Phase Minimization for Glottal Model Estimation

Gilles Degottex, *Member, IEEE*, Axel Roebel, *Member, IEEE*, and Xavier Rodet

***Abstract*—In glottal source analysis, the phase minimization criterion has already been proposed to detect excitation instants. As shown in this paper, this criterion can also be used to estimate the shape parameter of a glottal model (ex. Liljencrants–Fant model) and not only its time position. Additionally, we show that the shape parameter can be estimated independently of the glottal model position. The reliability of the proposed methods is evaluated with synthetic signals and compared to that of the IAIF and minimum/maximum-phase decomposition methods. The results of the methods are evaluated according to the influence of the fundamental frequency and noise. The estimation of a glottal model is useful for the separation of the glottal source and the vocal-tract filter and therefore can be applied in voice transformation, synthesis, and also in clinical context or for the study of the voice production.**

***Index Terms*—Glottal model, glottal shape, glottal closure instants (GCIs), joint estimation, phase minimization, voice analysis.**

I. Introduction

In VOICE analysis, using the source–filter model of the voice production, the filter is often assumed to be excited by a flat amplitude spectrum. However, many models of glottal excitation have been proposed [1]–[3]. Obviously, these models have particular shapes in time and frequency domains. Among their spectral characteristics, the glottal formant and the spectral tilt are often cited [1], [2], [4]. Therefore, it is interesting to estimate the glottal model parameters and thus the spectral characteristics of the source. For example, such estimates enable separating the glottal source from vocal-tract influences [5], [6]. This separation is very attractive in voice transformation since it allows a means to manipulate independently the source excitation and the resonating properties of the vocal-tract. In this paper, we use the source–filter model which is made of three principal elements: the glottal source, the vocal-tract filter (VTF) and the radiation. The glottal source is assumed to be produced by the air flow modulated by the periodic opening and closing of the glottis. This source has a shape in time and spectral domains and this shape has a time position in a given period of voiced signal. Then, the vocal-tract filter transforms this source by means of resonances and anti-resonances. Finally, this transformed source is radiated into the environment through the lips and the nostrils adding one more filter effect. In this presentation, analyzing a window of voiced signal, we want to estimate the shape parameter of a glottal model (a shape model of one period of the glottal source) and its time position in the analysis window.

Our approach is, in the following, we focus on the phase properties of the source and the VTF. Indeed, the glottal source is a mixed-phase signal [7], [1]. Zeros exist outside of the unit circle in the glottal source z-transform. Since the voice production model is made of time convolutions, these zeros remain in the final voiced signal. On the other hand, the poles of the VTF are inside the unit circle because it is a stable filter [8]. Concerning the zeros created by the coupling of the nasal cavity with the oral cavity, they lie on the unit circle when the vocal-tract is assumed to be lossless [9]. In our investigation, we postulate that the losses move these zeros inside the unit circle because the poles obey this behavior between the lossless and the lossy cases. Consequently, we can assume that the VTF impulse response is a minimum-phase signal. The minimum-phase assumption is more general than the usual all-pole hypothesis [8] (often modeled by linear prediction (LP) [8] or discrete all-pole (DAP) [10]). The minimum-phase assumption does not exclude zeros which can occur in nasalized sounds but implies they are strictly inside the unit circle. In terms of source-filter separation, these phase properties have been already used in minimum/maximum-phase decomposition methods, i.e., Complex Cepstrum (CC) [11] and Zeros of the Z-Transform (ZZT) [12]. With this approach, the well-known closed-phase hypothesis is not necessary [13] and it is thus possible to broaden the diversity of voices to analyze. In this paper, we assume that the mixed-phase and minimum-phase properties enable us to estimate the parameters of a glottal model. To focus on these phase properties, the phase minimization criterion is used. This criterion is the following: the phase spectrum of a model is fitted to the phase spectrum of an observed signal. The error of fitting is computed through the *convolutive residual* (the deconvolution of the observed signal by the model). Therefore, the better the estimate of the model, the closer the convolutive residual is to a Dirac delta function. In terms of phase, the better the estimate of the model, the smaller the phase spectrum of the convolutive residual. This criterion has already been proposed to estimate glottal closure instants (GCIs) resulting in robust estimators [14], [15]. In these methods, assuming the source is a Dirac delta and the VTF is an all-pole filter, the phase spectrum of an LP residual is minimized. In our study, we propose to use the phase minimization criterion to estimate, not only the position of the excitation model, but also the shape parameter of a glottal model. We already proposed a first method [16] which jointly estimate the shape and the position. First, this paper refines and improves the argumentation of that previous publication. Additionally,

Manuscript received March 30, 2010; revised July 12, 2010; accepted August 19, 2010. Date of publication September 16, 2010; date of current version May 04, 2011. This work was supported in part by a grant from the Centre National de la Recherche Scientifique (CNRS), in part by the *Affective Avatar* ANR project, and in part by the *Respoken* FEDER project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Steve Renals.

The authors are with the Analysis/Synthesis Team, IRCAM, 75004 Paris, France (e-mail: gilles.degottex@ircam.fr).

Digital Object Identifier 10.1109/TASL.2010.2076806

similarly to the GCI detection method using the group-delay [15], we propose two other methods which take advantage of the difference operator with respect to the harmonics phase: one method balances the influence of the shape and the position on the error function and a last method eliminates completely the influence of the glottal model position on the shape estimate. Compared to current methods of source-filter decomposition, the proposed methods try to directly estimate the glottal parameters of a glottal model without estimating the glottal source. Indeed, the Iterative Adaptive Inverse Filtering method (IAIF) [17] first estimates the VTF and a spectral envelope of the glottal source. In the same way, minimum/maximum-phase decomposition methods (CC and ZZT-based methods) first estimate the maximum-phase contribution of the speech signal in order to retrieve the glottal source. In all of those approaches, the estimated glottal source is fitted by a glottal model in a second step. In the proposed methods, the glottal source is not explicitly computed and the VTF is jointly estimated with the glottal model parameters. One can thus expect more consistency between the VTF estimate and the glottal model estimate.

Even though the estimation of glottal parameters is a very active research field [7], [6], [18], [5], [17], the lack of ground truth makes the results of such estimators difficult to validate. A measurement of the glottal flow which is usually associated with the source of the source-filter model could be compared to glottal model estimates. However, the acoustic coupling (between the glottal flow and the vocal-tract) and the issues related to the measurement of this flow make this comparison difficult to establish. Nevertheless, in the context of voice transformation and synthesis, only the perception of the voice has to be manipulated. Therefore, recovering the glottal flow precisely may be not necessary for these applications. In current literature, the validation of analysis methods is usually avoided by proposing transformation and synthesis systems to support the analysis/synthesis processes [5], [19], [20], [17], [21]. However, because such a process using a glottal model is far from straightforward, forthcoming publications should address this problem and thus evaluate the significance of glottal model estimates in real applications. In this presentation, in order to evaluate the proposed methods compared to the state of the art of the source-filter separation methods (IAIF, CC and ZZT), we use synthetic signals and Electro-Glotto-Graphic (EGG) signals.

The following discussion consists of three main parts. To make the innovative theoretical ideas as clear as possible about Mean Squared Phase (MSP) and the phase difference operator, Part II discusses the estimation process and the mathematical derivations without taking into account the details related to the realization which are described in Part III. That more practical Part III presents the algorithm using the mean squared phase to jointly estimate the position and a shape parameter of the Liljencrants-Fant glottal model. Then, the computation of the phase difference is detailed. The last Part IV evaluates precision and robustness of the proposed methods with synthetic and EGG signals. The estimation of the glottal source by inverse filtering is discussed and two examples of real signals conclude the evaluation part.

II. Voice Production Model and Phase Minimization

The shape and position parameters of a glottal model are estimated in an optimization context by means of error minimization: given hypothetical parameters (θ, ϕ), the VTF is first computed according to the voice production model. Then, to represent the differences between the observed signal and the voice model, the convolutive residual is used. Finally, the error related to (θ, ϕ) is computed using the Mean Squared Phase (MSP) of the convolutive residual. Below, each step of the estimation process is described and the conditions of convergence are discussed. The presentation of the methods using the difference operator concludes this theoretical part.

