GLOTTAL CLOSURE INSTANT DETECTION USING LINES OF MAXIMUM AMPLITUDES (LOMA) OF THE WAVELET TRANSFORM

Nicolas Sturmel, Christophe d'Alessandro, Francois Rigaud

LIMSI/CNRS
BP 133, 91405 ORSAY CEDEX
FRANCE

ABSTRACT

The Lines Of Maximum Amplitude (LOMA) of the wavelet transform are used for glottal closure instant detection. Following Kadambe & al. (1992), the wavelet transform modulus maxima can be used for singularity detection. The LOMA method extends this idea. All the lines chaining maxima of a wavelet transform across scales are built. Then a backtracking procedure allows for selection of the optimal line for each pitch period, the top of which indicates the GCI. The LOMA method is then evaluated by comparing its results to the DYPSA (Naylor & al.) algorithm, with the option of using inverse filtering as preprocessing. The LOMA method compares favorably to DYPSA, particularly on accuracy. One of the advantage of the LOMA method is its ability to deal with variations in the glottal source parameters.

Index Terms— Wavelet, GCI, EGG, Pitch Marks

1. INTRODUCTION

Many speech processing applications require the knowledge of Glottal Closure instants (GCI): pitch synchronous analysis, overlap-add speech synthesis, time and frequency scaling and so on. Several types of methods have been proposed so far for GCI detection. These methods are based on different aspects of the GCI.

A first idea is to find directly the GCI using inverse filtering based on linear prediction (LPC): the LPC residual shows large peaks at GCI. However, Ananthapadmanabha & al. [1] showed that one could find significant differences between the GCI and the peaks in the LPC residual. Consequently, these authors proposed a method based on 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 & al. [2]). This method is refined and carefully assessed in the work by Naylor & al. [3] (the DYPSA method). This method includes a fair amount of post-processing, and leads to excellent results. Another LPC based method has been proposed by Moulines & al. [4] : short-term covariance LPC is used to compute changes in the statistical properties of the wave, corresponding to GCI. This method has not been thoroughly assessed, and it is known that covariance analysis lacks robustness. Some differences in the position of GCI are also noticeable in the method based on the maximum likelihood estimation of pitch period, proposed by Cheng et al. [5]

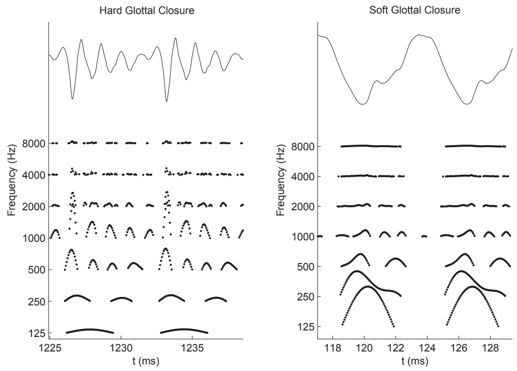

Fig. 1. Hard glottal closure (left) and soft glottal closure (right) wavelet analysis (positive part) and signal. Soft glottal closure lack of high frequency information, whereas hard glottal closure is located by the highest frequency scale and not the lowest.

As the CGIs are associated to singularities in the voice source signal, the theory developed by Mallat [6] for singularity detection has also been tried. The wavelet transform modulus maxima has been applied by Kadambe & Boudreaux Bartels [7], who used a dyadic wavelet transform for pitch estimation. First, the wavelet transform is computed on the 2 or 3 smallest scales (high frequencies). Then GCI are detected by local modulus maxima above a given threshold across two dyadic scales. This method works well when the speech signal contains sharp singularities at glottal closure, which is not always the case for voiced speech. For instance, a voiced onset often presents a quasi sinusoidal waveform, without sharp closure. Examples of sharp and soft glottal closures are given in figure 1. In the situation of quasi-sinusoidal voice, the method based on sharp variation of the speech signal at glottal

