Glottal inverse filtering: a new road-map and first results

Sandra Dias, Ricardo Sousa, and Aníbal Ferreira
Department of Electrical and Computer Engineering
University of Porto - Faculty of Engineering, Porto, Portugal
ajf@fe.up.pt

***Abstract*— In this paper we propose a frequency-domain approach to glottal inverse filtering that addresses typical problems in classic time-domain techniques, notably the lips/nostrils radiation and the group delay of the vocal tract filter. In particular, we use a new phase-related feature based on the Normalized Relative Delays (NRDs) of the harmonics representing the quasi-periodic component of the glottal source, in order to decouple the cumulative group delay effects due to the vocal tract filter and the quasi-periodic glottal excitation. While this topic is still on-going research, our paper details the global system architecture, demonstrates how to accurately reverse in the frequency domain the effects of lips/nostrils radiation, and demonstrates how the analysis-synthesis framework is able to implant any desired wave shape on a voiced sound, while preserving the original F0 fundamental frequency and time shift. Next research steps are also addressed that involve analysis-by-synthesis paving the way to joint source and filter estimation.**

I. Introduction

Glottal inverse filtering consists in a computational procedure that takes a discrete-time representation of a voiced sound (i.e., a sound whose generation involves vibration of the vocal folds), and that estimates the waveform produced at the glottis, i.e., the glottal source that excites acoustically the vocal tract filter. The estimation of the glottal source is important in many application areas which include the non-invasive quality assessment of the voice and especially of the vocal folds; the extraction of voice parameters denoting the idiosyncrasies of the speaker [1], [2], the emotional state of the speaker, or the phonation type (e.g., modal, breathy or pressed) [3]; and the extraction of musically relevant phonation parameters for biofeedback purposes, for example, in singing. The estimation of glottal source parameters is also important in applications involving synthesis or re-synthesis of speech, in order for example to reach higher naturalness in text-to-speech, and to implement voice/identity transformation or disguise.

In this paper we briefly revisit the source-filter model of voice production (section II), we address typical inverse filtering approaches (section III) and their pitfalls, and we present the main ideas for a new, frequency domain, approach to glottal inverse filtering (section IV). This approach focuses on signal integration implemented in the frequency domain, and decoupling of the effects due to the glottal source and the vocal tract filter. It is emphasized that this approach takes advantage of the spectral diversity existing in the voiced signal, takes into consideration the impact of sampling in the time and frequency domains, and takes advantage of a new phase-related feature (NRD) applied to harmonic signals and paving the way to the group-delay characterization of the vocal tract filter. While a complete functional algorithm is still on-going research, preliminary results are shown in section IV addressing signal integration in the frequency domain, and implantation of any periodic wave shape on a voiced signal. Section V addresses the next development steps and section VI concludes this paper.

II. The source-filter model

Glottal inverse filtering is strongly linked to the source-filter model of voice production proposed by Fant [4]. This model consists of three main processing stages: source excitation, vocal tract filtering, and lips and nostrils radiation. This model is schematically represented in Fig. 1. The source excitation results from air expelled by the lungs and that may give rise to two main types of sound when the air flows through the glottis, an opening (i.e., gap) between the vocal folds located at the larynx. If the air flow is interrupted periodically as a result of a quasi periodic cycle of opening and closing of the vocal folds, the source excitation has the form of glottal pulses which are filtered by the vocal tract giving rise to a voiced sound. If the vocal folds are (intentionally) adjusted is such a way as to not vibrate while the air flows continuously through the glottis, a constriction either at the glottis or downstream in the vocal tract (e.g., teeth or lips), turns the flow into a turbulent (noise) signal giving rise to an unvoiced sound. Thus, a voiced sound exhibits a clear periodic pattern while an unvoiced sound does not. In general, speech sounds involve a combination of both types of sound.

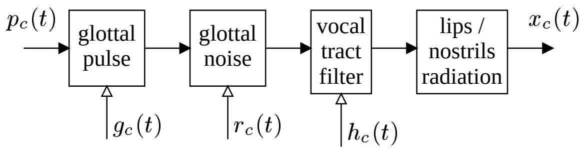

Fig. 1. Simplified source-filter model of voice production.

We admit in Fig. 1 that the noise component is produced at the glottis and is represented by $r_c(t)$, a random-like continuous function of time. Also, the periodic glottal pulses are obtained by convolving an ideal pulse train $p_c(t) = \sum_{\ell=-\infty}^{\infty} \delta(t - \ell T_0)$ with a prototype function of the glottal pulse which is represented by $g_c(t)$ (the subscript c denotes

This work was supported by the Portuguese Foundation for Science and Technology, an agency of the Portuguese Ministry for Science, Technology and Higher Education, under research project PTDC/SAU-BEB/14995/2008.

continuous time function). T_0 represents the period corresponding to a complete cycle of the vocal folds vibration, i.e., a pitch period. Its reciprocal, $F_0 = 1/T_0$ is the fundamental frequency of voicing. The impulse response of the vocal tract filter is represented by $h_c(t)$. In reality, this filter is slowly time-varying with respect to the pitch period. Typically, in the literature, vocal tract filtering is usually modeled as an all-pole filter that shapes the spectrum of the source according to the resonant frequencies (also known as formants) of the vocal tract.