A. Voice Production Model

Within a given window, the voiced signal is assumed to be stationary and periodic with a fundamental frequency f_0. Therefore, one can build a discrete spectrum S_k where the k-bins represent all the available k-harmonics in this window. Using this single period representation, we express the voice production model as follows:

$$S_k = e^{jk\phi} \cdot G_k \cdot C_{k-} \cdot L_k \tag{1}$$

where $e^{jk\phi}$ represents the time position ϕ of the glottal shape in the period. G_k is a mixed-phase spectrum representing the shape of the glottal source. In the following, G_k^θ will represent a glottal model where its shape is parametrized by θ. C_{k-} is a minimum-phase filter corresponding to the VTF (the minimum-phase property is denoted by the negative sign). Finally, L_k is the filter corresponding to the lips radiation. This filter can be modeled with a time derivative and therefore $L_k = jk$ [8], [22].

B. Estimation Process

First, we define $\mathcal{E}_-(\,\cdot\,)$ as a function computing the minimum-phase version of a given spectrum:

$$X_{k-} = \mathcal{E}_-(X_k) = \exp(\mathcal{F}(\hat{x}_{n-}))$$

where $\mathcal{F}(\,\cdot\,)$ is the Discrete Fourier Transform and the minimum-phase cepstrum $\hat{x}_{n-}$ is computed from the power cepstrum $\hat{x}_n$ corresponding to the spectrum X_k [23]:

$$\hat{x}_{n-} = \begin{cases} 0, & n < 0 \\ 2\hat{x}_n, & n > 0 \\ \hat{x}_n, & n = 0 \end{cases}$$

and

$$\hat{x}_n = \mathcal{F}^{-1}(\log |X_k|)$$

where $\mathcal{F}^{-1}(\,\cdot\,)$ is the inverse Discrete Fourier Transform. Note that $\mathcal{E}_-(\,\cdot\,)$ has no linear-phase component since X_{k-} is minimum-phase. Additionally, this function is multiplicative (i.e., $\mathcal{E}_-(A \cdot B) = \mathcal{E}_-(A) \cdot \mathcal{E}_-(B)\ \forall |A|, |B| \geq 0$). Then, using $\mathcal{E}_-(\,\cdot\,)$ and the voice production model (1), by inverse filtering in the frequency domain, one can derive an expression of the VTF which depends on the shape parameter θ of a given glottal model:

$$C_{k-}^\theta = \mathcal{E}_-\left(\frac{S_k}{G_k^\theta \cdot jk}\right) \tag{2}$$

Note that this VTF model does not represent the real VTF because this representation is reduced to harmonic frequencies. This VTF expression (2) can be replaced in the voice production model (1) to derive the convolutive residual $R_k^{(\theta,\phi)}$, the ratio of the observed spectrum by the model spectrum:

$$R_k^{(\theta,\phi)} = \frac{S_k}{e^{jk\phi} \cdot G_k^\theta \cdot \mathcal{E}_-(S_k/G_k^\theta \cdot jk) \cdot jk} \tag{3}$$

In the first proposed method of this paper, the Mean Squared Phase (MSP) of this convolutive residual is minimized to obtain the optimal parameters which best fit the observed spectrum:

$$\text{MSP}(\theta,\phi,N) = \frac{1}{N}\sum_{k=1}^{N}\left(\angle R_k^{(\theta,\phi)}\right)^2 \tag{4}$$

The multiplicative property of $\mathcal{E}_-(\,\cdot\,)$ allows us to write (3) as:

$$R_k^{(\theta,\phi)} = e^{-jk\phi} \cdot \frac{S_k}{\mathcal{E}_-(S_k)} \cdot \frac{\mathcal{E}_-(G_k^\theta)}{G_k^\theta} \cdot \frac{\mathcal{E}_-(jk)}{jk} \tag{5}$$

One can see that the calculation of the convolutive residual flattens the amplitude spectrum of S_k, G_k^θ and jk by their respective minimum-phase versions. $R_k^{(\theta,\phi)}$ is thus all-pass for any chosen glottal model and its parameters: $|R_k^{(\theta,\phi)}| = 1\ \forall k\ \forall\theta\ \forall\phi$. Consequently, an error of the parameters affects only the phase spectrum of $R_k^{(\theta,\phi)}$. Additionally, $R_k^{(\theta,\phi)}$ tends to a Dirac delta when its phase spectrum is minimized because the Dirac delta has a flat amplitude spectrum and a zero phase spectrum. Therefore, the smaller the phase spectrum of the convolutive residual, the closer the model is to the observed spectrum. Using a Liljencrants-Fant (LF) glottal model ([5], p. 19), [24], [3] parametrized by the single Rd shape parameter [2], Fig. 1(a) shows an example of MSP$(Rd,\phi,12)$ computed on a synthetic speech signal (see IV for more details on the synthesis).

C. Conditions of Convergence

In this section, we assume that the shape of the real glottal source G_k can be correctly represented by our chosen glottal model $G_k^{\theta^*}$ with an optimal parameter θ^*. In this context, it is important to known which properties of the glottal model are necessary to ensure the convergence of (θ,ϕ) to the optimal parameters (θ^*,ϕ^*). In the computation of the convolutive residual (3), the observed spectrum S_k can be replaced by the voice production model (1) with optimal parameters:

$$R_k^{(\theta,\phi)} = e^{jk(\phi^*-\phi)} \cdot \frac{G_k^{\theta^*} \cdot C_{k-}^*}{G_k^\theta \cdot \mathcal{E}_-\left(G_k^{\theta^*} \cdot C_{k-}^*/G_k^\theta\right)} \tag{6}$$

Then, by distributing $\mathcal{E}_-(\,\cdot\,)$ to the terms of its argument, the VTF terms cancel from the previous equation because $\mathcal{E}_-(C_{k-}^*) = C_{k-}^*$. Therefore, (6) can be rewritten as

$$R_k^{(\theta,\phi)} = \underbrace{e^{jk(\phi^*-\phi)}}_{\text{position error}} \cdot \underbrace{\frac{G_k^{\theta^*} \cdot \mathcal{E}_-\left(G_k^\theta\right)}{G_k^\theta \cdot \mathcal{E}_-\left(G_k^{\theta^*}\right)}}_{\text{shape error}} \tag{7}$$

First, according to (7), note that the error function of (4) is periodic with respect to ϕ since the *position error* term is periodic. Therefore, looking for an optimal position in the interval $[-\pi;\pi[$ is sufficient.

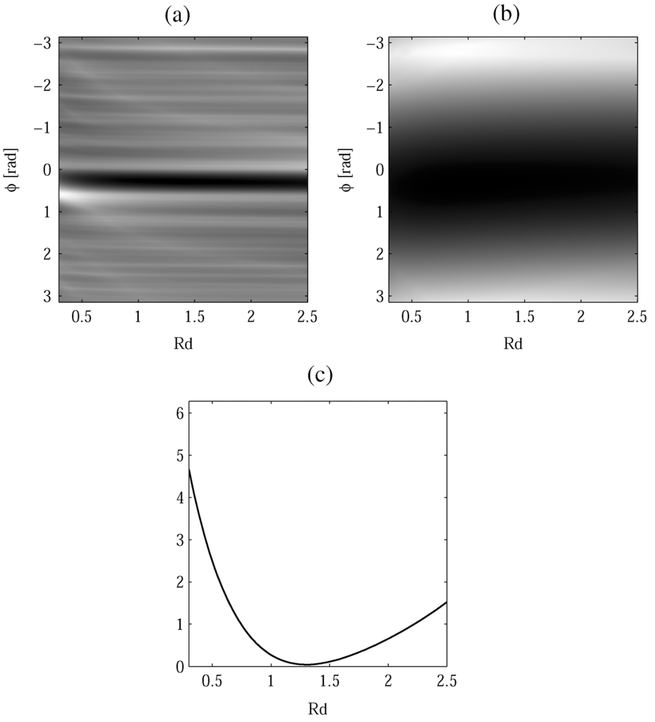


Fig. 1. Examples of Mean Squared Phase (MSP) computed on a synthetic signal. In the upper plots, the darker the color, the smaller the error. (a) $\text{MSP}(Rd,\phi,12)$. (b) MSPD$(Rd,\phi,12)$. (c) MSPD$^2(Rd,12)$.

