Phase-Based Methods for Voice Source Analysis

Christophe d'Alessandro, Baris Bozkurt, Boris Doval, Thierry Dutoit,
Nathalie Henrich Bernardoni, Vu Ngoc Tuan, Nicolas Sturmel

Phase-based methods for voice source analysis

Christophe d'Alessandro (1), Baris Bozkurt (2), Boris Doval (1), Thierry Dutoit (3), Nathalie Henrich (4), Vu Ngoc Tuan (1), Nicolas Sturmel (1)

(1) LIMSI-CNRS Orsay, France
(2) Izmir Institute of Technology, Izmir, Turkey
(3) TCTS-FPMs, Mons, Belgium
(4) DPC-GIPSA-Lab Grenoble
cda@limsi.fr, barisbozkurt@iyte.edu.tr, boris.doval@limsi.fr, thierry.dutoit@fpms.ac.be, Nathalie.Henrich@gipsa-lab.inpg.fr, vnt@limsi.fr, sturmel@limsi.fr

Abstract.

Voice source analysis is an important but difficult issue for speech processing. In this talk, three aspects of voice source analysis recently developed at LIMSI (Orsay, France) and FPMs (Mons, Belgium) are discussed. In a first part, time domain and spectral domain modelling of glottal flow signals are presented. It is shown that the glottal flow can be modelled as an anticausal filter (maximum phase) before the glottal closing, and as a causal filter (minimum phase) after the glottal closing. In a second part, taking advantage of this phase structure, causal and anticausal components of the speech signal are separated according to the location in the Z-plane of the zeros of the Z-Transform (ZZT) of the windowed signal. This method is useful for voice source parameters analysis and source-tract deconvolution. Results of a comparative evaluation of the ZZT and linear prediction for source/tract separation are reported. In a third part, glottal closing instant detection using the phase of the wavelet transform is discussed. A method based on the lines of maximum phase in the time-scale plane is proposed. This method is compared to EGG for robust glottal closing instant analysis.

1 Introduction

Voice source analysis is an important issue for speech and voice processing, with many applications such as source tract decomposition, formant estimation, pitch synchronous processing, low-rate speech coding, speaker characterisation, singing, speech synthesis, phonetic and prosodic analyses, voice pathology and voice quality evaluation, etc.

However, voice source analysis is also a difficult issue for speech processing. There is generally no measurable reference to the "true" source and vocal tract components. Speech and voice signals are rapidly time-varying, and subject to large individual and inter-subject variations. Finally source tract interactions are not well known to date, but they may render voice source decomposition questionable in the situations where strong interactions are likely to occur (e.g. source-tract adjustments in singing).

The aim of this tutorial is to present some aspects of the authors' recent works in the domain of voice source analysis. A common feature of this line of research is the specific attention paid to the phase structure of the voice source signal. Two aspects of the voice source phase are explored: the spectral phase of the glottal pulse itself and cross-scale instantaneous phases in a time-scale space. This paper presents the concepts without much technical details. The general reader will get the main ideas out of this paper. As the most significant references to published literature are pointed out at the beginning of each section, the interested reader will easily find more detailed presentations of the material described herein.

Definitions of phase

"Phase" is a highly polysemic word in the general language. In signal processing also, the meaning of "phase" is manifold (for a review, see Alsteris & Paliwal, 2007). Starting from the more basic periodic signals, sine waves, phase is defined as the argument of the sinusoidal function. This first definition of phase, in time domain, or "instantaneous phase" is useful for dealing with waveforms. For instance maxima of sine waves are located at phases $\pi/2$ modulo 2π, Mathematically, instantaneous phase and instantaneous envelopes are defined using the Hilbert Transform. They are useful for describing the time evolution of signals. This first definition of phase can be extended to time-frequency or time-scale representations. This will be used below for time-scale analysis of the glottal closings instants of the voice source.

As spectral representation is a decomposition of the signal on a basis of complex exponentials (i.e. sine and cosine waves), a second definition of phase is the argument of the complex spectrum. This "spectral phase" or "phase spectrum" is often difficult to deal with. On the one hand, spectral phases computed using the Fourier transform are obtained modulo 2π and must be unwrapped. On the other hand, even the smallest delay in the signal changes dramatically the phase spectrum, because it introduces a linear component (note that the phase spectrum derivative in frequency, or group delay, is a more robust representation (Yegnanarayana & Murthy, 1992). This second definition of phase will be used below for glottal flow analysis and synthesis.

Phase structure of the glottal pulse

Let's write the linear speech production model of voiced speech, in the time domain and in the spectral domain:

$$s(t) = e(t) * v(t) * l(t) = (\sum_n \delta(t - nTo) * g(t)) * v(t) * l(t) \quad \textbf{(1)}$$

$$S(f) = E(f) \times V(f) \times L(f) = (\sum_n \delta(f - nFo) \times G(f)) \times V(f) \times L(f) \quad \textbf{(2)}$$

Where: t represents times, f frequency, s the speech signal, e the voiced excitation component, v the vocal tract impulse response, l the lip radiation component, T_0 (F_0) the fundamental period (frequency) and g the glottal flow component.

Depending on the application, the time domain (1) or spectral domain (2) model is preferred. But it goes generally unnoticed that time domain and spectral domain approaches may not be equivalent, because of a different underlying phase structure implicitly assumed for the glottal flow component. Let us examine this point in more detail.

In the early years of the source-filter theory of speech production, the effect of the voice source was mainly studied in the spectral domain, like in Equation (2): the glottal flow signal being considered as the output of a low-pass system to an impulse train. For instance, in a transmission line analogue (Fant, 1970) four poles (*sr1, sr2, sr3, sr4*) on the negative real axis were used, with $|sr1| = |sr2| = 2\pi 100Hz$, and $|sr3| = 2\pi 2000Hz$, $|sr4| = 2\pi 4000Hz$. Note that two poles are low pass (we shall interpret these poles later on in terms of "glottal formant"), and that two poles (*sr*3 and *sr*4) are fixed (we shall interpret these poles later on in terms of "spectral tilt"). This simple form entailed important practical consequences, because it has been used (for discrete time signals) for deriving the linear prediction equations (see for instance Markel and Gray, 1976). In this later case, only two poles are used, because the linearity of this acoustic model only holds for frequencies below about 4000 Hz.

$$G(z) = 1/(1 - pz^{-1})(1 - p^* z^{-1}) \quad \textbf{(3)}$$

The corresponding impulse response, magnitude and phase spectra of are plotted in Fig 1.

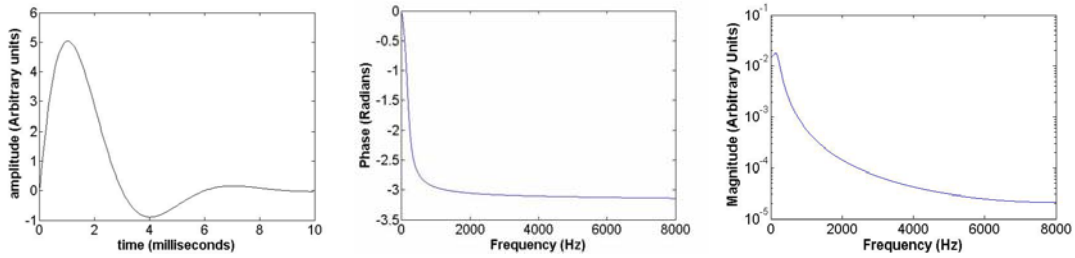

Fig. 1. All pole glottal flow model, as assumed by the Linear Prediction synthesis model. From left to right: glottal waveform, spectral phase and spectral magnitude.

On the other hand, for e.g. formant synthesis, the time domain model like in equation (1) is generally preferred, using time domain models of the glottal flow component. A neglected dimension of this glottal flow component is its phase structure. In time-domain models, glottal flow models are generally not viewed as filters or linear systems, but are rather described by ad hoc equations (based on polynomials or trigonometric functions). An example of such a model is the KLGLOTT88 model (Klatt & Klatt, 1990), described by the following equations (when there is no additional spectral tilt component):

$$Ug(t) = \begin{cases} at^2 - bt^3 & 0 < t < OqTo \\ 0 & OqTo < t < To \end{cases} \quad (4)$$

The corresponding impulse response, magnitude and phase spectra of are plotted in Fig 2.

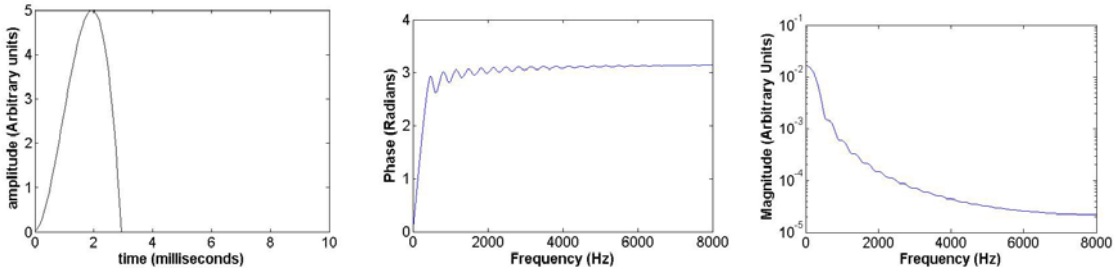

Fig. 2. KLGLOTT88 glottal flow model, as assumed by the Klatt synthesizer (Klatt & Klatt, 1990). From left to right: glottal waveform, spectral phase and spectral magnitude.

Note that, as far as spectral magnitude is concerned, Figure 1 and Figure 2 are very close (the same glottal flow parameters being used). However, both waves are reversed in time, or equivalently, their phases are opposed in sign (corresponding to a symmetry relative to the glottal closing instant (GCI)). Then it is argued in the following that the specific phase structure of glottal flow models can be used for source/tract separation and for designing new linear glottal flow models. Section 2 give details on the spectrum of glottal flow models, and presents a new model: the Causal-Anticausal Linear model. In Section 3 a new method that takes advantage of this causal-anticausal model is presented, along with some application to voice source analysis.