The radiation effect of the lips/nostrils is typically modeled by taking the derivative of the air flow [5]. This operation converts volume velocity of the air flow into sound pressure variation which is the physical quantity captured by a microphone in the form on an electric signal. Thus, according to the source-filter model, this continuous-time signal may be modeled as

$$x_c(t) = \frac{d}{dt} \left\{ h_c(t) * \left[r_c(t) + g_c(t) * p_c(t) \right] \right\} \ . \qquad (1)$$

Simply stated, glottal inverse filtering involves estimating $g_c(t)$ from $x_c(t)$. Of course, in addition to reversing the effect of differentiation (i.e., in addition to integration), glottal inverse filtering requires that the vocal tract filter $h_c(t)$ be estimated so as to obtain the inverse of its transfer function. This approach presumes the assumption that the source and filter are independent of each other. Although this assumption is not realistic, more complex models taking explicitly into consideration the source-filter interaction, do not provide substantially better results than if independence is assumed [6].

As commonly assumed in the literature, it is very convenient to model the vocal tract filter as an all-pole filter because

- simple algorithms (such as the Levinson-Durbin recursion) within the framework of Linear Predictive Coding (LPC) and relying only on the short-time magnitude spectrum of the voice signal (or, equivalently, the autocorrelation function), may be used in the estimation of a smooth spectral envelope model (i.e., an AR model [7]) of that magnitude spectrum, and because
- the inverse transfer function of an all-pole filter is an all-zero filter, i.e., an FIR filter which is inherently stable.

However, all-pole modeling is not truly representative since many voice sounds, e.g. nasalized or fricative sounds, exhibit clear nulls (i.e., zeros) in their magnitude spectrum which denotes that the transfer function of the vocal tract filter, in order to be realistic, should include both zeros and poles.

III. Approaches to glottal inverse filtering

In the literature, the three main processing stages (source signal excitation, vocal tract filtering and lips/nostrils radiation) illustrated in Fig. 1, are implemented in the discrete-time domain as an additive combination between a quasi-periodic source component and a noise source component, convolution between the combined signal and the impulse response of an all-pole filter by means of a difference equation realizing this filtering operation and, finally, first-order discrete-time differentiation of the filtered signal. This operation is just a simple approximation to differentiation in the continuous-time domain and is implemented using the difference equation $y[n] = x[n] - x[n-1]$ which, in the Z-domain, corresponds to $Y(z) = X(z) \left[1 - z^{-1}\right]$.

Although in the literature enhanced analytical approaches have been developed with good results, including ARMA modeling (and not just AR modeling) of the vocal tract filter during the closed phase of the glottis (see [6] for an overview), they also present several difficulties. For example, the assumption of complete closure of the glottis is not realistic in female or child voices given the high fundamental frequency, and is even less realistic in the case of dysphonic or pathological voices. For theses reasons and taking into consideration the typical high computational complexity of those analytical approaches, as well as their typical low robustness since they depend on accurate techniques identifying individual pitch pulses and glottal closure instants, for the purpose of our discussion in this paper, we do not mention them specifically. Instead, we focus on those aspects that are common to the techniques widely used and implemented in publicly available software, e.g., [2].

Most techniques [5], [8], [2], [3], [6] for inverse filtering typically comprise

1) estimation of the all-pole model of the vocal tract filter using LPC techniques and presuming local stationarity,
2) filtering of the discrete-time voice signal with the all-zero filter resulting from inverting the all-pole filter (the purpose of this processing is to cancel the spectral coloration of the signal due to the effect of the vocal tract filter), and
3) filtering the resulting signal with a first-order integrator.

The integrator implements a transfer function which only approximates the inverse of the differentiator, such as to avoid the pole at $z = 1$: $Y(z) = X(z) / \left[1 - \alpha z^{-1}\right]$, where α is a constant close to 1.0 [5].