Second, we need to express the condition which, if satisfied, ensures that the shape parameter influences the *shape error*: The zeros inside the unit circle in $G_k^{\theta^*}$ and G_k^θ are always canceled by their corresponding $\mathcal{E}_-(\,\cdot\,)$ expressions. However, a zero outside of the unit circle in $G_k^{\theta^*}$ can be canceled only by G_k^θ. Consequently: *θ influences $R_k^{(\theta,\phi)}$ if θ influences at least one zero outside of the unit circle in G_k^θ.*

Finally, we need to express the condition which has to be satisfied to ensure that the shape and the position do not offset each other, at least theoretically: it is sufficient to ensure that the *shape error* has no linear-phase component. G_k^θ has a linear-phase which depends on the zero-time reference given by the definition of the glottal model. Therefore, if θ influences that linear-phase component, a residual linear-phase exists in $G_k^{\theta^*}/G_k^\theta$ which biases the position error. To have no offset effect, the condition is: *θ does not influence the linear-phase component of the glottal model.* Note that, using the Liljencrants-Fant model, this condition is satisfied if the zero-time reference is set to the t_e instant ([5], p. 19), [24], [3].

The next two paragraphs describe the main biases which affect the presented methods as well as voice analysis in general:

1) Vocal-Tract Filter Reconstruction: One of the main issues is the sampling of the VTF frequency response by the harmonic structure of the excitation source. To ensure (2) gives a good approximation of the VTF, the cepstral coefficients of the VTF above the quefrency $1/f_0$ have to be negligible. Therefore, one can expect a rough approximation with high fundamental frequency (see Section IV.A.1). Additionally, the lips radiation creates a zero at zero-frequency in the z-transform of the voiced signal. In (2), the DC of the argument of $\mathcal{E}_-(\,\cdot\,)$ is undefined. Therefore, this DC value has to be extrapolated from the lowest harmonics. Here, we simply used $C_{0-}^{\theta} = |C_{1-}^{\theta}|$.

2) Minimum-Phase Reconstruction: In order for (2) to give a good minimum-phase estimate, the Nyquist frequency has to be as high as possible. Additionally, in a real signal, the noise level exceeds the harmonic level in high frequencies. The number of available harmonics in an observed spectrum can thus be drastically reduced. In this paper, we assume that the lack of harmonics in high frequencies, due to the Nyquist frequency or the noise level, does not influence significantly the lowest harmonics of the convolutive residual.

D. Difference Operator for Phase Distortion Measure

To detect Glottal Closure Instants (GCI), it has been shown that the group-delay can be used instead of the phase [15]. In this section, we propose to apply this idea to the estimation of the shape parameter. Since we used only harmonics in (3), we use the difference operator with respect to the phase of these harmonics to approximate the frequency derivative of the phase

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

The corresponding objective function to minimize is the mean squared phase difference (MSPD)

$$\mathrm{MSPD}(\theta,\phi,N) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta\angle R_k^{(\theta,\phi)}\right)^2. \tag{8}$$

Applying the difference operator to (7) leads to

$$\Delta\angle R_k^{(\theta,\phi)} = (\phi^* - \phi) + \Delta\angle\left(\frac{G_k^{\theta^*}\cdot\mathcal{E}_-\left(G_k^{\theta}\right)}{G_k^{\theta}\cdot\mathcal{E}_-\left(G_k^{\theta^*}\right)}\right). \tag{9}$$

Compared to (7), one can see that the linear-phase error is no longer weighted by the harmonic number k. Moreover, this conditioning is also promising for estimating the shape parameter because it represents linearly the time shifting of a given frequency. Using the LF model, Fig. 1(b) shows an example of $\mathrm{MSPD}(Rd,\phi,12)$. Although the influence of Rd and ϕ seems better balanced compared to 1(a), the two parameters are highly dependent on each other. Indeed, the position error in (9) can fit the average value of the phase distortion of the shape error. Without the difference operator, the harmonic number k weights the MSP error function and constrains ϕ on its ideal value. The example of Fig. 1(a) shows that the optimal ϕ value is not affected by Rd (a straight horizontal trench is visible at $\phi \approx 0.3$).

Finally, using the second-order phase difference (Δ^2), the position parameter ϕ can be removed from the convolutive residual

$$\Delta^2\angle R_k^{\theta} = \Delta^2\angle\left(\frac{G_k^{\theta^*}\cdot\mathcal{E}_-\left(G_k^{\theta}\right)}{G_k^{\theta}\cdot\mathcal{E}_-\left(G_k^{\theta^*}\right)}\right). \tag{10}$$

However, the first-order frequency derivative representation which emphases the phase distortion by the shape error has to be retrieved. The anti-difference operator (Δ^{-1}) is thus used and the corresponding objective function to minimize is

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

where R_k^{θ} is computed using (3) ignoring the linear-phase term. Note that, by minimization of MSPD^2, the shape parameter can be optimized whatever the position of the glottal model. Fig. 1(c) shows an example of $\mathrm{MSPD}^2(Rd,12)$.

III. Methods

First, the spectrum of a voiced segment is computed with a *blackman* window and the discrete Fourier transform (DFT). A window of only one period would estimate the complex coefficients S_k directly. However, such a duration is not suitable since the convolutive effect of the window in the spectral domain has to be negligible compared to the harmonic amplitudes and phases of the underlying signal we need to represent. Therefore, four periods are used and a harmonic model is built from the DFT of these periods [25]. The amplitude and phase of the kth-harmonic are obtained using the amplitudes and phases of the neighbor bins of $k\cdot f_0$ in the DFT. A parabola is fitted to the amplitudes of the bins to estimate the harmonic amplitude and the harmonic phase is obtained by linear interpolation.

To synthesize $G_k^{\theta}\cdot jk$, the Liljencrants–Fant (LF) glottal model is used [5, p. 19], [24], [2], [3]. The time and amplitude scaling parameters are the fundamental frequency f_0 and the excitation amplitude E_e, respectively. We assume f_0 to be known *a priori*. Numerous methods can be used to compute it from the voiced signal directly (ex. YIN [26], Swipep [27], harmonic matching [28]). Moreover, regardless of the E_e value, the amplitude spectrum of the convolutive residual is always equal to one. Therefore, the proposed methods estimate the shape parameter independently of this value. The shape of the LF model is controlled by three parameters (O_q,α_m,t_a). It can be interesting to estimate these three shape parameters. However, in terms of error minimization of an AutoRegressive model with eXogenous input (ARX), it has already been shown that the effect of O_q can be partially offset by α_m [29]. Additionally, the same has been shown from measurements of the first two harmonics [30]. Such a relation creates ambiguities between pairs of parameters raising serious estimation issues. In our experiments with the phase minimization, we encounter the same issues. Consequently, in this paper we focused on the methods to estimate glottal parameters. Investigating the existing glottal models and the estimation of their multiple parameters should be the subject of a dedicated study. Accordingly, we used the relaxing shape parameter Rd to reduce the shape parameter space to a single meaningful curve [2], [5]. When Rd tends to big values, the time-derivative glottal model approaches a period of a sinusoid. If Rd tends to small values, it approaches roughly a negative Dirac delta. In the context of this presentation, it is important to note that this drastic reduction of the LF model shape space implies that the methods, as presented in this paper, can be applied to a reduced number of voices.