Glottal pulse phases in the time-scale space

A second noticeable aspect of the phase of the voice source signals is the instant of glottal excitation or glottal closing. According to Equation (1) the source component can be split in two parts: (1) a linear (and thus linearly predictable using a small set of preceding samples) glottal flow filter and (2) excitation by a train of Dirac pulses (which is not linearly predictable at all using a small set of preceding samples) at the GCI. GCIs correspond to singularities in the signal. Time-scale representation using the wavelet transform is well suited to the problem of singularity detection. However, in the case of glottal pulses, the phase structure of the glottal pulse also influences instantaneous phases in the time-scale domain. A specific method for following the phases across scales is proposed: the lines of maximum amplitude (LOMA) of the wavelet transform. This method is applied to GCI detection, and compared to direct GCI measurement using electroglottography (EGG) in Section 4. Finally, Section 5 summarizes the main results obtained.

2 Time-domain and spectral glottal flow models[1]

Glottal source

Several mathematical "glottal flow models", abbreviated as GFM hereafter, have been proposed over the past decades, such as the well-known LF model (Fant & Liljencrants, 1985), the KLGLOTT88 model (Klatt & Klatt, 1990), the Rosenberg's models (Rosenberg, 1971) or the R++ model (Veldhuis, 1998). Figure 3 gives a typical example of a glottal flow model and its time derivative.

[1] The main references for this Section are: Doval, d'Alessandro, Henrich, 2006; Doval, d'Alessandro, Henrich., 2003.

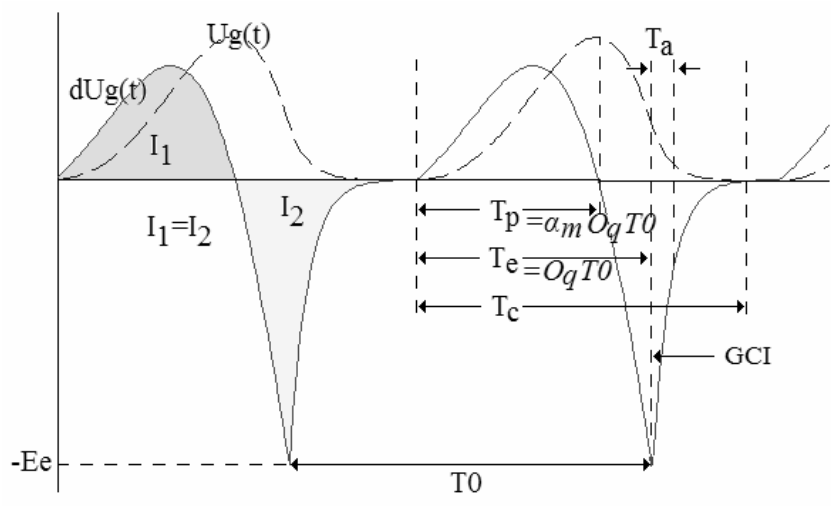

Fig. 3. Example of glottal flow model (top) and differential glottal flow mode (bottom). See text for explanation of the symbols (from Sturmel et al., 2007).

Generally, voice quality is better described by spectral parameters, such as the spectral tilt, the relative amplitude of the first harmonics (Hanson, 1997), the harmonic richness factor (Childers, 1991), the parabolic spectral parameter (Alku et al, 2002). A remarkable spectral feature of GFM is the spectral peak that can be observed on the glottal flow derivative spectrum in the region of the first harmonics. This peak has been coined the "glottal formant", although it is not a resonance, like vocal tract formants.

The link between spectral voice quality and glottal source parameters has not previously been addressed systematically. For answering this problem, one must study the position, variation and properties of the glottal formant and derive closed-form equations for relating time-domain glottal flow parameters to the glottal formant. This work has been conducted in (Doval et al., 2006), with the following aims:

- Studying the spectral behavior of the most common glottal flow models
- Deriving the relationships between time-domain parameters and spectral parameters
- Providing some hints for spectral estimation or modification of glottal flow parameters

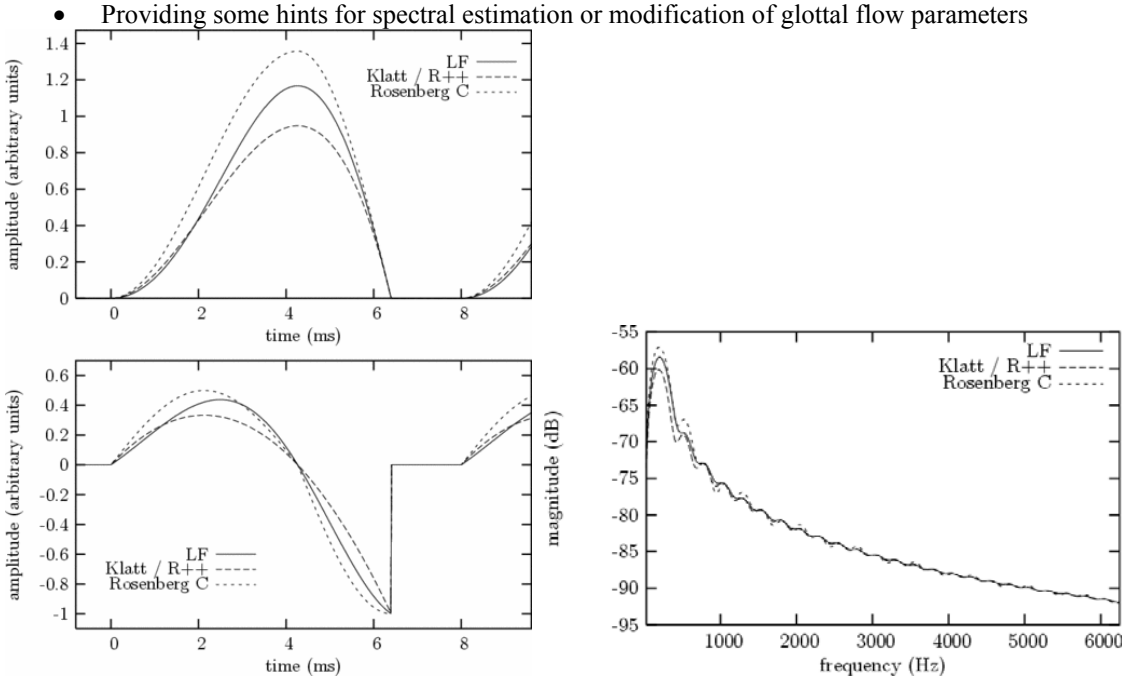

Fig. 4. Comparison of several glottal flow models in time and magnitude spectral domains. Left, top: GFM waveform. Left, bottom: GFM derivative waveform. Right: GFM Magnitude spectrum (from Doval et al., 2006).

Glottal flow models

Among the GFMs proposed in the literature, we have studied the following ones (the parameters mentioned in this paragraph are explained later in the paper):

- KLGLOTT88 model (Klatt & Klatt, 1990): the glottal flow is modelled by a third order polynomial which is possibly smoothed using the low-pass filter method. There are four parameters: A_v, T_0, O_q and TL which is the attenuation in dB of the low-pass filter at 3000 Hz. Notice that the asymmetry of the flow cannot be changed and is always: $\alpha_m = 2 / 3$.
- R++ model (Veldhuis, 1998): the glottal flow is composed of a fourth order polynomial for the open phase followed by an exponential return phase. There are five parameters: K (an amplitude coefficient), T_0, T_e, T_p and T_a. The glottal flow is computed so that it returns exactly to 0 at the end of the cycle.
- Rosenberg C (Rosenberg, 1971): the glottal flow is composed of two sinusoidal parts. The 4 parameters are: A_v, T_0, $T_p$and $T_n = T_e - T_p$. Noticed that the smooth closure case is not handled.
- LF model (Fant & Liljencrants, 1985): the glottal flow derivative is modelled by an exponentially increasing sinusoid followed by a decreasing exponential. There are five parameters: $E_e = E$ (the maximum excitation), T_0, T_e, T_p, T_a. The glottal flow is computed so that it returns exactly to 0 at the end of the cycle. For that, two implicit equations must be solved.

Figure 4 shows an example of the four GFMs (left, top) and their derivatives (left, bottom) with abrupt closure and with a common set of parameters: $T_0 = 8ms$, $O_q = 0.8$, $\alpha_m = 2 / 3$ and $E = 1$. KLGLOTT88 and R++ models are identical for this parameter set. Note that A_v differs between models when E and the other parameters are fixed. However these differences are hardly audible. All GFMs share some common time-domain features:

- the glottal flow is always positive or null
- the glottal flow and its derivative are quasi-periodic
- during a fundamental period, the glottal flow is bell-shaped: it increases, then decreases, then becomes null
- during a fundamental period, the glottal flow derivative is positive, then negative, then null.
- the glottal flow and its derivative are continuous and differentiable functions of time, except in some situations at the glottal closing instant.

Furthermore, GFMs are described in terms of phases in the time domain:

- the opening phase: the glottal flow increases from baseline at time 0 to its maximum amplitude A_v also called "amplitude of voicing" at time T_p.
- the closing phase: the glottal flow decreases from A_v to a point at time T_e where the derivative reaches its negative extremum E. T_e is the glottal closing instant (GCI) and E is called the "maximum excitation".
- the open phase: it consists of the opening and closing phases, characterized by the open quotient $O_q = T_e / T_0$. The ratio between the opening phase duration and the open phase duration is called "asymmetry coefficient" and noted α_m.
- the closed phase: in the situation of "abrupt closure" there is a discontinuity in the glottal flow derivative which instantaneously reaches 0 after maximum excitation. The glottal flow is null between O_qT_0 and T_0. In the situation of "smooth closure" the glottal flow derivative is continuous and exponentially returns to 0 at time T_c. This phase is called "return phase" and the exponential time constant is noted T_a. It can also be characterized by the relative parameter $Q_a = T_a / [(1 - O_q)T_0]$ which takes its value between 0 and 1. The smooth closure case can be modelled in two different ways: either a time-domain decreasing exponential (leading to the return phase as described above) noted "return phase method" or a low-pass first (or second) order filter applied to the whole open phase noted "low-pass filter method".

Spectral properties of glottal flow models: the glottal formant

Since the early years of the source-filter theory of speech production, it is well known that the effect of the glottal flow in the spectral domain can be approximated by a low-pass system. When considering the GFM derivative spectrum, one can show that it behaves like a band-pass filter, and a spectral peak appears in low frequencies, the so-called "glottal formant". Figure 4 shows the magnitude spectra of the four GFM derivatives (phase spectra are also similar for the four models). Closed-form equations for the GFM spectra and the GFM derivative spectra are given in (Doval et al., 2006).

