INTERSPEECH 2009 BRIGHTON

Towards Robust Glottal Source Modeling

Javier Pérez, Antonio Bonafonte

Department of Signal Theory and Communication
TALP Research Center
Technical University of Catalonia (UPC), Barcelona, Spain
{javierp,antonio}@gps.tsc.upc.edu

Abstract

We present here a new method for the simultaneous estimation of the derivative glottal waveform and the vocal tract filter. The algorithm is pitch-synchronous and uses overlapping frames of several glottal cycles to increase the robustness and quality of the estimation. Two parametric models for the glottal waveform are used: the KLGLOTT88 during the convex optimization iteration, and the LF model for the final parametrization. We use a synthetic corpus using real data published in several studies to evaluate the performance. A second corpus has been specially recorded for this work, consisting of isolated vowels uttered with different voice qualities. The algorithm has been found to perform well with most of the voice qualities present in the synthetic data-set in terms of glottal waveform matching. The performance is also good with the real vowel data-set in terms of resynthesis quality.

Index Terms: speech synthesis, speech analysis, speech processing, glottal modeling

10.21437/Interspeech.2009-15

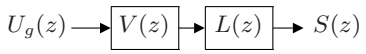
1. Introduction

One of the main goals of researchers on voice quality is the automatic acquisition of reliable voice source measures connected with the human production system proposed by Fant [1]. According to this model, the speech $S(z)$ is produced when the waveform glottal source $U_g(z)$ excites the vocal tract $V(z)$, and is radiated by the lips $L(z)$ (fig. 1 (a)). The model can be simplified by approximating $L(z)$ as a first-order differentiator, and by modeling the $V(z)$ using an all-pole filter. Since all the filters are linear, we can reorder them and work with the derivative glottal volume-velocity waveform instead: $G(z) = U_g(Z) \cdot L(z)$. Then, the simplified source-filter equation is (fig. 1 (b)):

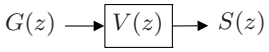
$$S(z) = G(z)\,V(z) = G(z)\,\frac{1}{A(z)} = G(z)\,\frac{1}{1-\sum_{k=1}^{N} a_k\, z^{-k}} \tag{1}$$

In recent years there have been several methods aiming at the automatic estimation of the glottal source and the vocal tract. Most efforts focus into the independent estimation of the vocal tract, and then obtain the glottal waveform by inverse filtering the speech signal [2]. These methods often need to work using only speech segments corresponding to the glottis closed-phase, thus resulting in inaccurate estimations since it can often be very short (or non-existing). Hence, recent research focuses on the joint estimation of both the voice source and the vocal tract [3, 4]. The method we propose belongs to this second category and it is built on the basis of our previous work [5], based on convex optimization techniques as previously proposed for singing speech in [6]. The algorithm operates on pitch-synchronous, frame-based basis, using overlapping frames of several glottal cycles to increase the robustness and quality of the estimation.

The paper is organized as follows. Section 2 presents a brief review of the source-filter model. The proposed algorithm is explained in detail in Section 3. The experimental setup and results are reported in Section 4, and the paper ends with the conclusions and directions for future work (Section 5).



(a) Original



(b) Equivalent

Figure 1: *Block diagram of the production model.*

2. Source-filter model

In this work, we use two models for the derivative glottal waveform $G(z)$: the KLGLOTT88 [7] model, which has a simple mathematical formulation suitable for the core optimization algorithm, and the Liljencrants-Fant (LF) model [8], a well established and more complete model, used for the final parametrization. Figure 2 shows a glottal cycle of each model.

The KLGLOTT88 waveform g_{kl} consists of a basic Rosenberg-Klatt waveform g_{rk}, followed by a first-order low-pass filter controlling the smoothness of the glottis closure (i.e., spectrum tilt $TL(z) = \frac{1}{1-\mu z^{-1}}$):

$$\begin{aligned} g_{kl}(n) &= g_{rk}(n) + \mu\, g_{kl}(n-1), \\ g_{rk}(n) &= \begin{cases} b\, n\,(2n_c - 3n) & ,0 \le n < O_q\, T_0 \\ 0 & ,O_q\, T_0 \le n < T_0. \end{cases} \end{aligned} \tag{2}$$

where O_q is the duration of the open phase (%), T_0 is the glottal cycle length, b controls the amplitude, and $n_c = T_0\, O_q$ is the glottal closure instant.

