Glottal source shape parameter estimation using phase minimization variants

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

[1]Sound Analysis/Synthesis Team, STMS IRCAM-CNRS-UPMC, 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. Then we propose an adaption of the standard R_d parameter regression which enables us to coherently assess the normal and 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, that is the vibration mode 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 human speech signal. Much effort has been conducted by the speech research community 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 source is still 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 properties of the phase spectra of the glottal source and vocal tract filter models. We propose three phase minimization methods extending the method proposed in [4, 5] to estimate the glottal shape parameter R_d [6] describing the parameters of the glottal source model LF [7]. 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 evaluation errors calculated by the different phase error methods. The objective of this paper is to identify the methods achieving the most reliable R_d estimation. For the usage of the normal and upper R_d range we propose to adapt the equations defining the predicted waveshape R_{*p} parameter set for the regression of the glottal shape parameter R_d. Additionally be propose to extent the R_d parameter range. The experimental findings show that the methods are as well beneficial to estimate R_d for abducted phonations to describe with the upper R_d range breathy voice qualities at word or speaking pause boundaries.

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 are explained. The adaptation and extention of the R_d parameter regression is explained in Section 4. The confusion matrices of the different phase minimization methods are evaluated in Section 5. Section 6 presents an objective evaluation validating the improvement for one method.

2. Voice production model

The human voice production model $S(\omega)$ as in [5] consists 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 $H(w, f_0, D)$ parameterized by the fundamental frequency f_0 and the delay between pulse sequence and frame center in terms of the phase delay D of the fundamental:

$$S(\omega) = G(\omega) \cdot C(\omega) \cdot L(\omega) \cdot H(w, f_0, D) \qquad (1)$$

Following Eq. 1, we contruct a discrete spectrum S_k of a single period as in [4] with each bin k representing a single quasi-harmonic sinusoidal partials k. These partials are estimated 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, parameterized 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 [8] 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, calculated by using the real cepstrum [9].

3. Glottal shape parameter estimation

The VTF expression C_k^{Rd} of Eq. 3 is inserted into the voice production model of Eq. 2 to form the mathematical basis for the definition of the convolutive residual $R_k^{(\theta,\phi)}$ in Eq. 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 [5].

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

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 [5]. 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 $R_k^{(\theta,\phi)}$ is its dependency on the pulse position ϕ. As shown in [4] 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 in the complex plane. This removes the linear-phase component of the observed phase spectrum and removes therefore the dependency to ϕ. Only the deviation from a linear phase trend remains. 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 of Eq. 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 can be applied

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

that invertes the high-pass filter without re-establishing 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. The different objective functions described as [MSPD2I0, MSPD2I1, MSPD2I2] present a different and not necessarily correlated error surface.

The phase slope is set to zero for each method as a result of the preceeding differentiation operations in order to be independent to the position of the glottal pulse with respect to the window position in time. Each integration step leads to a different weighting of the phase errors of the convolutive residual. The emphasis of the phase distortion by the shape error optimizes the shape parameter.

Objective function MSPD2I0: The objective function to minimize the results of Eq. 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 the second order phase difference of Eq. 5 performs an integration according to Eq. 6 to retrieve again the first order frequency derivative representation.

The results of Eq. 8 are evaluated by the corresponding objective function named MSPD^2 in [4, 5]. In this study we refer to this state-of-the-art baseline method by MSPD2I1 to be consistent with our naming convention:

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

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

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

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

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

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.

Objective function MSPD2IX: It might be beneficial to combine error surfaces of different objective functions by means of

$$\text{MSPD2IX(w0,w1,w2)} = \text{w0} \cdot \text{MSPD2I0} + \text{w1} \cdot \text{MSPD2I1} + \text{w2} \cdot \text{MSPD2I2} \quad (12)$$

In this paper we will demonstrate that the weighting w0=w1=w2=1/3 slightly improves the robustness of the method. Not presented are the results of an investigation showing that more refined variations of the weighting sequence do not lead to major improvements. Therefore we will present only results obtained with equal weighting and denote this objective function as MSPD2IX.

