Glottal source shape parameter estimation using phase minimization variants

Stefan Huber, Axel Roebel[1], *Gilles Degottex*[2]

[1]Sound Analysis/Synthesis Team, IRCAM-CNRS-UPMC STMS, 75004 Paris, France
[2]Computer Science Department, University of Crete, 71409 Heraklion, Greece
stefan.huber@ircam.fr, axel.roebel@ircam.fr, degottex@csd.uoc.gr

Abstract

The glottal shape parameter R_d provides a one-dimensional parameterisation of the Liljencrants-Fant (LF) model which describes the deterministic component of the glottal source. In this paper we first propose to estimate the R_d parameter by means of extending a state-of-the-art method based on the phase minimization criterion. The utilization of a recently proposed adaption of the standard R_d parameter regression enables us to assess supplementary to the normal R_d range as well the upper R_d range. By evaluating the confusion matrices depicting the error surfaces of the involved different R_d parameter estimation methods and by objective measurement tests we verify the overall improvement of one new method compared to the state-of-the-art baseline approach.

Index Terms: glottal excitation source, shape parameter, voice quality, confusion matrices, R_d regression

1. Introduction

The voice quality of human speech production is related to the glottal source and its vibration of the vocal folds. The convolution of the glottal excitation waveform with the impulse responses of the vocal-tract filter (VTF) and the filters defining the radiation at the lips and nostrils level results in the complex human speech signal. Much effort has been conducted by the speech research community over the last decades to establish a reliable, robust and efficient method to extract the deterministic source from a recorded speech signal. Various algorithms have been proposed for this challenging task, as summarized in [1]. Due to the complexity of the problem, the robust estimation of the glottal excitation source still raises an open research question.

Similar to the minimum/maximum-phase decomposition paradigm, like Complex Cepstrum (CC) [2] or Zeros of the Z-Transform (ZZT) [3], we exploit the different phase properties assumed in our employed model for glottal source and vocal tract filter. We propose three continuative phase minimization methods extending the methods of [4, 5, 6] to estimate the glottal shape parameter R_d [7] describing the parameters of the glottal source model LF [8]. The first two proposed methods extent the phase minimization paradigm by applying different differentiation-integration schemata. The third proposed method achieves a more robust estimation of the glottal shape parameter R_d by means of superimposing the error residuals calculated by the different phase error methods employed. The objective of this paper is to identify the best performing method to distinct between fitting and mismatching LF model parameters. The experimental findings are as well promising to utilize the proposed methods for the assessment of the upper R_d range $R_d > 2.7$ for abducted phonation to describe breathy voice qualities at word or speaking pause boundaries. The usage of the normal and upper R_d range follows a recently proposed adapted parameter regression for the glottal shape parameter R_d [9]. There the improvement of two methods being proposed in this paper is validated by objective measurements on natural speech.

The article is organized as follows. In section 2 the model for the human speech production is introduced. It is utilized in section 3 in which the baseline and the different proposed extentions for the glottal pulse parameter estimation methods based on extended phase minimization are explained. A proof-of-concept investigation which examines confusion matrices of the parameter space for each method is carried out in section 4. An objective evaluation validating the improvement for one method is presented in section 5.

2. Voice production mode

The human voice production model $S(\omega)$ as in [6] consists of the characteristics of the acoustic excitation at the glottis level $G(\omega)$, the resonating filter of the vocal tract $C(\omega)$, the nasal and lip radiation $L(\omega)$ and the harmonic excitation $H(w, f0, D)$ parameterized by the fundamental frequency $f0$ and the delay between pulse sequence and frame center in terms of the phase delay D of the fundamental:

$$S(\omega) = G(\omega) \cdot C(\omega) \cdot L(\omega) \cdot H(w, f0, D) \qquad (1)$$

The discrete spectrum S_k as in [4, 5] with the bins k represents all estimated quasi-harmonic sinusoidal partials k from a Fourier transform of a windowed speech signal. The voice production model of the deterministic component of the speech signal is expressed by:

$$S_k = e^{jk\phi} \cdot G_k^{Rd} \cdot C_{k-} \cdot L_k \qquad (2)$$

The linear-phase term $e^{jk\phi}$ defines the time position of the glottal pulse in the period. G_k^{Rd} represents the LF glottal model, parameterised by the R_d parameter. The vocal-tract filter C_{k-} is assumed to be minimum-phase. The term L_k represents the radiation at the lips and nostrils level. According to [10] the filter L_k can be approximated by a time derivative and is thus set to $L_k = jk$.

