Sādhanā Vol. 36, Part 5, October 2011, pp. 601–622. © Indian Academy of Sciences

Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude

CHRISTOPHE D'ALESSANDRO* and NICOLAS STURMEL

LIMSI-CNRS, B.P. 133, F-91403, Orsay, France
e-mail: cda@limsi.fr; Nicolas@sturmel.com

Abstract. [1]Time-scale representation of voiced speech is applied to voice quality analysis, by introducing the Line of Maximum Amplitude (LoMA) method. This representation takes advantage of the tree patterns observed for voiced speech periods in the time-scale domain. For each period, the optimal LoMA is computed by linking amplitude maxima at each scale of a wavelet transform, using a dynamic programming algorithm. A time-scale analysis of the linear acoustic model of speech production shows several interesting properties. The LoMA points to the glottal closure instants. The LoMA phase delay is linked to the voice open quotient. The cumulated amplitude along the LoMA is related to voicing amplitude. The LoMA spectral centre of gravity is an indication of voice spectral tilt. Following these theoretical considerations, experimental results are reported. Comparative evaluation demonstrates that the LoMA is an effective method for the detection of Glottal Closure Instants (GCI). The effectiveness of LoMA analysis for open quotient, amplitude and spectral tilt estimations is also discussed with the help of some examples.

Keywords. Voice source analysis; glottal closure instants; voice open quotient; voicing amplitude; voice spectral tilt; wavelet analysis.

1. Introduction

This article presents a multi-scale phase-based framework for analysing voice source features in speech data. Robust voice source analysis still remains a challenging and important issue for studying voice quality and vocal expression, vocal function and disfunction, and for many

[1]Portions of this work were presented in N. Sturmel, C. d'Alessandro, F. Rigaud, Glottal closure instant detection using Lines of Maximum Amplitudes (LoMA) of the wavelet transform, *Proc. IEEE-ICASSP'09*; Vu Ngoc Tuan, C. d'Alessandro Glottal Closure Detection using EGG and the Wavelet Transform Proc. Workshop Adv. in Objective Laryngoscopy, *Voice and Speech Res.* 2000. Vu Ngoc Tuan, C. d'Alessandro. Robust glottal closure detection using the wavelet transform, *Proc. ISCA-Eurospeech'99.*

*For correspondence

speech processing applications. Several time-domain and spectral features of the glottal source are important for speech perception and processing: the instant of maximum vocal tract excitation or glottal closure instant (GCI, correlated to fundamental frequency (F_0) and perceived pitch), the amplitude of excitation and the source spectral richness (correlated to vocal effort), the glottal open quotient (correlated to the press-lax vocal dimension) (Childers & Lee 1991; Fant 1993; Fant 1997; Alku *et al* 1997; d'Alessandro 2006; Gobl & Chasaide 2003). These voice source features are considered in a new framework, based on the lines of maximum amplitude (LoMA) in a time-scale representation. The LoMA concept has been introduced in our previous work (Tuan & d'Alessandro 1999, 2000; Sturmel *et al* 2009). This method has been successfully applied to the problem of GCI detection in speech. The main contribution of this article is to further develop the application of the LoMA method to GCI analysis and extend it to other voice source features such as glottal flow amplitude, spectral tilt and open quotient.

1.1 *Auditory-based wavelet representation*

Time-scale analysis entertains close relationships with auditory perception. According to the tonotopic organisation of the inner ear, the first stages of auditory perception can be modelled using a non-uniform filterbank. These auditory filters are organized according to some sort of psychological scale, such as third-octave, Bark, Mel or Equivalent Rectangular Bandwidth (ERB) scales. The filtering process is considered to be a simulation of the response of the basilar membrane to sound stimulation. This is the 'place coding' or tonotopic theory of auditory perception. After filtering, patterns of neurone firings at different places in the auditory nerve give birth to trains of spikes that are (statistically) synchronized with maxima in the basilar membrane vibration pattern. Timing and synchrony of neuron firing patterns allow for 'time coding' of auditory events. These principles, i.e., filtering and maximum amplitude detection at the output of the filterbank, have been implemented in various auditory models, sometimes called 'cochleograms' (see several approaches for computing cochleograms in (Cooke & Beet 1993)). Along these lines, Patterson (1987) proposed the 'pulse ribbon model' of auditory perception, and showed that this model was able to explain many aspects of phase perception for periodic signal. This model is essentially an auditory filterbank followed by a maxima detection module. The output of the model is a time–place 'ribbon' of pulses.

Auditory frequency scales are close to logarithmic, at least above about 1000 Hz. Then, auditory filtering is somewhat analogous to time-scale analysis. Speech representation in the time-scale domain with the help of an auditory-inspired wavelet transform has been explored by several authors (Irino & Kawahara 1993; d'Alessandro 1993). The auditory wavelet filterbank can be used for visualization of speech signals in the form of 'auditory-wavelet' spectrograms (or 'scalograms'). An example of such a representation is given in figure 1. This picture represents the output of a Bark scale zero-phase filterbank (implemented using a Hamming window and short-term Fourier transform filtering). The bottom panel represents the filters outputs. The top panel displays only positive part of the filter outputs. Characteristic 'tree' patterns are obtained for voiced speech, as a result of the multi-scale analysis of this quasi-harmonic signals. It should be noted that in this example, the fundamental (first harmonic) is hardly visible on the bottom panel, although it appears clearly in the top panel.

1.2 *Previous work on voice singularities detection using wavelets*

The analogy between auditory models and the wavelet transform provides an extended mathematical framework for signal analysis along these principles. If one defines a set of fixed

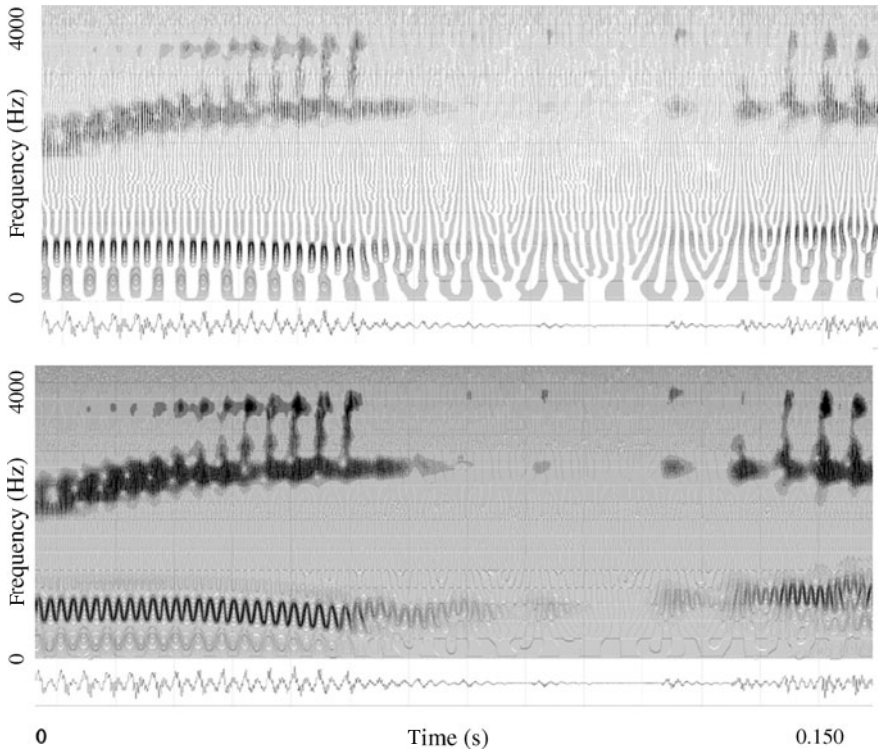

Figure 1. Output of a Bark scale zero-phase filterbank. The bottom panel represents the filters outputs. The top panel displays only positive part of the filter outputs. 'We hear' spoken by female voice.

scales, the wavelet transform can be interpreted as a non-uniform filterbank. As for voice source analysis, the wavelet transform has been mainly applied to singularities detection. Mallat & Zhong (1992) and Mallat & Hwang (1992) showed that wavelet transform modulus maxima are organized as edges in the time-scale representation, and are pointing to the singularities of the signal. An application of this feature to pitch detection was first proposed by Kadambe & Boudreaux-Bartels (1992). More recently, Bouzid & Ellouze (2007, 2009) revisited Kadambe and Boudreaux-Bartels' method, using a multi-scale product of the wavelet transform, inspired by image processing works. These analyses are based on the dyadic wavelet transform computed only for two or three scales encompassing two or three octaves above the average F_0. Then, GCIs are detected by locating local maxima of the transform using a multi-scale product. Local minima of the multi-scale product are associated with glottal opening instants. The voice open quotient (ratio of the difference between opening and closure over the fundamental period) can be derived from these measures. Comparison with an electroglottographic (EGG) reference showed good agreement.

