Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 26 (2012) 20–34
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A comparative study of glottal source estimation techniques☆

Thomas Drugman[a,*], Baris Bozkurt[b], Thierry Dutoit[a]

[a] TCTS Lab, University of Mons, 31 Boulevard Dolez, 7000 Mons, Belgium
[b] Department of Electrical & Electronics Engineering, Izmir Institute of Technology, Gulbahce Koyu 35430, Urla, Izmir, Turkey

Received 25 March 2010; received in revised form 8 March 2011; accepted 10 March 2011
Available online 8 April 2011

Abstract

Source-tract decomposition (or glottal flow estimation) is one of the basic problems of speech processing. For this, several techniques have been proposed in the literature. However, studies comparing different approaches are almost nonexistent. Besides, experiments have been systematically performed either on synthetic speech or on sustained vowels. In this study we compare three of the main representative state-of-the-art methods of glottal flow estimation: closed-phase inverse filtering, iterative and adaptive inverse filtering, and mixed-phase decomposition. These techniques are first submitted to an objective assessment test on synthetic speech signals. Their sensitivity to various factors affecting the estimation quality, as well as their robustness to noise are studied. In a second experiment, their ability to label voice quality (tensed, modal, soft) is studied on a large corpus of real connected speech. It is shown that changes of voice quality are reflected by significant modifications in glottal feature distributions. Techniques based on the mixed-phase decomposition and on a closed-phase inverse filtering process turn out to give the best results on both clean synthetic and real speech signals. On the other hand, iterative and adaptive inverse filtering is recommended in noisy environments for its high robustness.

Keywords: Source-tract separation; Glottal flow estimation; Inverse filtering; Mixed-phase decomposition; Voice quality

1. Introduction

Speech results from filtering the glottal flow by the vocal tract cavities, and converting the resulting velocity flow into pressure at the lips (Quatieri, 2002). In many speech processing applications, it is important to separate the contributions from the glottis and the vocal tract. Achieving such a *source-filter deconvolution* could lead to a distinct characterization and modeling of these two components, as well as to a better understanding of the human phonation process. Such a decomposition is thus a preliminary condition for the study of glottal-based vocal effects, which can be segmental (as for vocal fry), or be controlled by speakers on a separate, supra-segmental layer (as it is the case for the voice quality modifications mentioned in Section 5). Their dynamics is very different from that of the vocal tract contribution, and requires further investigation. Glottal source estimation is then a fundamental problem in speech processing, finding applications in speech synthesis (Cabral et al., 2008), voice pathology detection (Drugman et al., 2009b), speaker recognition (Plumpe et al., 1999), emotion analysis/synthesis (Airas and Alku, 2006), etc.

☆ This paper has been recommended for acceptance by Bernd Moebius.
* Corresponding author. Tel.: +32 65374749.
E-mail address: thomas.drugman@umons.ac.be (T. Drugman).

0885-2308/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2011.03.003

In this paper, we limit our scope to the methods which perform an estimation of the glottal source contribution directly from the speech waveform. Although some devices such as electroglottographs or laryngographs, which measure the impedance between the vocal folds (but not the glottal flow itself), are informative about the glottal behaviour (Henrich et al., 2004), in most cases the use of such apparatus is inconvenient and only the speech signal is available for analysis. This problem is then a typical case of blind separation, since neither the vocal tract nor the glottal contribution are observable. This also implies that no quantitative assessment of the performance of glottal source estimation techniques is possible on natural speech, as no target reference signal is available.

As one of the basic problems and challenges of speech processing research, glottal flow estimation has been studied by many researchers and various techniques are available in the literature (Walker and Murphy, 2007). However, the diversity of algorithms and the fact that the reference for the actual glottal flow is not available often leads to the questionability about relative effectiveness of the methods in real life applications. In most of studies, tests are performed either on synthetic speech or on a few recorded sustained vowels. In addition, very few comparative studies exist (such as (Sturmel et al., 2007)). In this paper, we compare three of the main representative state-of-the-art methods: closed-phase inverse filtering, iterative and adaptive inverse filtering, and mixed-phase decomposition. For testing, we first follow the common approach of using a large set of synthetic speech signals (by varying synthesis parameters independently), and then we examine how these techniques perform on a large real speech corpus. In the synthetic speech tests, the original glottal flow is available, so that objective measures of decomposition quality can be computed. In real speech tests the ability of the methods to discriminate different voice qualities (tensed, modal and soft) is studied on a large database (without limiting data to sustained vowels).

The paper is structured as follows. In Section 2 the main state-of-the-art methods for glottal source estimation are reviewed, and the three techniques compared in this study are detailed. Section 3 discusses how the resulting glottal signal can be parameterized both in time and frequency domains. The three methods are evaluated in Section 4 through a wide systematic study on synthetic signals. Their robustness to noise, as well as the impact of the various factors that may affect source-tract separation, are investigated. Section 5 presents decomposition results on a real speech database containing various voice qualities, and shows that the glottal source estimated by the techniques considered in this work conveys relevant information about the phonation type. Finally Section 6 draws the conclusions of this study.

2. Glottal source estimation

Glottal flow estimation mainly refers to the estimation of the voiced excitation of the vocal tract. During the production of voiced sounds, the airflow arising from the trachea causes a quasi-periodic vibration of the vocal folds (Quatieri, 2002), organized into so-called opening/closure cycles. During the *open phase*, vocal folds are progressively displaced from their initial state due to the increasing subglottal pressure. When the elastic displacement limit is reached, they suddenly return to this position during the so-called *return phase*. Fig. 1 displays the typical shape of one cycle of the glottal flow (Fig. 1(a)) and its time derivative (Fig. 1(b)) according to the Liljencrants–Fant (LF) model (Fant et al., 1985). It is often preferred to gather the lip radiation effect (whose action is close to a differentiation operator) with the glottal component, and work in this way with the glottal flow derivative on the one hand, and with the vocal tract contribution on the other hand. It is seen in Fig. 1 (bottom plot) that the boundary between open and return phases corresponds to a particular event called the glottal closure instant (GCI). GCIs refer to the instances of significant excitation of the vocal tract (Drugman and Dutoit, 2009). Being able to determine their location is of particular importance in so-called pitch-synchronous speech processing techniques, and in particular for a more accurate separation between vocal tract and glottal contributions.

The main techniques for estimating the glottal source directly from the speech waveform are now reviewed. Relying on the speech signal alone, as it is generally the case in real applications, allows to avoid the use of intrusive (e.g video camera at the vocal folds) or inconvenient (e.g. laryngograph) device.

Such techniques can be separated into two classes, according to the way they perform the source-filter separation. The first category (Section 2.1) is based on inverse filtering, while the second one (Section 2.2) relies on the mixed-phase properties of speech.

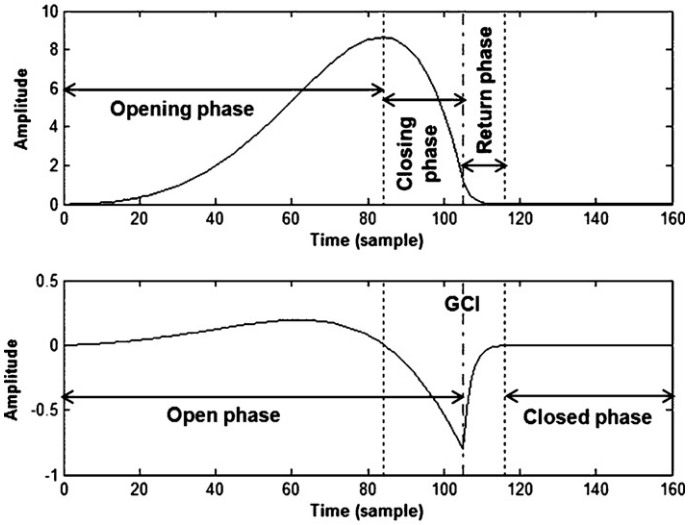

Fig. 1. Typical waveforms, according to the Liljencrants–Fant (LF) model, of one cycle of: (top) the glottal flow, (bottom) the glottal flow derivative. The various phases of the glottal cycle, as well as the glottal closure instant (GCI) are also indicated.