Figure 5 shows the magnitude spectrum of a GFM derivative and a straight line stylization of this spectrum. The glottal formant is clearly visible in this log-log representation. Again, "glottal formant", does not refer in this case to a resonance in the speech apparatus but only to a maximum in the spectrum.

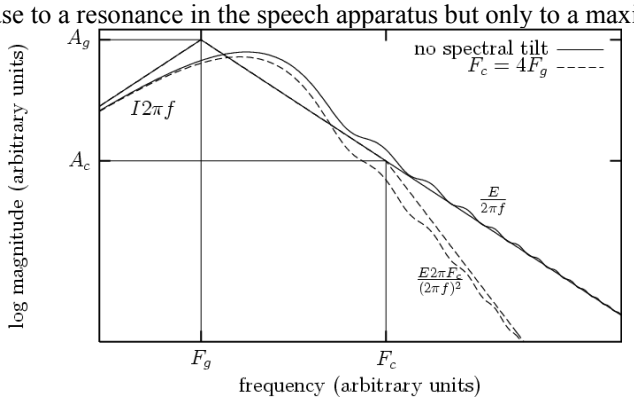

Fig. 5. Magnitude spectrum of a GFM derivative (in a log-log representation), and straight line stylization (from Doval et al., 2006).

Figure 5 shows that the glottal formant frequency is slightly higher than the asymptotes crossing point. Its amplitude is also different from the crossing-point amplitude. However, straight line stylisation gives a good approximation of the glottal formant position, and it can be computed from the glottal flow parameters (see Doval et al., 2006).

Effects of open quotient and asymmetry coefficient on the glottal formant

The glottal formant frequency is mainly inversely proportional to the open quotient. It depends only marginally on α_m. The glottal formant is roughly found between the first and the 4th harmonics. Its relative amplitude depends mainly on α_m An example of synthetic speech is given in Figure 6.

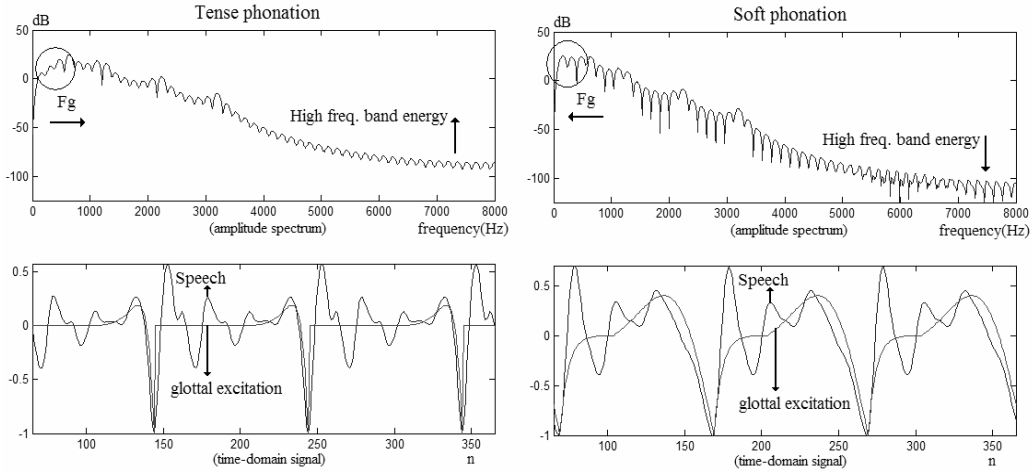

Fig. 6. Glottal formant and open quotient (synthetic speech). Top row: magnitude spectrum of the speech signals. Bottom row: time domain signals, with glottal excitation superimposed to speech (arrows are showing the corresponding waves). Left column: small open quotient (tense phonation); right column: large open quotient (lax phonation) (from Bozkurt, 2005).

Figure 7 (right panels) shows the influence of O_q (4th column right) and α_m (3rd column right) on the GFM (3rd row) and the GFM derivative (4th row) spectra. Several points may be observed:

- the mid and high frequency spectral energy is not much modified by α_m and O_q variations
- O_q mainly changes the glottal formant frequency
- α_m mainly changes the glottal formant amplitude (or rather its bandwidth)

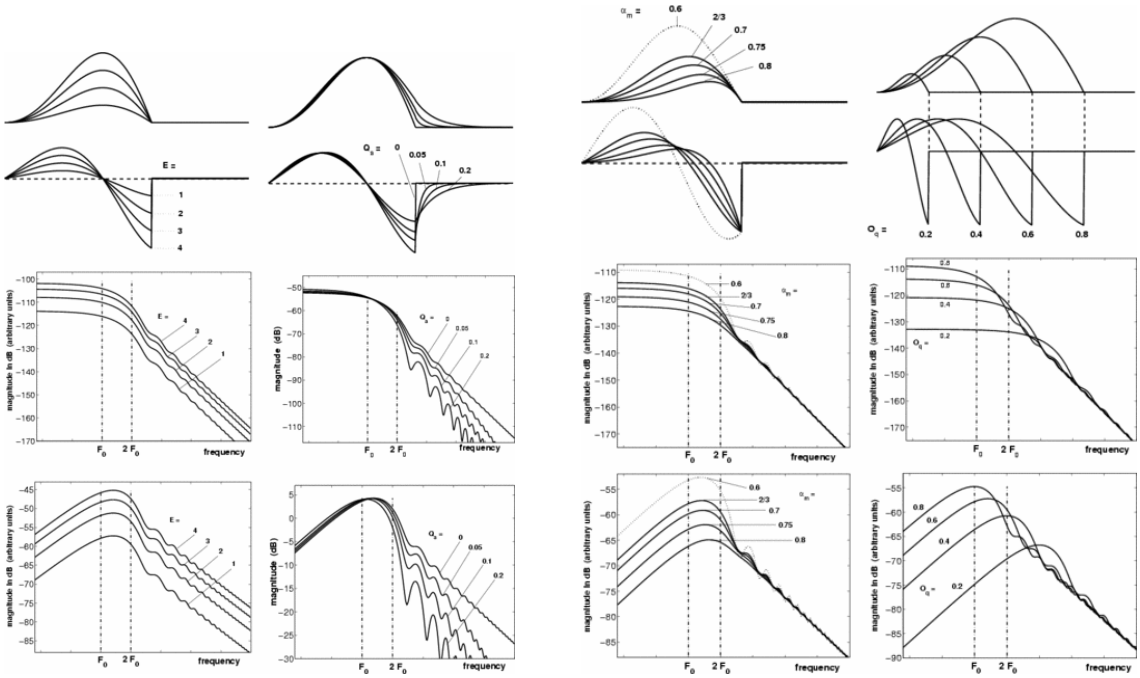

Fig.7. Effect of amplitude of voicing, spectral tilt, asymmetry and open quotient on the waveform and spectra of GFM and GFM derivatives. Top row: GFM; Second row; GFM derivative; Third row GFM log-log magnitude spectrum; Bottom row: GFM derivative log-log magnitude spectrum. First column: effect of amplitude of voicing; Second column: effect of the return phase and spectral tilt; Third column: effect of GFM asymmetry; Last column: effect of open quotient (from Doval et al., 2006).

Spectral tilt

The voice spectral tilt describes the GFM spectral profile in mid to high frequencies. It is related to the return phase in the case of smooth closure of the vocal folds. From a modelling point of view, two methods can be applied: either a time-domain decreasing exponential (leading to the return phase as described above) noted "return phase method" or a low-pass first (or second) order filter applied to the whole open phase noted "low-pass filter method". For example, R++ and LF are using the "return phase method" while KLGLOTT88 is using the low-pass filter method. The spectral tilt parameter is Q_a. Its main effect is to add an additional -6dB/oct attenuation above the cut-off frequency F_c . Figure 7 (second column) illustrates the effect of Q_a. The main vocal quality effect of spectral tilt is voice loudness: a low or null Qa corresponds to a minimum spectral tilt and a loud voice. Conversely, a high (close to 1) Q_a corresponds to a high spectral tilt and a weak voice.

Causal-anticausal glottal flow model and application to speech synthesis

The preceding discussion shows that the source log magnitude spectrum can be stylized by three linear segments with +6dB/octave, -6dB/octave and -12dB/octave (or sometimes -18dB/oct) slopes, respectively, like in Figure 5. The two breakpoints in the spectrum correspond to the glottal spectral peak and the spectral tilt cut-off frequency.

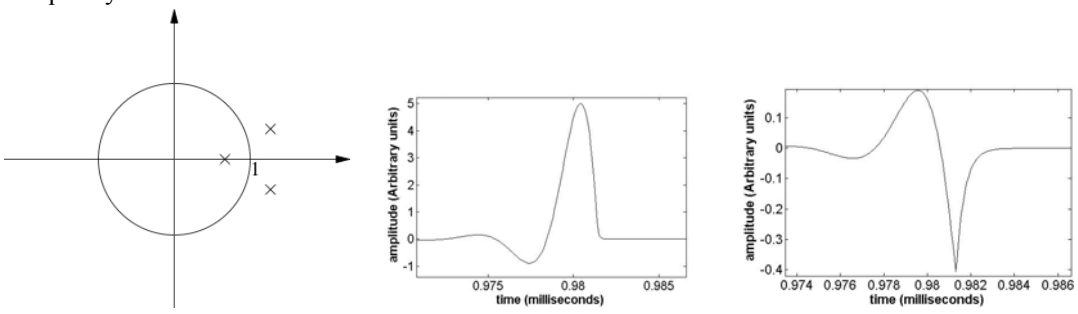

Fig. 8. poles pattern (left).for the Causal-Anticausal Linear Model (left). Corresponding: GFM (middle) and GFM derivative (right).