1.3 *Voice source analysis using LoMA*

Inspired by the tree patterns observed in figure 1, the concept of lines of maximum amplitude across scales in the wavelet transform domain has been introduced (Tuan & d'Alessandro 1999).

LoMA makes use of all the scales for analysis of the tree patterns. This concept has been introduced for GCI detection. The results obtained were compared to an EGG reference (Tuan & d'Alessandro 2000) and the DYPSA (DYnamic programming projected Phase-Slope Algorithm) method (Sturmel *et al* 2009).

It appears that the LoMA contains much more information than only singularities positions. Application of LoMA to voiced speech analysis is further extended in this article.

Section 2 introduces the LoMA concept and the algorithm developed for LoMA analysis. The LoMA patterns in the time-scale domain are related to the linear acoustic model of speech production in section 3. The principle for voice source feature analysis using LoMA is theoretically derived in this section. The remaining sections report on experiments using this theory. GCI detection using LoMA is discussed in section 4. The LoMA-derived GCI are compared to an EGG reference, to videoendoscopy, and to data obtained with the help of the DYPSA method (Naylor *et al* 2007).

Observations of auditory spectrograms and section 3 show that LoMA morphology, and particularly their length and their shape, vary as a function of voice quality. Voice quality features estimation with LoMA is experimentally studied in section 5. It is shown that the voice open quotient, the amplitude of voicing and the voice spectral tilt can be derived from LoMA. Section 6 concludes this article.

2. Lines of maximum amplitude in the time-scale domain

The lines of maximum amplitudes in the time-scale domain are built on the output of a non-uniform filterbank. A zero-phase filterbank is better suited for the purpose of speech analysis. The global response of the filterbank should be flat. The wavelet transform gives a simple and elegant solution for implementing a non-uniform zero-phase filterbank.

2.1 *The wavelet transform as a zero-phase filterbank*

Non-uniform filtering is commonplace since a long time in acoustic signal processing. Nowadays, non-uniform filtering can be considered in the framework of the wavelet transform. The continuous wavelet transform (WT) can be considered as the convolution between the signal and a dilated/compressed mother wavelet. Let $s(t)$ be the speech signal, its WT $y_i(t)$ at the i^{th} scale s_i is given by:

$$y_i(t) = s(t) * h\left(\frac{t}{s_i}\right) = s(t) * h_i(t), \tag{1}$$

then $h_i(\frac{t}{s_i})$ can be interpreted as a filter impulse response and $y_i(t)$ as the response of this filter to signal $x(t)$. Many possible choices are possible for h. In this study, we choose a simple Gaussian pulse:

$$h(t) = -\cos(2\pi f_{\text{w}} t) \exp\left(-\frac{t}{2\tau^2}\right), \tag{2}$$

where f_{w} represents the centre frequency of the smallest scale, and $\tau = \frac{1}{2f_{\text{w}}}$. Because of the minus sign in the cosine, the wavelet analysis will have a maximum response to negative peaks. This is because glottal closures are in principle corresponding to negative peaks in the speech signal.

The filters impulse responses are not causal, because of the zero-phase condition. Then, there is no phase delay between the signal and filter outputs. An impulse, for instance, gives a synchronized response at each scale.

The number of filters used depends on the application. For analysis, a dyadic WT with $s_i = 2^i$; $i = 0, 1, 2, 3, 4, 5$ seems sufficient. In the experiments, the signals were sampled at $f_e = 8$ kHz or $f_e = 16$ kHz, and then $f_w = 4$ kHz or $f_w = 8$ kHz. $y_i(t)$ represents one of the outputs of a six band filterbank, centered at frequencies f_i: 4000, 2000, 1000, 500, 250, 125 Hz, with -3 dB bandwidths of $\approx 0.5 \times f_i$. For visualization, a larger number of filters is needed (for instance three to six filters by octave) because it makes the lines more visible.

Scaling of the impulse responses results in scaling of the filter gains, according to:

$$h_i(t) = h\left(\frac{t}{s}\right) \xrightarrow{\text{FT}} s\widehat{h}(sf) = \widehat{h_i}(f), \tag{3}$$

where $\widehat{h}$ represents the Fourier transform (FT) of h.

Figure 2 shows the output of a zero-phase wavelet filterbank (in this case using 35 filters) to a periodic train of pulses. Only amplitudes above a threshold ε, $0 < \varepsilon \ll 1$ are displayed. This shows that lines of positive amplitudes are converging to the peak position. The lines corresponding to maximum amplitudes for each period and each scale are straight lines, pointing towards the signal singularities. The next section discusses construction of these lines.

2.2 *Lines of maximum amplitudes of the wavelet transform*

For each speech period, lines of maxima across scales are searched. The amplitude maxima of the wavelet transform are computed at each scale. These local maxima are defined as any point η at scale i such that $y_i(\eta) > y_i(t)$ when t belongs to the right or left neighbourhood of η.

For analysing events in the time-scale domain, the next step of the method aims at organizing amplitude maxima into lines of maximum amplitude across scales. A single characteristic line is associated with each pitch period. This is achieved using a dynamic programming algorithm. Dynamic programming is a two-step process: in the first step, all the lines of maximum amplitude from the smaller scale down to the larger scale are built. Then backtracking is used for finding the optimal line, i.e., the line cumulating the maximum amplitudes.

Let $M_a(j, i)$ represent the j^{th} amplitude maxima at scale i. For $f_e = 8$ kHz, scales are ordered from 0 (centre frequency 4000 Hz) to 5 (centre frequency 125 Hz). Let $L_m(j, i)$ be the LoMA that is searched. The search begins at scale 0, and the LoMA is built up to scale 5. Accumulated amplitudes $A_c(j, i)$ along LoMA are computed using the following local equations:

$$A_c(j, i) = \max \begin{cases} A_c(jl, i+1)/lw + M_a(j, i) \\ A_c(j, i+1) + M_a(i, j) \\ A_c(jr, i+1)/rw + M_a(j, i) \end{cases}, \tag{4}$$

where jl (resp. jr) is the index of the amplitude maximum in the left (resp. right) neighbourhood of $M_a(j, i)$ at scale $i + 1$, and where lw (resp. rw) is a weighting factor, taken as the absolute value of the difference between the maxima positions in time, j and jl (resp. jr): $lw = |j - jl|$ (resp. $rw = |j - jr|$).