2.1. Methods based on inverse filtering

Most glottal source estimation techniques are based on an inverse filtering process. These methods first estimate a parametric model of the vocal tract, and then obtain the glottal flow by removing the vocal tract contribution via inverse filtering. The methods in this category differ by the way the vocal tract is estimated. In Section 2.1.1 this estimation is computed during the glottal closed phase, while in Section 2.1.2 an iterative and/or adaptive procedure is used. A more extended review of the inverse filtering-based process for glottal waveform analysis can be found in Walker and Murphy (2007).

2.1.1. Closed phase inverse filtering

Closed phase refers to the timespan during which the glottis is closed (see Fig. 1). During this period, the effects of the subglottal cavities are minimized, providing a better way for estimating the vocal tract transfer function. Therefore, methods based on a closed phase inverse filtering (CPIF) estimate a parametric model of the spectral envelope, computed during the estimated closed phase duration (Wong et al., 1979). The main drawback of these techniques lies in the difficulty in obtaining an accurate determination of the closed phase. Several approaches have been proposed in the literature to solve this problem. In Veeneman and Bement (1985), authors use information from the electroglottographic signal (which is avoided in this study) to identify the period during which the glottis is closed. In Plumpe et al. (1999), it was proposed to determine the closed phase by analyzing the formant frequency modulation between open and closed phases. In Alku et al. (2009), the robustness of CPIF to the frame position was improved by imposing some dc gain constraints. Besides this problem of accurate determination of the closed phase, it may happen that this period is so short (for high-pitched voices) that not enough samples are available for a reliable filter estimation. It was therefore proposed in Brookes and Chan (1994) a technique of multicycle closed-phase LPC, where a small number of neighbouring glottal cycles are considered in order to have enough data for an accurate vocal tract estimation. Finally note that an approach allowing non-zero glottal wave to exist over closed glottal phases was proposed in Deng et al. (2006).

In this study, the CPIF-based technique that is used is based on a discrete all pole (DAP, (Jaroudi and Makhoul, 1991)) inverse filtering process estimated during the closed phase. In order to provide a better fitting of spectral envelopes from discrete spectra (Jaroudi and Makhoul, 1991), the DAP technique aims at computing the parameters of an autoregressive model by minimizing a discrete version of the Itakura–Saito distance (Itakura, 1970), instead of the time squared error used by the traditional LPC. The use of the Itakura–Saito distance is justified as it is a spectral distortion measure arising from the human hearing perception. The closed phase period is determined using the Glottal Opening and Closure Instants (GCIs and GOIs) located by the algorithm detailed in Drugman and Dutoit (2009). This algorithm has been shown to be effective for reliably and accurately determining the position of these events on a

large corpus containing several speakers. For tests with synthetic speech, the exact closed phase period is known and is used for CPIF. Note that for high-pitched voices, two analysis windows were used as suggested in Brookes and Chan (1994); Yegnanarayana and Veldhuis (1998) and Plumpe et al. (1999). In the rest of the paper, speech signals sampled at 16 kHz are considered, and the order for DAP analysis is fixed to 18 ($=F_s/1000+2$, as commonly used in the literature). Through our experiments, we reported that the choice of the DAP order is not critical in the usual range, and that working with an order comprised between 12 and 18 leads to sensibly similar results.

2.1.2. Iterative and/or adaptive inverse filtering

Some methods are based on iterative and/or adaptive procedures in order to improve the quality of the glottal flow estimation. Fu and Murphy (2006) proposed to integrate, within the autoregressive exogenous (ARX) model of speech production, the LF model of the glottal source. The resulting ARXLF model is estimated via an adaptive and iterative optimization (Vincent et al., 2005). Both source and filter parameters are consequently jointly estimated. The method proposed by Moore and Clements (2004) iteratively finds the best candidate for a glottal waveform estimate within a speech frame, without requiring a precise location of the GCIs. Finally a popular approach was proposed by Alku et al. (1992) and called iterative adaptive inverse filtering (IAIF). This method is based on an iterative refinement of both the vocal tract and the glottal components. In Alku and Vilkman (1994), the same authors proposed an improvement, in which the LPC analysis is replaced by the discrete all pole modeling technique (Jaroudi and Makhoul, 1991), shown to be more accurate for high-pitched voices.

As a representative technique of this category, the IAIF method proposed by Alku et al. (1992) is considered in the rest of this paper. More precisely, we used the implementation of the IAIF method (Airas, 2008) from the toolbox available on the TKK Aparat website (Online, 2008), with its default options.

2.2. Mixed-phase decomposition

The methods presented in this section rely on the mixed-phase model of speech (Bozkurt and Dutoit, 2003). According to this model, speech is composed of both minimum-phase (i.e causal) and maximum-phase (i.e. anticausal) components. While the vocal tract impulse response and the glottal *return phase* of the glottal component can be considered as minimum-phase signals, it has been shown in Doval et al. (2003) that the glottal *open phase* of the glottal flow is a maximum-phase signal. Besides it has been shown in Gardner and Rao (1997) that mixed-phase models are appropriate for modeling voiced speech due to the maximum-phase nature of the glottal excitation. They showed that the use of an anticausal all-pole filter for the glottal pulse is necessary to resolve magnitude and phase information correctly. The key idea of mixed-phase decomposition methods is then to separate minimum from maximum-phase components of speech, where the latter is only due to the glottal contribution.

A crucial issue in mixed-phase separation is the weighting window that is applied to the speech signal for short-term analysis. Indeed, since the decomposition is based on phase properties, windowing may have a dramatic influence. It has been shown that GCI-synchronization, as well as the respect of some constraints on the window length and function, are essential for guaranteeing a correct decomposition (Drugman et al., 2009a; Bozkurt et al., 2005). Throughout the rest of this study, we use an appropriate GCI-centered two pitch period-long Blackman window satisfying these conditions.

In previous works, we proposed two approaches achieving such a decomposition: a technique based on the Zeros of the Z-transform (ZZT, (Bozkurt et al., 2005)), and one based on the complex cepstrum decomposition (CCD, (Drugman et al., 2009a; Quatieri, 2002)). Both techniques are briefly presented in Sections 2.2.1 and 2.2.2 and depicted in Fig. 2. Finally, the methods are shown to be functionnaly equivalent in Section 2.2.3.