For synthesis in the spectral domain, it is possible to design an all-pole filter which is comparable to e.g. the LF model. This filter is a 3rd order low-pass filter, with a pair of conjugate complex poles, and a simple real pole. The simple real pole is given directly by the spectral tilt parameter. It is mainly effective in the medium and high frequencies of the spectrum. The pair of complex-conjugate poles is used for modelling the glottal formant (see Figure 8). If one wants to preserve the glottal pulse shape, and thus the glottal flow phase spectrum, it is necessary to design an anticausal filter for this pole pair. The spectral model is then a Causal (spectral tilt) Anti-causal (glottal formant) Linear filter Model (CALM, Doval et al. 2003). This model is computed by filtering a pulse train by a causal second order system, according to the frequency and bandwidth of the glottal formant, whose response is reversed in time to obtain an anti-causal response. Note that if one wants to preserve the finite duration property of the glottal pulse, it is necessary to truncate the impulse response of the filter. Otherwise, the decay of the filter response may continue longer than a single period and get mixed with the next period. Spectral tilt is introduced by filtering this anti-causal response by the spectral tilt component of the model. The waveform is then normalized in order to control accurately the intensity parameter *E*. The CALM has been recently used successfully for real-time gesture-controlled voice synthesis (D'Alessandro et al., 2007).

3 Zero of the Z-Transform (ZZT) Representation of Speech [2]

Principle of the ZZT

ZZT is a new representation of signals. ZZT means Zeros of Z-Transform and is defined by the set of roots of the Z-transform of any signal frame. Mathematically speaking, if *x(n)*, *n*=0…*N*-1 is a signal frame, its ZZT is the set of *N* complex roots (or zeros) Z_m of its Z-transform X(z):

$$X(z) = \sum_{0}^{N-1} x(n)\, z^{-n} = x(0) z^{-N+1} \prod_{1}^{N-1} (z - Z_m) \qquad (5)$$

The ZZT is then an all-zero representation of the Z-Transform. It can be represented on the complex plane either in the classical Cartesian coordinates (see Figure 9c) or in the more readable polar coordinates (see Figure 9b).

To compute the ZZT, an algorithm for polynomial root extraction is needed, like such as the "root" function in Matlab™, since it is known that there is no general closed-form expression to calculate the roots of a polynomial whose degree is larger than 5 from its coefficients (Galois' theorem).

[2] The main references for this Section are: Bozkurt, 2005, Bozkurt, Doval, d'Alessandro, Dutoit, 2005, Sturmel, d'Alessandro, Doval, 2007.

ZZT and the linear speech production model

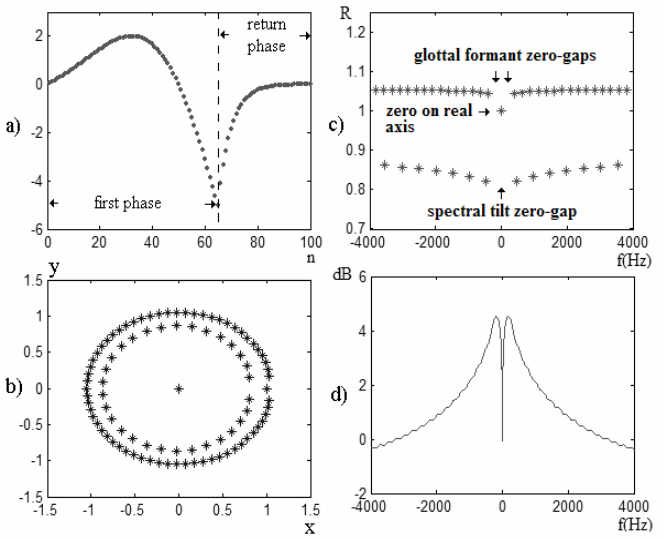

Fig. 9. ZZT, applied to a differential glottal flow signal. a) signal; d) corresponding magnitude spectrum; c) Zeros of the Z transform in Cartesian coordinates; b) Zeros of the Z transform in polar coordinates (from Bozkurt et al, 2005).

The ZZT and the source-filter model have strong relationships. The ZZT shows different patterns for each contribution of the source/filter model, and particularly for each part of the glottal flow signal; the speech signal is simply the union of the ZZT of each contribution. These results indicate that ZZT is well-suited for source/filter deconvolution, and especially for the study of the source parameters. Let us consider the LF GFM, with equations:

$$g(t) = E_0 e^{\alpha t} \sin(\omega_g t), 0 \le t \le t_e \quad \textbf{(6)}$$

$$g(t) = -\frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right], t_e \le t \le t_c \le T_0 \quad \textbf{(7)}$$

$$g(t) = 0, t_c \le t \le T_0 \quad \textbf{(8)}$$

The corresponding ZZT patterns are displayed in Figure 9. Two rows of zeros are obtained, one for the causal part, and the other for the anticausal part. Gaps appear in each row, corresponding to the poles of the CALM model.

According to Equation 1, a voiced speech frame *s(n)* can be written as the convolution product of an impulse train (excitation signal) *e(n)* by the differential* glottal flow waveform *g(n)* followed by the vocal tract filter with impulse response *v(n)* . As usual for source-filter model, the lip radiation contribution is approximated as a derivation and is incorporated into the source contribution as a differentiation. Therefore it is the differential glottal flow rather than the glottal flow itself which is represented. An example is given in Figure 10.

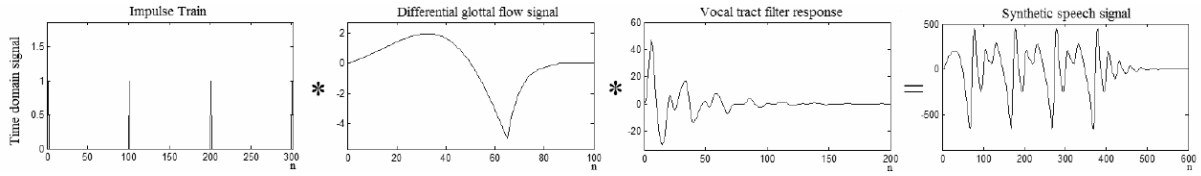

Fig. 10. A pictorial view of the speech production model of Equation 1 (from Bozkurt, 2005)

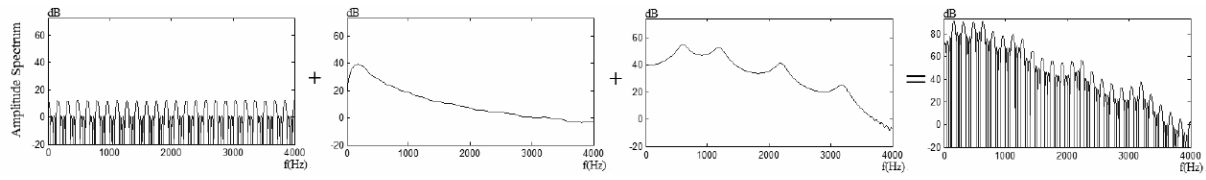

Fig. 11. A pictorial view of the speech production model of Equation 2 (from Bozkurt, 2005)

It is well known that the magnitude spectral representation (in dB) transforms the convolution into a sum :

$$\log(|S(\nu)|) = \log(|E(\nu)|) + \log(|V(\nu)|) + \log(|L(\nu)|) \qquad \textbf{(9)}$$

For instance, this is the first step of cepstrum source/filter deconvolution. An example is given in Figure 11. In contrast, the ZZT representation transforms the convolution into a union:

$$ZZT\{S\} = ZZT\{E\} \bigcup ZZT\{V\} \bigcup ZZT\{L\} \qquad \textbf{(10)}$$

This can be clearly seen on Figure 12, where the sets of zeros of each contribution (left three plots) are simply "copy-pasted" in the speech signal ZZT (right plot).

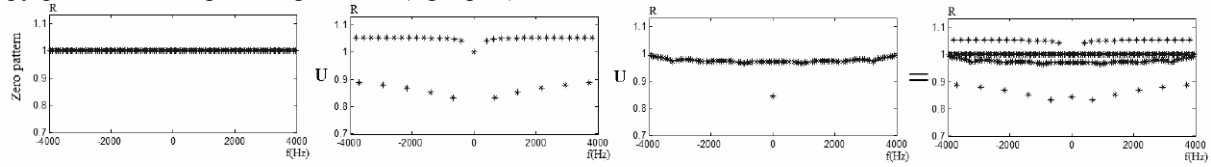

Fig. 12. A pictorial view of the speech production model in terms of ZZT (from Bozkurt, 2005)

Let us describe more precisely the different ZZT patterns encountered in each part of the source-filter model (like in Figure 12):

- For the impulse train (left plot), the ZZT pattern is a set of regularly spaced zeros on the unit circle, with a gap at each multiple of the fundamental frequency. Since the magnitude spectrum is the modulus of the Z-Transform taken on the unit circle, one can observe the effect of these zeros as spikes on the spectrum between the harmonics (Figure 11, left).
- For the differential glottal flow (left middle plot), the ZZT pattern is the union of a row of zeros lying outside the unit circle and of a row of zeros inside the circle. The outside row shows a gap between the zero which is on the real axis and the others. This corresponds to the glottal formant which can be seen as a local maximum in the low-frequency region on the spectrum representation (Figure 11, left middle plot). On the other hand, the zero gap on the inside row corresponds to the "spectral tilt" which can be seen as a global slope on the spectrum representation.
- For the vocal tract (right middle plot), the ZZT pattern is a row of zeros lying inside the unit circle. Gaps appear in this row, each one corresponding to a (vocal tract) formant.

Finally, the ZZT pattern of the speech signal (right plot) is the union of all the zeros that appear in each part of the source-filter model. This is the key to source/filter deconvolution using separation of the zeros into two subsets.

Source parameter estimation using the ZZT

An interesting property of ZZT decomposition is that it can distinguish between the minimum and maximum phase parts of a signal. Note that the glottal formant is maximum phase: this can be seen on the group delay because the corresponding peak is negative, or on the ZZT because the corresponding zeros are outside the unit circle (Bozkurt et al., 2004), or on the CALM model where the corresponding poles are outside the unit circle. Since all other speech components are minimum phase (or causal), the outside zeros belong to the glottal flow component. Therefore they can be extracted to estimate the glottal flow component and the glottal formant frequency. Using this position, the open quotient can be deduced using the theory exposed in Section 2 (Henrich et al., 2001).