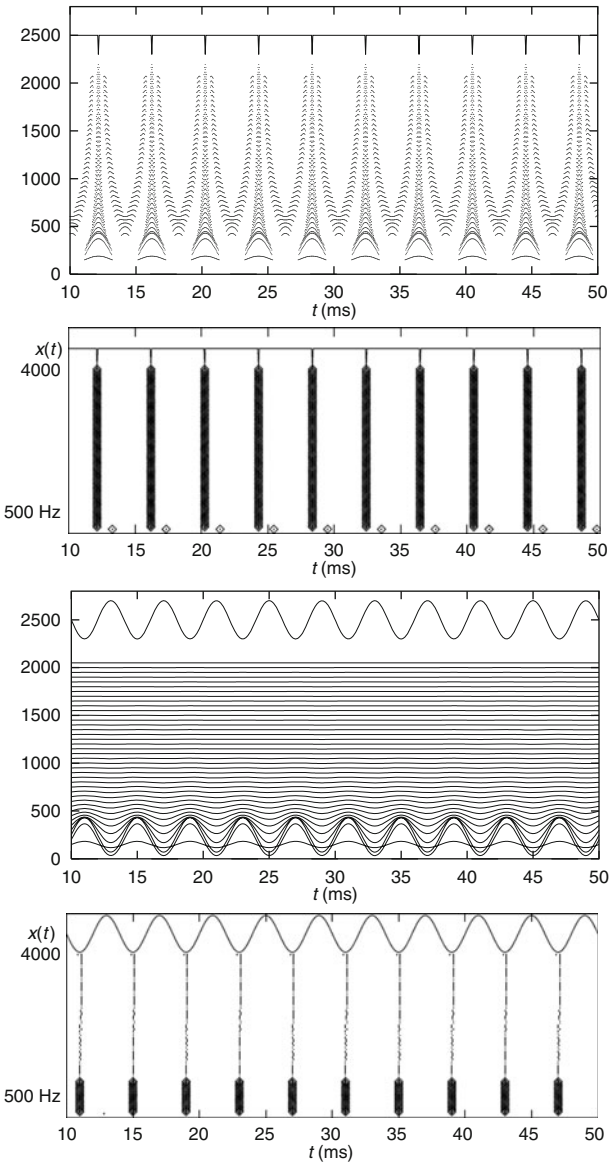

Figure 2. Output of the wavelet filterbank for a Dirac pulse train (top) and a sinusoid (second from bottom) with fundamental frequency: 200 Hz. Thirty-five filters in the range of 100–2100 Hz are used. Only the positive values of the responses are plotted for the Dirac pulse train. The optimal LoMA detected is also plotted (second from top for the Dirac pulse train, bottom for the sinusoid).

Accumulated amplitudes are computed from the smaller scale to the larger scale. At the larger scale, the LoMA are searched from the array of accumulated amplitudes using back-tracking. Following the standard dynamic programming procedure, maxima are chained back from large to small scales to build the lines. The dynamic programming procedure used here for building LoMA is close to the classical dynamic time warping procedure used in speech recognition.

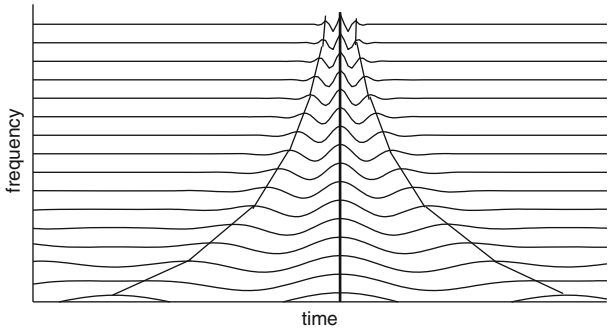

Figure 3. LoMA found for a Dirac pulse corresponding to figure 2.

The number of LoMA in the time-scale domain depends on the larger scale chosen: there is one LoMA for each maximum at this scale. Each LoMA is characterized by its chain of maxima and by its weight (accumulated amplitude maxima along the line).

Examples of LoMA for a Dirac pulse train and a sinusoid are displayed in figure 2. Figure 3 displays in more detail the analysis of a Dirac pulse. Three lines linking maxima are built. The optimal line, cumulating the maximal amplitudes (thick line), is a straight line, because all the zero-phase filter responses are synchronized.

An example for a voiced speech segment is presented in figure 4. In contrast to figure 3, because of the phase delays of the signal for different scales, LoMA are not straight lines. In this example, seven scales are used.

The lines are organized in packets, or trees, as can be seen in figure 4. Several separated lines at a smaller scale are merging and form a common line at the scale immediately larger. Ideally, for voiced speech, exactly one tree is expected for each voicing period.

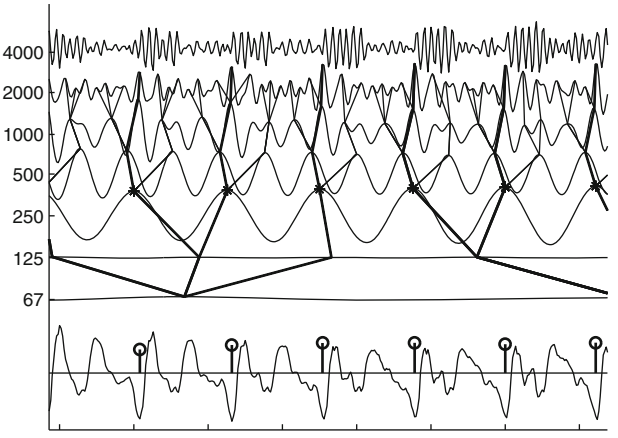

Figure 4. LoMA trees and optimal LoMA for a voiced speech segment.

3. LoMA and voiced speech production

3.1 *Linear model of speech production*

In this section, the linear model for voiced speech production (Fant 1960; Flanagan 1972) is reviewed, in order to interpret the LoMA and to link features observed on the LoMA to voice source features.

The linear model of voiced speech production writes:

$$s(t) = \sum_n \delta(t - nT_0) * g(t) * v(t) * l(t) \tag{5}$$

$$= \sum_n \delta(t - nT_0) * d_g(t) * v(t), \tag{6}$$

where $s(t)$ is the speech signal, $v(t)$ is the impulse response of the vocal tract system, $T_0 = 1/F_0$ the fundamental period and $g(t)$ the glottal flow component. The lip radiation component l can be approximated by a first-order high-pass filter, close to a derivative filter. Then, it is usually combined with the glottal pulse component to form the glottal flow derivative (GFD) d_g.

In the spectral domain the model is:

$$\widehat{s}(\omega) = |\,\widehat{s}(\omega)\,|\, e^{j\theta(\omega)} \tag{7}$$

$$= \frac{1}{T_0} \sum_n \delta(2\pi f - 2n\pi F_0) \widehat{d_g}(\omega) \widehat{v}(\omega)$$

$$= \frac{1}{T_0} \sum_n \delta(2\pi f - 2n\pi F_0)$$

$$|\,\widehat{d_g}(\omega)\,|\, e^{j\theta_{d_g}(\omega)}\, |\,\widehat{v}(\omega)\,|\, e^{j\theta_v(\omega)}. \tag{8}$$

For voiced speech, one can assume that the vocal tract v is an all-pole filter, with N pairs of poles $\hat{z}_i$ et $\hat{z}_i^*$ corresponding to spectral formants:

$$\widehat{v}(\omega) = \frac{K e^{-jN\omega}}{\prod_{i=1}^{N} (1 - \hat{z}_i e^{-j\omega})(1 - \hat{z}_i^* e^{-j\omega})}. \tag{9}$$

The poles can be expressed as:

$$z_i, z_i^* = \exp[-\pi B_i T \pm 2j\pi f_i T], \tag{10}$$

where K is a gain constant, B_i represents the formant -6dB bandwidth and f_i the formant centre frequency. It is a minimum phase system.

3.2 *Glottal flow derivative*

The glottal flow derivative component is generally described in time domain using pulse-like waveforms. According to a recent study (Doval *et al* 2006) comparing the main GFD models proposed in the literature, this component can be described by the following parameters.

(i) The fundamental period T_0, or pulse duration.
(ii) The maximum excitation E, or maximum of the GFD.
(iii) The open quotient O_q, or ratio of the open phase over the pulse duration. The GCI is located at $O_q T_0$ if the pulse begins at time 0.

(iv) The asymmetry coefficient, or ratio of the opening and closing phases. In some models, this ratio is fixed (e.g. to 0.66 in the KLGLOTT88 model (Klatt & Klatt 1990)).
(v) The glottal spectral tilt, often implemented as a low-pass filter. This parameter is linked in the LF (Liljencrants–Fant) model (Fant 1993, 1997) to the smoothness of the wave after the GCI (or return phase, from maximum excitation to 0).

In the spectral domain, the GFD is equivalent to a low-pass filter. If the GCI is taken as the reference point, it is a causal/anticausal low-pass filter (Doval *et al* 2006). This filter exhibits a low-frequency spectral peak, the so-called glottal formant, and an additional attenuation at medium or high frequency corresponding to the source spectral tilt. The maximum excitation is highly correlated to sound pressure level (Gauffin & Sundberg 1989). The glottal formant frequency is mainly controlled by the open quotient while its bandwidth is mainly controlled by the asymmetry coefficient. The glottal formant roughly takes place between the first and the fourth harmonic. The spectral tilt part of the spectrum behaves similar to a first order (or second order) low-pass filter which results in a -12 dB/oct (or -18 dB/oct) slope in the GFD spectrum. Only a summary can be given here, the interested reader will find more details in Doval *et al* (2006).

