6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ROBUST GLOTTAL CLOSURE DETECTION USING THE WAVELET TRANSFORM.

Vu Ngoc Tuan & Christophe d'Alessandro

LIMSI-CNRS, BP 133, F91403 Orsay, France.
E-mail: vnt@limsi.fr cda@limsi.fr

10.21437/Eurospeech.1999-619

ABSTRACT

In this work, a time-scale framework for analysis of glottal closure instants is proposed. As glottal closure can be soft or sharp, depending on the type of vocal activity, the analysis method should be able to deal with both wide-band and low-pass signals. Thus, a multi-scale analysis seems well-suited. The analysis is based on a dyadic wavelet filterbank. Then, the amplitude maxima of the wavelet transform are computed, at each scale. These maxima are organized into lines of maximal amplitude (LOMA) using a dynamic programming algorithm. These lines are forming "trees" in the time-scale domain. Glottal closure instants are then interpreted as the top of the strongest branch, or trunk, of these trees. Interesting features of the LOMA are their amplitudes. The LOMA are strong and well organized for voiced speech, and rather weak and widespread for unvoiced speech. The accumulated amplitude along the LOMA gives a very good measure of the degree of voicing.

Keywords Glottal closure analysis, pitch tracking, wavelets

1. INTRODUCTION

The glottal closure instants carry very important information on the speech signal. Prosodic parameters like the degree of voicing (in the simplest form, a voiced/unvoiced decision) and the frequency of voicing (F0) can be derived from glottal closure instants (GCI). For speech synthesis also, the GCIs are needed in methods based on pitch period or pitch synchronous processing.

Glottal closure are often points of sharp variations, or singularities in the speech signal. According to Mallat [6, 7], the wavelet transform demonstrated excellent capabilities for detection of singularities in signals. This feature has been applied to pitch detection by Kadambe and Boudreaux-Bartels [4, 5]. Their work is based on the the dyadic wavelet transform of the speech signal. This transform is computed only for 2 or 3 small scales (high frequencies), typically 2^4, 2^5 and 2^6. Then, GCIs 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 sharp singularities at glottal closure. However, this is not always the case for voiced speech. For instance systematic falls of the vocal effort at the end of sentences result in quasi-sinusoidal glottal vibration, where there is no more sharp signal variation at glottal closure. It this situation the method based on sharp variation of the speech signal at glottal closure does not work.

Another pitch detection algorithm using the dyadic wavelet transform has been proposed by Montrésor and Baudry [8, 9]. In this method, the wavelet transform is considered as a zero-phase filterbank. It is then possible to track the fundamental frequency or its harmonics by computing the instantaneous frequency of the signal viewed by each filter. Contrary to the method of Kadambe and Boudreaux-Bartels, the wavelet transform is computed only for 2 or 3 large scales (low frequencies), typically 2^1, 2^2 and 2^3. This method is not suited to GCI detection, but is well-suited to quasi-sinusoidal voice.

In [1, 2] we explored speech representation in the time-scale domain with the help of an auditory-based wavelet transform. A wavelet filterbank was used for visualization of speech signals. Characteristic patterns were obtained for voiced and unvoiced speech. Kind of "tree" patterns were obtained for voiced speech, as a result of the multiscale analysis of quasi-harmonic signals.

This idea is exploited in this paper. We present a new algorithm for GCI detection with the help of the wavelet transform. The main difference with previous work is that all the scales are used for analysis. Like in other wavelet-based detector, the high frequency features related to abrupt closures are analyzed with accuracy. But low-pass speech signals can also be analyzed with accuracy, and the algorithm is also able to detect GCI when there are no strong singularities in the signal, e.g. for soft voices or low vocal effort.

For cross-scale analysis, the concept of concept of lines of maximum amplitude (LOMA) is introduced. These lines are linking the points of the time-scale space that are carrying most of the signal energy across scales. In addition to GCIs, the algorithm gives a measure of the energy carried by each LOMA, which is an indication of the degree of voicing.

In Section 2, the types of signal corresponding to glottal closure are reviewed. The detection algorithm and application to speech signals are described in Section 3. Some remarks conclude this work.

2. GLOTTAL CLOSURES IN TIME AND FREQUENCY

Several modes of glottal activity exist in the voice source. At least four mechanisms (fry, modal, falsetto, whistle) can be used. Of course, the speech signal corresponding to these different mechanisms may show much variability, and glottal flow models are able to simulate signals with abrupt glottal closure, or signals with soft glottal closure, or even without any glottal closure. Therefore, the GCI does not always correspond to a strong peak in the speech wave. For breathy or soft voice, for instance, GCI is only the minimum glottal flow.