The LF model is more complex and can be formulated as:

$$g_{lf}(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t) & ,0 \le t \le t_e, \\ -\frac{E_e}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)} & ,t_e < t \le t_c, \\ 0 & ,t_c < t \le T_0, \end{cases} \tag{3}$$

where t_p is the zero-crossing instant, t_e the time instant of the minimum in the derivative, t_a is defined as the point where the tangent to the exponential return phase crosses 0, t_c the moment when the return phase reaches 0 (assumed T_0 here [8]) and E_e

as the absolute value of the minimum of the derivative. t_a relates to the abruptness of the glottal closure, by means of the equivalent parameter $F_a = 1/(2\pi t_a)$ [8], the cut-off frequency at which -6dB/octave are added to the source spectrum. The remaining parameters are computed derived from these [8]. In this work we use the normalized parameters: $R_a = t_a/T_0$, $R_k = t_e/t_p - 1$ and $R_0 = t_e \cdot T_0$.

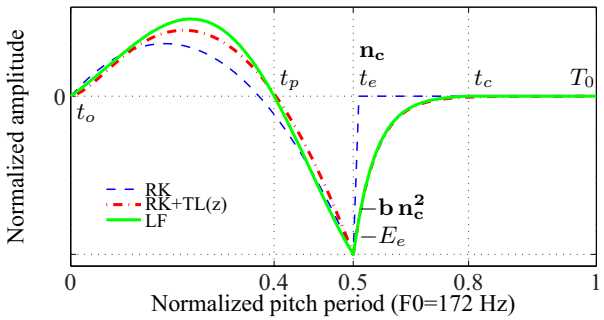
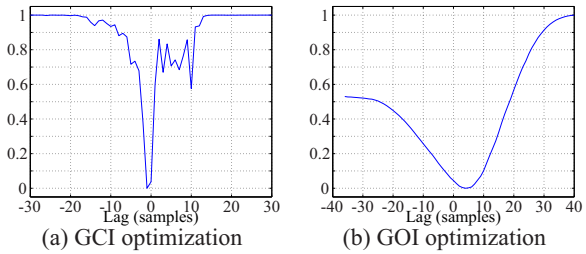


Figure 2: Sample pitch cycle of KLGLOTT88 and LF models

3. Proposed algorithm

3.1. Glottal epochs location

The algorithm needs to be provided with an estimation of the glottal opening (GOI) and closure (GCI) instants. We obtain an initial estimation for the GOIs/GCIs from the laryngograph signal [5], which are optimized (first the GCIs and then the GOIs) by searching for the candidate around the initial estimates giving minimum error from the convex algorithm explained in next subsection (Sec. 3.2, Eq. 5). In Fig. 3 we see that the error surface has only one minimum at the optimal GOI point. It follows that if we obtain the GCIs from the speech signal using e.g., the DYPSA algorithm [9]), the GOIs can then be estimated without need of laryngograph signal by searching over a standard range for the open quotient (OQ) $[0.4 - 0.85]$.



(a) GCI optimization (b) GOI optimization

Figure 3: *Normalized distortion [0,1] for a local search around original (lag 0) (a) GCI, and (b) GOI.*

3.2. Source-filter decomposition using convex optimization

From eq. 1, we see that given a set of $N+1$ filter coefficients (N for the vocal tract, and 1 for μ), the speech signal $s(n)$ can be inverse-filtered to obtain an approximation of the glottal waveform:

$$g_{if}(n) = s(n) - \sum_{k=1}^{N+1} \hat{a}_k s(n-k). \quad (4)$$

As we model it using the KLGLOTT88 model, we want to obtain the parameters that minimize the parametrization error $e(n) = g_{rk}(n) - g_{if}(n)$ for each cycle in the analysis frame:

$$e(n) = \begin{cases} b_1\, C_1(n) + \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n) & OP_1 \\ \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n) & CP_1 \\ b_2\, C_2(n) + \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n) & OP_2 \\ \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n) & CP_2 \\ \vdots & \vdots \\ b_M\, C_m(n) + \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n) & OP_M \\ \sum_{k=1}^{N+1} \hat{a}_k s(n-k) - s(n) & CP_M \end{cases} \quad (5)$$

where we have simplified the notation using $C_i(n) = n(2n_{ci} - 3n)$, and OP_i and CP_i are the open and closed phases for glottal cycle i inside the analysis frame. M is the number of cycles in the analysis frame. Since the error is linear w.r.t. our unknown variables, we can rewrite eq. 5 in matrix form, as $\mathbf{e} = \mathbf{Fx} - \mathbf{y}$, where $\mathbf{x} = [b_1\, b_2 \cdots b_M\ \hat{a}_1 \cdots \hat{a}_{N+1}]^T$ is the vector of variables to be estimated, $\mathbf{y} = [s(1) \ldots s(P)]^T$ contains the speech samples of the analysis frame, and $\mathbf{F}$ is:

$$\mathbf{F} = \begin{pmatrix} C_1(1) & 0 \cdots & 0 & s(0) & \cdots & s(-N) \\ \vdots & \vdots \ddots & \vdots & s(1) & & s(-N+1) \\ C_1(n_{c1}) & 0 \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 \cdots & C_k(1) & \vdots & & \vdots \\ \vdots & \vdots \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 \cdots & C_k(n_{ck}) & \vdots & & \vdots \\ \vdots & \vdots \ddots & \vdots & s(P-2) & & s(P-N-2) \\ 0 & 0 \cdots & 0 & s(P-1) & \cdots & s(P-N-1) \end{pmatrix}$$

where n_{ck} is the GCI of the kth cycle inside the analysis frame of length P. The convex optimization using these new matrices is exactly as in our previous work [5]. The L_2 norm of the error between g_{if} and g_{kl} (Eq. 5) is chosen as the minimization objective. Since the problem is convex, we obtain an optimal set of $N+1$ filter coefficients and g_{rk} parameters. To build g_{kl}, the largest real root of the estimated filter can be assigned to μ, and the remaining N to the vocal tract $V(z)$. To improve the robustness of the estimation, we will estimate μ with a order 1 LPC analysis of the pre-emphasized speech frame [10]. The convex estimation uses only N coefficients and works over the *untilted* (effect of μ removed) speech frame.

3.3. LF parametrization

The last step is to parametrize g_{if} using the LF model. We estimate the LF parameters t_e, t_a, t_p and E_e for each pitch period (t_o and t_c are initially set to 0 and $T_0 = 1/F_0$ and are not modified). First we obtain an initial estimate, and then we use a Sequential Quadratic (SQ) algorithm [5] to refine the initial LF values.

The initialization step is crucial to obtain good estimates. Instead of relying on traditional direct estimation methods prone to errors due to noise [5, 11], we map g_{kl} onto the LF space: t_p is set to the zero-crossing point of g_{kl}, $t_e = n_c$, and $E_e = -g_{kl}(n_c)$. An elegant and efficient initial estimate for t_a

is to set F_a to the cut-off frequency F_c of the KLGLOTT88 tilt filter $TL(z)$, and then compute $t_a = 1/(2\pi F_a)$ [8].

Once we have this initial estimation, the LF parameters need to be optimized. We have found that it is not necessary to further optimize t_e. We perform a simultaneous optimization of t_a, t_p and E_e using a SQ algorithm, and then E_e is reoptimized using a Golden Search method. This improves the robustness and the convergence of the algorithm. The function error is again the L_2 norm between g_{if} and g_{lf}. Convolution with a length 7 Blackman window is used to low-pass filter the pulses and reduce the noise before error computation [11].

4. Results

To validate the proposed algorithm, we have created a two different test corpus. The first one is a synthetic corpus using realistic LF data (Table 1). Each configuration is used to create reference glottal source signals by concatenating the same pulse 30 times. We add white Gaussian noise at several SNR levels (from 5dB to 20dB, in 5dB increments), amplitude-modulated using a Hanning window placed at the GCIs (this is motivated by turbulence noise theory [6]). These source signals are then filtered with a vocal tract filter constructed using 6 formants (12 coefficients) extracted from a vowel */o/* uttered in isolation.