For the sake of illustration, the KLGLOT88 model (Klatt & Klatt 1990) is considered in the following. The KLGLOT88 glottal flow derivative is given by:

$$d_{\mathrm{g}}(t) = \begin{cases} 2a(t+T) - 3b(t+T)^2 & -T \le t \le 0 \\ 0 & 0 \le t \le T_0 - T \end{cases} \tag{11}$$

with $T = O_{\mathrm{q}} T_0$, $a = \frac{27}{4}\frac{AV}{O_{\mathrm{q}}^2 T_0}$ and $b = \frac{27}{4}\frac{AV}{O_{\mathrm{q}}^3 T_0^2}$.

It should be noted that in Eq. 11, the GCI is taken as time reference ($t = 0$), in contrast to the original presentation of the model. The advantage of the GCI-centred formulation is that one can interpret the GFD model as a mixed phase (causal–anticausal) linear system (in the original presentation, it is rather a non-linear excitation waveform). Then the source d_g and filter v components of Eq. 8 can be combined in a global vocal linear filter x:

$$\widehat{x}(\omega) = \widehat{d_{\mathrm{g}}}(\omega)\widehat{v}(\omega) = \mid \widehat{x}(\omega) \mid e^{j\theta_{\mathrm{x}}(\omega)}. \tag{12}$$

3.3 *Phase delay and group delay*

It is interesting to consider the phase delay and group delay of this filter. They are derived from the phase spectrum according to:

$$\tau_\phi = \frac{-\theta_{\mathrm{x}}(\omega)}{\omega} \tag{13}$$

$$\tau_g = \frac{-d\theta_{\mathrm{x}}(\omega)}{d\omega}. \tag{14}$$

The amplitude spectra, phase delay and group delay of one period of the GFD according to the models in Eqs. 11 and 12 are plotted in figure 5, using the KLGLOTT88 model ($T_0 = 10$ ms, $O_{\mathrm{q}} = 0.3$, no spectral tilt filter (abrupt closure)). This source component is filtered by a /a/ vowel (f_i, B_i = (800,150); (1200, 270); (2600,180); (3500, 200); (5000, 280)).

The phase delay line and group delay are plotted with frequencies on the y-axis and time delay on the x-axis. With this representation, the phase delay and group delay lines are vertical, as is the LoMA for one period.

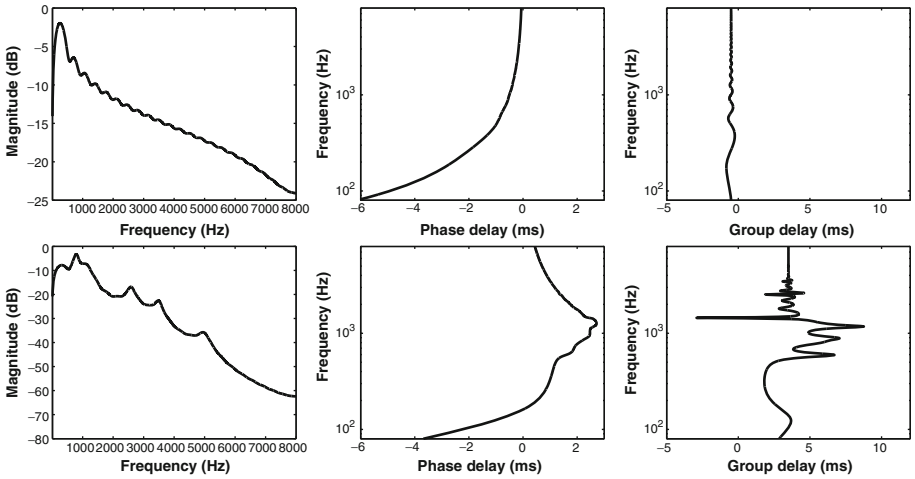

Figure 5. Magnitude spectrum, phase delay and group delay for a GFD (top row), and for a 5 formants synthetic vowel /a/, using the same GFD as source component.

Consider first the top row of figure 5. The phase delay is larger for low frequencies, giving the typical slanted shape of the LoMA. The group delay is oscillating around a constant delay, with a small negative delay for the GFD (remember that the GCI is taken at time 0, then the source is anticausal, and precedes the GCI), and a larger, positive, group delay after vocal tract filtering. The bottom row of figure 5 shows the effects of vocal tract filtering on the phase delay and group delay. Additional deviations result from the formant phase spectra.

It is interesting to study the phase and group delays of the signal at the output of the wavelet filterbank. When observed through the wavelet filterbank, the signal writes:

$$s(t) = \sum_i h_i(t) * \left(\sum_{\mathrm{n}} \delta(t - nT_0) * x(t) \right) \tag{15}$$

$$= x(t) * \left(\sum_i \sum_{\mathrm{n}} \delta(t - nT_0) * h_i(t) \right). \tag{16}$$

As the filter bank is zero-phase, $\widehat{h_i}(\omega) = \mid \widehat{h_i}(\omega) \mid$:

$$\widehat{s}(\omega) = \sum_i \widehat{h_i}(\omega) \sum_{\mathrm{n}} \delta(2\pi f - 2n\pi F_0) \widehat{x}(\omega) \tag{17}$$

$$= \mid \widehat{d_{\mathrm{g}}}(\omega) \widehat{v}(\omega) \mid e^{j(\theta_{d_{\mathrm{g}}}(\omega) + \theta_{\mathrm{v}}(\omega))} \sum_i \sum_{\mathrm{n}} \delta(2\pi f - 2n\pi F_0) \mid \widehat{h_i}(\omega) \mid . \tag{18}$$

3.4 *Interpretation of low frequency phase delay*

Consider first the situation for low frequencies (large scales). At the scale containing the fundamental frequency component, the wavelet analysis is narrow band, and then it selects generally

only one spectral component (see for instance the first and second harmonics in the first part of figure 1):

$$\sum_{\mathrm{n}} \delta(2\pi f - 2n\pi F_0) \mid \widehat{h_i}(\omega) \mid = |\widehat{h_i}(2\pi F_0)|\delta(2\pi(f - F_0)), \tag{19}$$

then in this band the signal is:

$$\begin{aligned} \widehat{s_i}(\omega) &= \mid \widehat{d_{\mathrm{g}}}(\omega)\widehat{v}(\omega) \mid e^{j(\theta_{d_{\mathrm{g}}}(\omega)+\theta_{\mathrm{v}}(\omega))} \\ & \quad |\widehat{h_i}(2\pi F_0)|\delta(2\pi(f - F_0)) \end{aligned} \tag{20}$$

in the time domain:

$$s_i(t) = x(t) * |\widehat{h_i}(2\pi F_0)| \cos(2\pi F_0 t), \tag{21}$$

then $s_i(t)$ corresponds to the output of the filter $x(t)$ driven by a sinusoidal signal. In this situation, as the sinusoidal envelope is constant, the output signal:

$$s_i(t) \simeq \mid \widehat{x}(2\pi F_0) \mid |\widehat{h_i}(2\pi F_0)| \cos(2pi F_0(t - \tau_\phi) + \theta_x(2\pi F_0)). \tag{22}$$

It should be noted that only the phase delay appears: the group delay does not appear because the sinusoidal signal has a constant amplitude (the reader is reffered to (Papoulis 1984, p. 124) for details).

Considering only the source component, the LoMA at F_0 gives a direct measurement of the phase delay τ_ϕ of the speech signal. As the source component dominates the speech signal for low frequencies, the LoMA phase delay at F_0 can be used for estimation of the source phase delay, despite the phase delay of the vocal tract.

