 2$.

A complete pole-zero model of an LF model pulse can therefore be expressed:

$$\begin{aligned} G'_{LF}(z) &= H_{cp}(z) H_{op}(z) \\ &= \frac{1 + b_{T_e+1} z^{-(T_e+1)} + b_{T_e+2} R(z)}{1 - 2e^{\alpha} \cos \omega_g z^{-1} + e^{2\alpha} z^{-2}} \end{aligned} \tag{13}$$

From the above derivation, it follows that the open phase filter parameters H_{op} can be estimated given an approximation of the filter H_{cp}. However, due to the complexity of the filter given by (13), in order to produce a tractable approximation, the approximation developed here is based on the simpler transfer function of the abruptly closing LF model, whose truncating filter is given by (6).

A filter which approximates the behaviour of the impulse response of H_{cp} near the point of glottal closure is a filter of the form:

$$\hat{H}_{cp}(z) = 1 + \beta z^{-1} \tag{14}$$

where $\beta = \frac{b_{T_e+2}}{b_{T_e+1}}$. This ratio can be expressed as:

$$\beta = \frac{e^{\alpha}}{\frac{\sin \omega_g (T_e - 1)}{\sin \omega_g T_e} - 2 \cos \omega_g} \tag{15}$$

The range of values of β are dependent upon the variables ω_g, α and T_e. Surveying the range of β values given the physically reasonable limits of these parameters indicates that a good estimate for β is -1.01. The similarity between a glottal pulse transfer function given by $G_{lf}(z) = \frac{H_{cp}(z)}{H_{op}(z)}$ and the approximation $\hat{G}_{lf}(z) = \frac{1-1.01z^{-1}}{H_{op}(z)}$ can be seen in Fig. 3.

From this approximation, it follows that the coefficients describing the glottal formant open phase filter can be estimated by a two-pole autoregressive analysis on the glottal source that has been pre-emphasised by the inverse of the filter given by (14). However, as β will be less than -1, this filter would constitute an unstable filter. In order to avoid this issue, the stable equivalent of this filter, i.e. where its pole is reflected about the unit circle, can be as effective once the signal is time reversed. For the analysis algorithm, the signal frames are pre-emphasised by the following filter:

$$H_{pe}(z) = \frac{1}{1 + \frac{1}{\beta} z^{-1}} \tag{16}$$

IV Experiment

An experiment was carried out in order to test the efficacy of the above approximation to determine the glottal formant parameters. At a sampling frequency of 10kHz, a broad range of different LF model pulses were generated, the parameters of which are listed in Table 1. In order to minimise the effect of the spectral tilt on the estimation of the glottal formant, each analysis pulse was first decimated to 2000 Hz.

Parameter	Range
f_0	80 : 20 : 200 (Hz)
O_q	0.3 : 0.05 : 0.9
α_m	0.67 : 0.05 : 0.9
Q_a	0 : 0.05 : 1

Table 1: All LF model parameter configurations used for synthetic testing.

In order to cancel the effects of H_{cp}, the frame was then time-reversed and pre-emphasised using the filter H_{pe}, with β chosen to be -1.01. A Hann window was then applied to the pulse, which then underwent second-order discrete all-pole (DAP) analysis [14], yielding an IIR filter.

The roots of the filter polynomial were calculated in order to locate the poles on the Z-plane. The coordinates of the poles underwent a Cartesian to polar conversion, yielding the estimates of the glottal formant center frequency $\hat{\omega}_g$ and the exponential of its bandwidth $\hat{\alpha}$. As DAP analysis will generally find stable poles, i.e. within the unit circle, the value of $\hat{\alpha}$ is multiplied by -1 in order to determine the pole's maximum-phase reflection.

The errors of $\hat{\omega}_g$ and $\hat{\alpha}$ are compared to their original values by the following formulae:

$$E_{\omega_g} = \frac{\hat{\omega}_g - \omega_g}{\omega_g} \qquad E_{\alpha} = \frac{e^{\hat{\alpha}} - e^{\alpha}}{e^{\alpha}} \tag{17}$$

The error associated with is calculated as the percentage error of the pole radius, rather than the percentage error of $\hat{\alpha}$. This is because the small values of α tends to exaggerate even slight deviations from its true value.