	R_a	R_k	R_0	F_0		R_a	R_k	R_0	F_0
1	4.1	37.1	51	170	21	10.0	44.8	84	200
2	7.6	35.7	81	170	22	2.0	37.7	54	126
3	1.3	47.9	59	132	23	5.0	51.7	71	246
4	3.5	42.9	81	144	24	2.6	42.5	61	102
5	13.1	27.6	65	281	25	5.1	41.9	76	190
6	11.7	29.0	85	340	26	1.5	45.0	56	131
7	0.3	30.0	52	110	27	4.2	48.0	71	250
8	0.6	50.0	69	110	28	9.9	32.1	87	288
9	2.0	51.0	82	110	29	3.0	48.7	65	360
10	1.1	25.0	41	110	30	2.7	40.7	69	129
11	1.8	37.0	54	110	31	10.5	57.1	81	249
12	3.5	43.0	65	110	32	1.9	45.0	57	127
13	2.1	30.6	64	106	33	3.7	51.2	68	258
14	2.5	34.0	71	127	34	1.6	37.7	49	132
15	1.5	33.3	68	154	35	1.9	52.2	64	257
16	0.8	28.6	63	84	36	4.6	51.0	65	131
17	0.5	25.0	25	45	37	8.1	48.3	79	254
18	13.3	35.1	77	344	38	1.3	39.5	41	128
19	4.3	43.6	89	213	39	3.2	49.9	71	261
20	6.8	41.7	68	137					

Table 1: *LF parameters and associated voice qualities used for the synthetic data set: 1–6 are obtained from our own research, 7–12 are taken from [14], 12–21 from [10] and 22–39 from [15] (R parameters in % and F_0 in Hz.)*

The second corpus consists of a small data set was recorded for these experiments, consisting of the 5 different Spanish vowels (/a/, /e/, /i/, /o/, /u/) uttered in isolation by a female professional speaker, with different voice phonations: normal (*modal*), with a lower F_0 (*low*), and with increased F_0 (*high*). Each utterance was roughly 2–3 seconds long and was recorded using a high-quality microphone in a professional studio. The signals were digitalized at a sampling rate of 96kHz, in raw PCM, 24 bits/sample, and then converted to 16kHz, 16 bits/sample. Care was taken not to introduce any distortion, specially in the low-frequency range, due to its extreme importance when retrieving the true glottal waveform.

We use the averaged perceptual error (APE) [5] to evaluate the quality of the LF estimations: $PE = |\hat{P} - P|/P$, where $\hat{P}$ is the estimated valueLF parameter and P is the reference parameter used to generate the utterance. The table below shows the mean of the individual APE (standard deviation in parenthesis), for each LF parameter at different SNR levels used in the generation of the test data:

	05dB	10dB	15dB	20dB	clean
R_a	0.79 (0.61)	0.55 (0.48)	0.39 (0.48)	0.34 (0.42)	0.35 (0.51)
R_k	0.19 (0.19)	0.13 (0.14)	0.09 (0.09)	0.08 (0.08)	0.07 (0.09)
R_0	0.05 (0.07)	0.04 (0.05)	0.03 (0.04)	0.03 (0.05)	0.03 (0.04)
F_0	0.04 (0.04)	0.03 (0.03)	0.02 (0.02)	0.02 (0.03)	0.01 (0.02)
E_e	0.92 (0.02)	0.92 (0.02)	0.91 (0.02)	0.92 (0.02)	0.94 (0.02)