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

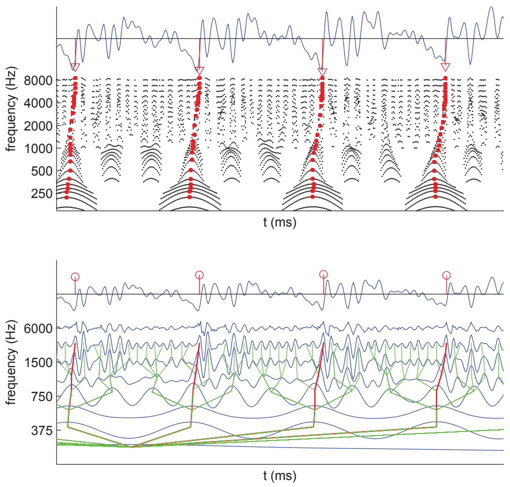

Fig. 2. Top: illustration of the tree shapes visible on a wavelet decomposition, maxima leading to the GCI are marked in red. Only positive parts of the filter responses are given. Bottom: actual wavelet decomposition used in the algorithm, and the lines constructed by following the maxima.

closure does not work. Another method using an inter-scale maxima product has been proposed by Bouzid et al. [8]. As the product maximum depends much on the harmonics relative amplitudes and phases, the low frequency component can change significantly the estimated GCI locations. We believe that the GCI corresponds to the line of maximum amplitude and not only to the amplitude products. Depending on phases all the maxima are not time aligned (see figure 3 right panel), although the product ideally assumes that all the maxima are occurring at the same instant.

Building on the wavelet modulus maxima theory, a new algorithm for GCI detection with the help of the wavelet transform has been presented [9]. Contrary to Kadambe & al.'s work, all the scales are used for analysis. Then, the high frequency features observed in abrupt closures as well the low-frequency features of quasi-sinusoidal speech can be analyzed with accuracy. A dynamic programming algorithm builds the lines of maximum amplitude (LOMA), which are chaining amplitude maxima across scales in the wavelets transform domain. GCI are then derived by selecting the optimal line within a pitch period.

In this paper we present an improved scheme for GCI detection using LOMA. Contrary to [9] the improved algorithm builds all the LOMA, starting from all the amplitude maxima observed at a given small scale, and then uses a backtracking algorithm for selection of the optimal LOMA within a pitch period (section 2). In a second part of the paper, a comparative evaluation of LOMA, DYPSA and an electroglottographic (EGG) reference is reported (section 3) and discussed (section 4). Section 5 concludes.

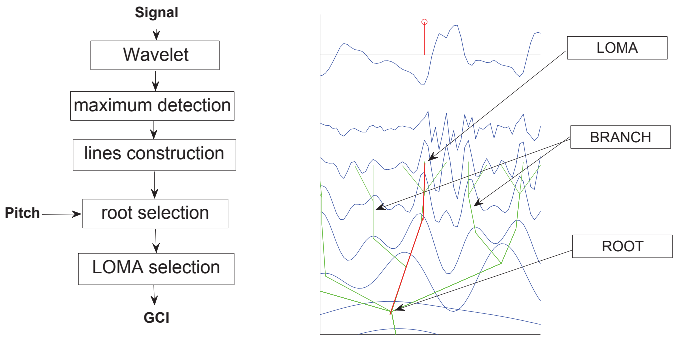

Fig. 3. Algorithm for LOMA glottal closure instant detection

2. LOMA ALGORITHM FOR GCI DETECTION

The algorithm is presented in figure 3.

1. Compute a dyadic (octave-band) wavelet transform, using 8 scales (or bands) between 62 Hz and 8000 Hz. Due to the dominant negative peak in the speech signal, the wavelet is chosen to have a negative maxima so that only positive maxima of each scale are kept for LOMA detection. Output of the wavelet filter bank are displayed on Figure 1

2. Select the smallest scale (usually the 4 kHz band) with significant maxima. Detect all the maxima at this scale. Maxima are defined as the local maxima between two zero crossings.

3. For each maximum previously detected, find the optimal line starting from this maximum, and descending the scales down to the lowest frequency scale. Maxima are chained across scales: for each maximum at a given scale i, the closest maximum at the scale $i-1$ is selected and cumulated amplitude for the line is computed. Maxima are chained down to the lowest scale forming the so-called branches on figure 3.

4. Using prior information on the average pitch, the band containing the fundamental is determined. No fine pitch detection is required, the aim being only to find the lowest analysis band. Then the optimal LOMA for each pitch period (maximum in the lowest band, called the tree root in figure 3) is selected.