Because of the sound radiation component in the acoustic theory of speech production [3], the general shape of the speech wave is given by the glottal flow derivative, rather than the glottal flow signal itself. Two typical situations must therefore be considered: sharp and soft glottal closures.

2.1. sharp glottal closures

A sharp glottal folds closure results in a peak in the glottal flow signal derivative, and in the speech wave. The prototype signal for sharp closure is the impulse. For a sharp closure, the GCI is located at the position of the impulse. As a consequence, the GCI is well localized in time, and the signal contains high frequency components. This is illustrated in Figures 1. The output of a wavelet (non-uniform) filterbank to an impulse train is displayed.

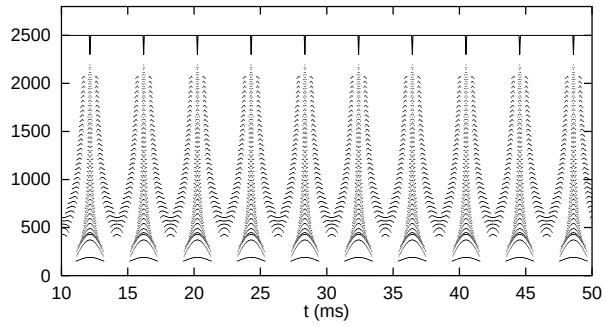

Figure 1. Output of the wavelet filterbank for a Dirac pulse train with fundamental frequency: 200 Hz. 35 filters are used, in the range 100-2100 Hz. Only the postive values of the responses are plotted.

Only the positive amplitudes are plotted, to enhance the lines of maximum responses. These lines are long, and are pointing to the peak due to the impulses across scales (or frequency). It should then be possible to detect the GCI by a suitable processing of these lines.

In the frequency domain, sharp glottal closures correspond to a low spectral tilt. In theory, this spectral tilt cannot be less than -6dB/oct. This corresponds to the derivative of a Dirac impulse, i.e. a step signal (a signal with a discontinuity). This type of peak can be found, in speech. Therefore the situation of figure 1 (corresponding to a 0dB/oct spectral tilt) cannot be observed in speech, but its derivative can be found (corresponding to a -6 dB/oct spectral tilt, see Figure 4, later in the paper).

2.2. soft glottal closures

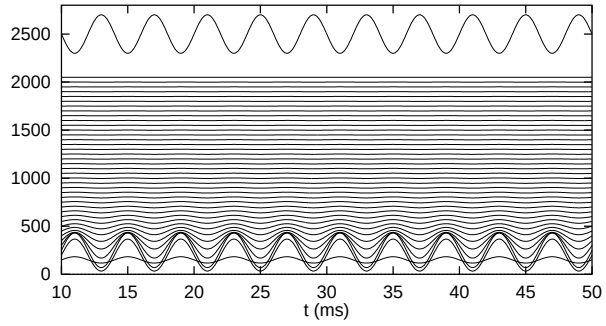

Figure 2. Output of the wavelet filterbank for a sinusoid with fundamental frequency: 200 Hz. 35 filters are used, in the range 100-2100 Hz.

Glottal folds closure can also be soft, for low vocal effort. In this case, the speech signal is quasi-sinusoidal. The prototype signal for soft closure is the sinusoidal signal. For a soft closure, the GCI is located at the minimum of the glottal signal, but there is no sharp peak in this signal, or in the speech signal. This is illustrated in Figures 2. The output of a wavelet (non-uniform) filterbank to a sinusoid is displayed. This signal contains mainly low-frequency components. Its response is not well localized in time. There are only rather short lines in the time-scale domain and the GCI should be searched for at the extrema of the responses at low frequencies, along these lines.

In the frequency domain, it is common to find speech signals with quasi-sinusoidal waveforms. Therefore the situation of figure 2 is realistic for speech analysis (see Figure 3, later in the paper).

3. GCI DETECTION ALGORITHM

The GCI detection algorithm is based on time-scale analysis. The main idea is to take advantage of the lines of maximum amplitudes (LOMA) in the time-scale domain to detect glottal closures. Pitch tracking and the degree of voicing are also obtained as a by-product of the LOMA analysis. Pitch is given by the inter-GCI distance. Amplitude carried by the LOMA is an indication of the degree of voicing.