2.2.1. Zeros of the Z-transform

For a series of N samples $(x(0), x(1), ..., x(N-1))$ taken from a discrete signal $x(n)$, the *ZZT* representation is defined as the set of roots (zeros) $(Z_1, Z_2, ...Z_{N-1})$ of the corresponding Z-transform $X(z)$:

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} \tag{1}$$

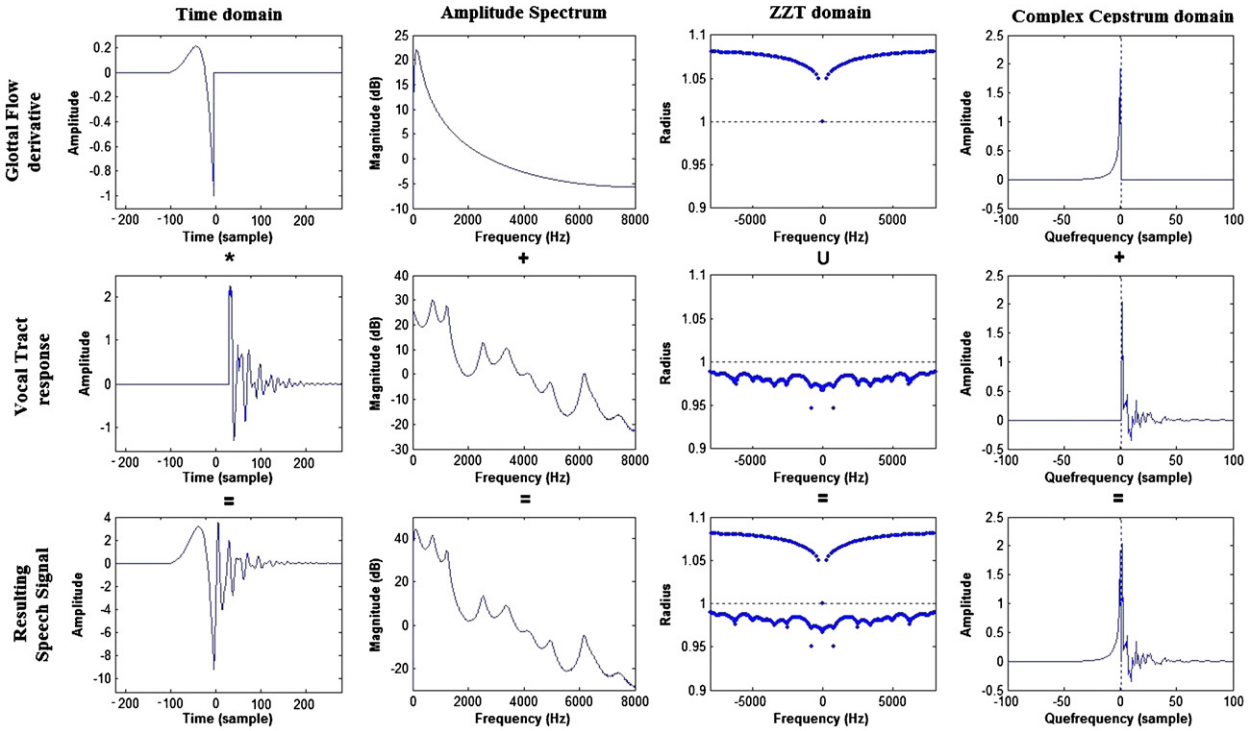

Fig. 2. Illustration of mixed-phase decomposition. Rows respectively exhibit the following signals: the glottal flow derivative (top), the vocal tract response (middle), and the resulting speech signal (bottom). Each column corresponds to a domain of representation of these signals: time domain (first column), amplitude spectrum (second column), ZZT representation in polar coordinates (third column), and complex cepstrum domain (fourth column). Interestingly, convolution in the time domain corresponds to the union operator in the ZZT domain and to the addition operator in the complex cepstrum domain. The ZZT and CC domains are suited for achieving mixed-phase decomposition since minimum and maximum-phase components become linearly separable. In the ZZT domain, the unit circle is used as a discriminant boundary, while the quefrency origin is used as a boundary in the complex cepstrum domain.

$$= x(0)z^{-N+1}\prod_{m=1}^{N-1}(z - Z_m) \tag{2}$$

$$= x(0)z^{-N+1}\prod_{k=1}^{M_o}(z - Z_{\max,k})\prod_{k=1}^{M_i}(z - Z_{\min,k}) \tag{3}$$

To achieve the ZZT-based decomposition of speech, speech frames are first weighted by a specific window (see above). When computing the ZZT of this signal as in Eq. (3), some roots $Z_{\max,k}$ fall outside the unit circle. These are due to the maximum-phase (i.e anticausal) component of speech, and are consequently only related to the glottal open phase. On the opposite, roots located inside the unit circle $Z_{\min,k}$ are due to the minimum-phase component of speech, i.e mainly to the vocal tract impulse response. Mixed-phase decomposition can then be easily achieved in the ZZT domain, using the unit circle as a discriminant boundary (see Fig. 2, third column).

2.2.2. *Complex cepstrum decomposition*

The complex cepstrum (CC) $\hat{x}(n)$ of a discrete signal $x(n)$ is defined by the following equations (Oppenheim and Schafer, 1989):

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \tag{4}$$

$$\log[X(\omega)] = \log(|X(\omega)|) + j\angle X(\omega) \tag{5}$$

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[X(\omega)]e^{j\omega n}d\omega \tag{6}$$

where Eqs. (4)–(6), respectively correspond to a discrete-time Fourier transform (DTFT), a complex logarithm and an inverse DTFT (IDTFT). Decomposition in the CC domain arises from the fact that the complex cepstrum $\hat{x}(n)$ of an anticausal (causal) signal is zero for all n positive (negative). Retaining only the negative indexes of the CC makes then it possible to estimate the glottal contribution. The separation in the complex cepstrum domain using the quefrency origin as a discriminant boundary is clearly seen in Fig. 2, fourth column.

2.2.3. Equivalence between ZZT and CCD

If $X(z)$ is written as in Eq. 3, it can be easily shown that the corresponding complex cepstrum can be expressed as (Oppenheim and Schafer, 1989):

$$\hat{x}(n) = \begin{cases} |x(0)| \, for \, n = 0 \\ \displaystyle\sum_{k=1}^{M_o} \frac{{Z_{\max,k}}^n}{n} \, for \, n < 0 \\ \displaystyle\sum_{k=1}^{M_i} \frac{{Z_{\min,k}}^n}{n} \, for \, n > 0 \end{cases} \tag{7}$$

This equation shows the narrow link between the ZZT and the CCD techniques. These two methods can then be seen as two different ways of performing the same operation: separate the minimum and maximum-phase components from a given Z-transform $X(z)$. Nevertheless, although functionally equivalent, it has been shown (Pedersen et al., 2010; Drugman et al., 2009a) that CCD performs much faster (speed is increased between 10 and 100 times for a sampling rate of 16 kHz, depending on the pitch period). This may be explained by the fact that it only relies on FFT and IFFT operations while ZZT requires the factoring of high-order polynomials.

As a method achieving mixed-phase separation, the CCD is considered in the rest of this paper for its higher computational speed. To guarantee good mixed-phase properties (Drugman et al., 2009a), GCI-centered two pitch period-long Blackman windows are used. For this, GCIs were located on real speech using the technique we proposed in Drugman and Dutoit (2009). CC is calculated as explained in Section 2.2.2 and FFT is computed on a sufficiently large number of points (typically 4096), which facilitates phase unwrapping.

3. Glottal source parametrization

Once the glottal signal has been estimated by any of the aforementioned algorithms, it is interesting to derive a parametric representation of it, using a small number of parameters. Various approaches, both in the time and frequency domains, have been proposed to characterize the human voice source. This section gives a brief overview of the most commonly used parameters in the literature, since some of them are used in Sections 4 and 5.

3.1. Time-domain features

Several time-domain features can be expressed as a function of time intervals derived from the glottal waveform (Alku, 1992). These are used to characterize the shape of the waveform, by capturing for example the location of the primary or secondary opening instant (Laukkanen et al., 1996), of the glottal flow maximum, etc. The formulation of the source signal in the commonly used LF model (Fant et al., 1985) is based on time-domain parameters, such as the Open Quotient O_q, the asymmetry coefficient α_m, or the voice speed quotient S_q (Doval and d'Alessandro, 2006). However, in most cases these instants are difficult to locate with precision from the glottal flow estimation. Avoiding this problem and prefered to the traditional open quotient, the quasi-open quotient (QOQ) was proposed as a parameter describing the relative open time of the glottis (Hacki, 1989). It is defined as the ratio between the quasi-open time and the quasi-closed time of the glottis, and corresponds to the timespan (normalized to the pitch period) during which the glottal flow is above 50% of the difference between the maximum and minimum flow. Note that QOQ was used in Laukkanen et al. (1996) for studying the physical variations of the glottal source related to the vocal expression of

Table 1
Table of synthesis parameter variation range.