As we can see, the estimations of R_k, R_0 and F_0 are quite good, and in general the accuracy improves when the level of additional noise at the input diminishes. We need to take a closer look into estimation of the amplitude of g_{lf}, E_e, since it is being consistently underestimated, although this has no effect in the resulting synthetic speech (the vocal tract compensates to produce the same output level. The higher APE in R_k (asymmetry coefficient,i.e., ratio of the opening phase to the closing phase of the glottis) with respect to R_0, can be explained by the fact that it is fixed in g_{rk} $(2/3)$. According to our regression data, when R_k falls outside of the $[35\ 50]$ range, g_{kl} cannot properly approximate g_{lf}, but this has only shown to be a problem in very few cases, since the algorithm has proved robust enough to recover the proper glottal waveform nonetheless. The results for R_a for lower SNR values are somewhat disappointing, noticeably improving as the SNR increases. This can be explained by the different ways in which the two parametric glottal models implement the smoothness of the return phase, since the KLGLOTT88 model is quite simple in comparison to the LF model. For this reason, we will estimate μ with a order 1 LPC analysis of the pre-emphasized speech frame [10].

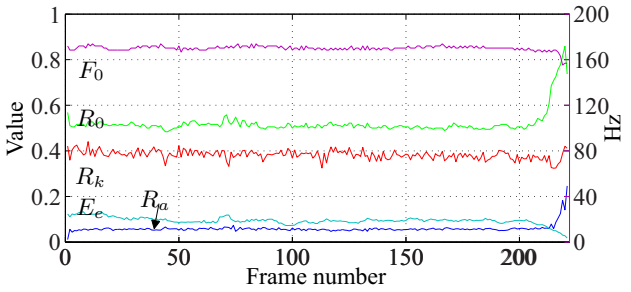


Figure 4: Estimated R parameters, E_e amplitude (left axis) and fundamental frequency F_0 in HZ (right axis) for the vowel /e/ in modal phonation.

It is well known that evaluating a glottal extraction algorithm in a real world scenario is difficult, due to the lack of reference [12]. One possible way is to compute the averaged SNR (dB) between g_{kl} and g_{if}, since this gives an idea of how well it approximates a idealized glottal waveform. We also use the Group Delay (GD) function of the glottal waveform, since it has been shown to perform well [13]. We chose to minimize the variance of the GD, as ideally it should be close to zero if all the formants have been removed in the inverse-filtering process. In order to select the optimal filter order for the vocal tract, the algorithm (without LF parametrization) was run using several filter orders (from $N = 8$ until $N = 24$, even orders only). We found that when the filter order increased, the SNR between the estimated and the parametrized glottal waveforms also increased (Fig. 5, bottom part), and seemed to stabilize

from 16 onwards. Visual inspection of the resulting waveforms showed that orders higher than 18 produced sub-optimal glottal waveforms, since the return phase (defined by $TL(z)$) was being over-estimated. This resulted in non-existing closed phases, which should not happen for modal voices. By observing the box-plots for the variance of the GD, we then found that the optimal filter order was 16. The length of the OLA window was set to 5, with independent amplitudes for each cycle, after observing that this resulted in a higher continuity of the estimated glottal parameters. Figure 4 shows an example of the smooth evolution of the estimated LF parameters for the vowel */e/* in modal phonation, as expected since it was being sustained. The table below presents the SNR results (between g_{if} a g_{lf}) for the real data set, using the optimal filter order and OLA lengths determined before.:

	modal	high	low
/a/	9.36 (0.81)	12.51 (2.97)	10.14 (0.74)
/e/	10.57 (0.71)	13.58 (0.95)	9.15 (0.76)
/i/	10.98 (1.06)	9.18 (1.62)	6.52 (0.79)
/o/	9.49 (0.58)	11.72 (2.02)	10.87 (1.12)
/u/	9.72 (0.87)	11.31 (1.89)	10.04 (1.15)

As we can see, the algorithm performs well in all the cases, even with higher values of F_0 (around 350Hz). This is an advantage of our method over traditional closed-phase inverse-filtering methods, since these often degrade due to the small amount of speech samples. An initial, informal evaluation of the resynthesis capabilities of the model was carried out as part of this work. We found that the quality of the synthetic speech was already quite good, considering that we are not yet using the parametrization error to model the aspiration noise. We are currently working on the extraction and parametrization of the aspiration noise in order to increase the naturalness of the synthetic speech.