3.5 *Interpretation of high frequency phase and group delay*

For small scales, the wavelet filters are wide-band. Then several harmonics are merged in the periodic excitation signal. These harmonics are beating, with an envelope modulation $A_i(t)$, a carrier frequency ω_i, and a phase ϕ_i:

$$s_i(t) = x(t) * \sum_{n=I_0}^{n=I_1} |\widehat{h_i}(2n\pi F_0)| \cos(2n\pi F_0 t) \tag{23}$$

$$= x(t) * A_i(t) \cos(\omega_i t + \phi_i) \tag{24}$$

assuming that the signal has a slowly changing amplitude envelope $A_i(t)$, relative to the change of phase ω_i of the sinusoid, this yields (Papoulis 1984, p. 124):

$$s_i(t) \simeq |\widehat{x(\omega_i)}| A_i(t - \tau_{\mathrm{g}}) \cos(\omega_i(t - \tau_\phi)). \tag{25}$$

As the phase delay decreases in $1/f$, it becomes small for small scales. The group delay is almost constant for high frequencies (small scales), as seen in figure 5. This means that the time-domain maxima for small scales, close to the maxima of A_i are synchronized. They are close to the GCI. As noted earlier, this property has already been exploited in a number of studies for GCI detection using multi-scale analysis.

The phase delay of the first harmonic relative to the GCI can be estimated on the LoMA. Considering the delay between the GCI and the local maximum at the fundamental frequency, the LoMA phase delay factor Lpf is defined by:

$$\mathrm{Lpf} = \tau_{\mathrm{p}}(F_0) - \mathrm{GCI}. \tag{26}$$

In turn the Lpf can be used for estimation of the voice open quotient O_q, because O_q can be related to the phase (and then the phase delay) of the first harmonic, using closed form expressions for the spectrum of glottal flow models (Doval *et al* 2006). Estimation of open quotient with the help of Lpf is discussed in section 5.

3.6 *LoMA amplitude and maximum excitation*

The LoMA for each period in the time-scale domain provides an elegant representation of the relative strength of each period. The total amplitude on the LoMA of one period is an estimation of the total excitation amplitude in one period P, because it is the sum of the maximum filter responses for all scales. A LoMA amplitude factor Laf is defined as the amplitude accumulated across scales along the LoMA as:

$$\text{Laf} = \sum_i \max_{t \in P}(|h_i(t) * x(t))|). \tag{27}$$

This factor is proportional to the amplitude of excitation in the vicinity of the GCI, and then gives an indication of the voiced period maximum excitation. This is discussed in section 5.

3.7 *Spectral tilt and LoMA centre of gravity*

The repartition of the energy along the LoMA is also a measure of interest. Considering voice spectral tilt as a low-pass filter attenuating the GFD, the amplitude along the LoMA decreases for small scales when spectral tilt increases.

A 'long' LoMA, indicates a spectrally rich voicing period, and conversely, a 'short' LoMA indicates a spectrally poor period. This is illustrated in figure 2 for a sinusoid and a Dirac pulse train, and in figure 6, for a voiced speech segment (voiced fricative to vowel transition). A measure of this repartition of energy along the LoMA is given by the LoMA centre of gravity:

$$\text{Lcg} = \frac{\sum_i \max_{t \in P} |h_i(t) * x(t))| f_i}{\sum_i f_i}, \tag{28}$$

where P is the time interval corresponding to the period, and f_i the wavelet filter centre frequencies.

This factor should give an indication of the voiced period strength or spectral tilt. This is discussed in section 5.

The effect of the vocal tract formant structure is illustrated in figure 5. The amplitude and spectral centre of gravity factors are influenced by the formant structure, and are likely to vary according to the vowel. Then, these measures will give a global indication of maximum excitation and spectral tilt, rather than an analysis of the source only.

4. Glottal closure detection using LoMA

In this section, GCI detection using LoMA is presented and evaluated. Several types of methods have been proposed for GCI detection, and it is important to compare the proposed method to previous studies. LoMA GCI detection is compared to the DYPSA method (Naylor *et al* 2007).

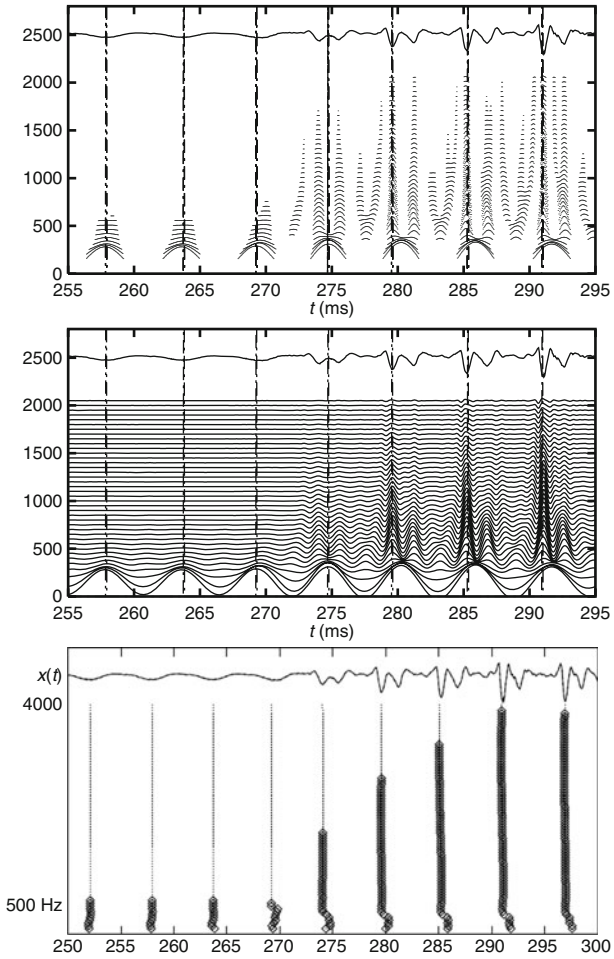

Figure 6. Output of the wavelet filterbank for a speech signal (voiced fricative/vowel transition, male voice). Thirty-five filters are used, in the range of 100–2100 Hz. To enhance the lines of maximum amplitude, only the positive values of the responses are plotted in the top panel.

4.1 *Algorithm for GCI detection*

GCIs are computed as the extremities of the optimal LoMA (smallest scale) for each pitch period (this is illustrated in figure 4). The algorithm for GCI detection can be described as follows:

(i) Pre-processing: estimate F_0 for the segment studied. A very accurate estimation is not needed: coarse estimation (within an octave) is sufficient, because the aim is only to define the largest scale, i.e., the filter containing the fundamental frequency.

(ii) Compute a dyadic wavelet transform (octave-band). The basic wavelet is chosen in such in a way that the transform is equivalent to a zero-phase filterbank. Each filter is a band-pass filter with a band-width proportional to its centre frequency.

(iii) Select the smallest scale (usually centred around the 4 kHz band) with significant amplitude. Detect all the time-domain maxima at this scale (local maxima between two zero-crossings).
(iv) For each of these maxima, build a LoMA according to the local dynamic programming equations, descending the scales down to the largest scale.
(v) Using prior pitch information, the scale containing the first harmonic is determined. The LoMA for each pitch period at this scale are selected.
(vi) The optimal LoMA for each period is determined using backtracking. Then the GCI for this period is defined as the time position along the optimal LoMA at the smallest scale.
(vii) Post-processing. Most of the errors observed are GCIs in excess, i.e., situations where two optimal LoMA are detected for the same pitch period. Two heuristics are employed for sorting out this type of errors: a period-to-period change in pitch of more than 30%, or a change in accumulated amplitude of more than 50% between two LoMA.

For experiments, discrete-time speech signals are analysed. Then, the GCI cannot be determined with an accuracy greater than the sampling period (the sampling period is 125 μs at 8 kHz). Parabolic interpolation can be used for increasing the GCI estimation accuracy. Near a GCI, a parabola passing by this maximum and two adjacent points is computed. The GCI is taken at the parabola maximum.

4.2 *Evaluation of GCI detection using an electroglottograph*

In this part, the results obtained with the LoMA algorithm and an electroglottographic (EGG) reference are compared. These results were first partly presented in Tuan & d'Alessandro (2000). GCI are detected by two methods. DEGG GCI represent GCIs detected on the derivative of the EGG (DEGG) signal using peak detection and thresholding (Henrich *et al* 2004). LoMA GCI represent GCIs detected on the speech signal using LoMA.