Source			Filter	Noise
Pitch (Hz)	*Oq*	α_m	Vowel type	SNR (dB)
100:5:240	0.3:0.05:0.9	0.55:0.05:0.8	14 vowels	10:10:80

stress and emotion. In Airas and Alku (2007) various variants of *Oq* have been tested in terms of the degree by which they reflect phonation changes. QOQ was found to be the best for this task.

Another set of parameters is extracted from the amplitude of peaks in the glottal pulse or its derivative (Gobl and Chasaide, 2003). The normalized amplitude quotient (NAQ) proposed by Alku et al. (2002) turns out to be an essential glottal feature. NAQ is a parameter characterizing the glottal closing phase (Alku et al., 2002). It is defined as the ratio between the maximum of the glottal flow and the minimum of its derivative, normalized with respect to the fundamental period. Its robustness and efficiency to separate different types of phonation was shown in Alku et al. (2002) and Airas and Alku (2007). Note that a quasi-similar feature, called *basic shape parameter*, was proposed by Fant (1995), where it was qualified as *"most effective single measure for describing voice qualities"*.

In Plumpe et al. (1999), authors propose to use 7 LF parameters and 5 energy coefficients (defined in 5 subsegments of the glottal cycle) respectively for characterizing the coarse and fine structures of the glottal flow estimation. Finally some approaches aim at fitting a model on the glottal flow estimate by computing a distance in the time domain (Plumpe et al., 1999; Drugman et al., 2008).

3.2. Frequency-domain features

In the frequency domain, the LF model presents a low-frequency resonance called the *glottal formant* (Doval and d'Alessandro, 2006) (see the amplitude spectrum of the glottal flow derivative in Fig. 2, row 1, column 2). Some approaches characterize the glottal formant both in terms of frequency and bandwidth (Drugman et al., 2009a). By defining a spectral error measure, other studies try to match a model to the glottal flow estimation (Ling et al., 2005; Fant, 1995; Drugman et al., 2008). This is also the case for the parabolic spectrum parameter (PSP) proposed in Alku et al. (1997).

An extensively used measure is the H1–H2 parameter (Fant, 1995). This parameter is defined as the ratio between the amplitudes of the magnitude spectrum of the glottal source at the fundamental frequency and at the second harmonic (Klatt and Klatt, 1990; Titze and Sundberg, 1992). It has been widely used as a measure characterizing voice quality (Hanson, 1995; Fant, 1995; Alku et al., 2009).

For quantifying the amount of harmonics in the glottal source, the harmonic to noise ratio (HNR) and the harmonic richness factor (HRF) have been proposed in Murphy and Akande (2005) and Childers and Lee (1991). More precisely, HRF quantifies the amount of harmonics in the magnitude spectrum of the glottal source. It is defined as the ratio between the sum of the amplitudes of harmonics, and the amplitude at the fundamental frequency (Childers, 1999). It was shown to be informative about the phonation type in Childers and Lee (1991) and Alku et al. (2009).

4. Experiments on synthetic speech

The first experimental protocol we opted for is close to the one presented in Sturmel et al. (2007). Decomposition is achieved on synthetic speech signals (sampled at 16 kHz) for various test conditions. The idea is to cover the diversity of configurations one can find in continuous speech by varying all parameters over their whole range. Synthetic speech is produced according to the source-filter model by passing a known sequence of Liljencrants–Fant glottal waveforms (Fant et al., 1985) through an auto-regressive filter extracted by LPC analysis (with an order of 18) from real sustained vowels uttered by a female speaker. As the mean pitch during these utterances was about 180 Hz, it can be considered that fundamental frequency should not exceed 100 and 240 Hz in continuous speech. For the LF parameters, the open quotient *Oq* and asymmetry coefficient α_m are varied through their common range (see Table 1). For the filter, 14 types of typical vowels are considered. Noisy conditions are modeled by adding a white Gaussian noise to the speech signal, from almost clean conditions (SNR = 80 dB) to strongly adverse environments (SNR = 10 dB). Table 1 summarizes all

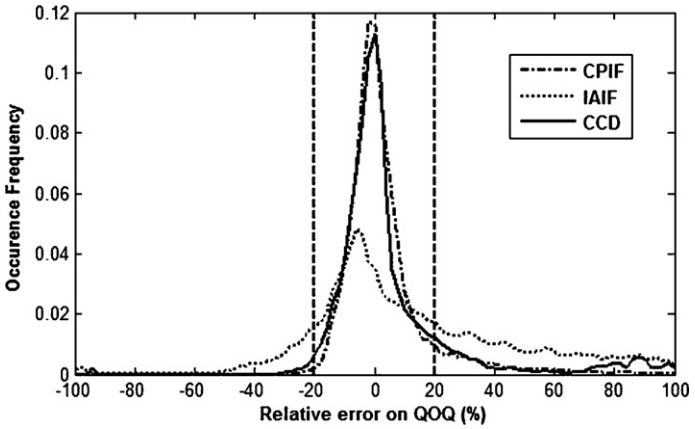

Fig. 3. Distribution of the relative error on *QOQ* for the three methods in clean conditions (SNR = 80 dB). The *error rate* is defined as the percentage of frames for which the relative error is higher than a given threshold of 20% (indicated on the plot).

test conditions, which makes a total of slightly more than 250,000 experiments. It is worth mentioning that the synthetic tests presented in this section focus on the study of non-pathological voices with a regular phonation. Although the glottal analysis of less regular voices (e.g. presenting a jitter or a shimmer; or containing an additive noise component during the glottal production, as it is the case for a breathy voice) is a challenging issue, this latter problem is not addressed in the present study.

The three source estimation techniques described in Section 2 (CPIF, IAIF and CCD) are compared. In order to assess their decomposition quality, two objective quantitative measures are used (and the effect of noise, fundamental frequency and vocal tract variations to these measures are studied in detail in the next subsections):

- *Error rate on NAQ and QOQ*: An error on the estimation of NAQ and QOQ after source-tract decomposition should be penalized. An example of distribution for the relative error on QOQ in clean conditions is displayed in Fig. 3. Many attributes characterizing such a histogram can be proposed to evaluate the performance of an algorithm. The one we used in our experiments is defined as the proportion of frames for which the relative error is higher than a given threshold of $\pm 20\%$. The lower the error rate on the estimation of a given glottal parameter, the better the glottal flow estimation method.
- *Spectral distortion*: Many frequency-domain measures for quantifying the distance between two speech frames x and y arise from the speech coding literature. Ideally the subjective ear sensitivity should be formalised by incorporating psychoacoustic effects such as masking or isosonic curves. A simple and relevant measure is the spectral distortion (SD) defined as (Nordin and Eriksson, 2001):

$$SD(x, y) = \sqrt{\int_{-\pi}^{\pi} (20\log_{10}|\frac{X(\omega)}{Y(\omega)}|)^2 \frac{d\omega}{2\pi}} \tag{8}$$

where $X(\omega)$ and $Y(\omega)$ denote both signals spectra as a function of normalized angular frequency. In Paliwal (1993), authors argue that a difference of about 1 dB (with a sampling rate of 8 kHz) is hardly perceptible. In order to take this point into account, we used the following measure between the spectra of the estimated and reference glottal signals:

$$\text{SD(Estimated,Reference)} \approx \sqrt{\frac{2}{8000}\int_{20}^{4000} (20\log_{10}|\frac{S_{\text{Estimated}}(f)}{S_{\text{Reference}}(f)}|)^2 \,\text{df}} \tag{9}$$

An efficient technique of glottal flow estimation is then reflected by low spectral distortion values.