In this paper we argue that this basic discrete time-domain approach to inverse source inverse filtering leads to poor estimates of the glottal source since

1) the modeling of a combination of continuous-time processing steps as a combination of corresponding but independent discrete-time processing steps, introduces errors and artifacts, notably at the integration,
2) full-bandwidth discrete-time processing does not allow to exploit the spectral diversity of the quasi-periodic and noise components of the glottal source with the consequence that the latter (i.e., the noise component) strongly disturbs and even corrupts the estimation of the former (i.e., the quasi-periodic component),
3) the use of an all-pole model for the vocal tract is not sufficiently representative since, as mentioned in the previous section, it has an inherent difficulty dealing with zeros of nasalized sounds for example, and is also not appropriate to model female or child speech sounds

since the high fundamental frequency (i.e., F0) in these cases makes that the spectral envelope modeling of the all-pole filter is ‘locked’ to the harmonic frequencies of F0,

4) the overall approach is not flexible and general enough, notably in the case of mildly dysphonic voices since the assumption of smooth and stable spectral organization of the quasi-periodic component due to the glottal source is simply not realistic.

IV. A NEW APPROACH TO GLOTTAL INVERSE FILTERING

We propose an approach to glottal inverse filtering that comprises frequency-domain signal analysis and synthesis, that relies on an accurate frequency-domain modeling of the multiplicative effects of filtering, including differentiation, and that takes advantage of the spectral diversity of the quasi-periodic and noise components of the glottal source.

While some of these aspects will be detailed in the following sub-sections, here we focus on the frequency-domain characterization of the source-filter model, and we address the implications of sampling in the time and frequency domains.

The Fourier transform of eq. (1) is obtained as

$$X_c(\Omega) = j\Omega \left\{ H_c(\Omega) \left[R_c(\Omega) + G_c(\Omega) P_c(\Omega) \right] \right\} \ , \qquad (2)$$

where the multiplication in the frequency domain by $j\Omega$ denotes differentiation in the (continuous) time domain, and the Fourier pair involving the infinite pulse train is assumed:

$$p_c(t) = \sum_{\ell=-\infty}^{\infty} \delta(t - \ell T_0) \overset{\mathcal{F}}{\longleftrightarrow} P_c(\Omega) \ , \qquad (3)$$

where

$$P_c(\Omega) = \frac{2\pi}{T_0} \sum_{\ell=-\infty}^{\infty} \delta\left(\Omega - \ell\frac{2\pi}{T_0}\right) = \Omega_0 \sum_{\ell=-\infty}^{\infty} \delta\left(\Omega - \ell\Omega_0\right) \ .$$

Assuming that $|G_c(\Omega)| = 0$, $|\Omega| > L\Omega_0$, where L is a positive integer, $X_c(\Omega)$ (eq. (2)) may be also written as

$$j\Omega H_c(\Omega) \left[R_c(\Omega) + \Omega_0 G_c(\Omega) \sum_{\ell=-L}^{L} \delta\left(\Omega - \ell\Omega_0\right) \right] \ .$$

In order to simplify notation, we define the spectral coefficient c_ℓ as

$$c_\ell = F_0 H_c(\ell\Omega_0) G_c(\ell\Omega_0) \ . \qquad (4)$$

This spectral coefficient may be seen as the result of sampling the Fourier transform of $h_c(t) * g_c(t)$ at $\Omega = \ell\Omega_0$ which leads to a Fourier Series describing a periodic signal. It should be noted however that a synthesis based on the Fourier series $\sum_{\ell=-L}^{L} c_\ell e^{j\ell\Omega_0 t}$ does not reconstruct $h_c(t) * g_c(t)$ but instead a time aliased version by folding all its replicas delayed by T_0, as one would expect from eq. (1). However, if means are provided to cancel in the frequency domain the effect of $H_c(\ell\Omega_0)$, the synthesis based on the Fourier Series coefficients correctly reconstructs $g_c(t)$ without aliasing.

Using (4), eq. (2) reduces to

$$X_c(\Omega) = j\Omega H_c(\Omega) R_c(\Omega) + j2\pi \sum_{\ell=-L}^{L} \ell\Omega_0 c_\ell \delta\left(\Omega - \ell\Omega_0\right) \ . \qquad (5)$$

This equation reveals that the Fourier representation of $x_c(t)$ consists of two components. The first component reflects the noise (i.e., the incoherent or random part) in the signal, and the second component reflects the sinusoidal structure of the signal. These two components are very informative regarding the underlying spectral diversity in the signal:

1) the noise component is continuous in Ω and quite likely does not exhibit strong local spectral peaks in the magnitude spectrum,
2) the sinusoidal structure of the signal consists of several peaks in the magnitude spectrum that are harmonically related, and that for normal and healthy voices should be well above the noise floor for a significant spectral range.

Given that the noise and sinusoidal parts are spectrally diverse but presumably both are influenced by the vocal tract filter, it should be possible to estimate $H_c(\Omega)$ using both parts. Then, it should be possible to estimate $g_c(t)$ after implementing integration in the frequency domain using the sinusoidal part of the spectrum, and after canceling the effect of $H_c(\Omega)$ in the spectrum. Thus, our processing framework is frequency-domain based and includes signal analysis and synthesis as illustrated in Fig. 2.