A database of speech, including various types of voice production (vocal fry, modal and falsetto voices, spontaneous and read speech, male and female voices) has been recorded. 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). Electroglottographic signals were recorded simultaneously, using a two-channel electroglottograph (EG2). The data were recorded (one channel for the acoustic signal and the other one for the EGG signal) using a DAT-recorder (PORTADAT PDR1000, 16 bits/16 kHz)).

Four subjects have been recorded (2 males and 2 females). The speakers were asked to read three 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 (figure 7).

The GCIs obtained with the DEGG signals are considered as true GCIs and taken as reference. The GCIs 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 (excepted the time-varying vocal tract group delay). To compare the GCI detected by the two algorithms, the DEGG and speech analyses must be resynchronized by delaying the DEGG. This is achieved by maximization of the correlation between the GCI trains obtained by DEGG and LoMA, as a function of the delay. Figure 8 displays the EGG, DEGG, speech and wavelet filterbank signals synchronized according to this procedure.

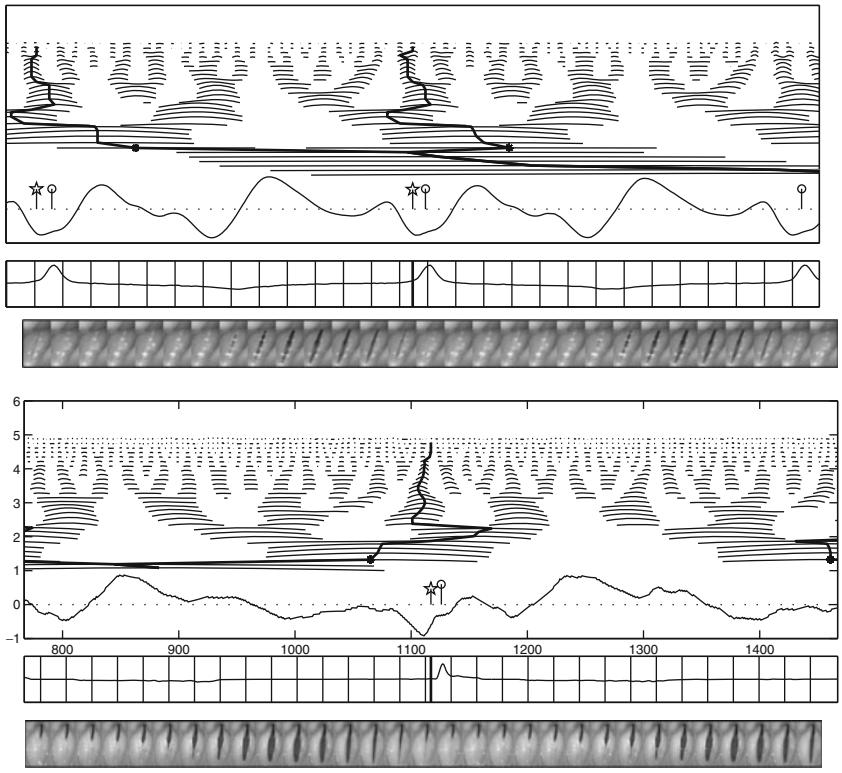

Figure 7. From bottom to top, high speed glottal imaging, DEGG, speech signal and LoMA, for a high (top panel) and a low (bottom panel) F_0.

DEGG and LoMA GCI analyses are compared in table 1 for sustained vowels. In this table, Dur is the segment duration (in s); Del: the estimated delay between sound and EGG (ms); Mdiff: average difference (in μs) between GCI detected by the two methods; N EGG: number of GCI detected on the EGG; N LoMA: number of GCI detected using LoMA; % diff: percentage of difference between the last two numbers. N EGG and N LoMA are in general very close, except for one condition (7.8% of differences). The mean value of Mdiff is −28 μs (but with standard deviation 300 μs). One can conclude that the LoMA GCIs and EGG GCIs are very close together.

4.3 *Comparison with DYPSA using EGG*

For further evaluation of GCI detection using LoMA, a comparative assessment using the DYPSA method (Naylor *et al* 2007) is presented. The DYPSA method is based on analysis of the spectral phase. In the speech production model, the speech signal is assumed to be minimum phase. Then the linear phase component observed in the speech spectrum reflects the delay between the GCI and the analysis window. Zero-crossing of the average phase slope are a good indication of the GCI (Smits & Yegnanarayana 1995; Naylor *et al* 2007).

The quality of GCI detection is assessed using the time delay between DEGG-CGI and the closest LoMA-GCI or DYPSA-GCI. Two other measures of quality are the rate of false alarms (a GCI is detected but is not present in the DEGG) and the miss rate (a DEGG GCI is not detected by the method).

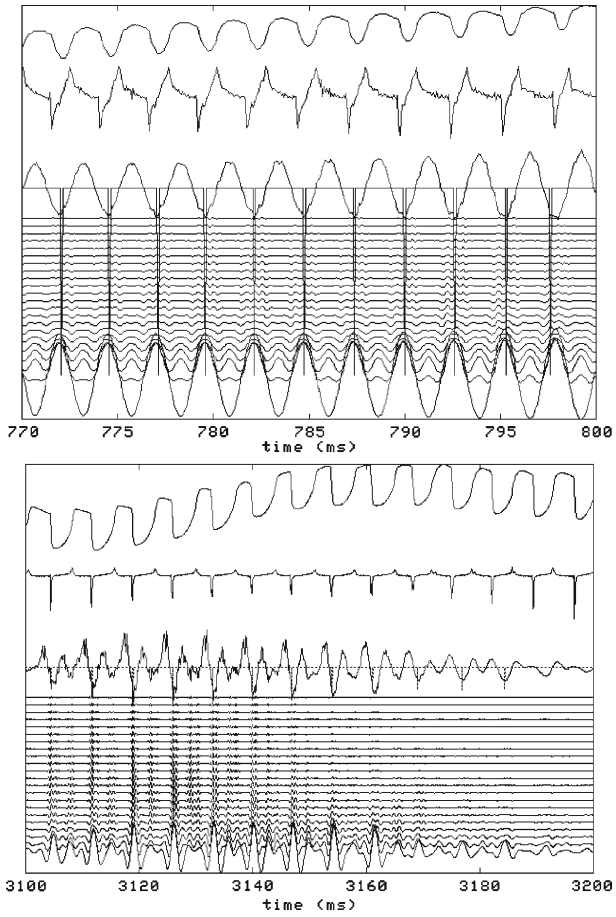

Figure 8. Examples of synchronous EGG and LoMA GCI detection. From top to bottom: EGG signal, DEGG signal, speech signal and wavelet filterbank. GCI–LoMA are indicated by vertical lines. EGG and speech time delay are compensated.

Again, prior to comparison, the EGG and acoustic signals are aligned in order to compensate for acoustic propagation delay between the glottis and the microphone. In an additional test condition, the speech signal is inverse filtered using LPC (Linear Predictive Coding) (18 coefficient,

Table 1. Comparison of GCI detection by EGG and LoMA (see text).

Dur (s)	Del (ms)	Mdiff (μs)	N DEGG	% diff	N LoMA
1.3	0.4	−39	206	1.9	210
2.5	2.2	11	349	0.8	346
1.8	1.4	12	175	0.5	176
3.0	3.2	−3	474	1.2	480
1.8	0.6	−38	380	0.2	379
1.5	0.1	−10	282	7.8	306
3.0	1.8	−132	517	2.1	506

autocorrelation), for the sake of removing the effects of the vocal tract prior to LoMA analysis (Ananthapadmanabha & Yegnanarayana 1979). A set of 20 sentences of read French extracted from the corpus (9 sentences from a male speaker, 11 sentences from a female speaker, overall about 2 minutes and 50 seconds of signal) are used for assessment.

The results are displayed in figure 9. Each panel presents the histogram of the delay between the GCI detected by a given method, and the GCI detected using DEGG. Ideally, the histogram should be an impulse at delay 0 (meaning that all the GCI are detected with 0 delay). These histograms are characterized mainly by their dispersions, measured by standard deviations. The top panels show the global results, the middle panels the results for the female speaker, and the bottom panels the results for the male speaker.

The standard deviation of the whole corpus is similar in DYPSA and LPC-LoMA. Histograms vary according to the gender of the speaker (i.e., average F_0), as indicated in figure 9, with standard deviations between 406 and 600μs.