Figure 13 gives an example of open quotient estimation using ZZT. This example is a natural vowel changing from lax to pressed voice quality. It has been recorded together with the EGG signal in order to extract the open quotient Oq [3]. For a pressed voice, the open quotient is low. Figure 13 shows the estimated glottal formant frequency *Fg*, together with a curve proportional to 1/*Oq* (the value of *k* has been chosen to fit the first part of the *Fg* curve) and the fundamental frequency (which is approximately constant). *Fg* follows well the curve in

[3] The EGG signal is proportional to the electrical current across the vocal fold. When the glottis is open this current is relatively weak, conversely when the glottis is closed this current is relatively strong. The variation of the EGG current is more important at the GCI than at the opening instant. In the derivative of the EGG current, a large peak indicates the closing instant and a smaller peak indicates the opening instant (Henrich et al. 2004). Notice that these two peaks have opposite signs.

1/*Oq*, which shows that ZZT is a promising means of estimating *Oq*. Experiments on synthetic signals have shown that *Fg* can be estimated with good precision if it does not coincide with the first formant (Sturmel et al. 2006).

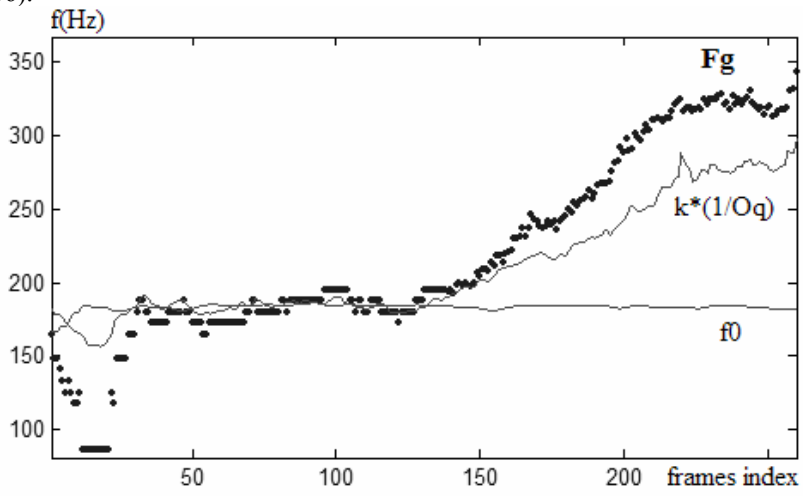

Fig. 13. Open quotient estimation using the ZZT (from Bozkurt, 2005).

Source-tract deconvolution using the ZZT and comparative evaluation with inverse filtering

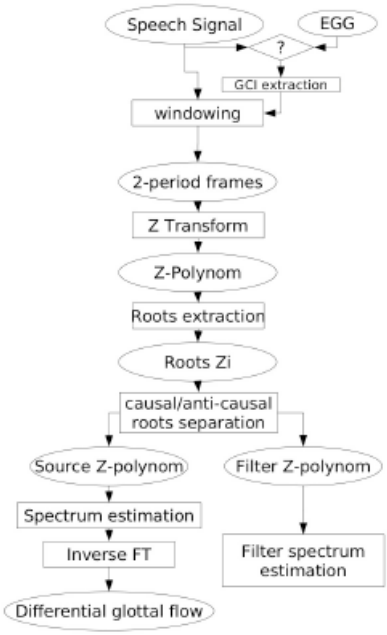

Fig. 14. Source-tract decomposition using ZZT (from Sturmel et al., 2007)

Source-filter deconvolution using ZZT is explained in Figure 14 (Bozkurt et al. 2004, Sturmel et al. 2006). The principle is to compute the ZZT of a speech frame, to separate its zeros in two sets according to their radius, and then to compute two components from these sets of zeros. These two components correspond to the "source dominated" and "vocal tract dominated" signals.

According to ZZT theory, the zeros outside the unit circle correspond to the anticausal part of the speech signal. The source-filter model predicts that the anticausal part corresponds to the glottal formant. However, spectral tilt corresponds to a causal signal (a decreasing exponential in the time-domain). Then the ZZT decomposition method separates the glottal formant contribution from the vocal tract and spectral tilt contributions. Figure 15 shows an example of decomposition.

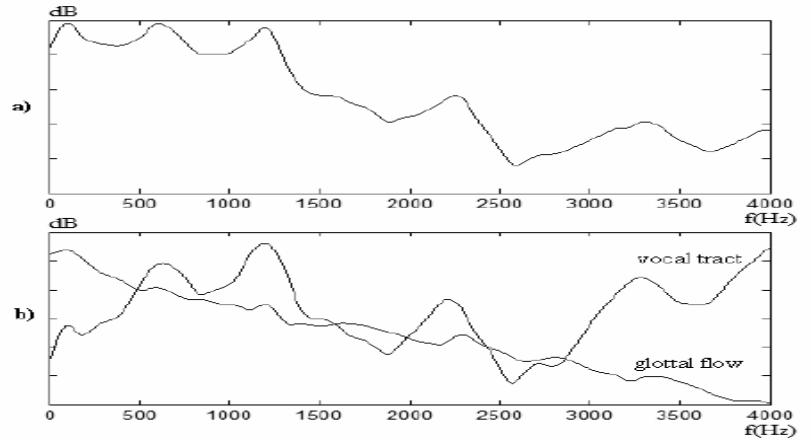

Fig. 15. Source tract decomposition using the ZZT (from Bozkurt, 2005).

The top plot shows the spectrum of a speech frame where formants can be seen at around 600Hz, 1200Hz, 2300Hz, 3300Hz and 4000Hz. The bottom plot shows the spectra of the two signals obtained from ZZT decomposition: one mainly corresponds to the vocal tract response showing the formants, the other corresponds to the glottal flow spectrum showing the glottal formant and a global spectral slope.

It must be pointed out that ZZT decomposition and inverse filtering are based on different principles. Inverse filtering requires the vocal tract filter identification, and is based on passing the speech signal through the inverse filter. Inverse filtering achieves source-filter decomposition mainly on the basis of properties of the filter. On the contrary, the ZZT representation is based on a very specific property of the source, i.e. its mixed-phase nature. Therefore, source tract decomposition using ZZT is not another inverse filtering method.

Four state-of-the-arts commonly used inverse filtering methods have been compared to ZZT for source filter separation (Sturmel et al., 2007). All methods are based on LP (Markel & Gray, 1976), the last one requiring additional processing steps. All these methods are well documented in the literature: (1) Linear prediction, autocorrelation method (unlike ZZT, Autocorrelation is an asynchronous method), (2) linear prediction, covariance (like ZZT, covariance LP is a pitch synchronous method); (3) Linear prediction, lattice filter (asynchronous method based on the Burg's algorithm, which ensures a stable lattice filter; this is an asynchronous method); (4) Iterative Adaptive Inverse Filtering (IAIF, asynchronous method; Alku, 1992).

As the source signal is unknown in natural speech, the evaluation procedure is mainly based on analyzing synthetic speech in a first part, and then on simultaneous recording of natural speech and electroglottographic signals (that give partial information on the voice source). The synthetic speech database contains a large number of test signals. Automatic procedures for comparisons of synthetic and estimated voice sources are also proposed.

More formal evaluation, with the help of a spectral distance were conducted, using a large number of experimental conditions

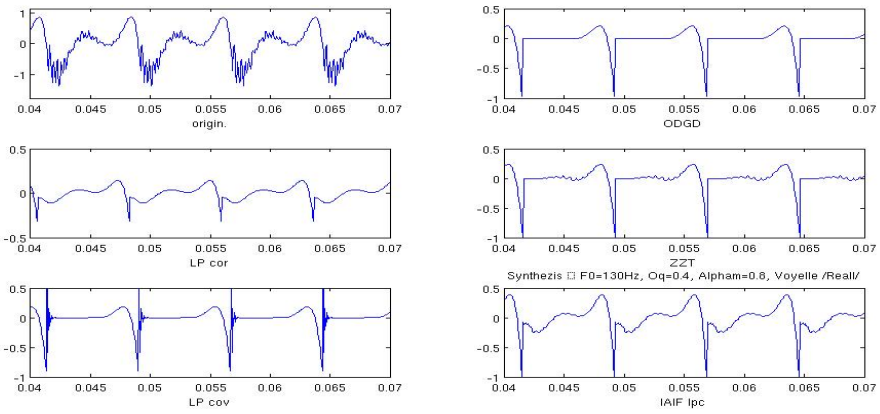

Fig. 16. Comparison of ZZT and LP inverse filtering for source-tract decomposition of a synthetic speech signal. Top Left, synthetic speech. Top right, synthetic differential glottal flow. Middle left, LP correlation, Middle right: ZZT; bottom left: LP covariance, bottom right IAIF (from Sturmel et al., 2007).

Source estimation examples are presented in figures 16 and 17. Figure 16 presents the estimated source waveforms for a synthetic speech signal /a/. Note that the original synthetic source is known, and can be compared directly to the estimated source waveforms. Figure 17 presents source estimations for a real speech signal. Both glottal flow and its derivative are shown for each method. An electroglottographic reference is available for this example, showing that the open quotient is about 0.5 (i.e. the closed phase of the source is about half of the period). In the example, the ZZT is the only method giving a closed phase of about 0.5.

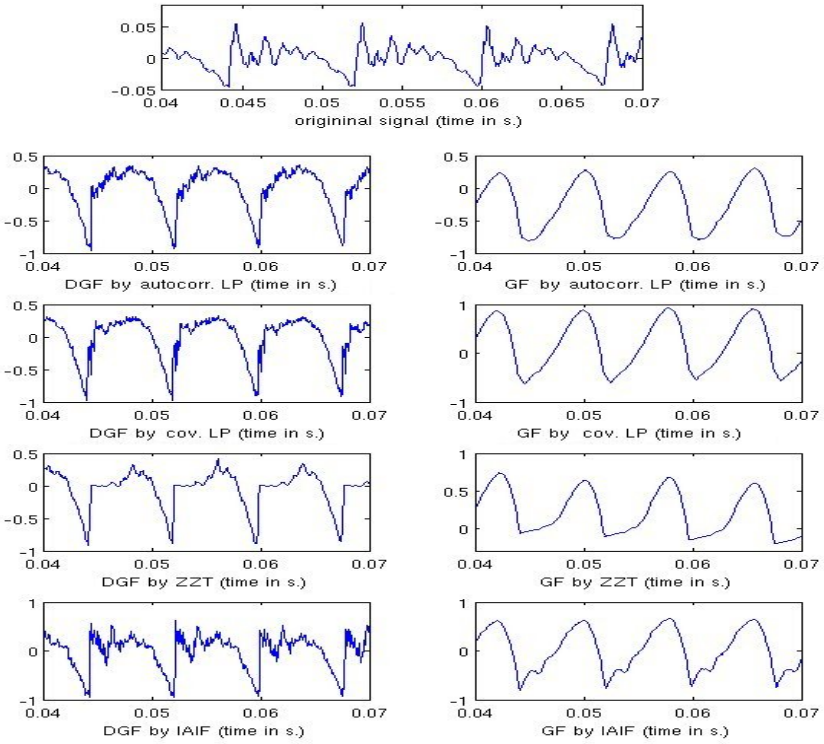

Fig. 17. Comparison of ZZT and LP inverse filtering for source-tract decomposition of a natural speech signal. Top center: natural speech (vowel /a/). Left column: estimated differential glottal flow. Right column: estimated glottal flow.