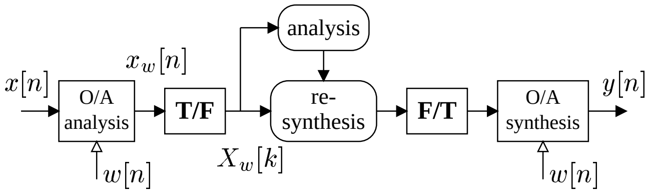

Fig. 2. Analysis-synthesis processing framework. T/F denotes Odd-DFT transformation and F/T denotes Inverse Odd-DFT.

This figure highlights that

1) the processing is frame-based using a segmentation based on a window $w[n]$ whose length (or time support) is N samples, and using 50% overlap between adjacent frames at the analysis, and 50% overlap-and-add between adjacent frames at the synthesis,
2) the time-frequency transformation is based on the Odd-frequency Discrete Fourier transform (Odd-DFT) for the reasons discussed in section IV-A,
3) using an appropriate window $w[n]$ and in the absence of spectral modification, the system is perfect reconstructing, i.e., $y[n] = x[n - n_d]$ where n_d is a constant system delay [9].

The main building blocks in Fig. 2 are the analysis and resynthesis. The former is responsible for estimating the sinusoidal components in the signal and their parametrization (frequency, magnitude, and phase), and for removing them (i.e., subtracting them) from the spectrum giving rise to the noise residual. The latter is responsible for selectively synthesizing the noise residual only, the sinusoidal components only, or any magnitude/phase modification affecting a specific sinusoidal or noise component. Preliminary experimental evidence of these capabilities will be illustrated in sections IV-A and IV-D.

The processing framework illustrated in Fig. 2 implies time sampling and frequency sampling. Here we address the impact of these operations on the spectrum represented by eq. (5).

Sampling in the time domain allows to obtain $x[n]$ from $x_c(t)$: $x[n] = x_c(nT_s)$, where T_s denotes the sampling period, the reciprocal of the sampling frequency F_s. From the theory of sampling, it is known that the associated spectra are related by $X\left(e^{j\omega}\right) = F_s \sum_{k=-\infty}^{\infty} X_c\left(F_s(\omega - k2\pi)\right)$, where the relation between unnormalized and normalized frequency variables is considered: $\Omega = \omega F_s$ [10].

We admit that the bandwidth of $X_c(\Omega)$ is limited to the Nyquist frequency (πF_s) due to the use of an appropriate analog anti-aliasing filter. Furthermore, we also admit that $L\Omega_0 < \pi F_s$, i.e., the bandwidth of $G_c(\Omega)$ is limited to πF_s. In this context, there is no aliasing in the range $\omega \in [-\pi,\ \pi[$ and, therefore, $X\left(e^{j\omega}\right) = F_s X_c\left(F_s\omega\right)$ in this range.

Since the constant F_s is canceled out in the discrete-to-analog signal reconstruction due to the effect of the anti-imaging filter, we omit it in order to simplify notation, hence, using (5)

$$\begin{aligned} X\left(e^{j\omega}\right) = X_c(F_s\omega) = j\omega F_s H_c(\omega F_s) R_c(\omega F_s) + \\ j2\pi \sum_{\ell=-L}^{L} \ell\omega_0 F_s c_\ell \delta\left(F_s(\omega - \ell\omega_0)\right) \ , \end{aligned} \qquad (6)$$

where we have considered $\Omega_0 = \omega_0 F_s$, and $c_\ell = F_0 H_c(\ell\omega_0 F_s) G_c(\ell\omega_0 F_s)$. Taking into consideration the normalization of the frequency axis due to sampling, we simplify eq. (6) further by using $H\left(e^{j\omega}\right) = H_c(\omega F_s)$, $R\left(e^{j\omega}\right) = R_c(\omega F_s)$, and $G\left(e^{j\omega}\right) = G_c(\omega F_s)$:

$$\begin{aligned} X\left(e^{j\omega}\right) = j\omega F_s H\left(e^{j\omega}\right) R\left(e^{j\omega}\right) + \\ j2\pi \sum_{\ell=-L}^{L} \ell\omega_0 F_s c_\ell \delta(\omega - \ell\omega_0) \ , \end{aligned} \qquad (7)$$

where $c_\ell = F_0 H\left(e^{j\ell\omega_0}\right) G\left(e^{j\ell\omega_0}\right)$.