False alarm and miss rates for the three methods are reported in table 2. These rates are comparable to the percentage of differences between DEGG and LoMA GCI reported in table 1, for a more difficult and more realistic task (sustained vowels vs. running speech in sentences). The LoMA GCI detection method compares favourably with DYPSA for female (high pitched) voices. Conversely DYPSA compare favourably with LoMA GCI for male (low pitched) voices. On an average, LoMA GCI seems slightly better than DYPSA as far as the miss rates are concerned. False alarm rates are low for all methods.

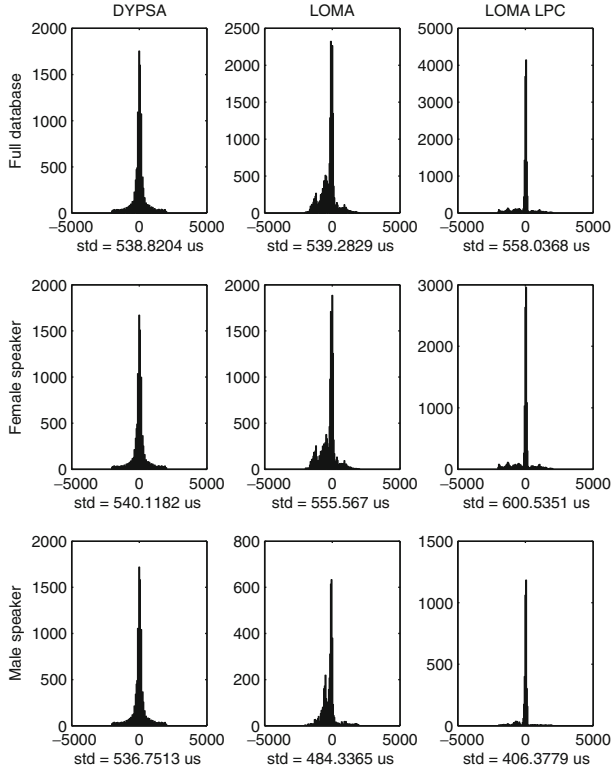

Figure 9. Distribution results–analysis of real speech corpus. 100μs steps.

Table 2. Miss rates (MR) and False Alarm rates (FA) for Male (M), Female (F) and Total (T) voices. EGG compared to DYPSA (DYP), LoMA (LOM) and LPC-LoMA (LPC).

Method	MR T	MR M	MR F	FA T	FA M	FA F
LPC	12.95%	10.25%	13.84%	0.53%	0.60%	0.50%
DYP	4.25%	1.33%	5.21%	0.52%	0.63%	0.48%
LOM	2.88%	3.03%	2.83%	0.50%	0.59%	0.47%

Prior LPC analysis does not improve GCI detection using LoMA. The GCI position are less widespread, but miss rates are much higher with LPC rather than without LPC. A possible explanation is that LPC often emphasizes the second harmonic relative to the first harmonic. The LoMA method is prone to errors in this situation.

For the sake of illustration of the LoMA and DEGG GCI detection methods, figure 7 displays LoMA analysis together with the EGG and glottal images obtained by high speed videoendscopy.[2] Images are sampled at a rate of 4000 frames/s. In this figure, the differences between the EGG GCI (indicated by a circle) and the LoMA GCI (indicate by a star) are less than one frame.

5. Voice source analysis using LoMA

In this section estimation of voice quality features using LoMA is discussed. Three parameters are estimated: the voice open quotient, the amplitude of voicing and the voicing strength.

5.1 *Open quotient*

The tense/lax voice quality dimension is mainly linked to the voice open quotient O_q. O_q is close to 1 for a lax or breathy voice and may be as low as 0.3 for a very pressed or tense voice. As discussed in section 3, the open quotient is linked to the low frequency behaviour of the glottal flow. In the spectral domain, O_q is closely related to the centre frequency of the glottal formant (see a detailed discussion in Doval *et al* 2006).

In the context of LoMA, the open quotient can be estimated using the phase delay at large scales, and particularly the phase delay of the fundamental frequency compared to the GCI. This measurement is based on the spectral resolution of the wavelet analysis in the vicinity of the voice first harmonic.

Using GFD models, one can compute analytically the phase delay at F_0 as a function of O_q. It is also possible to derive empirical values of the phase delay as a function of O_q using GFD models. This approach is illustrated in figure 10. The LoMA are displayed for two conditions

[2]For the purpose of illustration, a datafile was kindly provided by Nathalie Henrich. It contains a 500 ms long recording of high-speed images and the corresponding EGG signal in the case of a nonpathological male phonation. This recording was made in the Department of Voice, Speech and Hearing Disorders at the University Medical Center of Hamburg-Eppendorf in 2004, by Nathalie Henrich, Frank Müller, Götz Schade and Markus Hess. The subjects were Robert Expert and Cédric Gendrot. Data were processed by Nathalie Henrich et Sevasti-Zoi Karakozoglou.

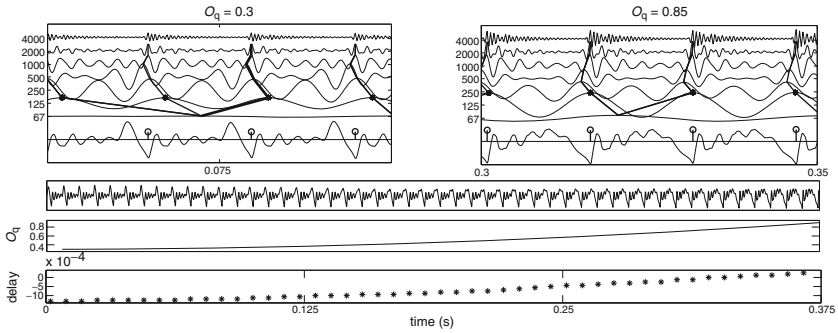

Figure 10. Estimation of open quotient using the difference between phase delay at F_0 and GCI. Top: two examples of LoMA analyses; second from top: speech signal, bottom: phase delay for each signal period; second from botom: open quotient estimated using DEGG.

of O_q (relatively high and low O_q) in the top panel for synthetic speech. The Lpf is used for O_q estimation in the bottom panel. The EGG derived O_q estimation is compared to Ldp. Both measures are highly correlated.

Open quotient measured using DEGG and LoMA are compared in figure 11, for a male voice sustained vowel. The speaker made a glottal abduction–adduction–abduction vocal gesture, resulting in a high–low–high open quotient pattern. Both measures are highly correlated. These results indicate that the Ldf is a promising measure of open quotient variation.

5.2 *Amplitude of voicing*

The LoMA is constructed as the optimal line, then it contains maximum amplitudes for each voicing period at each scale. It can be characterized by the total amplitude accumulated and by the repartition of amplitudes at different scales.

The total amplitude carried by the LoMA, as measured by the Laf, is linked to the amplitude of voicing in the corresponding period. Considering only the glottal flow derivative component, without vocal tract, the Laf represents the maximum excitation amplitude E, and then is strongly correlated to the sound pressure level (Gauffin & Sundberg 1989). The vocal tract component modify this amplitude: then the Laf is a measure of E multiplied by the vocal tract gain. Therefore, the Laf is proportional to E, and depends also on the vowel produced. The Laf provides a global measure of voicing amplitude.

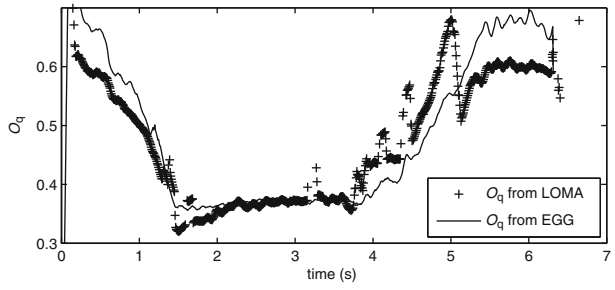

Figure 11. Estimation of open quotient using the difference between phase delay at F_0 and GCI.