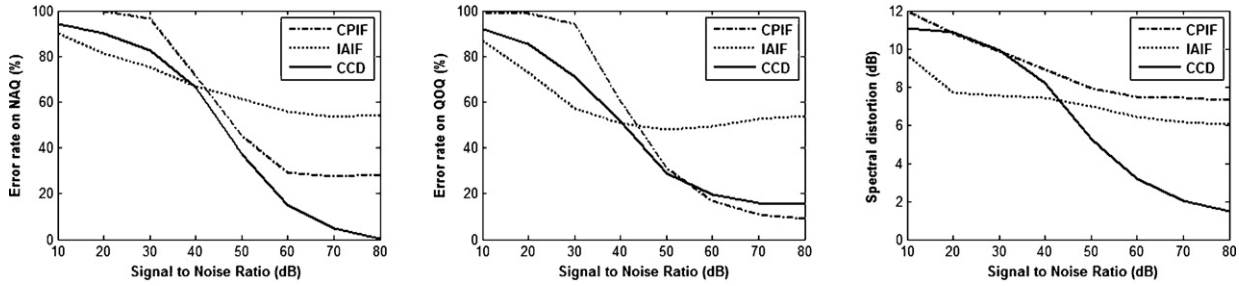

Fig. 4. Evolution of the three performance measures (error rate on *NAQ* and *QOQ*, and spectral distortion) as a function of the signal to noise ratio for the three glottal source estimation methods.

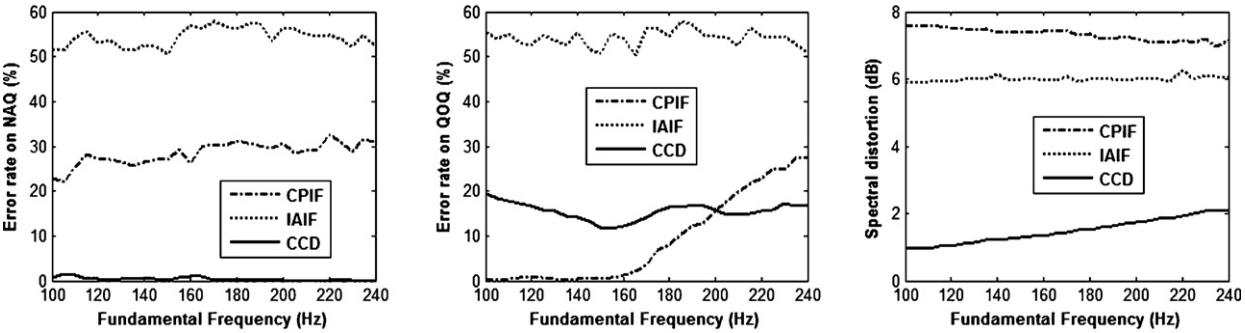

Fig. 5. Evolution of the three performance measures as a function of the fundamental frequency for the three glottal source estimation methods.

Based on this experimental framework, we now study how the glottal source estimation techniques behave in noisy conditions, or with regard to some factors affecting the decomposition quality, such as the fundamental frequency or the vocal tract transfer function.

4.1. Robustness to additive noise

As mentioned above, white Gaussian noise has been added to the speech signal, with various SNR levels. This noise is used as a (weak) substitute for recording or production noise but also for every little deviation to the theoretical framework which distinguishes real and synthetic speech. Results according to our three performance measures are displayed in Fig. 4. As expected, all techniques degrade as the noise power increases. More precisely, CCD turns out to be particularly sensitive. This can be explained by the fact that a weak presence of noise may dramatically affect the phase information, and consequently the decomposition quality. The performance of CPIF is also observed to strongly degrade as the noise level increases. This is probably due to the fact that noise may dramatically modify the spectral envelope estimated during the closed phase, and the resulting estimate of the vocal tract contribution becomes erroneous. On the contrary, even though IAIF is, in average, the less efficient on clean synthetic speech, it outperforms other techniques in adverse conditions (below 40 dB of SNR). One possible explanation of its robustness is the iterative process it relies on. It can be indeed expected that, although the first iteration may be highly affected by noise (as it is the case for CPIF), the severity of the perturbation becomes weaker as the iterative procedure converges.

4.2. Sensitivity to fundamental frequency

Female voices are known to be especially difficult to analyze and synthesize. The main reason for this is their high fundamental frequency which implies to process shorter glottal cycles. As a matter of fact the vocal tract response has not the time to freely return to its initial state between two glottal sollication periods (i.e. the duration of the vocal tract response can be much longer than that of the glottal closed phase). Fig. 5 shows the evolution of our three performance measures with respect to the fundamental frequency in clean conditions. Interestingly, all methods maintain almost the same efficiency for high-pitched voices. Nonetheless an increase of the error rate on QOQ for CPIF, and an increase

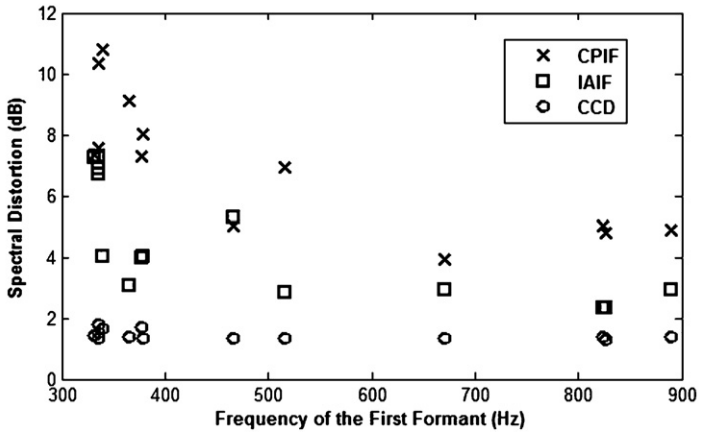

Fig. 6. Evolution, for the 14 vowels, of the spectral distortion with the first formant frequency F_1.

of the spectral distortion for CCD can be noticed. It can be also observed that, for clean synthetic speech, CCD gives the best results with an excellent determination of NAQ and a very low spectral distortion. Secondly, despite its high spectral errors, CPIF leads to an efficient parametrization of the glottal shape (with notably the best results for the determination of QOQ).

4.3. Sensitivity to vocal tract

In our experiments, filter coefficients were extracted by LPC analysis on sustained vowels. Even though the whole vocal tract spectrum may affect the decomposition, the first formant, which corresponds to the dominant poles, generally imposes the longest contribution of its time response. To give an idea of its impact, Fig. 6 exhibits, for the 14 vowels, the evolution of the spectral distortion as a function of the first formant frequency F_1. A general trend can be noticed from this graph: it is observed for all methods that the performance of the glottal flow estimation degrades as F_1 decreases. This will be explained in the next section by an increasing overlap between source and filter components, as the vocal tract impulse response gets longer. It is also noticed that this degradation is particularly important for both CPIF and IAIF methods, while the quality of CCD (which does not rely on a parametric modeling) is only slightly altered.

4.4. Conclusions on synthetic speech

Many factors may affect the quality of the source-tract separation. Intuitively, one can think about the *time interference* between minimum and maximum-phase contributions, respectively related to the vocal tract and to the glottal open phase. The stronger this interference, the more important the time overlap between the minimum-phase component and the maximum-phase response of the next glottal cycle, and consequently the more difficult the decomposition. Basically, this interference is conditioned by three main parameters:

- the pitch F_0, which imposes the spacing between two successive vocal system responses,
- the first formant F_1, which influences the length of the minimum-phase contribution of speech,
- and the glottal formant F_g, which controls the length of the maximum-phase contribution of speech. Indeed, the glottal formant is the most important spectral feature of the glottal open phase (see the low-frequency resonance in the amplitude spectrum of the glottal flow derivative in Fig. 2). It is worth noting that F_g is known (Doval and d'Alessandro, 2006) to be a function of the time-domain characteristics of the glottal open phase (i.e of the maximum-phase component of speech): the open quotient O_q, and the asymmetry coefficient (α_m).