We address now the implication of sampling in the frequency domain. As illustrated in Fig 2, signal segmentation implies windowing before the time-frequency transformation. Therefore

$$x_w[n] = x[n] \cdot w[n] \overset{\mathcal{F}}{\longleftrightarrow} \frac{1}{2\pi} X\left(e^{j\omega}\right) * W\left(e^{j\omega}\right) \ , \qquad (8)$$

and thus

$$\begin{aligned} X_w\left(e^{j\omega}\right) = \frac{j}{2\pi}\left[\omega F_s H\left(e^{j\omega}\right) R\left(e^{j\omega}\right)\right] * W\left(e^{j\omega}\right) + \\ j \sum_{\ell=-L}^{L} \ell\omega_0 F_s c_\ell W\left(e^{j(\omega-\ell\omega_0)}\right) \ . \end{aligned} \qquad (9)$$

Time-frequency transformation is achieved by taking the Odd-DFT [11] of $x_w[n]$:

$$X_w[k] = \sum_{n=0}^{N-1} x_w[n]\, e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n} \ , \qquad (10)$$

where N is the length of the transform. This means the spectral information we have access to for analysis and synthesis purposes, is discrete (i.e., for $k = 0, 1, \ldots, N-1$):

$$X_w[k] = X_w\left(e^{j\omega}\right)|_{\omega=(k+\frac{1}{2})\frac{2\pi}{N}} \ . \qquad (11)$$

It is a well-known signal processing fact that a discrete frequency-domain representation presumes that the associated time-domain signal is periodic with period N, which implies for example that a product in the frequency domain corresponds to a circular convolution in the time-domain. Provided however that a correct interpretation and modification of the sampled spectrum represented by (11) is performed with respect to the associated continuous-frequency version (represented by eq. (7)), signal resynthesis is possible without artifacts. This will be illustrated in sections IV-A and IV-D.

A. Handling lip and nostrils radiation

It has been mentioned previously that reversing the effect of differentiation, due to lips and nostrils radiation, using first-order discrete-time integration, motivates two sources or approximation errors. A first error is due to the first-order differentiator itself whose difference equation is $y[n] = x[n] - x[n-1]$. In fact, its frequency response differs significantly from the frequency response of an ideal differentiator, especially at high frequencies. A second source of error is introduced in the inverse transfer function of the first-order differentiator, in order to avoid the pole at $z = 1$ and to insure stability. A practical difference equation commonly used to implement the integrator is $y[n] = (1-\alpha)x[n] + \alpha y[n-1]$ [5]. This solution not only introduces signal distortions as it is vulnerable to noise and DC. In order to illustrate this, Fig. 3 represents the derivative of an ideal glottal pulse according to the LF model [12]. Two versions are represented: clean and contaminated by additive white noise at 9 dB SNR. The corresponding power spectral density is represented in Fig. 4. By using the above time-domain integrator with $\alpha = 0.98$, and taking as input the waveforms displayed in Fig. 3, the signals represented in Fig. 5 are obtained. It can be concluded that when the input is a clean signal, the closed phase of the glottal cycle shows an obvious artifact (in the form of an rising slope instead of being flat) which, interestingly, is quite similar to results of inverse filtering in other papers, e.g., Figure 1 in [8]. When the input is the noisy signal in Fig. 3, the output exhibits a strong distortion and fluctuation.

We now consider the alternative of accurate signal integration using frequency-domain processing presuming the frame-based environment illustrated in Fig 2, and signal modification using $X_w[k]$ as in eq. (11). Since signal modification and resynthesis uses only the sinusoidal components

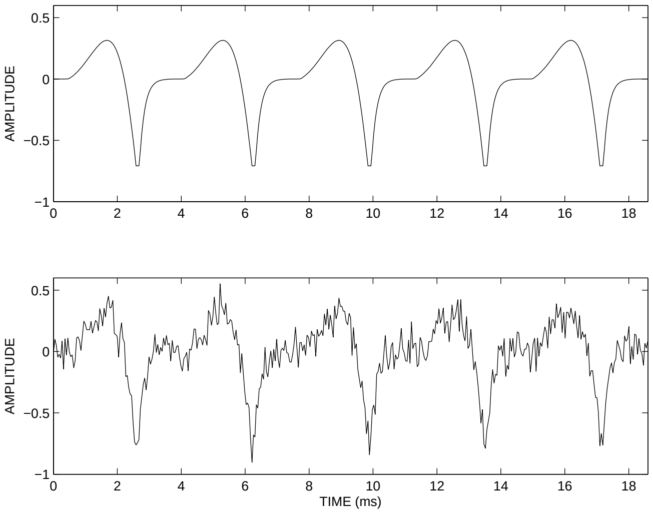

Fig. 3. Time representation of the derivative of the ideal LF glottal pulse without noise (upper figure) and with white noise at 9 dB SNR (lower figure).