4. R_d regression adaptation and extention

The derivation of the predicted waveshape R_{*p} parameter set as in [6] describes the LF model from the glottal shape parameter R_d by means of a statistical regression. To derive the predicted waveshape R_{*p} parameters from an estimated R_d value we consider equations 2 to 4 of [6] for the normal R_d range R_d<2.7 and equations 8 to 11 for the upper R_d range R_d>2.7 of [10]. However, following the proposed equations defining the waveshape R_{*p} parameter set for the upper R_d range and joining them at the interconnection point R_d=2.7 with the waveshape R_{*p} parameter set for the normal R_d range results in a discontinuity and does not derive the expected contour of the waveshape parameters R_{kp} and R_{gp} as shown in Fig. 2 of [6]. Our proposed set of equations to define the adaptation of the waveshape parameter regression of R_d for an extended R_d range of [0.01 6] is:

$$R_{ap} = \begin{cases} 0 & \forall\, R_d{<}0.21 \\ (-1 + 4.8 \cdot R_d)/100 & \forall\, 0.21{\le}R_d{\le}2.70 \\ (32.3/R_d)\,/100 & \forall\, R_d{>}2.70 \end{cases} \quad (13)$$

$$OQ_{upp} = \; 1 - 1/(2.17 \cdot R_d) \quad \forall\, R_d{>}2.7 \quad (14)$$

$$R_{kp} = \begin{cases} (22.4 + 11.8 \cdot R_d)/100 & \forall\, R_d{\le}2.70 \\ (2 \cdot R_{gp} \cdot OQ_{upp}) - 1.04 & \forall\, R_d{>}2.70 \end{cases} \quad (15)$$

$$R_{gp} = \begin{cases} \frac{0.25 \cdot R_{kp}}{\frac{0.11 \cdot R_d}{0.5 + 1.2 \cdot R_{kp}} - R_{ap}} & \forall\, R_d{\le}1.85 \\ 0.00935 + \frac{596 \cdot 10^{-2}}{7.96 - 2 \cdot OQ_{upp}} & \forall\, R_d{>}1.85 \end{cases} \quad (16)$$

5. R_d confusion matrices

To understand the properties of the different objective functions we will show and discuss examples of their R_d parameter confusion matrices [5] which show the sensitivity of the objective functions with respect to the variation of R_d over its complete range. According to [11] the robustness of the R_d estimate depends mainly on the fundamental frequency f_0, the first formant F_1 and the glottal formant F_g. As experimental setup we simulate the first formant F_1 by convolving the synthetic glottal pulses G^{Rd} with a 2-pole filter having a pole position at 800 Hz and radius 0.98, with f_0 set to 80 Hz.

We build as in [5] 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 method under investigation is able to distinguish between the shape of a fitting or mismatching glottal formant of the synthetic model, under the influence of the first formant.

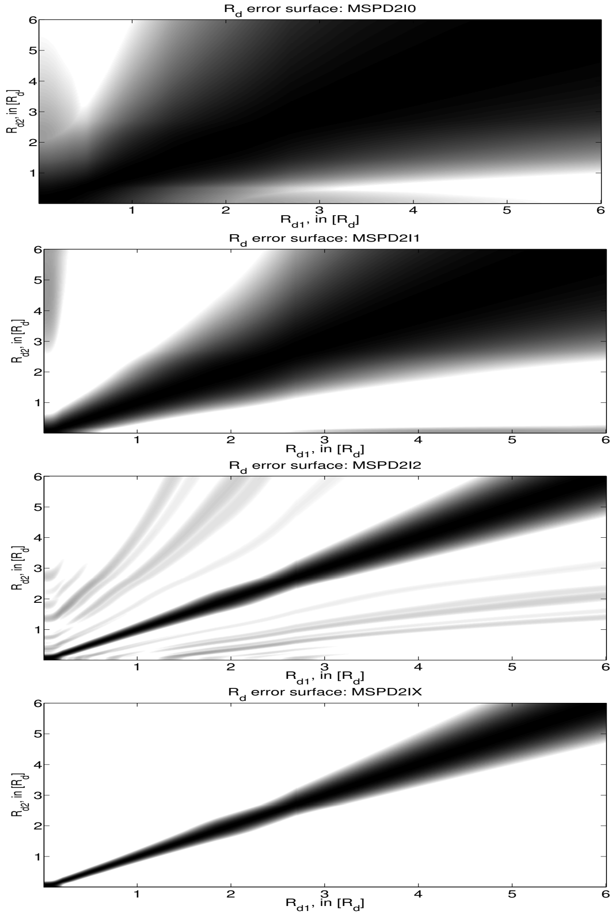

Figure 1: R_d confusion matrices for N=7 partials

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.