A. Iterative Algorithm Using MSP

To minimize $\mathrm{MSP}(Rd, \phi, N)$, since only two variables are estimated, only a small number of harmonics should be necessary to find the global minimum of the corresponding error surface. However, the glottal model definitely does not correspond perfectly to the real glottal pulse. Therefore, an average solution with the different contributions of all the available harmonics is preferable. N is therefore set to $\lfloor f_{\mathrm{lim}}/f_0 \rfloor$, where f_{lim} is the Nyquist frequency or a voiced/unvoiced frequency (VUF) [31]. As one can see in Fig. 1(a), the error function corresponding to a linear-phase deviation is a deep and narrow valley in a noisy neighborhood. In such an error surface, the search for a global minimum is difficult. However, the high-frequency behavior of the error function comes from the high frequencies of the convolutive residual. Therefore, to smooth down the error function, $R_k^{(Rd,\phi)}$ is first low-passed at the third harmonic ($N = 3$). Then, a preconditioned conjugate gradient (PCG) algorithm is used to find the nearest minimum of the error function from starting values. Then, N is increased one harmonic by one harmonic up to its maximum value while using the PCG algorithm at each incrementation to refine (Rd, ϕ) obtained at the preceding step (see Algorithm 1 and Fig. 2). In order to start this optimization method, initial values are necessary. Therefore, the results of this method depend on the choice of these initial values.

Algorithm 1 Iterative algorithm using $\mathrm{MSP}(Rd, \phi, N)$

```
Build $S_k$ using a sinusoidal model
Initiate $Rd$ and $\phi$ with rough estimates
for $N = 3$ to $\lfloor f_{\mathrm{lim}}/f_0 \rfloor$ do
    repeat
        Synthesize $G_k^{Rd} \cdot jk$ with LF model and $Rd$
        Compute the VTF $C_{k-}^{Rd}$ with (2)
        Compute convolutive residual $R_k^{(Rd,\phi)}$ with (3)
        Compute $\mathrm{MSP}(Rd, \phi, N)$ with (4)
    until PCG algorithm find a minimum of $\mathrm{MSP}(Rd, \phi, N)$
end for
```

B. Phase Difference Computation for MSPD and MSPD^2

To avoid any problems with the phase wrapping in a limited range (ex. $[-\pi; \pi[$), the phase difference operation of (8) is computed in the complex plane

$$\Delta \angle X_k = \angle \left(\frac{X_{k+1}}{X_k} \right).$$

From our observations with synthetic signals, the function $\mathrm{MSPD}(Rd, \phi, N)$ has always only one minimum. However, with real signals, since the glottal model does not always correspond to the real glottal pulse, more minima can exist. Algorithm 1 is not necessary in order to find the global minimum of $\mathrm{MSPD}(Rd, \phi, N)$, since the position error is not weighted by the harmonic number as in MSP. Instead, a regular preconditioned conjugate gradient method can be used with $N = \lfloor f_{\mathrm{lim}}/f_0 \rfloor$.

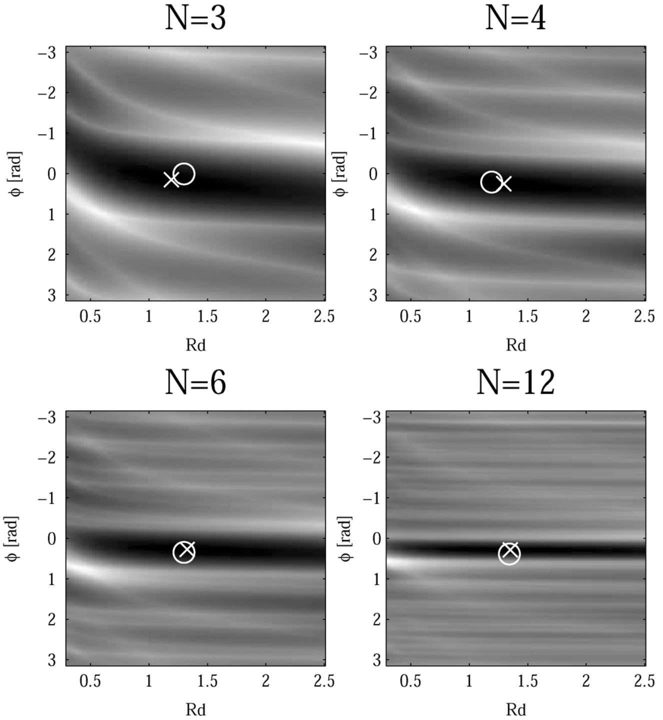


Fig. 2. Error surface corresponding to $\mathrm{MSP}(Rd, \phi, N)$ while increasing N. Starting values of each step are indicated with a circle and preconditioned conjugate gradient (PCG) final step with a cross.

To minimize $\mathrm{MSPD}^2(Rd, N)$, the second-order phase difference centered on each kth-harmonic is first computed in the complex plane

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

Then, applying the anti-difference operation, the previous equation leads to

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

Finally, like the MSPD error function, $\mathrm{MSPD}^2(Rd, N)$ has usually only one minimum with $N = \lfloor f_{\mathrm{lim}}/f_0 \rfloor$. A Brent's algorithm [32] is therefore used to find the global minimum of $\mathrm{MSPD}^2(Rd, N)$. Note that no initial values are necessary for this optimization method.

IV. Evaluation

First, the influence of the fundamental frequency on the proposed methods is evaluated using synthetic signals. Second, glottal and environment noise are used for the estimation of the reliability of the proposed methods. Then, the methods are compared to EGG signals with real speech signals. The GCIs estimated by the method using MSP are compared to reference GCIs computed from EGG signals. Since the EGG signals are close to a ground truth, we consider that the evaluation of the detected GCIs with synthetic signals is not necessary. The estimated shape parameters using MSP and MSPD^2 are also compared to the open quotient computed from EGG signals ($O_q = (t_e - t_0) \cdot f_0$ in [3]). Finally, the estimation of the glottal source by inverse filtering is discussed and two examples of parameter estimates on real speech recordings are shown.

In the evaluation tests with synthetic and EGG signals, three other methods are compared with the proposed ones.

1) *The Iterative Adaptive Inverse Filtering (IAIF)* [4], [17]: This method is designed to estimate the glottal source and not directly the parameters of a glottal model as in the proposed methods (see Fig. 7). In order to obtain parameter estimates of a glottal model, the LF model is fitted to the estimated source with a preconditioned conjugate gradient (PCG) algorithm by minimizing the mean squared error in the time domain. In its original implementation available in the *Aparat* toolkit [33], the three LF shape parameters are estimated as well as the excitation amplitude E_e. To obtain a valuable comparison with the proposed methods, that implementation has been replaced in order to estimate the Rd shape parameter (note that, E_e has to be estimated jointly with Rd because the mean squared error is sensitive to the glottal model amplitude). Additionally, the fitting process has been corrected according to the assumptions made by the PCG method. First, the error function has to be continuous. The position of the pulse is thus optimized using a linear-phase on the spectral representation of the glottal model and not with an integer shift of its time domain representation. Second, the dependency between the optimized parameters has to be as small as possible. Whereas the original implementation optimizes the time domain parameters (t_p, t_e) which are both dependent on the linear-phase of the glottal model, the implementation used in this presentation optimizes the linear-phase of the glottal model and the Rd parameter, which does not influence this linear-phase. Finally, in order to obtain a smooth influence of the time position of the glottal model on the error function, the mean squared error is weighted in time domain by a two-period Hanning window centered on t_e. Taking into account these considerations, the results of the IAIF have been significantly improved.
2) *Complex Cepstrum (CC) and ZZT*: The minimum/maximum-phase decomposition by means of the complex cepstrum has been already proposed to retrieve the maximum-phase component of the glottal pulse [11]. However, this method is known to be sensitive to the unwrapping of the phase spectrum involved in the computation of the complex logarithm. Bozkurt *et al.* [12] proposed to use the Zeros of the Z-Transform (ZZT) to obtain this decomposition but noise seems to also decrease the efficiency of this method [34]. Like the IAIF method, these two methods estimate the glottal source and not directly the parameters of the glottal model. Therefore, in this evaluation section, the LF fitting process discussed above for the IAIF method is used on the glottal source estimated by the CC and ZZT-based methods. In plots at the bottom of Fig. 7, one can see that the glottal pulse is damped to the left by the analysis window used in the decomposition algorithm. Therefore, during the fitting of the LF model, the same window is applied to the glottal model in order to reduce a possible bias between the observed pulse and its model. In this paper, the implementations of the CC and ZZT decomposition methods are the ones used in [35].