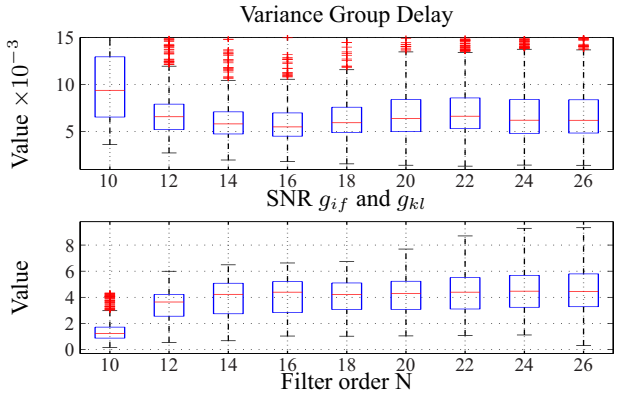


Figure 5: *Variance of the Group Delay (top) and SNR between g_{if} and g_{kl} (bottom) for various filter orders.*

5. Conclusions

We have presented here a new method for the simultaneous estimation of the derivative glottal waveform and the vocal tract filter. The algorithm operates on pitch-synchronous, frame-based basis, using overlapping frames of several glottal cycles to increase the robustness and quality of the estimation. The estimation is done by means of a convex optimization algorithm. The glottal opening (GOI) and closure (GCI) points are optimized as part of the main algorithm. The initialization of the final LF parametrization now uses the mapped KLGLOTT88 parameters as starting point. A test corpus has been designed using real data obtained in different studies, with amplitude-modulated, white Gaussian noise added at different SNR levels simulating aspiration noise. A second corpus has been specially recorded for this work, consisting of isolated vowels uttered with different voice qualities. The algorithm has been found to perform well with most of the voice qualities present in the synthetic data set, at the different SNR levels. The performance is also good with the real vowel data set in terms of resynthesis quality, although a more detailed analysis is needed here. We are currently studying the parametrization of the glottal estimation error, necessary to perform speech prosody modification using this source-filter paradigm as needed in voice conversion.

6. References

[1] G. Fant, *Acoustic theory of speech production*, 2nd ed. The Hague: Mouton, 1970.

[2] P. Alku, "Parameterisation methods of the glottal flow estimated by inverse filtering," in *VOQUAL*, Geneva, August 2003, pp. 81–87.

[3] M. Frölich, D. Michaelis, and H. W. Strube, "Sim – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *J. Acoust. Soc. Am.*, vol. 110, no. 1, pp. 479–488, July 2001.

[4] Q. Fu and P. Murphy, "Adaptive inverse filtering for high accuracy estimation of the glottal source," in *ITRW on Non-Linear Speech Processing (NOLISP 03)*, Le Croisic, France, May 2003.

[5] J. Pérez and A. Bonafonte, "Automatic voice-source parameterization of natural speech," in *Proc. of the EUROSPEECH*, Lisbon, Portugal, September 2005.

[6] H.-L. Lu, "Toward a high-quality singing synthesizer with vocal texture control," Ph.D. dissertation, Stanford University, 2002.

[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, no. 2, February 1990.

[8] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Båvegård, "Voice source parameters in continuous speech. transformation of lf-parameters," in *Proc. of the ICSLP*, Yokohama, 1994, pp. 1451–1454.

[9] P. N. A. Kounoudes and M. Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech," *ICASSP*, vol. 1, 2002.

[10] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[11] H. Strik, "Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2659–2669, 1998.

[12] T. Backstrom, M. Airas, L. Lehto, and P. Alku, "Objective quality measures for glottal inverse filtering of speech pressure signals," *ICASSP*, vol. 1, March 2005.

[13] P. Alku, M. Airas, T. Bäckström, and H. Pulakka, "Group delay function as a means to assess quality of glottal inverse filtering," in *Proc. of the EUROSPEECH*, Lisbon, Portugal, September 2005.

[14] R. van Dinther, R. Veldhuis, and A. Kohlrausch, "Perceptual aspects of glottal-pulse parameter variations," *Speech Communication*, vol. 46, no. 1, pp. 95–112, May 2005.

[15] I. Karlsson and J. Liljencrants, "Diverse voice qualities: models and data," *STL-QPSR*, vol. 2, pp. 143–146, 1996.