3.1. dyadic wavelet transform

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:

$$y_i(t) = x(t) * h(\frac{t}{s_i})$$

where $h_i(\frac{t}{s_i})$ can be interpreted as a filter impulse response and $y_i(t)$ as the response of this filter to signal $x(t)$. We choose:

$$h(t) = -cos(2\pi f_o t).exp(-\frac{t}{2\tau^2})$$

Note the minus sign in the cosine. Then the wavelet analysis will have a maximum response to negative peaks, which are likely to result of glottal closures, according to the acoustic theory. Note also that the filters impulse responses are not causal. 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.

For pitch analysis, it is sufficient to use a dyadic WT, i.e. a WT with with $s_i = 2^i; i = 0, 1, 2, 3, 4, 5$. In the experiments the signals were sampled at $8kHz$, and $f_o = 4000$ Hz ; $\tau = \frac{1}{2f_o}$.

$y_i(t)$ is then the signal obtained with a bank of 6 band-pass filters, centered at frequencies f_i: 4000, 2000, 1000, 500, 250, 125 Hz, and with -3dB bandwidths of $\approx 0.5 \times f_i$. For the figure, more filters (about 35 in the [100,2000 Hz] range) are used because it makes the lines more visible. But this is not necessary for analysis.

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

$$h(\frac{t}{s}) \xrightarrow{FT} s\widehat{h}(sf)$$

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

Figure 1 and 2 show the output of a wavelet filterbank (in this case with about 35 filters) to test signals.

3.2. wavelet transform amplitude maxima

The amplitude maxima of the wavelet transform are computed by analysis of each filter response. These 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 neighborhood of η. Amplitude maxima for test signals can be seen in Figures 1 and 2. In Figures 1, only the positive values of the responses are represented. This shows that lines of positive amplitudes are converging to the peaks position. The lines corresponding to maximum amplitudes for each period and each scale are straight line, pointing towards the signal singulatities. In Figure 2, the whole responses are displayed. For this type of signal there is no strong line. However, the signal extrema are synchronized with the maxima in the filters response. This is particularly clear for low frequencies, because the responses vanish quickly for high frequencies.

3.3. lines of amplitude maxima

The next step of the algorithm is to organize amplitude maxima into lines of amplitude maxima. These lines should link amplitude maxima across scales. Because of the differences in phase of the signal for different scales, the lines are not straight, and

some work must be done to link the amplitude maxima across scales. This is achieved with the help of a dynamic programming algorithm, as follows.

Let $M_a(i,j)$ represent the i^{th} amplitude maxima at scale j. Scales are ordered from 0 (center frequency 4000 Hz) to 5 (center frequency 125 Hz). Let $L_m(i,j)$ be the LOMA that is searched. Let start the search by scale 5, and try to built LOMA to scale 0. Accumulated amplitudes $A_c(i,j)$ along LOMA are computed using the following local equations:

$$A_c(i,j) = max \begin{cases} A_c(il, j+1)/lw + M_a(i,j) \\ A_c(i, j+1) + M_a(i,j) \\ A_c(ir, j+1)/rw + M_a(i,j) \end{cases}$$

where il (resp. ir) is the index of the amplitude maximum in the left (resp. right) neighborhood of $M_a(i,j)$ at scale $j+1$, and where lw (resp. rw) is a weighting factor factor of the absolute value of the difference between i and il (resp. ir): $lw = i - il$ (resp. $rw = ir - i$).

The indices il (resp. ir) are determined by comparing the weighted amplitude maxima in a left (resp. right) neighborhood Dl (resp. Dr) of i, which is inversely proportional to the center frequency of the scale:

$$max_{Dl}(M_a(il, j+1)/i - il)$$

The LOMA are built from the array of accumulated amplitudes using back-tracking. Finally, the time-scale domain is represented by a small number of LOMA. Furthermore, each LOMA is characterized by its weight (accumulated amplitude of the amplitude maxima along the line).

The lines are organized in packets, or trees, as can be seen on Figures 3 and 4. They are several separated lines for high frequencies, that are merging, and form a common line. Ideally, there is exactly one tree for each voicing period. The trunk of this tree is located at low frequencies, near the fundamental frequency. Each harmonic creates a new division on the tree, and then gives birth to a new system of branches.