A strong interference then appears with high pitch, and with low F_1 and F_g values. The previous experiments confirmed for all glottal source estimation techniques the performance degradation as a function of F_0 and F_1. Although

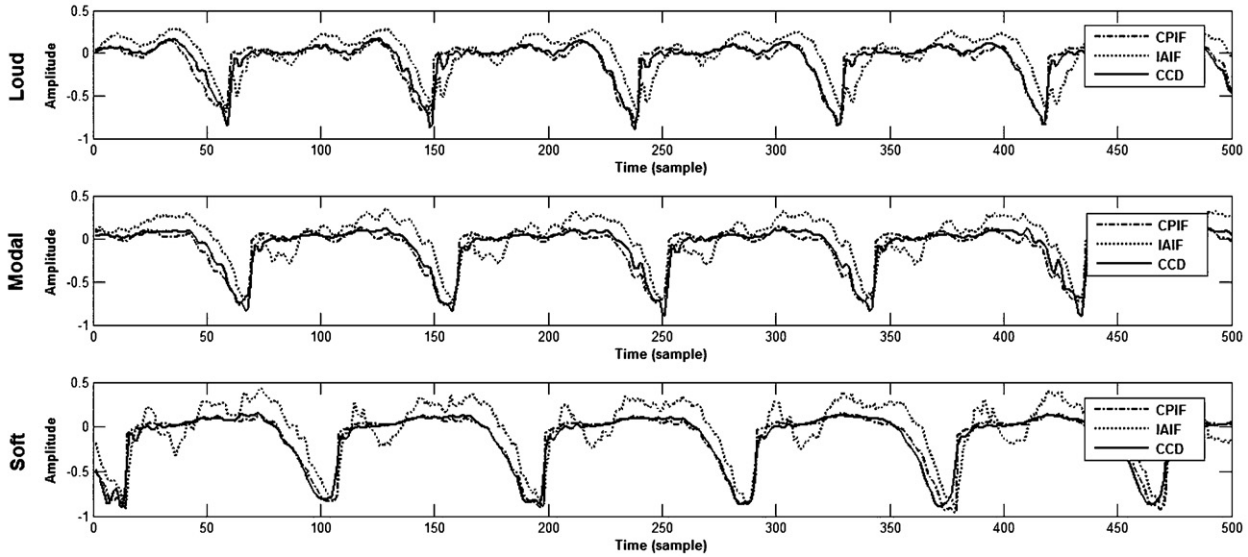

Fig. 7. Example of glottal flow derivative estimation on a given segment of vowel (*/aI/* as in "*ice*") for the three techniques and for the three voice qualities: (*top*) loud voice, (*middle*) modal voice, (*bottom*) soft voice.

we did not explicitly measure the sensitivity of these techniques to F_g in this manuscript, it was confirmed in other informal experiments we performed.

It can be also observed from Figs. 4 and 5 that the overall performance through an objective study on synthetic signals is the highest for the complex cepstrum-based technique. This method leads to the lowest values of spectral distortion and gives relatively high rates for the determination of both NAQ and QOQ parameters. The CPIF technique exhibits better performance in the determination of QOQ in clean conditions and especially for low-pitched speech. As for the IAIF technique, it turns out that it gives the worst results in clean synthetic speech but outperforms other approaches in adverse noisy conditions. Note that our results corroborate the conclusions drawn in Sturmel et al. (2007) where the mixed-phase deconvolution (achieved in that study by the ZZT method) was shown to outperform other state-of-the-art approaches of glottal flow estimation.

5. Experiments on real speech

Reviewing the glottal flow estimation literature, one can easily notice that testing with natural speech is a real challenge. Even in very recent published works, all tests are performed only on sustained vowels. In addition, due to the unavailability of a reference for the real glottal flow (see Section 1), the procedure of evaluation is generally limited to providing plots of glottal flow estimates, and checking visually if they are consistent with expected glottal flow models. For real speech experiments, here we will first follow this state-of-the-art experimentation (of presenting plots of estimates for a real speech example), and then extend it considerably both by extending the content of the data to a large connected speech database (including non-vowels), and extending the method to a comparative parametric analysis approach.

In this study, experiments on real speech are carried out on the De7 corpus, a diphone database designed for expressive speech synthesis (Schroeder and Grice, 2003). The database contains three voice qualities (modal, soft and loud) uttered by a German female speaker, with about 50 min of speech available for each voice quality (leading to a total of around 2h30). Recordings sampled at 16 kHz are considered. Locations of both GCIs and GOIs are precisely determined from these signals using the algorithm described in Drugman and Dutoit (2009). As mentioned in Section 2, an accurate position of both events is required for an efficient CPIF technique, while the mixed-phase decomposition (as achieved by CCD) requires, among others, GCI-centered windows to exhibit correct phase properties.

Let us first consider in Fig. 7 a concrete example of glottal source estimation on a given voiced segment (*/aI/* as in "*ice*") for the three techniques and for the three voice qualities. In the IAIF estimate, some ripples are observed as

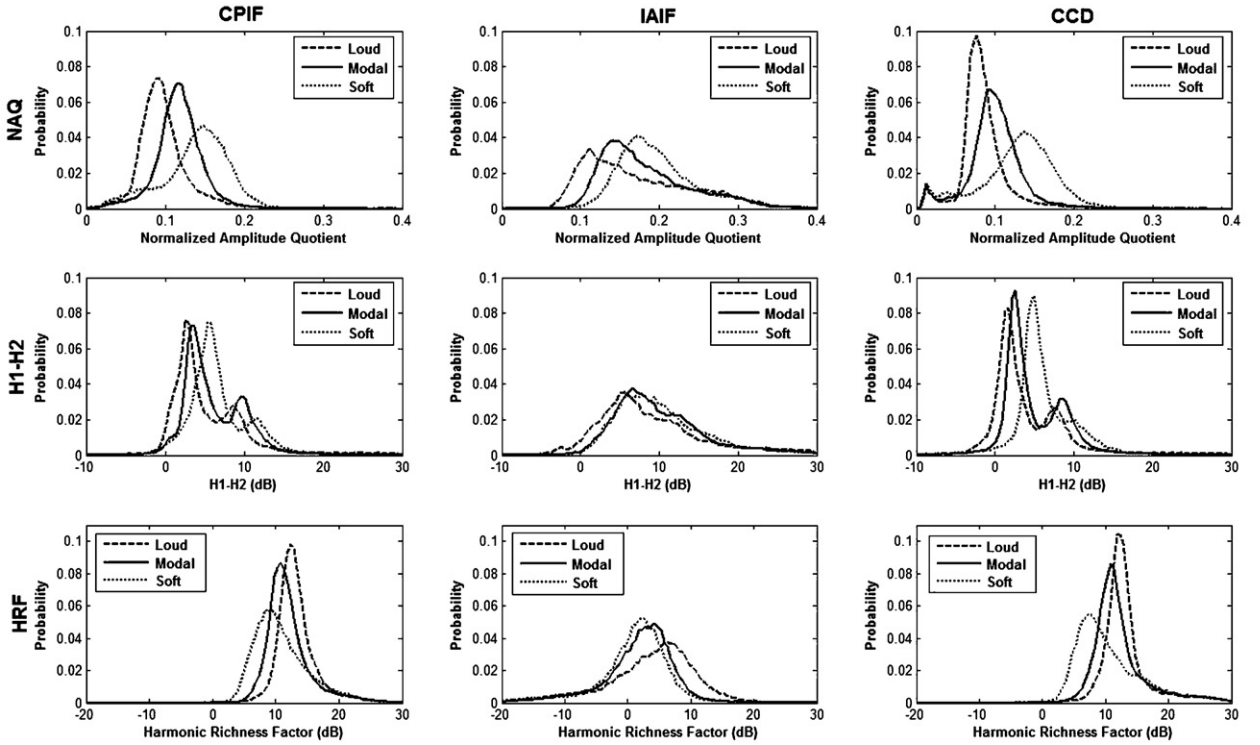

Fig. 8. Distributions, for various voice qualities, of three glottal features (from top to bottom: NAQ, H1–H2 and HRF) estimated by three glottal source estimation techniques (from left to right: CPIF, IAIF and CCD). The voice qualities are shown as dashed (loud voice), solid (modal voice) and dotted (soft voice) lines.