5. The GCI is detected using backtracking on the optimal LOMA, as the time position of the highest scale maximum of the highest energy LOMA for each pitch period tree. This process is illustrated in figure 2.

6. Post-processing. Most of the errors observed are LOMA in excess, i.e. situations where two LOMA are detected for the same pitch period. Two criteria are

applied 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.

3. EVALUATION AND COMPARISON WITH DYPSA

The GCI detected with LOMA and DYPSA are compared to the GCI extracted from the EGG derivative. The chosen comparison criterion is the time delay between an EGG-CGI and the closest LOMA-GCI or DYPSA-GCI. These delays are expressed in μs. Two other measures are the rate of false alarms and the rate of missing GCI. Prior to comparison, the EGG and acoustic signals are aligned in order to compensate for acoustic propagation delay between the glottis and the microphone. An additional test condition combines LOMA and LPC inverse filtering: the speech signal is inverse filtered using LPC [10], with the idea of removing the effects of the vocal tract and glottal pulse phase, before LOMA analysis.

Two corpora are used for evaluation:

1. a synthetic speech corpus containing 6 vowels and various glottal parameter settings. This corpus is designed to point out the strengths and weaknesses of the method.

2. a natural speech database, containing 20 sentences of read French, with simultaneous acoustic and EGG recordings (male voice: 9 sentences, female voice: 11 sentences). This corpus has a total duration of approximately 2 minutes and 50 seconds, sampled at 16 bits/16 kHz.

3.1. Results on synthetic signals

Two sets of signals are synthesized, consisting of a glottal (LF model) signal filtered through a vocal tract filter obtained from LPC analysis on real speech (sampling frequency: 16 kHz). The first set contains gliding vowels (/a/, /i/ and /u/), with pitch from 50 to 400 Hz and fixed glottal settings conditions (open quotient $O_q = 0.6$ and return phase quotient $Q_a = 0.05$).

Table 1. Results on synthetic signals, percentage of detection within 0.25 ms around the actual GCI

set	vowel	dypsa	LOMA LPC	LOMA
1	$\backslash a\backslash$	100%	97%	97%
	$\backslash i\backslash$	100%	87%	86%
	$\backslash u\backslash$	100%	97%	97%
2	$\backslash a\backslash$	78%	86%	93%
	$\backslash i\backslash$	94%	76%	80%
	$\backslash u\backslash$	82%	94%	94%

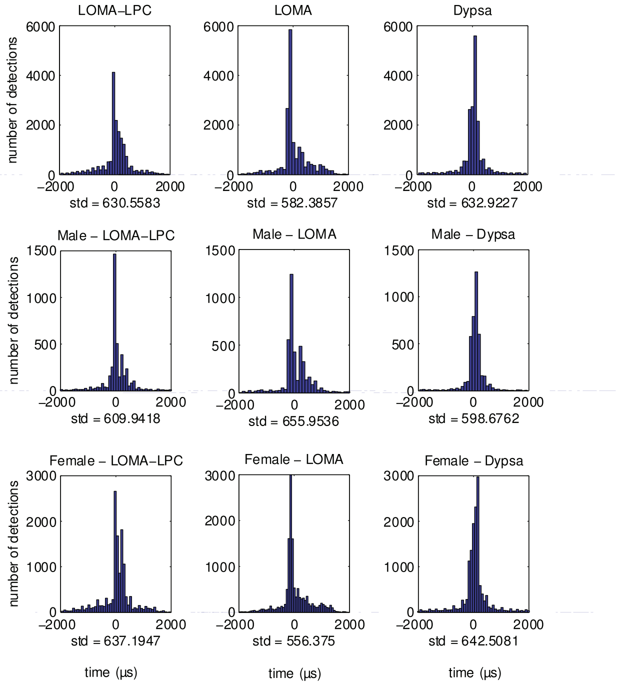

Fig. 4. Histogram distribution results ($100\mu s$ intervals) of the analysis of the real speech corpus. Number of GCI detected sorted by delay from EGG-GCI

The second set contains constant pitch vowels (/a/, /i/ and /u/, 120Hz), and varying glottal conditions from soft ($O_q = 0.6$ and $Q_a = 0.3$) to hard ($O_q = 0.1$ and $Q_a = 0.05$) glottal closure.