The VTF can be expressed with respect to the shape parameter R_d of the glottal model by division in the frequency domain:

$$C_k^{Rd} = \mathcal{E}_-\left(\frac{S_k}{G_k^{Rd} \cdot jk}\right) \tag{3}$$

The operator $\mathcal{E}_-(.)$ is the minimum-phase realization of its argument that is calculated by means of using the real cepstrum [11].

3. Glottal shape parameter estimation

The VTF expression C_k^{Rd} of equation 3 is inserted into the voice production model of equation 2 to form the mathematical basis for the computation of the convolutive residual $R_k^{(\theta,\phi)}$, which is defined in equation 4. The shape of the glottal pulse is denoted by θ, while ϕ refers to the position of the glottal pulse with respect to the fundamental period in the time domain [12].

$$R_k^{(\theta,\phi)} = \frac{S_k}{e^{jk\phi} \cdot G_k^{\theta} \cdot jk \cdot \mathcal{E}_-(S_k/G_k^{\theta} \cdot jk)} \tag{4}$$

The division of S_k, G_k^{θ} and jk by their respective minimum-phase versions flattens their amplitude spectrum. The remaining convolutive residual $R_k^{(\theta,\phi)}$ is thus all-pass for any chosen glottal model. Its modulus is of unit amplitude: $|R_k^{(\theta,\phi)}| = 1\ \forall k, \theta, \phi$. Therefore, a mismatch of the model parameters to describe the observed speech signal affects only the phase spectrum of $R_k^{(\theta,\phi)}$. The result is that the better the estimate of the fitted voice model S_k, the closer is the convolutive residual $R_k^{(\theta,\phi)}$ to a Dirac delta function with a flat amplitude and zero phase spectrum. Hence, the smaller the phase spectrum of $R_k^{(\theta,\phi)}$ the closer is the R_d-value utilized to synthesize the glottal model G_k^{θ} to the true glottal shape contained in the observed signal [6]. This solution is unique as long as the glottal pulse that is present in the speech signal is covered by the R_d parameter space.

The main problem with the convolutive residual of 4 is its dependency on the pulse position phi. As shown in [5] we can remove this dependency by means of applying a 2^{nd} order difference operator

$$\Delta^2 \angle X_k = \angle \frac{X_{k+1} \cdot X_{k-1}}{X_k^2} \tag{5}$$

centered on each of the harmonics k of the convolutive residual $R_k^{(\theta,\phi)}$ in the complex plane. This removes the linear-phase component of the observed phase spectrum and removes therefore the dependency to the position parameter ϕ. Only the deviation from a linear phase trend will remain. To find the optimal R_d parameter the phase of the convolutive residual can be compared to the optimal target value 0.

Note, however, that the difference operator 5 not only removes the linear phase. It also applies a high pass filter to the phase difference that will be used to determine the optimal R_d parameter. To compensate this high pass filter a phase integration operator can be applied

$$\Delta^{-1} X_k = \angle \prod_{n=1}^{k} X_k \tag{6}$$

that invertes the high pass filter without reestablishing the linear phase trend. The main objetive of the following experimental investigation is to determine the number of integration steps to be performed that creates the objective function leading to the most reliable R_d estimates.

For this we compare setups with L integrations with L being in the set [0,1,2]. These objective functions will be denoted MSPDaIb with a being the number of differentiations and b representing the number of integrations. All the different objective functions constructed as described [MSPD2I0, MSPD2I1, MSPD2I2] present a different and not necessarily correlated error surface.

Objective function MSPD2IX: Accordingly, it might be beneficial to combine error surfaces of different objective functions by means of MSPD2IX(w0,w1,w2) = w0 * MSPD2I0 + w1 * MSPD2I1 + w2 * MSPD2I2. In the present paper we will show that the weighting w0=w1=w2=1/3 slightly improves the robustness of the method, but that more refined variations of the weighting sequence does not lead to major improvements. Therefore we will present only results obtained with equal weighting and denote this objective function as MSPD2IX.