Spectral distance results and visual inspection of the waveforms lead to the following conclusions:

- The pitch synchronous covariance linear prediction seems the worst differential glottal wave estimator. Since it is performed on a very short signal segment, the autoregressive filter order may probably be too small for accurate estimation of the vocal tract filter. Nevertheless, the overall low frequency restitution of the glottal formant seems realistic.
- − The IAIF method seems the most robust one tested in the sense that it gives good results in almost every case : the adaptive part of the algorithm appears to be useful for fitting even the most difficult signals. However noise and ripples on the estimated differential glottal waveform make it hardly suited to parameter extraction or analysis.
- − The auto-correlation linear prediction is surprisingly the best LP-based source estimation in this benchmark. However, tests on signals are using long analysis windows, exploiting the time invariance assumed by the method. This is not always realistic for actual time-varying real speech signals. Furthermore, we observed that the worst cases are those where the pre-emphasis does not completely suppress the glottal formant :low Oq values leading to a glottal formant at two or three times $F0$, and low values for αm leading to a more resonant formant.
- − The ZZT inverse filtering outperforms LP-based methods both in spectral measurements and time-domain observations. The absence of ripples in the glottal closed phase together with the very good benchmark results are the strongest arguments in favour of this method. On real signals, it is the only one to present a clearly visible closed phase on glottal flow waveforms (figure 17). The low error values achieved during benchmark make ZZT the best choice for glottal parameter estimation by model fitting. However, the method relies heavily on precise glottal closing instants determination,

and it seems also relatively weak for low signal to noise ratio. Computational load is heavier than for LP based methods, because it is based on roots extraction from a high degree polynomial.

4 Lines of Maximum Amplitude of the wavelet transform[4]

Glottal closing instants for soft and strong voices

In this section another aspect of the phase of the glottal source is explored. The periodicity of the voice source signal is defined by the positions in time of the GCI. In addition to periodicity, another important prosodic parameter is the degree of voicing (in the simplest form, a voiced/unvoiced decision). For speech synthesis, the GCI are needed in methods based on pitch period or pitch synchronous processing. The preceding discussions also pointed out that the GCI is important for determining the "anticausal" and the "causal" parts of the glottal source signals.

Glottal closings are often points of sharp variations, or singularities in the speech signal. This is particularly the case for abrupt glottal closings, when there is no additional spectral tilt component. In this situation, GCI are corresponding to a discontinuity in the glottal flow derivative, and methods for discontinuity detection would be desirable.

On the other hand, for soft voices with low vocal effort, the glottal closing instants do not correspond to well-marked discontinuities in the glottal flow derivative. The waveform is smooth, and the spectral tilt is large. The waveform resembles a sine wave. Then instead of searching for discontinuities in the signal, it seems more important to follow the signal instantaneous phase.

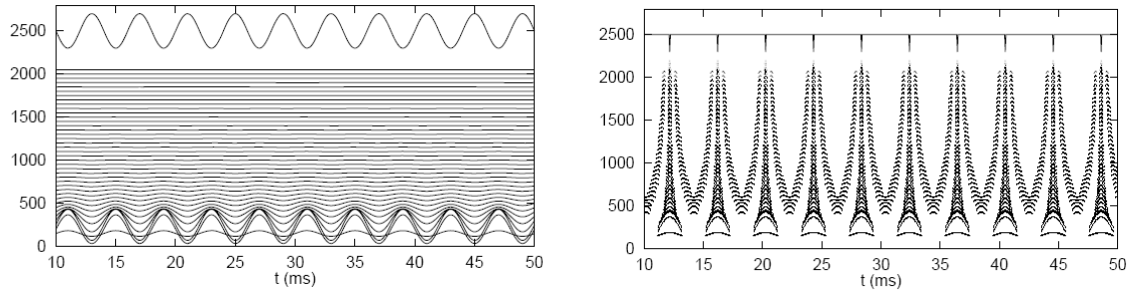

Fig. 18. Signal analysis using a non-uniform filterbank. Two extreme situations for glottal excitation signals (signal at the top of the Figures), and time-scale analysis. Left panel: sinusoidal excitation, without any singularity at glottal closing. Right panel: impulsive excitation.

The Wavelet Transform demonstrates excellent capabilities for detection of singularities in signals (Mallat & Wang, 1992). This feature has been applied to pitch detection (Kadambe and Boudreaux-Bartels, 1992). Their work is based on the dyadic wavelet transform of the speech signal. This transform is computed only for two or three small scales (high frequencies), typically 2^4, 2^5 and 2^6. Then, GCI are detected by locating the local maxima of the transform which are above a threshold level across two dyadic scales. This method works well when the speech signal contains singularities at glottal closing (Figure 18, right panel). This is not always the case, and the singularity detection is questionable for quasi-sinusoidal voice (Figure 18, left panel).

When voiced speech is seen using a non-uniform filterbank, characteristic tree-like patterns were obtained for voiced, as can be seen in Figure 18. Long lines pointing to the singularities are obtained for strong voices or signal containing singularities. On the contrary, only the first filters give a significant response to soft voice, or smooth signals. However both situations are actually encountered in speech. This is an indication that one could take advantage of the length of lines in the time-scale space for improving GCI detection.

A new algorithm for GCI detection with the help of the wavelet transform has been presented (Vu Ngoc Tuan & d'Alessandro 1999). Contrary to previous works, all the scales are actually used for analysis. Then, the high frequency features related to abrupt closures as well as low-pass quasi-sinusoidal speech signals of soft voices can be analyzed with accuracy. This is achieved by a new concepts, the lines of maximum amplitude (LOMA), which are linking amplitude maxima across scales in the wavelets transform domain.

[4] The main references for this section are: Vu Ngoc Tuan & d'Alessandro, 1999, 2000

Line of maximum amplitude of the wavelet transform

A wavelet filter-bank (6 band-pass filters centered on 4000, 2000, 1000, 500 250 and 125 Hz), with bandwidths proportional to center frequencies is used for signal decomposition. The purpose of the filter-bank is to detect the most important periodic peaks. Small peaks due to noise are present only in few high frequency (HF) filters, and are uncorrelated. On the contrary, large periodic peaks are likely to produce large amplitudes for filters at all scales. Lines of maximal amplitude (LOMA) are defined by following the amplitude maxima of the filter responses, starting at HF filters and ending at low frequency (LF) filters. The LOMA for voiced and unvoiced segments have rather different shapes. Unvoiced segments result in short HF lines. On the contrary, voiced segments are represented by long lines starting from HF filters and ending at the LF filters. For each voicing period, the LOMA are drawing a kind of tree pattern. The GCI for each period is associated to the position of the principal LOMA, taken in the highest filter. The analysis algorithm can be summarized as follows:

1. Compute a wavelet transform. The basic wavelet is chosen in such a way that the transform is equivalent to a zero-phase filter-bank. Each filter is a band-pass filter with a bandwidth proportional to its center frequency. The wavelet transform (WT) can be considered as the convolution between the signal and a dilated/compressed mother wavelet. Let x(t) be the speech signal, its WT $y_i(t)$ at scale i is given by:

$$y_i(t) = x(t) * h(\frac{t}{S_i}) \quad (11)$$

The mother wavelet is a band-pass impulse response in the form:

$$h(t) = -\cos(2\pi f_0 t)\exp(-\frac{t^2}{2\tau^2}) \quad (12)$$

Note the minus sign in the cosine. Then the wavelet analysis will have a maximum response to negative peaks. The filters impulse responses are not causal, because this is a zero-phase filter-bank. Thus, the signal and its response are in phase, and the phase of the signal can be read in the phases of the filters, at each scale. The filters frequency responses are displayed in Figure 19.

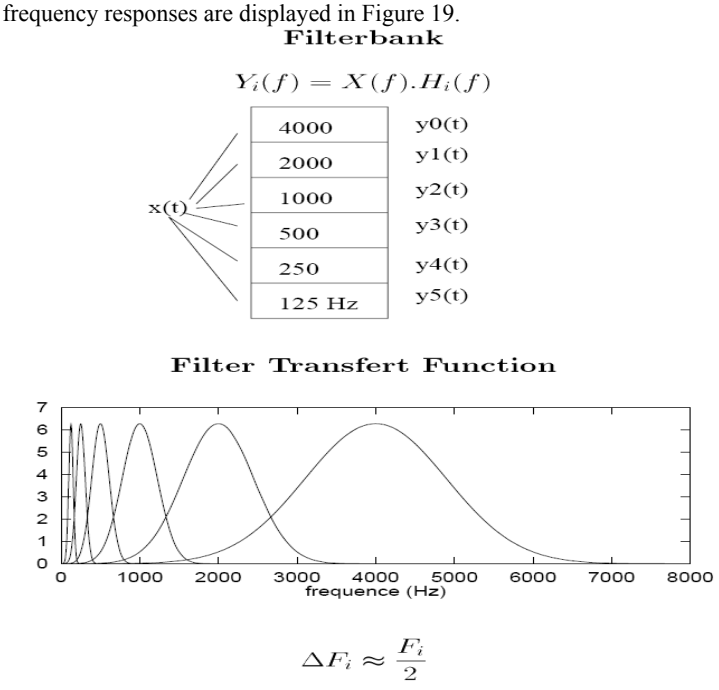

Fig. 19. Wavelet filterbank.