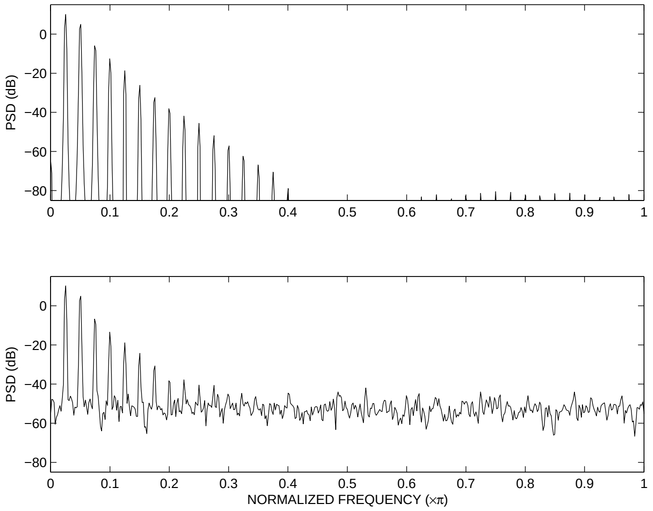

Fig. 4. Power spectral density of the derivative of the ideal LF glottal pulse without noise (upper figure) and with white noise at 9 dB SNR (lower figure).

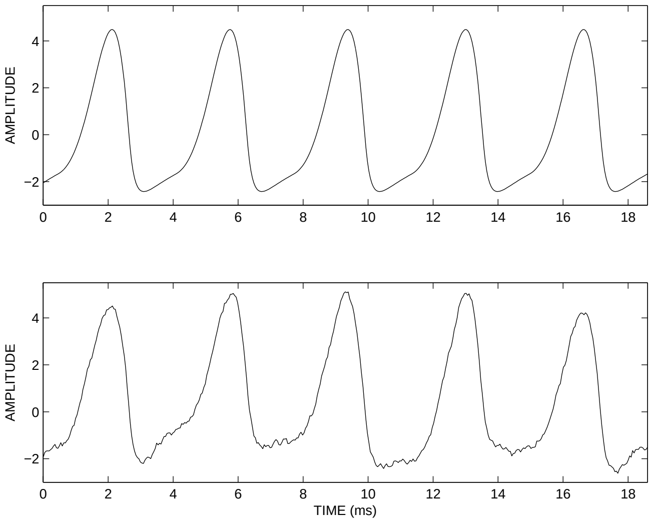

Fig. 5. Output results of a first-order time-domain integrator when the input signals are the waveforms represented in Fig. 3.

in the spectrum, using (11), eq. (9) reduces to

$$X_w[k] = j \sum_{\ell=-L}^{L} \ell\omega_0 F_s c_\ell W \left(e^{j(2\pi(k+0.5)/N - \ell\omega_0)} \right) . \quad (12)$$

After the frequency, magnitude and phase of all the identifiable sinusoids in the $X_w[k]$ spectrum are accurately estimated using the results in [13], the sinusoids are resynthesized in the discrete frequency domain generating a new spectrum vector $X_w[k]$. This operation explicitly insures high immunity to the existing noise and other signal components. The resynthesized spectrum vector is further modified to include the effect of integration:

$$Y_w[k] = \frac{X_w[k]}{jF_s 2\pi(k+0.5)/N}, \; k = 0, 1, \ldots, N-1 . \quad (13)$$

Contrarily to the plain DFT, the Odd-DFT has the advantage that all discrete frequency values are different from zero. Specifically, no division by zero exists for $\omega = 0$ rad. since this frequency value is not sampled by the Odd-DFT.

After inverse Odd-DFT and overlap-add (Fig. 2), the output signals represented in Fig. 6 are obtained when the input signals are the waveforms displayed in Fig. 3. In can be concluded that when the input is the clean derivative of the LF model, the output signal is a faithful reconstruction of the ideal LF glottal pulse. When the input signal is the noisy derivative of the LF model, the output still is a correct reconstruction of the LF glottal pulse. The rather subtle artifacts appearing during the closed phase of the glottal pulse are mainly due to the missing high-frequency components of the derivative of the glottal pulse, which have been ignored during the resynthesis because they are overwhelmed by the noise. Overall, a comparison between Fig. 5 and Fig. 6 clearly reveals that the signal integration using discrete Odd-DFT domain processing, is markedly superior to approximate integration in the time domain.

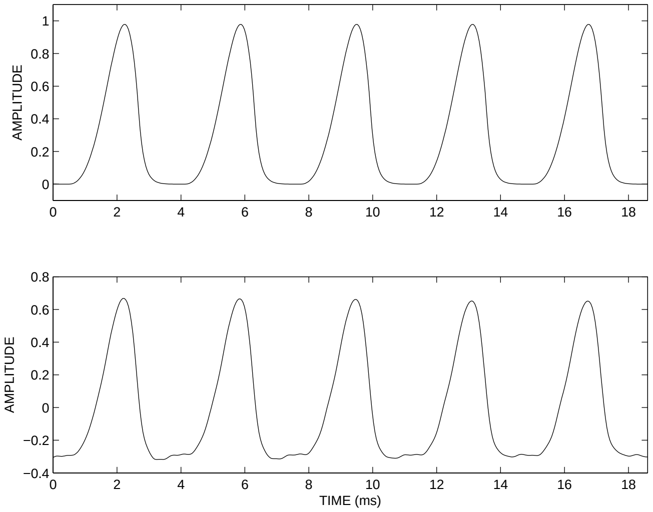

Fig. 6. Output results of time-integration using frequency-domain processing when the input signals are the waveforms represented in Fig. 3.