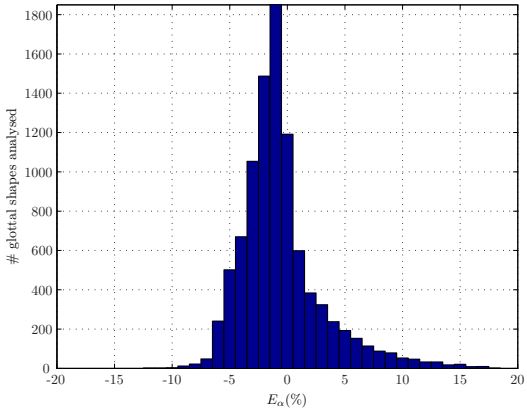

Fig. 4: A histogram displaying the results of the experiment described to estimate α.

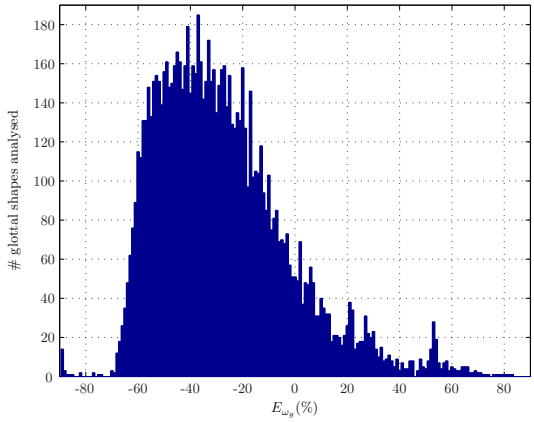

Fig. 5: A histogram displaying the results of the experiment described to estimate ω_g.

V Discussion of Results

The results of the experiment are given in Figs. 4 and 5. Each figure plots the percentage error associated with α and ω_g against the number of waveforms analysed exhibiting that error.

As can be seen in Fig. 4, the E_α is generally slightly underestimated. The accentuation of the low frequencies of the filter H_{pe} tends to "dull" spectral peaks of the glottal formant, reducing the bandwidth. This slight affect can be seen in Fig. 6, where increased return phase coefficients, which are known to exhibit broader glottal formants (see [12]), are shown to be correlate with slightly over-estimated bandwidth values.

Contrary to E_α, the values of E_{ω_g} exhibit a far wider degree of variation. The approximation tends to underestimate the glottal formant centre frequency.

The main source of error can most likely be attributed to the coarse approximation of the closed phase truncation filter H_{pe}. This filter necessarily boosts the low frequencies in order to compensate for the truncation of the open phase filter, but seems to be too crude in the cases where a more gradual return phase is required to be compensated. Indeed, even in cases of abrupt glottal closures, ω_g values are underestimated, see Fig. 7. Clearly, "one-size-fits-all" pre-emphasis is inadequate for the precise determination of the glottal formant parameters across the variety of voice source signals that may be encountered.

As can be seen in Fig. 7, more localised results are exhibited for glottal pulses that are characterised by a null return phase coefficient value. As the return phase coefficient grows larger, the results become more spread, consistently bias towards underestimating the true ω_g value. It is thought that this is due to the lack of a representation of the return phase within the approximation filter H_{pe}. Additionally, this filter is constructed

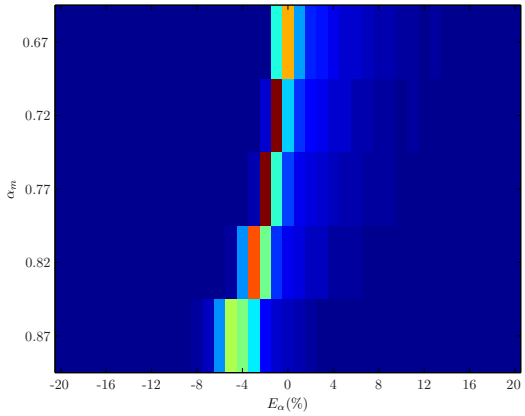

Fig. 6: The distribution of the E_α compared with the α_m parameters.

from an approximation to the signal discontinuity required to truncate the open phase expanding sine wave at the point of glottal closure. The neglect of the initial filter coefficient of H_{cp} would introduce some low frequency content which may also distort the subsequent auto-regressive analysis.