The frequency response of the mother wavelet is:

$$H(\nu) = |H(\nu)| = \frac{\tau}{\sqrt{1 + (2\pi(\nu - f_0)\tau)^2}} \quad (13)$$

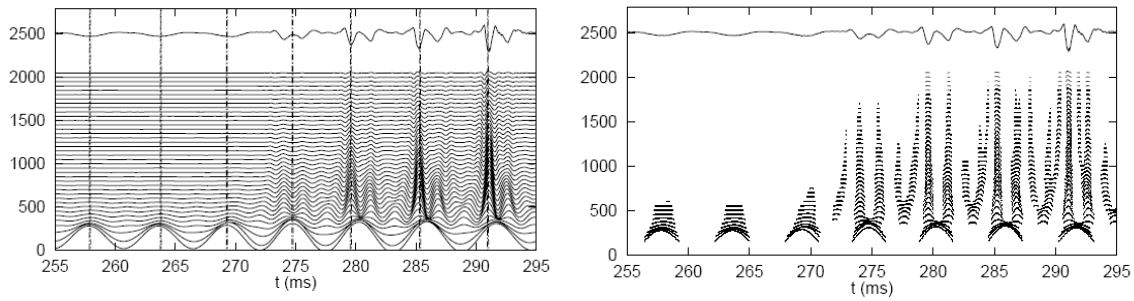

Fig. 20. Filterbank output (right panel: positive amplitudes only) for a consonant-vowel transition. More filters are used for the sake of display (from Vu Ngoc Tuan & d'Alessandro, 1999).

2. Signals at the output of these filters have local maxima (see Figure 21). These maxima are tracked across scales, starting from the highest frequency (HF=4000 Hz) towards the lowest frequencies (BF=125 Hz). Several LOMA are starting from the highest filter in a period (the number of LOMA at a given scale is roughly equal to the center frequency of this scale). However, all these lines are joined together at a unique instant in the lowest filter, for each voicing period, as there is one LOMA "tree" per period. Thus, one can gather these lines into groups, with only one group per period of voiced signal. In each group, the strength of each LOMA is computed by adding all the amplitudes along the lines. Optimal selection of maxima across scale is achieved using a dynamic programming algorithm. Then the set of optimal LOMA for a period is computed.

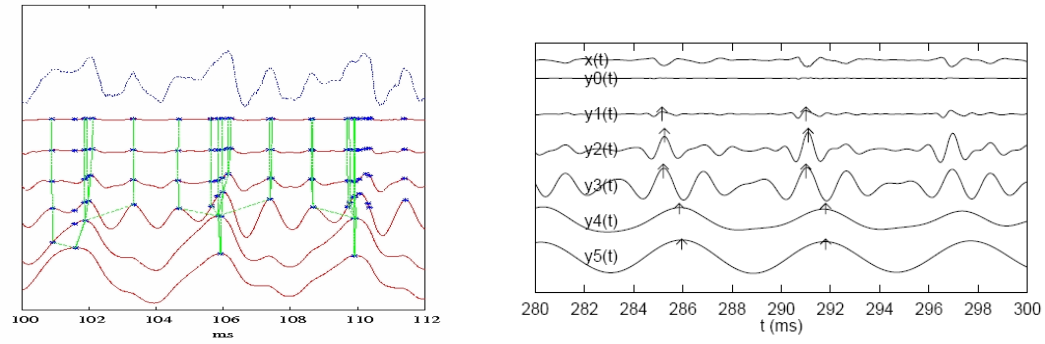

Fig. 21. Lines of Maximum Amplitude across scales. All the lines are displayed in the left panel. In the right panel, the optimal line for each period are shown.

3. The line with the highest cumulated amplitude is then chosen for representing the period. The GCI is then computed as the instant where the selected line starts at the smallest scale (highest frequency) (see Figure 21, right panel).

Comparison with electroglottography

For GCI detection algorithm evaluation, a reference is needed. The electroglottographic (EGG) reference is chosen, because it is an accurate and non-invasive GCI measurement method. A database of speech including various productions like vocal fry, modal and falsetto voices, spontaneous and read speech, male and female voices, has been recorded.

Most of the GCI peaks in the EGG derivative signal are well-defined, but some of them are too close: the time interval between two successive peaks is shorter than the period of the highest possible fundamental frequency. So we developed a simple algorithm for selection of the most prominent peaks that represent GCI. The EGG signal and the EGG derivative signal are represented in Figure 22.

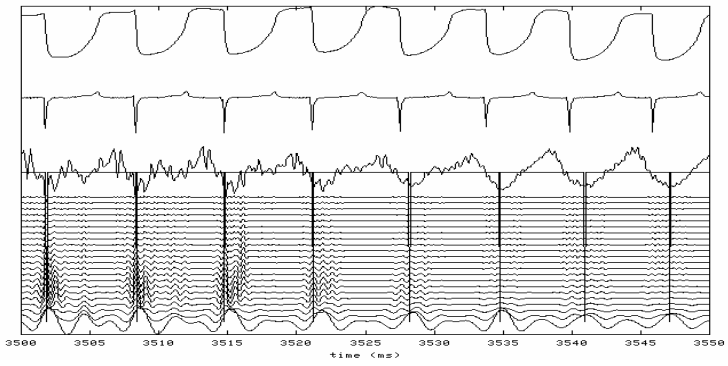

Fig. 22. Simultaneous EGG and acoustic signals analyses. EGG and DEGG (top panels), acoustic signal (middle panel), and response of the wavelet filterbank (from Vu Ngoc Tuan & d'Alessandro, 2000).

For experiments, the speech signal is sampled, and discrete-time signals are analyzed. Then the GCI cannot be determined with accuracy greater than the sampling period, and there is a slight difference between the discrete-time signal maximum and the corresponding continuous-time signal maximum. To increase time-domain accuracy, parabolic interpolation is used. Near a GCI, a parabola passing by this maximum and two adjacent points is computed. The GCI is taken at the parabola maximum. Parabolic interpolation is applied to both methods.

Acoustic signals were recorded in a sound-proof room, using a condenser microphone (Brüel & Kjær 4165) placed at 50 cm from the speaker's mouth, a preamplifier (Brüel & Kjær 2669) and a conditioning amplifier (Brüel & Kjær NEXUS 2690). EGG signals were recorded simultaneously, using a two-channel electroglottograph (EG2). The data were digitally recorded (one channel for the acoustic signal and the other one for the EGG signal). Four subjects have been recorded (two males and two females).The speakers were asked to read 3 short stories, with normal voices, then with a high pitch using falsetto, and then with a very low pitch using vocal fry. Sustained vowels and spontaneous speech (an informal conversation on daily life matters) were also recorded.

The data-base has been analyzed using both algorithms. The GCI obtained with the DEGG signals are taken as reference measures. As a matter of fact, visual inspection shows that this is true in almost all cases: when a pitch period (or a vocal pulse in case of vocal fry) is visible on the speech signal, then there is a peak in the DEGG signal.

The GCI obtained in the speech signal are delayed from the DEGG peaks, mainly because of the sound propagation time. This delay depends on the distance between the lips and the microphone, on the vocal tract group delay and on the electronic delay of the measurement apparatus. The delay is almost constant for each recording (except for the time-varying vocal tract group delay). For comparing the GCI detected by the two algorithms, the DEGG and speech analyses must be resynchronized by delaying the DEGG. This is achieved by the following procedure:

1. The DEGG signal is delayed by Td ms.
2. GCI are detected using DEGG.
3. GCI are detected using LOMA.
4. The mean difference between both sets of GCI is computed (Dm)
5. The procedure is repeated, varying Td.

The optimal delay Do between the DEGG and the speech signal is obtained for the value of Td corresponding to minimum Dm. The results of this analysis are summarized in Table 1 for 8 sustained vowels:

Tr (s)	Do (ms)	Dm (ms)	N Degg	% diff	N Speech
1.3	0.4	-0.039	206	1.9	210
2.5	2.2	0.011	349	O.8	346
1.8	1.4	0.012	175	0.5	176
3.0	3.2	-0.003	474	1.2	480
1.8	0.6	-0.038	380	0.2	379
1.5	0.1	-0.010	282	7.8	306

3.0	1.8	-0.132	517	2.1	506
2.3	0.2	-0.010	599	34.0	908

Table 1. Comparison of GCI detection using the EGG and LOMA (from Vu Ngoc Tuan & d'Alessandro, 2000).

Where Tr represents the sentence duration, Ndegg is the number of GCIs detected on the DEGG signal, Nspeech is the number of GCIs detected using wavelets, and %diff the percentage of difference between the two measures.

Except for the last example Ndegg and Nspeech are generally very close. In the last example, in many cases the second harmonic (octave) is much stronger than the first harmonic (fundamental frequency). In this situation, the wavelet algorithm tends to detect two peaks for each voicing period, instead of only one. When fundamental frequency is known (which is actually the case), these extra peaks are removed by a simple post-processing procedure. After this procedure the mean value of Dm is -0.028 ms (standard deviation 0.3 ms). This indicates that the LOMA method correctly detects the GCI, when peaks due to the second harmonic are removed by post processing.

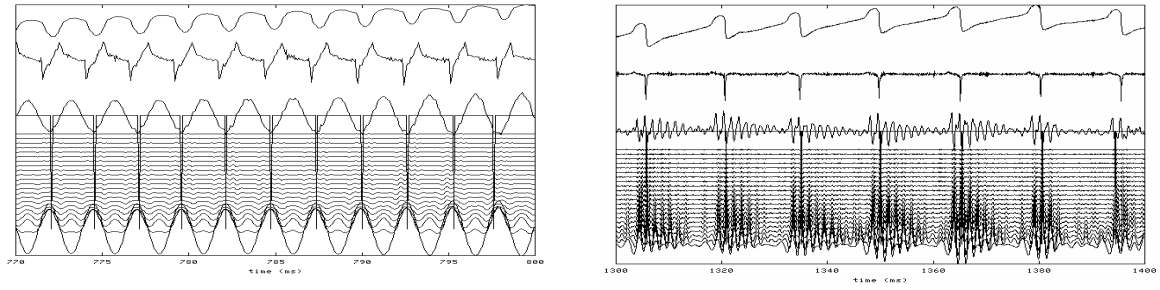

Fig. 23. Examples of EGG and wavelet GCI detection. Left panel: soft voice, "head" register, female voice, right panel, "chest" register, male voice (from Vu Ngoc Tuan & d'Alessandro, 2000).