3.4. glottal closure instants detection

GCI are computed by analyzing the system of LOMA. The idea is to built the pitch period trees, and then to find the GCI within each period, i.e. within each tree. All the LOMA of a same tree have different accumulated amplitude, depending on their path along the branches of the tree. The GCI is identified to the "best" line, i.e. the strongest branch or trunk in each tree. This trunk is determined as the line carrying the maximum accumulated amplitude.

LOMA are not necessarily vertical straight lines in the time-scale space. As a matter of fact the best LOMAs are vertical straight lines only for series of Dirac impulses. Thus, the modulus maxima are not located at the same point in times for a same LOMA across different scales.

Depending on the input signal, the LOMA, and then the trees, have also different length. An example is given in Figure 3. For a low frequency signal the trees are short, and for a signal with sharp peaks, they are long. For sharp glottal closures, as the lines have energy in small scales (high frequencies), the glottal closure instant is located on LOMA in the smaller scales. On the contrary for smooth glottal closure, the lines are "short", and the glottal closure instants are located on the LOMA at the larger scales.

In Figures 3 and 4 analysis of male and female voices are plotted. The GCI detected using the LOMA are also displayed. Figure 3 (male voice) is a transition between a voiced fricative (smooth glottal closure) and a back vowel (sharp glottal closure). When the voicing is low-pass (voiced fricative) the lines are short: above a given frequency the maxima amplitude is low. On the

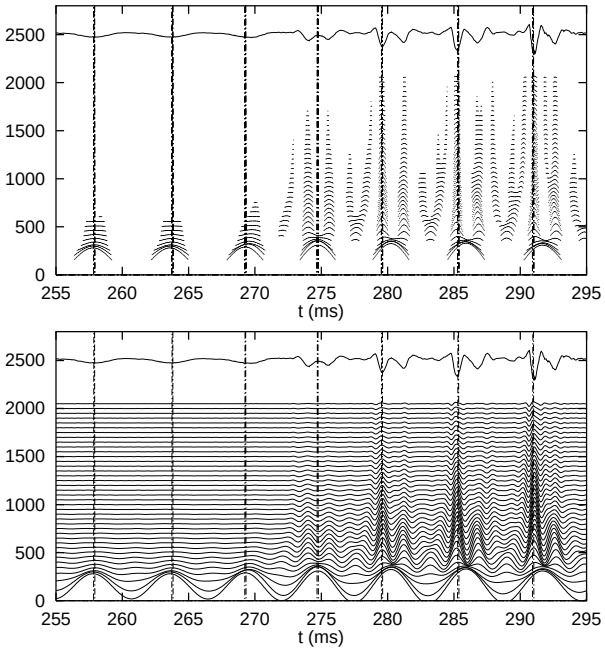

Figure 3. Output of the wavelet filterbank for speech signal (voiced fricative/vowel transition, male voice). 35 filters are used, in the range 100-2100 Hz. In order to enhance the lines of maximum amplitude, only the postive values of the responses are plotted in the top panel.

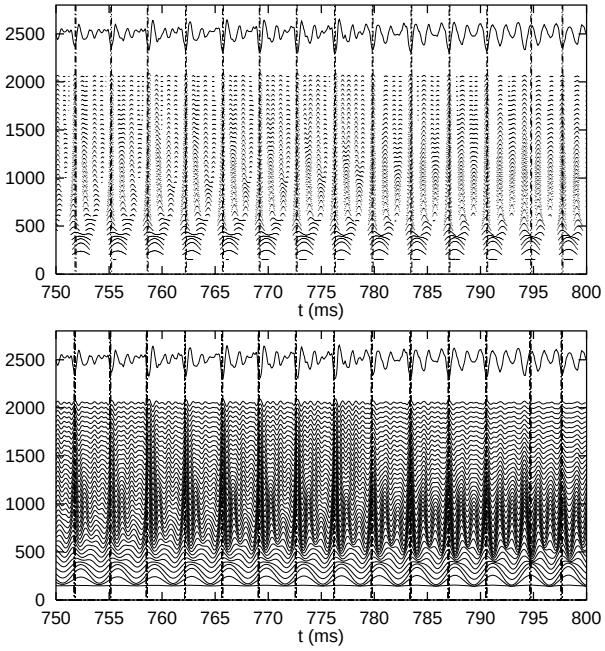

Figure 4. Output of the wavelet filterbank for speech signal (vowel, female voice). 35 filters are used, in the range 100-2100 Hz. In order to enhance the lines of maximum amplitude, only the postive values of the responses are plotted in the top panel.

contrary, for a signal with sharp closure (vowel), the lines are long. For each best line, the GCI is determined as the higher (in frequency) point on this line with significant amplitude. Figure 4 displays a vowel (female speech) with relatively sharp peaks, and high F0 (about 300 Hz).