In the following, for all of the compared methods, the analyzed signal is resampled to 16 kHz and the error measure is limited to a voiced/unvoiced frequency fixed to 2 kHz. For synthetic signals, this value is kept constant in order to have all of the compared methods equally affected by this limit. The influence of this value on the results of the methods is discussed for real signals in Section IV-B2.

A. Synthetic Signals

The synthetic signal (12) is controlled by the LF shape parameter Rd^*, the delay ϕ^* between the first GCI and the start of the signal, the known fundamental frequency f_0, one Gaussian noise $n^{\sigma_g}[n]$ called *glottal noise* of standard deviation σ_g added to the glottal source and one Gaussian noise $n^{\sigma_e}[n]$ called *environment noise* added to the voiced signal. Filters $C^{p}_{-}(\omega)$ are designed to model 13 different voiced phonemes p covering the vocalic triangle. Among these phonemes, four are nasalized. The transfer function of $C^{p}_{-}(\omega)$ is computed using the Maeda's digital simulator [36]. The main advantage of such VTF models compared to estimated frequencies and bandwidths of autoregressive models on real speech signals is the complete independence of the generated formants from the influence of the source. The following synthetic voiced signal can thus be generated:

$$E(\omega) = e^{j\omega\phi^*} \cdot G^{Rd^*}(\omega) \cdot \left[\sum_{l\in\mathbb{N}} e^{j\omega l/f_0}\right] + \mathcal{F}(n^{\sigma_g}[n])$$
$$s[n] = \mathcal{F}^{-1}(E(\omega) \cdot C^{p}_{-}(\omega) \cdot j\omega) + n^{\sigma_e}[n] \qquad (12)$$

where $\mathcal{F}(\,\cdot\,)$ is the discrete-time Fourier transform and $\mathcal{F}^{-1}(\,\cdot\,)$ its inverse. The amplitude of the Gaussian noise is set so as to control the signal-to-noise ratio (SNR) with either the glottal source or the voiced signal.

1) Error Related to the Fundamental Frequency: In this first test, knowing the issue raised by the sampling of the VTF frequency response by the harmonic structure of the source (see Section II-C), the influence of the fundamental frequency on the reliability of the estimators is evaluated. For each f_0 value, the estimation error of the compared methods is computed for the 13 VTFs and a random delay ϕ^*. For the methods using MSP and MSPD, the initial shape value is given by the method using MSPD^2 and the initial position is given by the ideal value ϕ^* delayed by a random variable in $[-0.1/f_0; 0.1/f_0]$ to simulate an initial error of position. The error is computed eight times with different initial positions in order to obtain a valuable statistical estimate of the mean and standard-deviation of the error. Finally, to focus on the influence of f_0, the noise signals are set to zero. Fig. 3(a) and (b) shows the mean and the standard-deviation of the estimation error.

As expected, the variance of the estimators increases with f_0 since the sampling of the VTF by f_0 does not provide enough information to reconstruct the VTF perfectly. The MSPD is the worst of the proposed methods because the position parameter can offset the shape error as discussed in Section II-D

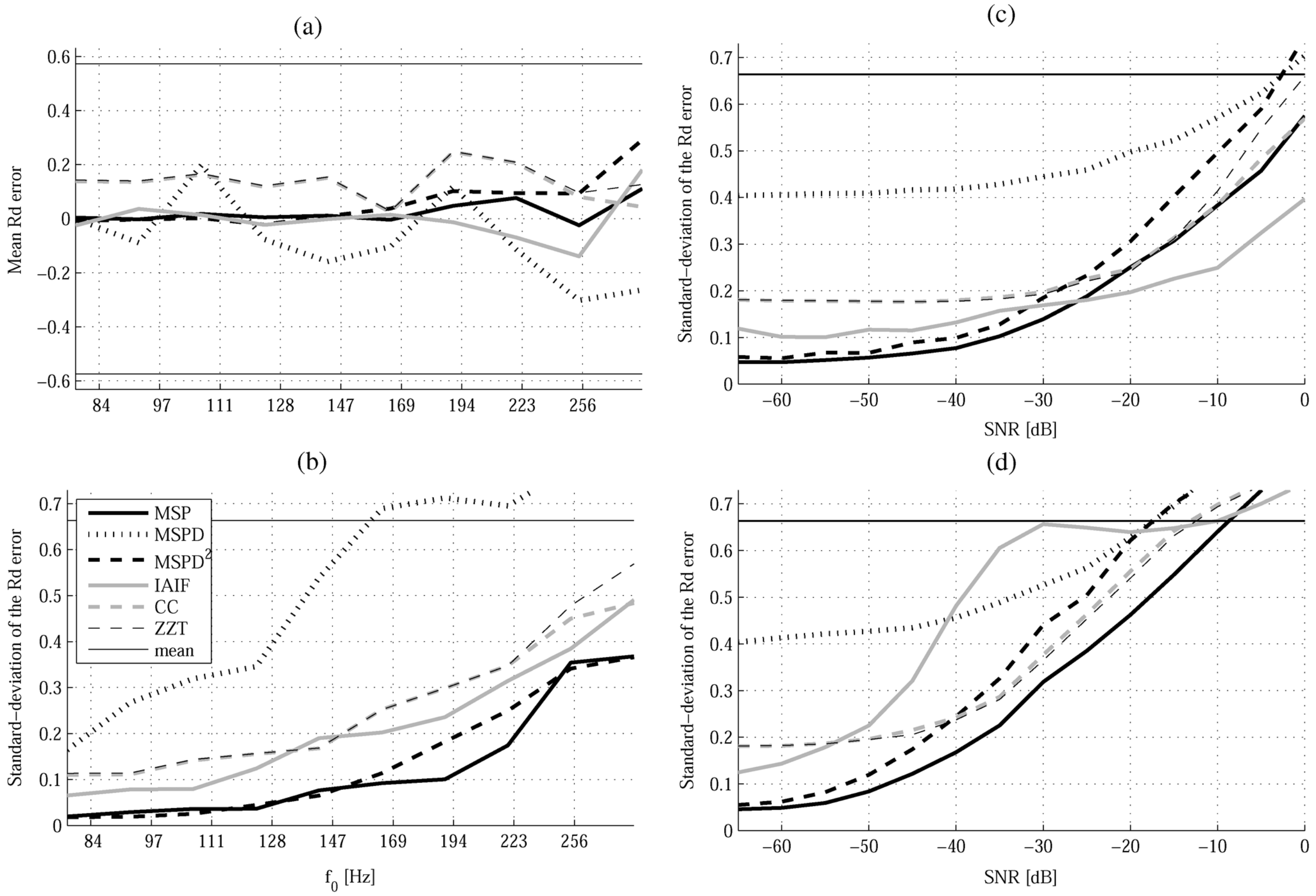


Fig. 3. (a) and (b) Mean and standard-deviation of Rd error with respect to f_0. (c) and (d) Standard-deviation of Rd error with respect to σ_g and σ_e. Theoretical limits are given through the *mean* estimator: in plots (b), (c), and (d) The *mean* method return the mean value of the Rd parameter range, without taking into account the input; in plot (a), the black thin lines represent the mean absolute error of the mean estimator. (a) Fundamental frequency. (b) Fundamental frequency. (c) Environment noise. (d) Glottal noise.

(This method is thus discarded in the evaluation using real signals). Additionally, the second-order phase difference of the MSPD and MSPD^2 removes the information provided by the average phase spectrum of the glottal model. Therefore, the method based on MSP is more precise than the two other proposed ones. Moreover, in the two last ones, the phase frequency derivative is approximated by the difference operator using discrete frequencies.

2) Error Related to the Noise Levels: This second test evaluates the influence of the noise levels σ_g and σ_e on the compared methods. To obtain a valuable statistical evaluation according to these levels, the error is computed 16 times for each σ value with the 13 different VTFs and a random position ϕ^*. To focus on the influence of the noises, f_0 is fixed to 128 Hz. In addition, when one noise is tested, the other one is set to zero. The results are shown in the right plots of Fig. 3.