if some part of the vocal tract filter contribution could not be removed. On the other hand, it can be noticed that the estimations from CPIF and CCD are highly similar and are very close to the shape expected by the glottal flow models, such as the LF model (Fant et al., 1985). It can be also observed that the abruptness of the glottal open phase around the GCI is stronger for the loud voice, while the excitation for the softer voice is smoother.

We investigated whether the glottal source estimated by these techniques conveys information about voice quality. Indeed the glottis is assumed to play an important part for the production of such expressive speech (d'Alessandro, 2006). As a matter of fact we found some differences between the glottal features in our experiments on the De7 database. In this experiment, the NAQ, H1–H2 and HRF parameters described in Section 3 are used. Fig. 8 illustrates the distributions of these features estimated by CPIF, IAIF and CCD for the three voice qualities. This Figure can be considered as a summary of the voice quality analysis using three state-of-the-art methods on a large speech database. The parameters NAQ, H1–H2 and HRF have been used frequently in the literature to label phonation types (Alku et al., 2002; Hanson, 1995; Childers and Lee, 1991). Hence the separability of the phonation types based on these parameters can be considered as a measure of effectiveness for a particular glottal flow estimation method.

For the three methods, significant differences between the histograms of the different phonation types can be noted. This supports the claim that, by applying one of the given glottal flow estimation methods and by parameterizing the estimate with one or more of the given parameters, one can perform automatic voice quality/phonation type labeling with a much higher success rate than by random labeling. It is noticed from Fig. 8 that parameter distributions are convincingly distinct, except for the IAIF and H1–H2 combination. The sorting of the distributions with respect to vocal effort are consistent and in line with results of other works ((Alku et al., 2002) and (Alku et al., 2009)). Among other things, strong similarities between histograms obtained by CPIF and CCD can be observed. In all cases, it turns out that the stronger the vocal effort, the lower NAQ and H1–H2, and the higher HRF.

Although some significant differences in glottal feature distributions have been visually observed, it is interesting to quantify the discrimination between the voice qualities enabled by these features. For this, the Kullback–Leibler

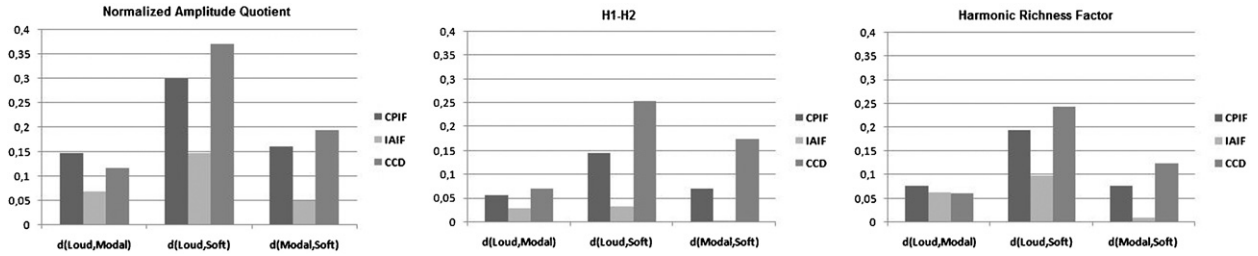

Fig. 9. Jensen-Shannon distances between two types of voice quality using (from left to right) the NAQ, H1–H2 and HRF parameters. For each feature and pair of phonation types, the three techniques of glottal source estimation are compared.

(KL) divergence, known to measure the separability between two discrete density functions A and B, can be used (Lin, 1991):

$$D_{KL}(A, B) = \sum_i A(i) \log_2 \frac{A(i)}{B(i)} \tag{10}$$

Since this measure is non-symmetric (and consequently is not a true distance), its symmetrised version, called Jensen-Shannon divergence, is often preferred. It is defined as a sum of two KL measures (Lin, 1991):

$$D_{JS}(A, B) = \frac{1}{2} D_{KL}(A, M) + \frac{1}{2} D_{KL}(B, M) \tag{11}$$

where M is the average of the two distributions ($M = 0.5 * (A + B)$). Fig. 9 displays the values of the Jensen-Shannon distances between two types of voice quality, for the three considered features estimated by the three techniques.

From this figure, it can be noted that NAQ is the best discriminative feature (i.e. has the highest Jensen-Shannon distance between distributions), while H1–H2 and HRF convey a comparable amount of information for discriminating voice quality. As expected, the loud-soft distribution distances are highest compared to loud-modal and modal-soft distances. In seven cases out of nine (three different parameters and three different phonation type couples), CCD leads to the most relevant separation and in two cases (loud-modal separation with NAQ, loud-modal separation with HRF) CPIF provides a better separation. Both Figs. 8 and 9 show that the effectiveness of CCD and CPIF is similar, with slightly better results for CCD, while IAIF exhibits clearly lower performance (except for one case: loud-modal separation with HRF).

6. Conclusion

This study aimed at comparing the effectiveness of the main state-of-the-art glottal flow estimation techniques. For this, detailed tests on both synthetic and real speech were performed. For real speech, a large corpus was used for testing, without limiting analysis to sustained vowels. Due to the unavailability of the reference glottal flow signals for real speech examples, the separability of three voice qualities was considered as a measure of the ability of the methods to discriminate different phonation types. In synthetic speech tests, objective measures were used since the original glottal flow signals were available. Our first conclusion is that the usefulness of NAQ, H1–H2 and HRF for parameterizing the glottal flow is confirmed. We also confirmed other works in the literature (such as Alku et al. (2002) and Alku et al. (2009)) showing that these parameters can be effectively used as measures for discriminating different voice qualities. Our results show that the effectiveness of CPIF and CCD appears to be similar and rather high, with a slight preference towards CCD. However, it should be emphasized here that in our real speech tests, clean signals recorded for text-to-speech (TTS) synthesis were used. We can thus confirm the effectiveness of CCD for TTS applications (such as emotional/expressive TTS). However, for applications which require the analysis of noisy signals (such as telephone applications) further testing is needed. We observed that in the synthetic speech tests, the ranking dramatically changed depending on the SNR and the robustness of CCD was observed to be rather low. IAIF has lower performance in most tests (both in synthetic and real speech tests) but shows up to be comparatively more effective in very low SNR values.

Acknowledgment

Thomas Drugman is supported by the Belgian Fonds National de la Recherche Scientifique (FNRS). Authors also would like to thank the reviewers for their fruitful comments.

References

Airas, M., 2008. Tkk aparat: an environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocology 33, 49–64.

Airas, M., Alku, P., 2006. Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. Phonetica 63, 26–46.

Airas, M., Alku, P., 2007. Comparison of multiple voice source parameters in different phonation types. In: Proceedings of Interspeech , pp. 1410–1413.