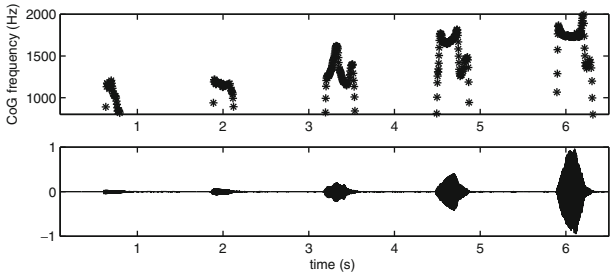

Figure 12. LoMA centre of gravity for six syllables with increasing vocal effort.

5.3 *Strength of excitation*

It is well established that vocal effort is not only related to the voicing amplitude but also to spectral richness. Another measure of the period strength is provided by the lengths of LoMA in the time-scale domain. Figure 2 illustrates variation of the LoMA lengths as a function of the time-scale energy distribution of the signal. A sinusoid excites only the larger scales of the filterbank. In contrast, an impulse excites all the scales. For a consonant to vowel transition (voiced fricative to vowel transition), the lines grow according to the increase of spectral richness across scales, as displayed in figure 6.

Spectral richness is a rather vague term, that can receive many interpretation. A simple form is the ratio of energy in higher frequency bands relative to lower frequency bands (Childers & Lee 1991). Another definition is 'spectral tilt' in the glottal flow, measured as the spectral attenuation at 3 kHz, relative to the fundamental frequency (Klatt & Klatt 1990). This parameter is related to the behaviour of the glottal flow near glottal closure (Doval *et al* 2006). Voicing strength can also be defined as the speech excitation impulsiveness (a recent review and discussion on the definition of voicing strength and its measurement can be found in Seshadri & Yegnanarayana 2009).

For measuring the voicing strength, in addition to Laf, the LoMA centre of gravity Lcg is proposed. This measure is illustrated in figure 12. The Lcg is plotted for syllables with increasing vocal effort. Figure 12 indicates that the spectral centre of gravity along the LoMA increases with increased vocal effort. Similar to the Laf, the Lcg is easy to interpret if one considers only the source component. Considering only the glottal flow derivative component, without vocal tract, the Lcg is a measure directly representing the voice spectral tilt. However, the vocal tract component modifies this amplitude: then the Lef is a measure of the source spectral tilt multiplied by the vocal tract gain. It is a global measure of speech spectral richness, and not only a voice source-related measure. For a same vowel, however, the Lcg varies according to the voice spectral tilt.

6. Conclusion

In this article, the tree patterns observed in time-scale representations of speech, such as cochleograms or auditory scalograms, are interpreted for analysis of voice quality features. The LoMA representation is proposed for this analysis. The first stage is the wavelet filterbank. In a second stage, time-domain maxima at each scale are detected. This representation is close in principle to auditory models such as the pulse-ribbon model (Patterson 1987). In the third stage, the specific LoMA method itself is introduced. Maxima in each band are linked across scales

to form tree patterns characterizing the signal in the time-scale domain. According to the linear model of speech production, and particularly considering spectral properties of glottal flow models, these patterns can be interpreted for voice source parameters estimation.

First, a unique system of branches appears for each voicing period. Then an optimal line, the LoMA, is searched using dynamic programming, for each period. The LoMA exhibits interesting properties:

(i) The line points to the GCI at smaller scales.
(ii) The phase delay at F_0 relative to the GCI is related to voice open quotient.
(iii) The accumulated amplitude along the LoMA is related to voicing amplitude.
(iv) The LoMA spectral centre of gravity is an indication of voice spectral tilt.

These properties are tested for voice quality analysis. The LoMA appears as an effective method for GCI detection, and compares favourably with EGG and DYPSA. Open quotient, amplitude and spectral tilt estimations provide promising results. In this article, they are illustrated only with the help of a few examples. However, they are currently systematically tested on large databases.

The work developed in the context of speech signals could also be extended to other types of highly structured monophonic signals, such as musical instrument signals, provided that specific models of the time-scale patterns for these signal are elaborated.

Contributions of Vu Ngoc Tuan and François Rigaud were instrumental in developing the present research, and they are gratefully acknowledged. The authors wish to thank Nathalie Henrich for providing the recordings displayed in figure 7.

References

Alku P, Bäckström T, Vilkman E 2002 Normalized amplitude quotient for parametrization of the glottal flow. *J. Acoust. Soc. Am.* 112(2): 701–710

Ananthapadmanabha TV, Yegnanarayana B 1979 Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust. Speech Signal Process.* 27(4): 309–318

Bouzid A, Ellouze N 2007 Open quotient measurements based on multiscale product of speech signal wavelet transform. *Res. Lett. Signal Process.* Article ID 62521

Bouzid A, Ellouze N 2009 Voice source parameter measurement based on multi-scale analysis of electroglottographic signal. *Speech Commun.* 51(9): 782–792

Childers DG, Lee CK 1991 Voice quality factors: Analysis, synthesis and perception. *J. Acoust. Soc. Am.* 90(5): 2394–2410

d'Alessandro C 1993 Auditory-based wavelet representation of speech. in M Cooke and S Beet (eds), *Visual representations of speech signals*. London: John Wiley & Sons, 131–138

d'Alessandro C 2006 Voice source parameters and and prosodic analysis. in S Sudhoff *et al* (eds), *Method in empirical prosody research*. Berlin: Walter de Gruyter, 63–87

Cooke M, Beet S (eds), 1993 *Visual representations of speech signals* (John Wiley & Sons)

Doval B, d'Alessandro C, Henrich N 2006 The spectrum of glottal flow models. *Acta Acustica united with Acustica* 92(6): 1226–1246

Fant G 1960 *Acoustic theory of speech production* (Den Hague: Mouton)

Fant G 1993 Some problems in voice source analysis. *Speech Commun.* 13: 7–22

Fant G 1997 The voice source in connected speech. *Speech Commun.* 22: 125–139

Flanagan JL 1972 *Speech analysis, synthesis and perception* (Berlin: Springer-Verlag)

Gauffin J, Sundberg J 1989 Spectral correlates of glottal voice source waveform characteristics. *J. Speech Hear. Res.* 32: 556–565

Gobl C, Chasaide AN 2003 Amplitude-based source parameters for measuring voice quality. *Proc. ISCA Tutorial and Research Workshop VOQUAL'03*, Geneva, 151–156

Seshadri G, Yegnanarayana B 2009 Perceived loudness of speech based on the characteristics of glottal excitation source. *J. Acoust. Soc. Am.* 126(4): 2061–2071

Henrich N, d'Alessandro C, Castellengo M, Doval B 2004 On the use of the derivative of elecroglottographic signals for characterization of non pathological phonation. *J. Acoust. Soc. Am.* 115(3): 1321–1332

Irino T, Kawahara H 1993 Signal reconstruction from modified auditory wavelet transform. *IEEE Trans. Signal Process.* 41(12): 3549–3554

Kadambe S, Boudreaux-Bartels GF 1992 Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. Inform. Theory* 38(2): 917–924

Klatt D, Klatt L 1990 Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* 87: 820–857

Mallat S, Liang Hwang W 1992 Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory* 38(2): 617–643

Mallat S, Zhong S 1992 Characterization of signals from multiscale Edges *IEEE Trans. Pattern Anal. Mach. Intel.* 14(7): 710–732

Naylor PA, Kounoudes A, Gudnason J, Brookes M 2007 Estimation of the glottal closure instant using the dypsa algorithm. *IEEE Trans. Speech Audio Process* 15: 34–43

Papoulis A 1977 *Signal analysis* (New Jersey: McGraw-Hill)

Patterson RD 1987 A pulse ribbon model of monaural phase perception. *J. Acoust. Soc. Am.* 82(5): 1560–1586

Smits R, Yegnanarayana B 1995 Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. Speech Audio Process* 3(5): 325–333

Sturmel N, d'Alessandro C, Rigaud F 2009 Glottal closure instant detection using lines of maximum amplitudes (LoMA) of the wavelet transform, *Proc. IEEE ICASSP'09*, Taiwan, 4517–4520

Tuan VN, d'Alessandro C 1999 Robust glottal closure detection using the wavelet transform, *Proc. ISCA Eurospeech'99*, Budapest, 2805–2808

Tuan VN, d'Alessandro C 2000 Glottal closure detection using EGG and the wavelet transform, *Proc. 4th Workshop Advances in Objective Laryngoscopy, Voice and Speech Research*, Jena