By visual inspection of Fig. 1 one can observe that each integration step leads to a more tiny error valley (black) being delimited by broader error hills (white). Broader error valleys appear more at the upper R_d range R_d>2.7 and may lead to unnatural broad steps when estimating R_d especially at word or pause boundaries of a continuous speech signal. MSPD2I1 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, while MSPD2I2 may be mislead by several appearing side minima. 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.

6. Evaluation

6.1. Synthetic f_0 and noise test

We conduct a similar test setup as in [4] by synthesizing 16 synthetic vowels using Maeda's digital simulator [12] at 10 different f_0 values within the range [80 293] Hz. Each vowel is convolved with a glottal formant parameterized by an R_d value within the range [0.1 6] and on a grid of step size 0.1. We add 5 Gaussian noise levels between -50 to -30 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 $[-0.5 \cdot T \;\; 0.5 \cdot T]$.

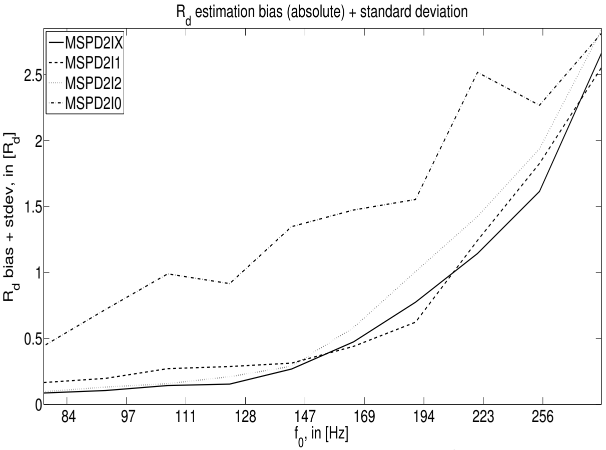

Figure 2: R_d estimation evaluation on f_0 and noise

MSPD2IX with a solid line in Fig. 2 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 only the combination of equations 9 and 11 does not perform better because the improvement by MSPD2IX is not achieved by adding up the different failures present but by suppressing the occuring side minima.

6.2. Spectral distortion effect

An explanation of the R_d estimation errors 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. 3 examines how many stable sinusoidal partials N_{harms} from the harmonic model are required to reliably construct the minimum-phase spectrum of the first N bins of S_k. 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 reasonably attenuated because the boundary effects that are introduced at the spectral border have sufficiently diminished.

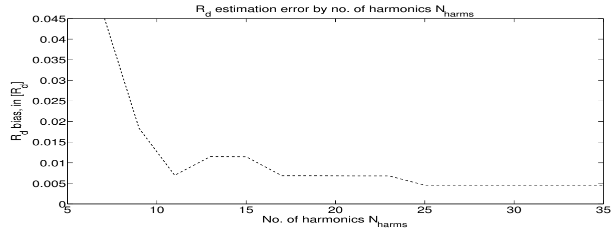

Figure 3: R_d estimation error by no. of harmonics

7. Conclusions

The results of Section 5 demonstrate a promising proof-of-concept which have partially been validated by the objective evaluation in Section 6. 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. In general it can be stated that the higher f_0 the more difficult is to evaluate the minimum-phase property of the vocal tract filter within a single fundamental period and accordingly the systematic errors of the R_d estimator will increase with f_0.

8. 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, "Phase minimization for glottal model estimation," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

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

[6] G. Fant, "The lf-model revisited. transformation and frequency domain analysis," *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH*, vol. 36, no. 2-3, pp. 119–156, 1995.

[7] 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.

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

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

[10] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Bvegrd, "Voice source parameters in continuous speech. transformation of lf-parameters," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP-94), Yokohama, Japan*, 1994, pp. 1451–1454.

[11] 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.

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