Glottal source shape parameter estimation using phase minimization variants

Stefan Huber, Axel Roebel1, Gilles Degottex2

1Sound Analysis/Synthesis Team, STMS IRCAM-CNRS-UPMC, 75004 Paris, France
2Computer 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 Liljencrantz-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 extend 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 $R_{dp}$ 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 phona-
tions 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 extensions for the glottal pulse parameter estimation methods are explained. The adaptation and extension 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:

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

(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:

$$S_k = e^{jk\phi} \cdot G_{kd}^R \cdot C_{kd} - L_k$$

(2)

The linear-phase term $e^{jk\phi}$ defines the time position of the glottal pulse in the period. $G_{kd}^R$ represents the LF glottal model, parameterized by the $R_d$ parameter. The vocal-tract filter $C_{kd}$ is assumed to be minimum-phase. The term $L_k$ represents the radiation at the lips and nos-
trils 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:

$$C_{\text{Rd}} = \mathcal{E}_- \left( \frac{S_k}{G_k^d \cdot jk} \right)$$

(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 $C_{\text{Rd}}$ 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 $\theta$, while $\phi$ refers to the position of the glottal pulse with respect to the fundamental period in the time domain [5].

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

(4)

The division of $S_k$, $G_k^d$, and $jk$ by their respective minimum-phase versions flattens the 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: $|R_k^{(\theta, \phi)}| = 1$ for all $k$, $\theta$, $\phi$. Therefore, a mismatch of the model parameters to describe the observed speech signal affects only the phase spectrum of $R_k^{(\theta, \phi)}$. This result suggests that the better 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 $G_k^d$ 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 phase angle $\phi$. As shown in [4] we can remove this dependency by means of applying a $2^{\text{nd}}$ order difference operator

$$\Delta^2 \angle X_k = \angle X_{k+1} - X_{k-1}$$

(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 $\phi$. 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

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

(6)

that inverts the high-pass filter without re-establishing the linear phase trend. The main objective 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 MSPD$^L_dB$ with $b$ 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 preceding 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:

$$\text{MSPD2I0}(\theta, N) = \frac{1}{N} \sum_{k=1}^{N} (\Delta^2 \angle R_k^0)^2$$

(7)

#### Objective function MSPD2I1:

An anti-difference operation ($\Delta^{-1}$)

$$\Delta^{-1} \angle X_k = \angle \prod_{n=1}^{k} X_{n+1} - X_{n-1}$$

(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:

$$\text{MSPD2I1}(\theta, N) = \frac{1}{N} \sum_{k=1}^{N} (\Delta^{-1} \Delta^2 \angle R_k^0)^2$$

(9)

#### Objective function MSPD2I2:

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

$$\Delta^{-2} \Delta^2 \angle X_k = \angle \prod_{n=2}^{N} X_{n+1} - X_{n-1}$$

(10)

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

$$\text{MSPD2I2}(\theta, N) = \frac{1}{N} \sum_{k=1}^{N} (\Delta^{-2} \Delta^2 \angle R_k^0)^2$$

(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
MSPD2IX(w0, w1, w2) = w0 · MSPD2I0 +
w1 · MSPD2I1 + w2 · MSPD2I2 (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 se-
quence 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 \(R_{sp}\) 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 \(R_{sp}\) 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 \(R_{sp}\)
parameter set for the upper \(R_d\) range and joining them
at the interconnection point \(R_d=2.7\) with the waveshape
\(R_{sp}\) 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 \(R_{up}\) 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:

\[R_{up} = \begin{cases} 0 & \forall R_d<0.21 \\ (-1 + 4.8 \cdot R_d)/100 & \forall 0.21 \leq R_d \leq 2.70 \end{cases} \]
\[(32.3/R_d)/100 & \forall R_d>2.70 \]

\[OQ_{upp} = 1 - 1/(2.17 \cdot R_d) & \forall R_d>2.7 \]

\[R_{kp} = \begin{cases} (22.4 + 11.8 \cdot R_d)/100 & \forall R_d \leq 2.70 \\ (2 \cdot R_{gp} \cdot OQ_{upp}) - 1.04 & \forall R_d>2.70 \end{cases} \]

\[R_{gp} = \begin{cases} 0.25 \cdot R_{kp} & \forall R_d < 1.85 \\ 0.00035 + 596 \cdot 10^{-2} \cdot OQ_{upp} & \forall R_d > 2.70 \] \]

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 sensi-
tivity of the objective functions with respect to the vari-
ation 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 ambigu-
ities of the functions for phase minimization by calcu-
lating 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 investi-
gation 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](image)

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). Broad
error valleys appear more at the upper \(R_d\) range \(R_d>2.7\)
and may lead to unnatural broad steps when esti-
mating \(R_d\) especially at word or pause boundaries of a
continuous speech signal. MSPD211 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 MSPD212 may be mislead
by several appearing side minima. The combinatorial
error surface of MSPD21X exhibits the least ambiguities, a
 quasi-ideal small error valley and not any significant sim-
ilarity 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 $n^{\text{gs}}[n]$ and as environmental noise $n^{\text{env}}[n]$ 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 $\phi^p$ covering the range $[-0.5 \cdot T, 0.5 \cdot T]$.

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 MSPD211. MSPD212 in dotted lines outperforms MSPD210 in dash-dotted lines only for lower frequencies up to 150 Hz. MSPD210 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 occluding 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 $N_{\text{harm}}$ 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 $N_{\text{harm}}$ and measure the mean error of the $R_d$ estimation. For $N_{\text{harm}}=11$ the error function is already reasonably attenuated because the boundary effects that are introduced at the spectral border have sufficiently diminished.

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