One can see that, for equivalent SNRs, the efficiencies of the estimators are less disturbed by environment noise than by glottal noise. Moreover, the efficiencies of all of the methods decrease rapidly when increasing the glottal noise level. This can raise a serious issue in the presence of turbulence noise in breathy vowels. Moreover, for low noise levels the most reliable methods are the proposed ones. However, for important environment noise, the IAIF method is the most robust although its efficiency reduces significantly with glottal noise. The results of the CC and ZZT methods are close to each other. These methods are outperformed by the other methods in low noise conditions whereas, in high noise conditions, their efficiencies are between those of methods using MSP and MSPD^2.

B. Comparison With Electro-Glotto-Graphic Signals

The EGG is a non-invasive tool used in phoniatry to retrieve features of the motion of the vocal-folds. Among these features, one can obtain the instants of closure of the glottis (GCI) using the SIGMA method [37]. Additionally, the open-quotient O_q can be estimated using the DECOM method [38]. Assuming high correlation between the glottal source and the motion of the vocal-folds, reference sets of GCIs and O_q parameters can be created and compared to the estimation of the parameters of the proposed methods (in the evaluation, more than 5000 comparison pairs are used). In the following, the initial values used by the Algorithm 1 using MSP are given by the MSPD^2 and the GCIGS [39] methods. The fundamental frequency f_0 is estimated using the YIN method [26]. Moreover, the evaluation is made on voiced segments only and these segments are computed from the EGG signal: a time in the EGG signal is defined voiced if there is a reference GCI closer than half a period.

1) Evaluation of GCI Estimates: The reference GCIs of the EGG signal are compared to the GCIs described by the LF model (t_e instant in [5, , p. 19], [24], and [3]). Additionally, due

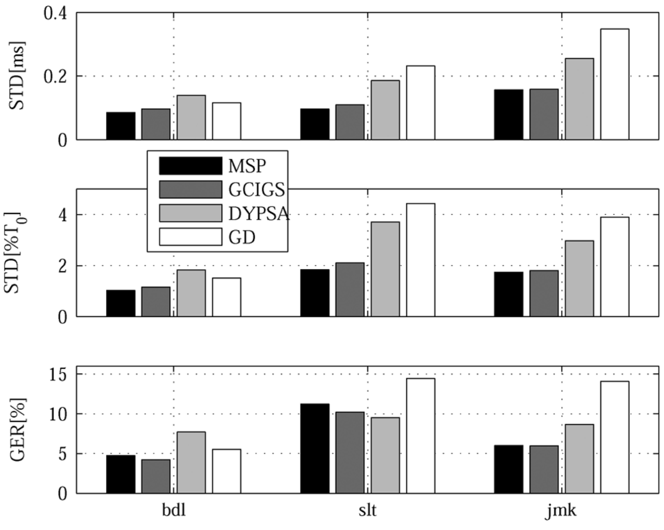


Fig. 4. Evaluation of GCI estimation methods with *Arctic* Databases. STD is the standard-deviation computed through the interquartile range of the duration between the reference and the detected GCIs, given in milliseconds [ms] and in percent of the period $[\%T_0]$. The gross error rate (GER) is the percent of that same durations$>0.1 \cdot T_0$.

to the propagation time between the EGG and the waveform, the reference GCIs and the detected GCIs are synchronized for each utterance by maximizing their correlation. Four methods are compared: the proposed method using MSP, the previously proposed method using a glottal shape estimate (GCIGS) [39], the DYPSA method [14] and another method based on group-delay (GD) [15], [18]. Fig. 4 shows the evaluation results on three *CMU Arctic* databases [40]. Note that each database is made of only one voice.

In conclusion, as expected, the method based on MSP slightly improves the precision of the GCIGS method. Indeed, by joint minimization of the shape and the position, the phase spectrum of the convolutive residual is closer to linear than without joint estimate. However, the GCIGS method assumes that a prominent peak exists in a period of the time derivative of the glottal source [39] whereas the method based on MSP assumes that the whole phase spectrum of the glottal source corresponds to the one of the LF model. The hypothesis of the GCIGS method is thus weaker than the hypothesis of the MSP-based method. With real signals, it can explain why the GCIGS method is more robust than the MSP-based method (less gross error). Finally, compared to the state of the art, the joint estimation of the shape and the position seems not to improve the results much more than the GCIGS method does. Removing the source amplitude when computing the VTF has much more impact on the results (has been done in both GCIGS and MSP) than using the phase spectrum of the LF model (has been done with MSP only).

2) Evaluation of the Shape Parameter Estimate: The open quotient O_q measured on EGG signals can be compared to the one predicted from the estimated Rd parameter (using the prediction formula in [2]). Note that the weighting of the error functions varies among the compared methods. The methods based on glottal source estimation (i.e. IAIF, CC, and ZZT) weight the mean squared error of the LF fitting in the spectral domain according to the estimated glottal source. The glottal formant around the first three harmonics is thus reinforced compared to the spectral tilt in high frequencies. Conversely, in the proposed methods, the weighting of the mean squared phase is uniform. In order to evaluate the influence of the weighting on the efficiencies of the estimators, Fig. 5 shows the O_q estimation error related to the number of harmonics taken into account in the error measure.

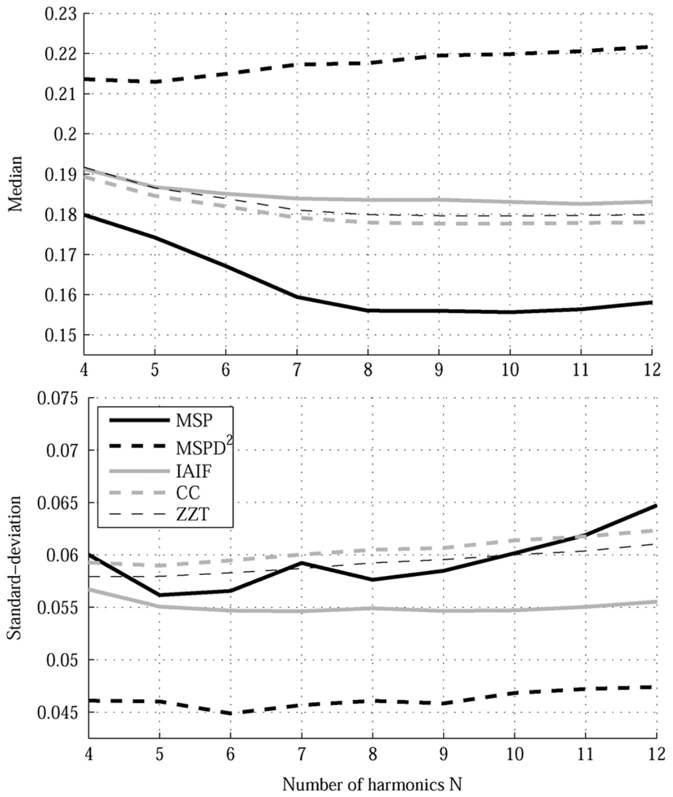


Fig. 5. Mean over the three databases of the median and standard-deviation of the O_q estimation error related to the number of harmonics taken into account in the error measure.

According to this figure, although it can be interesting to estimate the high-frequency properties of a glottal model (ex. spectral tilt), increasing the frequency band in the error measure seems to substantially decrease the efficiency of the MSP-based method. More generally, all of the methods have the same behavior except IAIF. Additionally, the method using MSPD^2 outperforms all of the compared methods. Note that, conversely to the evaluation with synthetic signals, the MSPD^2 outperforms the MSP-based method in this comparison with real signals. Although the optimization algorithm 1 might not find the optimum, a grid search algorithm has shown the same results for the MSP. Therefore, we assume that the following can explain this difference of results: The difference between the glottal model and the real glottal pulse introduces a distortion in the phase spectrum of the convolutive residual. With the method using MSP, ϕ and Rd can offset each other in order to minimize the error function. Conversely, the MSPD^2 can be systematically biased by this distortion but it has a smaller variance. In Fig. 5, the significant difference between MSP and MSPD^2 median values support this explanation.