Alku, P., 1992. An automatic method to estimate the time-based parameters of the glottal pulseform. In: Proceedings of ICASSP , pp. 29–32.

Alku, P., Backstrom, T., Vilkman, E., 2002. Normalized amplitude quotient for parametrization of the glottal flow. Journal of the Acoustical Society of America 112, 701–710.

Alku, P., Magi, C., Yrttiaho, S., Backstrom, T., Story, B., 2009. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. Journal of the Acoustical Society of America 125, 3289–3305.

Alku, P., Strik, H., Vilkman, E., 1997. Parabolic spectral parameter—a new method for quantification of the glottal flow. Speech Communication 22, 67–79.

Alku, P., Svec, J., Vilkman, E., Sram, F., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11, 109–118.

Alku, P., Vilkman, E., 1994. Estimation of the glottal pulseform based on discrete all-pole modeling. In: Third International Conference on Spoken Language Processing , pp. 1619–1622.

Bozkurt, B., Doval, B., d'Alessandro, C., Dutoit, T., 2005. Zeros of z-transform representation with application to source-filter separation in speech. IEEE Signal Processing Letters, 12.

Bozkurt, B., Dutoit, T., 2003. Mixed-phase speech modeling and formant estimation using differential phase spectrums. In: ISCA ITRW VOQUAL03 , pp. 21–24.

Brookes, D., Chan, D., 1994. Speaker characteristics from a glottal airflow model using glottal inverse filtering. Institute of Acoustics 15, 501–508.

Cabral, J., Renals, S., Richmond, K., Yamagishi, J., 2008. Glottal spectral separation for parametric speech synthesis. In: Proceedings of Interspeech , pp. 1829–1832.

Childers, D., 1999. Speech Processing and Synthesis Toolboxes. Wiley and Sons, Inc.

Childers, D., Lee, C., 1991. Vocal quality factors: analysis, synthesis, and perception. Journal of the Acoustical Society of America 90, 2394–2410.

d'Alessandro, C., 2006. Voice source parameters and prosodic analysis. In: Method in Empirical Prosody Research. Walter de Gruyter, pp. 63–87.

Deng, H., Ward, R., Beddoes, M., Hodgson, M., 2006. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. IEEE Transactions on Acoustics, Speech, and Signal Processing 14, 445–455.

Doval, B., d'Alessandro, C., 2006. The spectrum of glottal flow models. Acta Acustica united with Acustica 92, 1026–1046.

Doval, B., d'Alessandro, C., Henrich, N., 2003. The voice source as a causal/anticausal linear filter. In: ISCA ITRW VOQUAL03 , pp. 15–19.

Drugman, T., Bozkurt, B., Dutoit, T., 2009a. Complex cepstrum-based decomposition of speech for glottal source estimation. In: Proceedings of Interspeech.

Drugman, T., Dubuisson, T., d'Alessandro, N., Moinet, A., Dutoit, T., 2008. Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase. In: 16th European Signal Processing Conference.

Drugman, T., Dubuisson, T., Dutoit, T., 2009b. On the mutual information between source and filter contributions for voice pathology detection. In: Proceedings of Interspeech.

Drugman, T., Dutoit, T., 2009. Glottal closure and opening instant detection from speech signals. In: Proceedings of Interspeech.

Itakura, F., Saito, S., 1970. A statistical method for estimation of speech spectral density and formant frequencies. Electronics and Communications in Japan 53-A, pp. 36–43.

Fant, G., 1995. The lf-model revisited. transformations and frequency domain analysis. STL-QPSR 36, 119–156.

Fant, G., Liljencrants, J., Lin, Q., 1985. A four-parameter model of glottal flow. STL-QPSR 26, 1–13.

Fu, Q., Murphy, P., 2006. Robust glottal source estimation based on joint source-filter model optimization. IEEE Transactions on Audio, Speech, and Language Processing 14, 492–501.

Gardner, W., Rao, B., 1997. Noncausal all-pole modeling of voiced speech. IEEE Transactions on Speech and Audio Processing 5, 1–10.

Gobl, C., Chasaide, A., 2003. Amplitude-based source parameters for measuring voice quality. In: VOQUAL03 , pp. 151–156.

Hacki, T., 1989. Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. Folia Phoniatrica 41, 43–48.

Hanson, H., 1995. Individual variations in glottal characteristics of female speakers. In: Proceedings of ICASSP , pp. 772–775.

Henrich, N., d'Alessandro, C., Doval, B., Castellengo, M., 2004. On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation. Journal of the Acoustical Society of America 115, 1321–1332.

Jaroudi, A.E., Makhoul, J., 1991. Discrete all-pole modeling. IEEE Transactions on Signal Processing 39, 411–423.

Paliwal, K.B.A., 1993. Efficient vector quantization of lpc parameters at 24 bits/frame. IEEE Transactions on Speech Audio Processing 1, 3–14.

Klatt, D., Klatt, L., 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America 87, 820–857.

Laukkanen, A.M., Vilkman, E., Alku, P., Oksanen, H., 1996. Physical variations related to stress and emotional state: a preliminary study. Journal of Phonetics 24, 313–335.

Lin, J., 1991. Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37, 145–151.

Ling, Z., Hu, Y., Wang, R., 2005. A novel source analysis method by matching spectral characters of lf model with straight spectrum. Lecture Notes in Computer Science 3784, 441–448.

Moore, E., Clements, M., 2004. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. In: Proceedings of ICASSP.

Murphy, P., Akande, O., 2005. Quantification of glottal and voiced speech harmonics-to-noise ratios using cepstral-based estimation. In: Nonlinear Speech Processing Workshop , pp. 224–232.

Nordin, F., Eriksson, T., 2001. A speech spectrum distortion measure with interframe memory. In: Proceedings of ICASSP , pp. 717–720.

Online (2008). http://aparat.sourceforge.net/index.php/main_page. TKK Aparat Main Page.

Oppenheim, A., Schafer, R., 1989. Discrete-Time Signal Processing. Prentice-Hall.

Pedersen, C., Andersen, O., Dalsgaard, P., 2010. Separation of mixed-phase signals by zeros of the z-transform – a reformulation of complex cepstrum-based separation by causality. In: Proceedings of ICASSP.

Plumpe, M., Quatieri, T., Reynolds, D., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Transactions on Speech and Audio Processing 7, 569–586.

Quatieri, T., 2002. Discrete-Time Speech Signal Processing. Prentice-Hall.

Schroeder, M., Grice, M., 2003. Expressing vocal effort in concatenative synthesis. In: 15th International Conference of Phonetic Sciences , pp. 2589–2592.

Sturmel, N., d'Alessandro, C., Doval, B., 2007. A comparative evaluation of the zeros of z transform representation for voice source estimation. In: Proceedings of Interspeech , pp. 558–561.

Titze, I., Sundberg, J., 1992. Vocal intensity in speakers and singers. Journal of the Acoustical Society of America 91, 2936–2946.

Veeneman, D., Bement, S., 1985. Automatic glottal inverse filtering from speech and electroglottographic signals. IEEE Transactions on Acoustics, Speech, and Signal Processing 33, 369–377.

Vincent, D., Rosec, O., Chovanel, T., 2005. Estimation of lf glottal source parameters based on an arx model. In: Proceedings of Interspeech , pp. 333–336.

Walker, J., Murphy, P., 2007. A review of glottal waveform analysis. In: Progress in Nonlinear Speech Processing , pp. 1–21.

Wong, D., Markel, J., Gray, A., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27.

Yegnanarayana, B., Veldhuis, R., 1998. Extraction of vocal-tract system characteristics from speech signals. IEEE Transactions on Speech Audio Processing 6, 313–327.