Objective function MSPD2I0: The objective function to minimize the results of equation 5 is the proposed new method MSPD2I0:

$$\text{MSPD2I0}(\theta, N) = \frac{1}{N} \sum_{k=1}^{N} \left(\Delta^2 \angle R_k^{\theta}\right)^2 \tag{7}$$

Objective function MSPD2I1: An anti-difference operation (Δ^{-1})

$$\Delta^{-1} \Delta^2 \angle X_k = \angle \prod_{n=1}^{k} \frac{X_{n+1} \cdot X_{n-1}}{X_n^2} \tag{8}$$

applied to second order phase difference of equation 5 performs an integration according to 6 to retrieve again the first order frequency derivative representation, which emphasizes the phase distortion by the shape error.

The results of equation 8 are evaluated by the corresponding objective function named MSPD^2 in [5, 6]. In this study we refer to this state-of-the-art baseline method by MSPD2I1 to be consistent with our denomiation:

$$\text{MSPD2I1}(\theta, N) = \frac{1}{N} \sum_{k=1}^{N} \left(\Delta^{-1} \Delta^2 \angle R_k^{\theta}\right)^2 \tag{9}$$

MSPD2I1 optimizes the shape parameter independent of the window position in time to the glottal pulse [5].

Objective function MSPD2I2: Applying two anti-difference operators (Δ^{-2}) to the second order phase difference of equation 5 computes the twice differentiated and twice integrated phase term:

$$\Delta^{-2}\Delta^{2}\angle X_k = \angle \prod_{n=2}^{k}\prod_{n=2}^{k}\frac{X_{n+1}\cdot X_{n-1}}{X_n^2} \qquad (10)$$

The second differentation and second integration method emphasizes the phase slope of the convolutive residual while still being independent to the position of the glottal pulse with respect to the window position in time as a result of the preceeding differentiation operations.

The corresponding objective function to minimize the results of equation 10 is the proposed new method MSPD2I2:

$$\mathrm{MSPD2I2}(\theta, N) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta^{-2}\Delta^{2}\angle R_k^{\theta}\right)^2 \qquad (11)$$

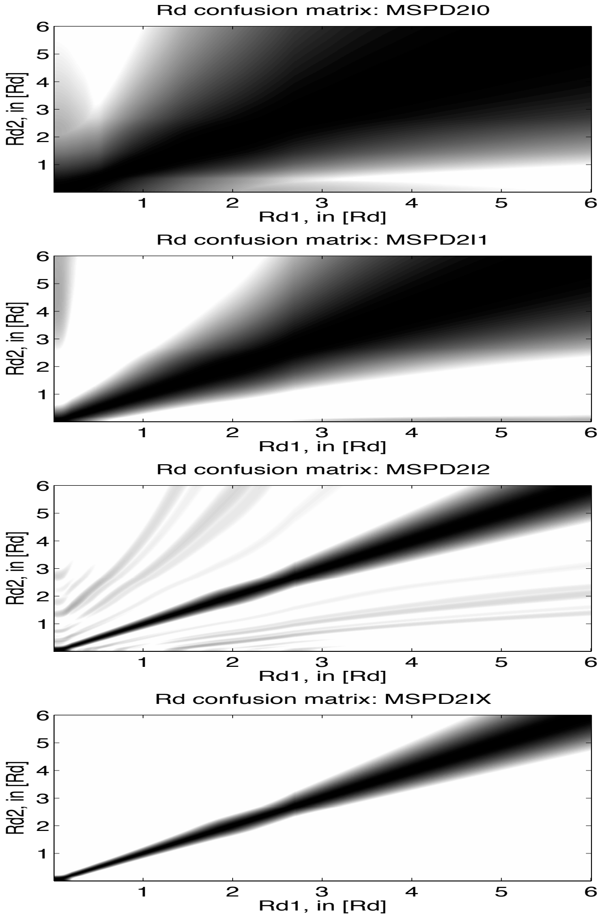
MSPD2I2 is the most selective and most distinctive among the different phase minimization methods and weights slight differences of the matched glottal model to the observed glottal source the most.