According to Fig. 5, one can select the number of harmonics implying the smallest variance for each method: 5 for MSP, CC

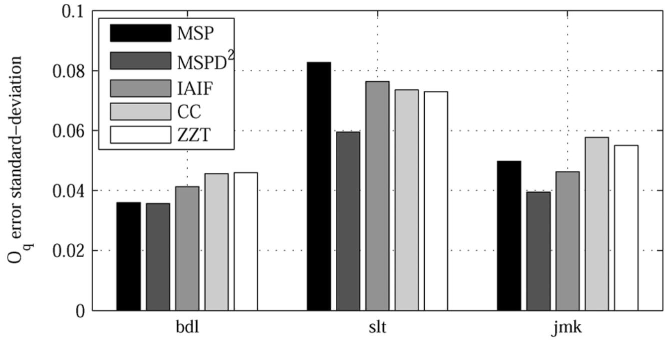


Fig. 6. Standard-deviation of the O_q estimation error (computed through the interquartile range).

and ZZT; 6 for MSPD^2; and 7 for IAIF. Fig. 6 shows the corresponding standard-deviations of the methods for each database separately. In conclusion, whereas the results of the MSP-based method vary significantly among the evaluated voices, the method using MSPD^2 clearly outperform all of the compared methods.

C. Glottal Source Estimation

The estimation of the glottal source is a straightforward application of the estimation of a glottal model. In this section, we will focus on the *radiated glottal source*, the time derivative of the glottal source. Conversely to IAIF, CC, and ZZT methods, the proposed methods do not estimate the glottal source explicitly before estimating the glottal model parameters. However, using the estimated parameters, the radiated glottal source $\tilde{G}_k$ can be retrieved through the VTF expression (2):

$$\tilde{G}_k = \frac{S_k}{C^{Rd}_{k-}} \quad \text{with } C^{Rd}_{k-} = \mathcal{E}_-\left(\frac{S_k}{G^{Rd}_k \cdot jk}\right).$$

Examples of $\tilde{G}_k$ are shown in the top of Fig. 7. Note that the function $\mathcal{E}_-(\cdot)$ changes only the phase spectrum of its argument. The amplitude spectrum is kept. Therefore, in terms of amplitudes, one can write the previous equation as

$$|\tilde{G}_k| = \frac{|S_k|}{|S_k|} \cdot \left|G^{Rd}_k \cdot jk\right| = \left|G^{Rd}_k \cdot jk\right|.$$

Consequently, the amplitude spectrum of the estimated radiated glottal source is the one of the radiated glottal model. Only the phase spectrum can reveal behaviors of the underlying real glottal pulse. Additionally, compared to the other methods, the proposed methods use a harmonic model for both the observed signal and the VTF estimate. Therefore, only a single period of the glottal source can be represented. Conversely, the IAIF method estimates an autoregressive filter (using the discrete all-pole method (DAP) [10]) in order to obtain a representation of the VTF which covers all the frequencies. The speech signal can thus be inverse filtered to retrieve multiple periods of the glottal source (four periods in Fig. 7). The glottal source estimated by CC or ZZT is made of two periods because the decomposition algorithm has to avoid zeros made by the periodicity of the speech signal [35]. Its anti-causal part contains the estimated maximum-phase component and the causal part remains to zero. One can see that the ripples in the estimated glottal source are more significant with the IAIF method because of the lack of precision of the DAP method. Conversely, the MSP-based method shows nearly no ripples because the VTF expression is based on the amplitudes of the harmonic model which can be estimated almost perfectly. Ripples made by the CC or ZZT methods are difficult to evaluate because the anti-causal part is damped by the analysis window and the causal part is set to zero by the decomposition algorithm.

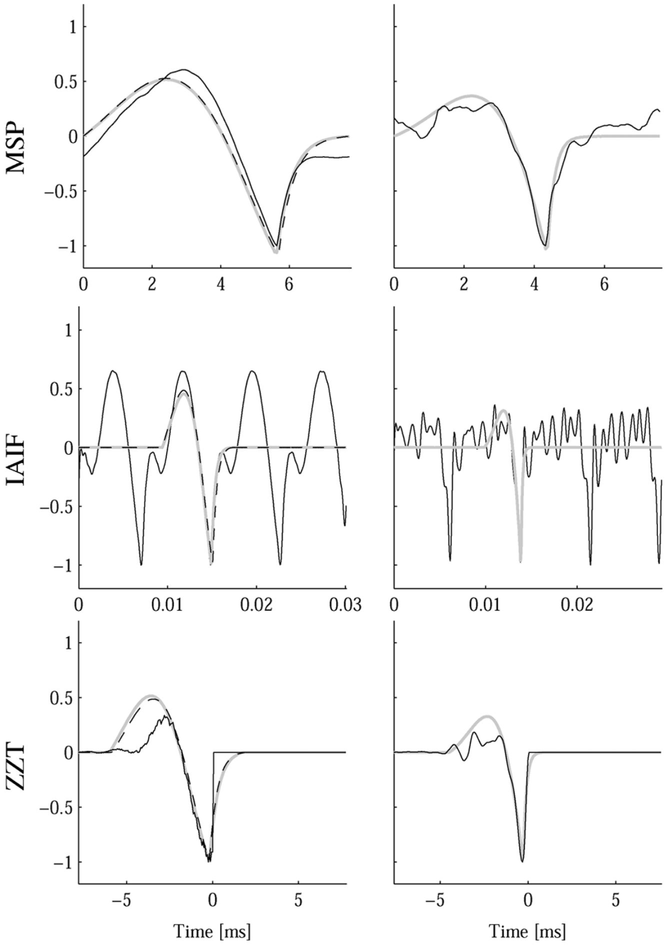


Fig. 7. Estimation of the glottal source in thin solid black line using MSP, IAIF, and ZZT. The synthetic glottal pulse is shown in dashed lines and the estimated LF pulse in thick gray line.

D. Examples on Recordings

Fig. 8 shows estimated Rd values of real recordings using MSP and MSPD^2. A sustained open /e/ from breathy to tense phonation is shown in the upper plot. As expected from the physiological behavior of the vocal folds, the estimated Rd value moves from a relaxed shape to a more tense shape. The bottom plot shows the start of the first utterance of the *Arctic bdl* database: “Author of the danger $[\cdots]$”. In the speech utterance, one can see that significant changes of the Rd parameter exist in short time intervals. We can see two different explanations. First, the harmonic model can be erroneous in transients (see time 0.35). Second, if the GCI is misestimated, the Rd estimate is also misestimated by the MSP-based method (time 0.74). However, one can see that the voice quality can vary inside a single phoneme (see interval [0.9; 1.1]).

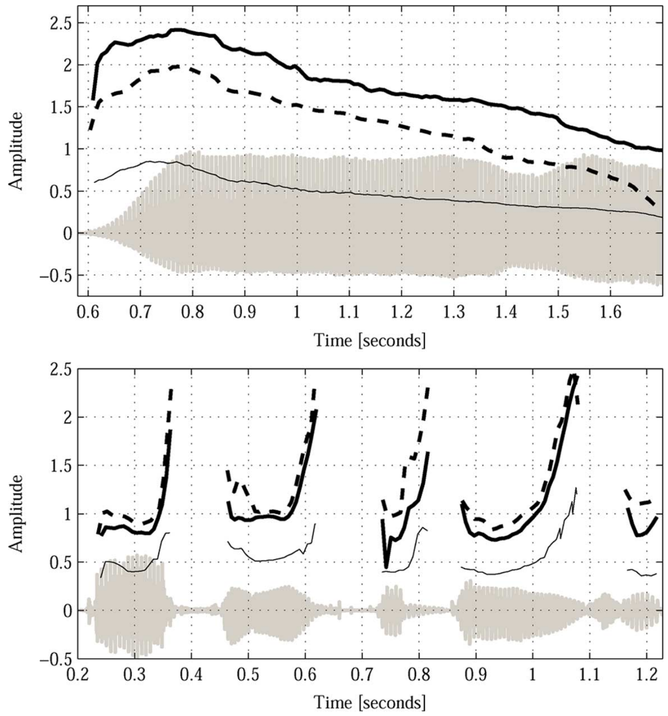


Fig. 8. Examples of Rd estimates on real recordings: The upper plot shows a sustained open /e/ from breathy to tense phonation. The bottom plot shows the analysis of the utterance "Author of the danger $[\cdots]$". The estimate using MSP in plain line and the estimate using MSPD^2 in dashed line. The O_q value computed from the EGG is shown in thin line.