The goal of this research is to develop a method of determining the glottal formant without reference to the time-domain shape of the signal, yet the placement of the source signal within the frame effects the results in a subtle fashion. The experiment above was repeated for different pulse shapes placements, and the results obtained were broadly similar.

VI Conclusion

This paper presents a method to estimate the glottal formant parameters, without the reliance on the specific time information of the signal. This was achieved first by developing a transfer function whose poles and zeros represent the LF model. The denominator and numerator of the transfer

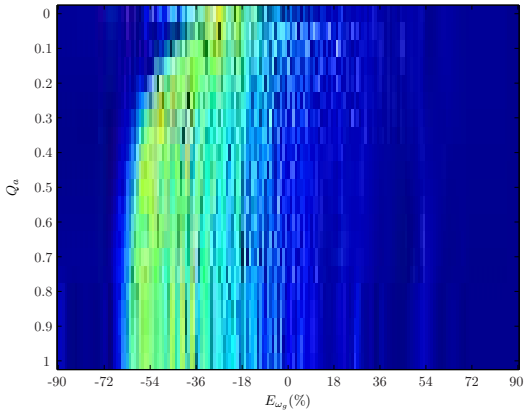

Fig. 7: The distribution of the E_{ω_g} compared with the Q_a parameters.

function represent separately the open and return/closed phase of a modified LF model respectively. In order to estimate the glottal formant parameters, characterised in the open phase of the LF model, a first order approximation to the numerator was developed, allowing an estimation of the glottal formant parameters.

The results of this method revealed that the estimations of the formant bandwidth are generally less than $\pm 5\%$, while the error is larger for the glottal formant centre frequency. It is believed that this is mainly due to a lack of generality in the first order approximation. An adaptive strategy using other available signal information (e.g. the amplitude difference between the first two harmonics which has been shown to be able to estimate O_q for moderate values [15]) may lead to better pre-emphasis filters.

Additionally, it is believed that this method to determine the glottal formant can be used to better inform an inverse filtering strategy regarding the contributions of the glottal source. A novel method of inverse filtering utilising such an approach is currently under development.

References

[1] Plumpe, M., et al., "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.: 7, pp. 569-586, 1999.

[2] Klatt, D.H. & L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acous. Soc. of America*, vol. 87, pp. 820-857, 1990.

[3] Vincent, D., O. Rosec, & T. Chonavel, "A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling," ICASSP, vol. 4, pp. 525528, 2007.

[4] Strik, H., "Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses," *The Journal of the Acoustical Society of America*, vol. 103(5), pp. 2659-2669, 1998.

[5] Drugman, T., Bozkurt, B., & Dutoit, T., "Chirp decomposition of speech signals for glottal source estimation", ISCA Workshop on Non-linear Speech Processing, 2009.

[6] Krishnamurthy, A. & D. Childers, "Two-channel speech analysis," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34(4), pp. 730-743, 1986.

[7] P. A. Naylor, A. Kounoudes, J. Gudnason, & M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech using the DYPSA Algorithm," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 15, pp. 3443, Jan. 2007.

[8] Akande, O., O.: Speech analysis techniques for glottal source and noise estimation in voice signals. Ph. D. Thesis, Uni. Limerick, 2004.

[9] Fant, G., *Acoustic Theory of Speech Production*. The Hague: Mouton, 1970.

[10] Fant, G., J. Liljencrants, & Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, pp. 1-13, 1985.

[11] Childers, D. G., Lee, C. K., "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90(5), pp.2394-2410, November 1991.

[12] Doval, B. and d'Alessandro, C., "The spectrum of glottal flow models." Notes et document LIMSI, num. 9907, 1999.

[13] Ó Cinnéide, A., Dorran, D., & Gainza, M., "On the Appearance of a Real Root at 0Hz in the Results of Glottal Closed Phase Linear Prediction", EUSIPCO 2010 (to appear).

[14] El-Jaroudi. A., Makhoul. J., "Discrete all-pole modeling," IEEE Trans. Signal Proc., vol. 39. pp. 411-423. Feb. 1991.

[15] Doval, B., d'Alessandro, C., & Diard, B., "Spectral methods for voice source parameters estimation", In EUROSPEECH-1997, 533-536.