4. R_d confusion matrices

As a first step to understand the properties of the different objective functions we will show and discuss their R_d parameter confusion matrices [6]. These confusion matrices show the sensitivity of the objective functions with respect to R_d variation over the complete R_d range. According to [13] the robustness of the R_d estimate depends mainly on the fundamental frequency f0, the first formant F_1 and the glottal formant F_g. Due to space constraints only a small number of confusion matrices can be shown in the present paper. To provide a general idea of the behavior of the objective functions we selected the following experimental setup for the creation of the confusion matrices. To simulate the first formant F_1 the synthetic glottal pulses G^{Rd} are convolved by a 2-pole filter having a pole position at 800Hz and radius 0.98. The fundamental frequency is selected to be 80 Hz.

We build as in [6] a confusion matrix to detect ambiguities of the functions for phase minimization by calculating each R_d value on a grid against all other R_d values on the same grid. The resulting error surface constitutes a proof-of-concept of how well the phase minimization method under investigation is able to distinct between the shape of a fitting or mismatching glottal formant of the synthetic model, under the influence of the first formant being simulated by the 2-pole filter.

An ideal error surface would have a tiny black error valley at the matching diagonal axis with the rest of the error surface in clear white colour indicating a complete mismatch. Since it is not predictable how many stable sinusoidal partials are observable from the speech signal for each frame, we present due to space constraints only the case of 7 partials as a realistic expectation before the harmonic content is masked by noise. Note that for other numbers of partials the results are qualitatively the same.


Rd confusion matrix: MSPD2I0

Rd confusion matrix: MSPD2I1

Rd confusion matrix: MSPD2I2

Rd confusion matrix: MSPD2IX

Figure 4: R_d confusion matrices for N=7 partials

By visual inspection of fig. 4 one can observe that MSPD2I0 has a broad black error valley at the lower normal R_d range $R_d <= 2.7$ being delimited by high error values (white) for higher values of R_d, but exhibits an even broader error valley at the upper R_d range $R_d > 2.7$ which may lead to unnatural broad steps when estimating R_d at word or pause boundaries of a continuous speech signal. MSPD2I1 exhibits a smaller error valley than MSPD2I0 but may suffer from ambiguities from the additional error valleys for low R_d values $R_d < 0.5$ versus higher R_d values $R_d > 3$ at the upper left and lower right. A more distinguishing error valley than MSPD2I0 and MSPD2I1 is achieved by MSPD2I2 with only very minor secondary minima present because MSPD2I2 is not only dependent on a remaining non-linear phase distortion as MSPD2I0 nor on a remaining non-linear phase distortion plus a linear-phase term represented as a constant as MSPD2I1. The combinatorial error surface of MSPD2IX exhibits the least ambiguities, a quasi-ideal small error valley and not any significant similarity for two or more R_d values.

5. Evaluation

5.1. Synthetic f0 and noise test

We conduct a similar test setup as in [5] by synthesizing 16 synthetic vowels using Maeda's digital simulator [14] at 10 different f0 values within the range [80 293] Hz and adding 5 Gaussian noise levels between -50 to -25 dB as glottal source noise $n^{\sigma_g}[n]$ and as environmental noise $n^{\sigma_e}[n]$ to the voiced signal to simulate acoustic turbulences present in real speech signals. A possible error introduced by different positions of the window with respect to the period in time is simulated by synthesizing each parameter set on a grid of 4 different delays ϕ^* covering the range $[-0.5 \cdot T0 \; 0.5 \cdot T0]$.