V. Conclusion

We argued that the main difference between the glottal source and the vocal-tract filter is their mixed-phase and minimum-phase property. Accordingly, we showed that this difference can be used in the estimation of the shape parameter of a glottal model. First, a method minimizing the mean squared phase (MSP) of the convolutive residual of a voice model has been proposed to jointly estimate the shape parameter and the time position of a glottal model. In order to estimate the parameters of a given glottal model with the proposed methods, we discussed the conditions which have to be satisfied by the glottal model and its parametrization. Second, to estimate the shape parameter only, we saw that the glottal model position can be ignored using the second-order phase difference with respect to the harmonics (leading to the method using MSPD^2).

Using synthetic and EGG signals the efficiencies of the proposed methods were evaluated. In terms of GCI detection, the method using MSP outperformed the compared methods and slightly improved the efficiencies of a previously proposed method. However, its robustness can be lower than the other methods because it is possible that the phase of the LF model does not correspond to the phase of the real source. To evaluate the shape parameter estimates, the proposed methods have been compared to the IAIF, Complex Cepstrum, and ZZT methods. The last methods estimate the glottal parameters after a separation of the Vocal-Tract Filter and the glottal source whereas the proposed methods jointly estimate the shape parameters of the glottal model and a representation of the VTF. Additionally, we saw that the weighting of the error functions involved in the different methods influences the efficiencies of all the methods.

In order to obtain the best efficiencies for each of the compared methods, the number of harmonics taken into account in the error functions must not exceed 6. In conclusion to the evaluations, whereas the method based on MSP seemed to imply more precise estimates using synthetic signals, evaluation with EGG signals showed that the method using MSPD^2 outperformed all the compared methods. Moreover, in addition to being independent of the glottal model position, another advantage of the MSPD^2 is that it does not need initial values. Finally, the estimated glottal source using the proposed method showed less ripples compared to the IAIF method and two examples on real recordings showed that the estimated Rd values are highly correlated to the breathy/tense voice quality.

References

[1] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," *VOQUAL*, 2003.

[2] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2–3, pp. 119–156, 1995.

[3] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[4] P. Alku, H. Tiitinen, and R. Naatanen, "A method for generating natural-sounding speech stimuli for cognitive brain research," *Clinical Neurophysiol.*, vol. 110, no. 8, pp. 1329–1333, 1999.

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

[6] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of LF glottal source parameters based on an ARX model," in *Proc. Interspeech*, 2005.

[7] T. Drugman, T. Dubuisson, A. Moinet, N. D'Alessandro, and T. Dutoit, "Glottal source estimation robustness," in *Proc. SIGMAP*, 2008.

[8] J. D. Markel and A. H. Gray*, Linear Prediction of Speech*. Berlin, Germany: Springer Verlag, 1976.

[9] I.-T. Lim and B. G. Lee, "Lossless pole-zero modeling of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 3, pp. 269–276, Jul. 1993.

[10] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, Feb. 1991.

[11] A. Oppenheim, R. Schafer, and T. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, no. 8, pp. 1264–1291, Aug. 1968.

[12] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE Signal Process. Lett.*, vol. 12, no. 4, pp. 344–347, Apr. 2005.

[13] D. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 350–355, Aug. 1979.

[14] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. ICASSP*, 2002, pp. I-349–I-352.

[15] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[16] G. Degottex, A. Roebel, and X. Rodet, "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness," in *Proc. ICASSP*, 2010, pp. 5058–5061.

[17] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[18] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, Mass. Inst. of Technol., Cambridge, 2004.

[19] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, 1999.

[20] T. Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems," Ph.D. dissertation, Nagoya Inst. of Technol., Nagoya, Japan, 2002.

[21] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, 2008.

[22] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Berlin, Germany: Springer Verlag, 1972.

[23] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, 2nd ed. Englewood Cliffs: Prentice-Hall, 1978.

[24] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: An analytic study," in *Proc. ICASSP*, 2000, pp. 1295–1298.

[25] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[26] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, Apr. 2002.

[27] A. Camacho, "SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music," Ph.D. dissertation, Univ. of Florida, Gainesville, USA, Dec. 2007.

[28] C. Yeh and A. Roebel, "A new score function for joint evaluation of multiple f0 hypothesis," in *Proc. DAFx*, Naples, Italy, Oct. 2004, pp. 234–239.

[29] D. Vincent, "Analyse et controle du signal glottique en synthese de la parole," (in French) Ph.D. dissertation, ENST, Paris, France, 2007.

[30] N. Henrich, C. d'Alessandro, and B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data," in *Proc. Eurospeech*, 2001.

[31] K. Hermus, H. Van Hamme, and S. Irhimeh, "Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score," *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 820–823, Nov. 2007.

[32] R. P. Brent, *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[33] M. Airas, H. Pulakka, T. Backstrom, and P. Alku, "Toolkit for voice inverse filtering and parametrization," in *Proc. Interspeech*, 2005, pp. 2145–2148.

[34] N. Sturmel, C. d'Alessandro, and B. Doval, "A comparative evaluation of the zeros of z transform representation for voice source estimation," in *Proc. Interspeech*, 2007.

[35] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proc. Interspeech*, 2009.

[36] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Commun.*, 1982.

[37] M. R. P. Thomas and P. A. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1557–1566, Nov. 2009.

[38] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. Acoust. Soc. Amer.*, vol. 115, no. 3, pp. 1321–1332, 2004.

[39] G. Degottex, A. Roebel, and X. Rodet, "Glottal closure instant detection from a glottal shape estimate," in *Proc. SPECOM*, 2009, pp. 226–231.

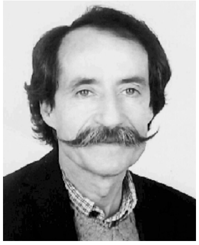
[40] J. Kominek and A. W. Black, "CMU arctic databases for speech synthesis," 2003.

Gilles Degottex (M'10) received the Diploma degree in computer science from the University of Neuchâtel, Neuchâtel, Switzerland, in 2003. He is currently pursuing the Ph.D. degree on the Analysis/Synthesis Team IRCAM, Paris, France, after a one-year specialization in signal processing at the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

His research interests include fundamental frequency tracking for musical instruments and the voice modeling using a glottal model for voice transformation and speech synthesis.

Axel Roebel (M'08) received the Diploma degree in electrical engineering from Hannover University, Hannover, Germany, in 1990 and the Ph.D. degree (*summa cum laude*) in computer science from the Technical University of Berlin, Berlin, Germany, in 1993.

In 1994, he joined the German National Research Center for Information Technology (GMD-First), Berlin, where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996, he became an Assistant Professor for digital signal processing in the Communication Science Department, Technical University of Berlin. In 2000, he was a Visiting Researcher at CCRMA Standford University, Stanford, CA, where he worked on adaptive sinusoidal modeling. In the same year, he joined the IRCAM to work on sound analysis, synthesis, and transformation algorithms. In summer 2006, he was Edgar-Varese Guest Professor for computer music at the Electronic Studio, Technical University of Berlin, and currently he is head of the Analysis-Synthesis Team at IRCAM. His current research interests are related to music and speech signal analysis and transformation.



Xavier Rodet (M'06) is currently an Emeritus Researcher on the Analysis/Synthesis Team, IRCAM, Paris, France. His research interests are in the areas of signal and pattern analysis, recognition, and synthesis. He has been working particularly on digital signal processing for speech, speech and singing voice synthesis, and automatic speech recognition. Computer music is his other main domain of interest. He has been working on understanding spectro-temporal patterns of musical sounds and on synthesis-by-rules. He has been developing new methods, programs, and patents for musical sound signal analysis, synthesis, and control. He is also working on physical models of musical instruments and nonlinear dynamical systems applied to sound signal synthesis.