In Figures 3 and 4 a rather large number of filters are used, for visual display. As a matter of fact, only 6 filters (corresponding to 6 dyadic scales) are actually used for GCI detection, as it is showed in Figure 5. This figure represents the response of the wavelet filterbank to a vowel signal. In each band, arrows are indicating the position of amplitude maxima. These amplitude maxima form the LOMA, which is pointing to the GCI for each pitch period.

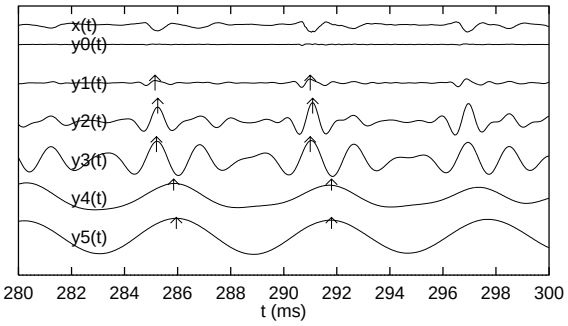

Figure 5. Wavelet filterbank response for few periods of a vowel signal. Arrows are indicating the position of amplitude maxima.

3.5. 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 are carrying 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 is carrying a large amount of the signal energy. The voiced/unvoiced decision is taken by using a simple threshold on the amplitude carried by the trunk of each tree. This simple measure is surprisingly robust.

3.6. pitch detection

Once the voiced trees and GCI have been determined, pitch is computed using as the inverse between successive GCI. For accurate pitch determination, it is necessary to refine the position of the GCI, by using an interpolation procedure near the first estimate of the corresponding amplitude maximum.

4. CONCLUSION

In this work, a time-scale framework for analysis of glottal closure instants is proposed. The analysis is based on a dyadic real wavelet transform. The amplitudes at the output of the filterbank are analyzed, and amplitude maxima are detected.

It is shown that the glottal signal gives birth to lines of maximum amplitude in the time-scale domain. These lines are organized in "tree" patterns, with exactly one such tree for each pitch period, when the signal is voiced. GCI are then interpreted as the top of the strongest branch, or trunk, of these trees. An interesting property of the LOMA is that they also give a good measure of voicing. The accumulated amplitude along the LOMA can by used to compute the degree of voicing.

The LOMA can also be useful for analysis of other types of glottal activity. For instance, it is possible to detect isolated periods, as can be found in vocal fry.

Future work will be devoted to a formal evaluation of this method for pitch tracking. As the analysis procedure is based on a linear filterbank, it would also be possible to use a similar framework for speech modification or speech synthesis. This is an interesting perspective that is currently under study.

REFERENCES

[1] C. d'Alessandro, D. Beautemps, (1991). "Justification perceptive du spectrographe auditif." Proc. XIIth ICPhS, Vol 5, pp. 86–89.

[2] C. d'Alessandro, 1993. "Auditory-based wavelet representation of speech." In Martin Cooke and Steve Beet, editors, *Visual Representations of Speech Signals*, John Wiley & Sons, chapter 8, pp. 131–138.

[3] G. Fant, 1960. *Acoustic theory of speech production.* (Mouton, La Hague).

[4] S. Kadambe, G.F. Boudreaux-Bartels, 1991. "A comparison of a wavelet functions for pitch detection of speech signals" proc. IEEE ICASSP'91, pp.449-452.

[5] 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.

[6] S. Mallat, Wen Liang Hwang, 1992. "Singularity detection and processing with wavelets" IEEE trans. on IT, vol.38. No.2. pp.617-643.

[7] S. Mallat, Sifen Zhong, 1992. "Characterization of signals from multiscale Edges" IEEE trans. on PAMI, vol.14. No.7. pp. 710-732.

[8] S. Montrésor & M. Baudry, 1990, "Représentation temps-échelle et détection de la fréquence fondamentale du signal de parole." proc. of XVIII JEP,, Montréal, 28-31 may 90, pp. 170-174.

[9] S. Montrésor & M. Baudry, 1991, "Quelques résultats sur l'utilité de la transformée en ondelettes pour l'analyse de la parole." proc. of SFA (French Acoustical Society) workshop "traitement et représentation du signal de parole", le Mans 3-4 June 91, pp. 74-79.