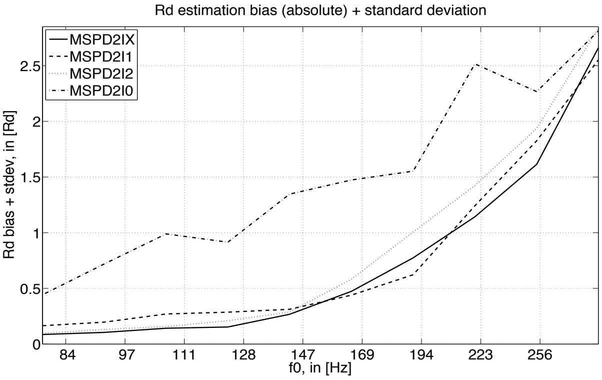


Figure 5.1: R_d estimation evaluation on f0 and noise

MSPD2IX with a solid line in fig. 5.1 exhibits the overall lowest error and is just slightly less performant for middle frequencies around 180 Hz compared to MSPD2I1. MSPD2I2 in dotted lines outperforms MSPD2I1 in dash-dotted lines only for lower frequencies up to 150 Hz. MSPD2I0 performs in general worse. Minimizing the combination of equations 8 and 10 does not perform better because the better performance of MSPD2IX is not achieved by adding up the different failures present but by suppressing the occuring side minima.

5.2. Spectral distortion effect

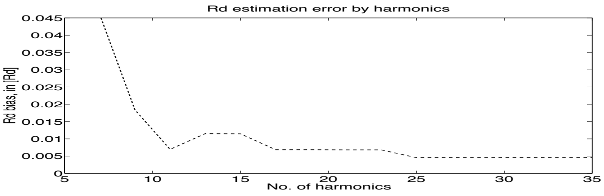


Figure 5.2: R_d estimation error by no. of harmonics

An explanation of the errors of the R_d estimation is given by the fact that the complete VTF cannot always be observed because some sinusoidal partials may be covered by noise. The evaluation shown in fig. 5.2 examines how many stable sinusoidal partials N_{harms} from the harmonic model are required to construct a minimum-phase spectrum to achieve a reliable R_d estimation with N partials. We choose N=7, vary the amount of N_{harms} and measure the mean error of the R_d estimation. For N_{harms}=11 the error function is already converged because a minimum-phase system has the property of a minimum group delay function with the main part of the energy being concentrated at time instant zero.

6. Conclusions

The results of section 4 provided a promising proof-of-concept which got partially validated by the objective evaluation measurements in section 5. This leads us to believe that the proposed objective function MSPD2IX improves the state-of-the-art R_d estimation method based on the phase minimization schemata. Note that MSPD2IX and MSPD2I2 showed in [9] significant improvements on natural speech compared to MSPD2I1.

7. References

[1] J. Walker and P. Murphy, "A review of glottal waveform analysis," *Progress in nonlinear speech processing*, pp. 1–21, 2007.

[2] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[3] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of z-transform (zzt) decomposition of speech for source-tract separation," in *Proc. ICSLP, International Conference on Spoken Language Processing, Jeju Island (Korea)*, 2004.

[4] G. Degottex, A. Roebel, and X. Rodet, "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA*, 2010, pp. 5058–5061.

[5] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[6] G. Degottex, *Glottal source and vocal tract separation*, Ph.D. thesis, IRCAM Paris, 2010.

[7] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Bvegrd, "Voice source parameters in continuous speech - transformation of lf parameters," in *Proceedings of the ICSLP-94*, 1994, pp. 1451–1454.

[8] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH*, vol. 26, no. 4, pp. 1–13, 1985.

[9] S. Huber and A. Roebel, "Glottal source shape parameter estimation for natural speech," in *Submitted to: Proc. Int. Conf. on Spoken Language Processing (Interspeech), Portland, Oregon, USA*, 2012.

[10] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, chapter 12, pp. 278–284, Communication and cybernetics. Springer Verlag, New York, 1976.

[11] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, PrenticeHall, 2nd edition, 1978.

[12] G. Degottex, A. Roebel, and X. Rodet, "Glottal closure instant detection from a glottal shape estimate," in *in Proc. 13th International Conference on Speech and Computer, SPECOM*, 2009, pp. 226–231.

[13] T. Drugman, T. Dubuisson, A. Moinet, N. D'Alessandro, and T. Dutoit, "Glottal source estimation robustness - a comparison of sensitivity of voice source estimation techniques," in *SIGMAP*, 2008, pp. 202–207.

[14] Shinji Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3-4, pp. 199–229, 1982.