B. A phase related feature: Normalized Relative Delay

Normalized Relative Delays (NRD) are phase-related features denoting the (normalized) delay between the harmonics and the fundamental frequency of a periodic signal. Their usefulness lies in the fact that, in addition to magnitude information, they help to characterize completely the time waveform of the periodic signal (a concept also described in the literature as ‘shape invariance’ [14]), independently of the overall time shift and of the fundamental frequency. This property is very useful for signal analysis, identification and transformation.

The NRD concept has been presented in [15] and first studies have been conducted assessing their discrimination potential regarding voice phonation type [15], and vowel and singer identification [16].

NRDs may also be seen as a very convenient phase-domain counterpart of the magnitude spectrum in the sense that a graphical representation of both features is stable for a stationary periodic signal and independently of the time shift. In [15] a few examples are illustrated.

C. Using magnitude and NRDs to model source and filter

In the context of the source-filter model, NRDs are very important because they have the potential to help decoupling the cumulative group delay contributions due to source (i.e., the glottal pulses) and due the vocal tract filter. In fact, in the frequency domain, the magnitude contributions due to the source excitation and due to the vocal tract filter, are combined in a multiplicative way to generate the spectrum of the radiated voice signal. On the other hand, provided that a correct phase compensation is implemented to reverse the effect of lips/nostrils radiation, the NRDs extracted from the radiated voice signal result from the additive combination between the NRDs of the excitation signal (i.e., the glottal pulses), and the group delay of the vocal tract filter. Thus, having good starting models of both the glottal source excitation and of the vocal tract filter for a given vowel, we conjecture that it should be possible to find the specific glottal excitation signal, and the specific vocal tract filter that best explain the exact phase and magnitude spectrum of the radiated voice signal.

D. Implanting the magnitude and NRD models on a periodic signal

In order to illustrate the exceptional flexibility of our frequency-domain analysis and resynthesis environment, we illustrate in Fig. 7 the resynthesis of a vowel uttered by a young female by implanting two desired periodic waveforms with the exact same F0: the sawtooth wave and the LF glottal pulse. Since the wave shape of these waveforms is completely determined by the relative magnitude between all relevant upper harmonics and the fundamental frequency, as well as the NRDs between all relevant upper harmonics and the fundamental frequency, complete independence is achieved with respect to the overall magnitude of the wave (which depends only on the magnitude of the fundamental frequency), with respect to the overall delay of the wave (which depends only on the phase of the fundamental frequency), and with respect to the fundamental frequency itself. Thus, by accurately estimating the frequency, magnitude and phase of the fundamental frequency of the voice signal, any wave prototype may be implanted (with the same F0) since it is completely defined by the relative magnitude and NRD of all relevant harmonics.

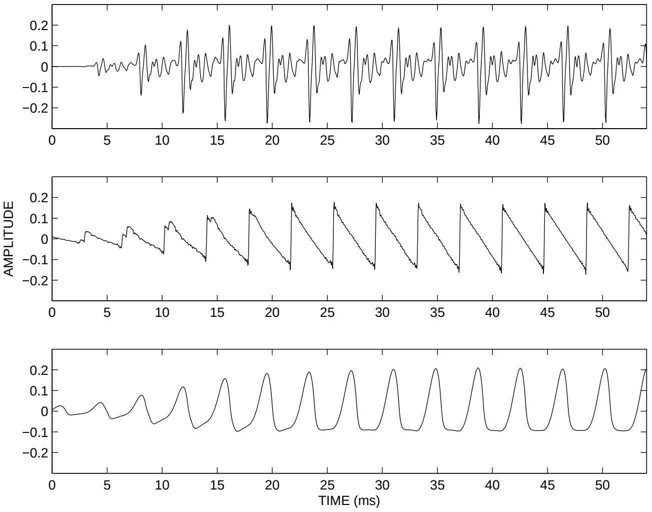

Fig. 7. Illustration of the resynthesis of a natural vowel sound (top panel) using the magnitude and NRD model of the first 30 harmonics describing a sawtooth wave (middle panel) and LF glottal pulse (bottom panel). These wave shapes may be implanted on any periodic signal using only the magnitude and phase of its fundamental frequency.