Figure 23 is presents short segments extracted from sentences in the data-base. All the figures are presented in the same way, from top to bottom: EGG signal, DEGG signal, speech signal, wavelet filter-bank output, with a line indicating the position of the GCI detected using LOMA. The right panel Figure 23 shows a male voice, in modal register (which is the normal register for this speaker). The algorithm takes advantage of small scales (high frequencies) for accurate GCI detection. Figure 23 shows another female speaker, using her normal ("head") voice register. In Figure 23 (left), GCI detection takes advantage of large scales (low-pass frequencies).

LOMA and glottal flow parameters

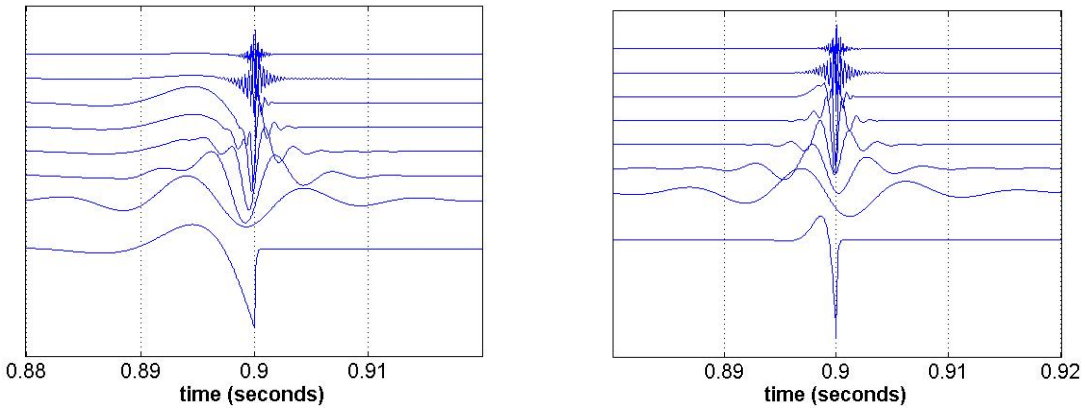

Fig. 24. Analysis of glottal pulses with different open quotients (left panel: Oq=0.8, right panel, Oq=0.3)

In general, the LOMA are not straight lines. Figure 24 shows 7-bands decomposition for a (derivative) glottal flow with open quotient 0.8 (left panel) and 0.3 (right panel). LOMA differ between these two conditions, and

then glottal flow parameter analysis using LOMA should in principle be possible. It is straightforward to obtain a first parameter: amplitude of voicing. Amplitude of voicing is computed using the energy carried by the best LOMA of each tree. When the signal is unvoiced, the lines carry very little energy: this is because in this situation, amplitude maxima are not well-organized in the time-scale space, and no strong and long lines are likely to occur. The energy is spread in a wide tree, with no strong trunk and many small branches. On the contrary, the best LOMA (i.e. the strongest trunk) corresponding to a period of voicing carries a large amount of the signal energy. The voiced/unvoiced decision can be carried out by using a simple threshold on the amplitude carried by the trunk of each tree. This simple measure is surprisingly robust.

In future work, LOMA will be used to investigate the shape of speech signal periods, particularly near GCI. It is well known that the shape of glottal closing is a strong correlate of spectral tilt and is important for studying voice quality. LOMA will also be used for speech signal modification (e.g. time and pitch scaling).

5 Conclusions

In this article recent work by the authors on voice source analysis and synthesis using methods based on the spectral phase or instantaneous phase are presented. The main results obtained are the following:

- detailed and careful consideration of the glottal flow shows that glottal flow models can be represented by causal-anticausal (mixed phase) filters
- the link between time-domain and spectral domain parameters can be worked out, equations are available for most time domain glottal flow models
- a new glottal model in the spectral domain is proposed for speech synthesis (Causal-Anticausal Linear Model, or CALM).
- Taking advantage of the mixed-phase nature of glottal flow models, a new speech representation method is proposed, the Zero of the Z Transform (ZZT). Speech analysis and synthesis using the ZZT is achieved using a simple (although computationally heavy) algorithm.
- Comparison of the ZZT and inverse filtering for source-tract decomposition indicates better performances for the ZZT, in terms of waveform and spectral distance.
- The ZZT is also useful for estimation of open quotient and asymmetry coefficient of the voice source
- Glottal closing instants correspond to specific patterns of instantaneous phases and amplitudes in the time-scale domain. These patterns can be analyzed in terms of lines of maximum amplitude (LOMA) across scales. LOMA are also providing information on the energy of the corresponding speech periods, and may be useful for further analysis of the glottal waveforms properties.
- A comparison of glottal closing detection using LOMA and EGG shows that the method performs reasonably well

References:.

1. P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering” *Speech Communication*, vol. 11, pp. 109–118, 1992.
2. P. Alku, T. Bäckström, & E. Vilkman. “Normalized amplitude quotient for parametrization of the glottal flow” *J. Acous. Soc. Am*., 112 (2):701--710, August 2002.
3. L. D. Alsteris, K. K. Paliwal. “Short-time phase spectrum in speech processing: A review and some experimental results”. *Digital Signal Processing* 17 (2007) 578–616
4. B. Bozkurt, B. Doval, C. d'Alessandro, T. Dutoit, “Improved differential phase spectrum processing for formant tracking” Interspeech 2004-ICSLP, 8th International Conference on Spoken Language Processing. Jeju Island, Korea, October 4-8, 2004 (4pages).
5. B. Bozkurt, B. Doval, C. d'Alessandro, T. Dutoit “Appropriate windowing for group delay analysis and roots of z-transform of speech signals”, EUSIPCO 2004, 12th European Signal Processing Conference, EURASIP, Vienna, Austria, September 6-10, 2004 (4 pages).
6. B. Bozkurt, B. Doval, C. d'Alessandro, T. Dutoit “A method for glottal formant frequency estimation”, Interspeech 2004-ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004 (4pages).
7. B. Bozkurt, B. Doval, C. d'Alessandro, T. Dutoit “Zeros of Z-Transform (ZZT) decomposition of speech for source-tract separation”, Interspeech 2004-ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004 (5 pages).
8. B. Bozkurt, “Zeros of z-transform(ZZT) representation and chirp group delay processing for analysis of source and filter characteristics of speech signals”, PhD Thesis, Université Polytechnique de Mons, Belgium and LIMSI-CNRS, France, October 2005.

9. B. Bozkurt, B. Doval, C. d'Alessandro, T. Dutoit "Zeros of Z-transform representation with application to source-filter separation in speech" *IEEE Signal Processing Letters,* p. 344-347, vol. 12, n°4, April 2005
10. D.G. Childers ,C.K. Lee. "Voice quality factors: analysis, synthesis and perception" *J.Acoust.Soc.Am.*Vol.90,No.5, pp.2394-2410, 1991.
11. N. d'Alessandro, B. Doval, S. Le Beux, P. Woodruff, Y. Fabre, C. d'Alessandro & Thierry Dutoit, "Realtime and Accurate Musical Control of Expression in Singing Synthesis," *Journal on Multimodal User Interfaces*, Volume 1, N°1, pp 31-39, March 2007
12. B. Doval, C. d'Alessandro, N. Henrich. "The voice source as a causal/anticausal linear filter". In proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop, pages 15--20, Geneva, Switzerland, August 2003.B.
13. Doval, C. d'Alessandro, and N. Henrich. "The spectrum of glottal flow models". *Acustica united with Acta Acustica,* 92:1026-1046, 2006.
14. G. Fant, Acoustic theory of speech production, Mouton De Gruyter; Revised edition (January 1970).
15. G. Fant, J. Liljencrants, and Q. Lin. "A four-parameter model of glottal flow". *STL-QPSR*, 4:1--13, 1985.
16. H. M. Hanson. "Glottal characteristics of female speakers : Acoustic correlates." *J. Acous. Soc. Am*., 101:466--481, 1997.
17. N. Henrich, C. d'Alessandro, and B. Doval. "Spectral correlates of voice open quotient and glottal flow asymmetry : theory, limits and experimental data". In Eurospeech 2001, Aalborg, Denmark, Sept. 2001.
18. N. Henrich, C. d'Alessandro, M. Castellengo and B. Doval "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation", *J. Acoust. Soc. Amer*., Vol. 115 (3), pp. 1321-1332, 2004.
19. S. Kadambe,G.F. Boudreaux-Bartels,1992. "Application of the wavelet transform for pitch detection of speech signals". *IEEE trans. on IT*,vol.38.No.2.pp.917-924.D.
20. D. Klatt and L. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers". *J. Acous. Soc. Am*., 87 (2):820--857, 1990.
21. S. Mallat, Wen Liang Hwang, 1992 "Singularity detection and processing with wavelets" *IEEE trans. on IT*, vol.38.no.2.pp.617-943.
22. J.D. Markel & A.H. Gray, Jr. (1976), Linear Prediction of Speech, Springer Verlag, Berlin.
23. E. Rosenberg. "Effect of glottal pulse shape on the quality of natural vowels". *J. Acous. Soc. Am*., 49:583--590, 1971.
24. N. Sturmel, C. d'Alessandro, B. Doval, "A spectral method for estimation of the voice speed quotient and evaluation using electroglottography" 7th Conference on Advances in Quantitative Laryngology, Groningen, The Netherlands, October 6-7, 2006
25. N. Sturmel, C. d'Alessandro, B. Doval, "A comparative evaluation of the Zeros of Z Transform representation for voice source estimation" Proceedings of Interspeech 2007, 27-31 august 2007, Antwerp, Belgium
26. R. Veldhuis. "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation". *J. Acous. Soc. Am*., 103:566--571, 1998.
27. Vu Ngoc Tuan & C. d'Alessandro "Robust Glottal closing Detection using the Wavelet Transform", Proceedings of Eurospeech 99 Budapest Budapest, Hungary (1999) vol.6 pages 2805-2808
28. Vu Ngoc Tuan & C. d'Alessandro "Glottal closing Detection using EGG and the Wavelet Transform" Advances in Quantitative Laryngoscopy, Voice and Speech Research Proceedings of the 4th International Workshop Jena, Germany (2000) pages 147-154
29. B. Yegnanarayana, H.A. Murthy, "Significance of group delay functions in spectrum estimation", *IEEE Trans. Signal Process*. 40 (9) (1992)