V. Next research steps

The next research steps will be determined by the approach depicted in Fig. 8 and that will be implemented to estimate the glottal source excitation and the vocal tract filter in an analysis-by-synthesis procedure.

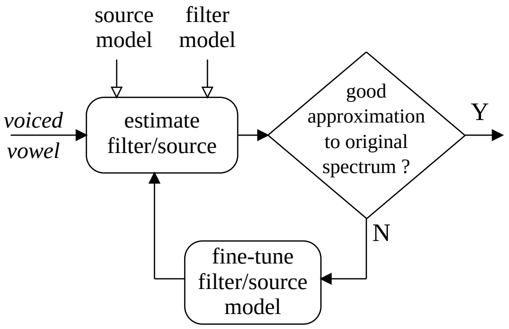
source model
filter model
voiced vowel
estimate filter/source
good approximation to original spectrum ?
Y
N
fine-tune filter/source model

Fig. 8. Analysis-by-synthesis approach to filter/source fine-tuning.

The approach with start with some default source and filter models, or, as described in section IV, with some first filter estimate resulting from the harmonic-noise decomposition of the voice signal. Then, an iterative algorithm using analysis-by-synthesis will find the best combination of source and filter spectra matching the voice spectrum, taking also into consideration the lips/nostrils radiation. In this process, and in order to make the algorithm able to deal with dysphonic voices, an adaptive evaluation must be made as to what assumption is stronger: either the source model or the filter model. This evaluation will govern the emphasis of the algorithm fine-tuning the filter or the source model.

While several general models are available for the the glottal source (e.g., the Rosenberg and the LF model), more realistic models than a simple LPC model do not exist for the vocal tract filter for a specific vowel. This will require specific experimentation with medical support in order to accurately extract (using two matched microphones) the transfer function between a point near the larynx, just after the vocal folds, and a point outside the mouth. This represents a very special test set up that for technical, safety and ethical reasons, must be conducted by an ORL professional.

VI. Conclusion

We have proposed a new frequency-domain approach to glottal inverse filtering that focuses on accurate modeling of the multiplicative effects of glottal source excitation, vocal tract filtering, and lips/nostrils radiation. Results have been presented regarding the cancellation of the radiation effect, and the ability to model and implant any periodic wave shape using a parametric characterization which is independent of F0 and time-shift. These capabilities will be implemented in an analysis-by-synthesis procedure jointly estimating the magnitude and group delay characteristics of the source excitation and vocal tract filter.

References

[1] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 7, no. 5, pp. 569–586, September 1999.

[2] Matti Airas, “TKK aparat: Enviroment for voice inverse filtering and parameterization,” *Logopedics Phoniatrics*, vol. 33, no. 1, pp. 49–64, 2008.

[3] Paavo Alku, “An automatic method to estimate the time-based parameters of the glottal pulseform,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1992, pp. II–29–II–32.

[4] G. Fant, *Acoustic Theory of Speech Production*, The Hague, 1970.

[5] Hector R. Javkin, Norma A. Barroso, and Ian Maddieson, “Digital inverse filtering fo linguistic research,” *Journal of Speech and Hearing Research*, vol. 30, pp. 122–129, 1987.

[6] Jacqueline Walker and Peter Murphy, “A review of glottal waveform analysis,” *Lecture Notes in Computer Science - Progress in Nonlinear Speech Processing*, vol. 4391, pp. 1–21, 2007, Springer-Verlag.

[7] L. Rabiner and B-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1993.

[8] Laura Lehto, Matti Airas, Eva Bjokner, Johan Sundberg, and Paavo Alku, “Comparison of two inverse filtering methods in parametrization of the glottal closing characteristics in different phonation types,” *Journal of Voice*, vol. 21, no. 2, pp. 138–150, 2007.

[9] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, 1993.

[10] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck, *Discrete-time Signal Processing*, Prentice-Hall Inc., 1998, 2nd ed.

[11] Maurice Bellanger, *Digital Processing of Signals*, John Willey & Sons, 1989.

[12] Gunnar Fant, “Glottal flow: models and interaction,” *Journal of Phonetics*, vol. 14, no. 3/4, pp. 393–399, 1986.

[13] Aníbal J. S. Ferreira, “Combined spectral envelope normalization and subtraction of sinusoidal components in the ODFT and MDCT frequency domains,” in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 51–54.

[14] Thomas. F. Quatieri and Robert J. McAulay, “Shape invariant time-scale and pitch modification of speech,” *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, March 1992.

[15] Ricardo Sousa and Aníbal Ferreira, “Importance of the relative delay of glottal source harmonics,” in *39th AES International Conference on Audio Forensics - practices and challenges*, 2010, pp. 59–69.

[16] Ricardo Sousa and Aníbal Ferreira, “Singing voice analysis using relative harmonic delays,” in *INTERSPEECH 2011*, 2011, accepted.