For this test, the quality measure is the percentage of detected GCI within $250\mu s$ around the actual synthetic GCI. Those percentages are presented in table 1 for the three tested methods.

Note that for the vowel /i/ the LOMA method performs inferiorly than DYPSA. A possible explanation is that the LOMA method always finds a local minimum on the signal. In some situations the GCI does not correspond to a signal minimum.

3.2. Spontaneous speech results

This corpus is composed of 20 sentences read from French newspapers, for a grand total of 18949 GCI detected via EGG. The results are displayed in figure 4. Each panel presents the delay between the GCI detected by a given method, and the GCI detected using EGG. Ideally, the results should be an impulse at delay 0 (meaning that all the GCI are detected with 0 delay). These distributions are characterized mainly by their dispersions, measured by their standard deviations. The top panels show the global results, the middle panel the male speaker results, and the bottom panel the female speaker results.

The standard deviation of the whole corpus is very similar for DYPSA and LPC-LOMA (about $640\mu s$). LOMA performed about 10 % better with a standard deviation of about $580\mu s$. Results also vary with the gender of the speaker, as indicated in figure 4, with standard deviations varying from 550 to $660\mu s$. The false alarm rates are 4.1% DYPSA, 7.35% LPC-LOMA and 6.55% for LOMA. The miss rates are 7.8% DYPSA, 7.6% LPC-LOMA and 6.37% for LOMA.

4. DISCUSSION

It seems that LPC brings no improvement in GCI detection using LOMA. It is known that LPC performs poorly for estimation of the first formant, because of its interaction with the glottal pulse spectrum maxima (the glottal formant). Therefore, LPC does not help in correcting the phase alignment at those frequencies. This phase shift, especially between the second and third harmonics, seems to be the main cause of GCI misdetection by LOMA. However, our results show that LPC preprocessing brings more problems than it solves, emphasizing noise and phase distortion.

The tested methods give comparable results on both synthetic and real speech signals. DYPSA performed perfectly on gliding vowels, whereas LOMA makes some errors, mainly octave errors, when pitch crosses two analysis scales. On the contrary, for varying glottal settings, LOMA performs better than DYPSA, because it is better fit to low vocal effort (quasi-sinusoidal voices).

The same general tendencies are also observed for natural speech results. The different methods are giving the same distribution standard deviation. It must be pointed out that the highest peak around ($0\mu s$) are obtained for the LOMA, with a more concentrated distribution. However, even if the LOMA seems slightly more accurate than DYPSA, it seems also less reliable regarding false detections.

5. CONCLUSION

A time-scale framework for analysis of glottal closure instants is proposed and evaluated. The analysis is based on a dyadic real wavelet transform. It is shown that the glottal signal gives birth to lines of maximum amplitude in the time-scale domain. GCI are found at the top of LOMA for each voicing period.

LOMA is compared to DYPSA, a state-of-the-art method for GCI detection. The results are very similar, showing that LOMA is a good candidate method for GCI detection. Analysis of the results for test signals indicates that LOMA is more robust than DYPSA against voice quality variation (open quotient, vocal effort), particularly for low vocal effort. On the contrary, LOMA seems less robust against large pitch variations. Provided that pitch information is given to the algorithm, one can expect a robust detection of glottal closure instants.

In addition to GCI detection, LOMA could also be exploited for other purposes in voice source analysis. The cumulated amplitudes along a LOMA seems a good indication of voicing : it could be used to compute the degree of voicing. Using a voice source model, open quotient, glottal pulse asymmetry and loudness can be linked to the LOMA properties. Finally, as the wavelet analysis procedure can be considered as a linear filter bank, it would also be possible to use a similar framework for implementing various speech modification schemes. All these interesting perspectives are currently under study.

6. ACKNOWLEDGMENT

The authors also wish to thank Pr. Vu Ngoc Tuan and Pr. Nathalie Delprat for their contribution and active discussion about wavelet and glottal event detection.

7. REFERENCES

[1] T. V. Ananthapadmanabha S. and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE trans. on ASSP*, vol. 27, no. 4, pp. 309–318, 1979.

[2] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation on speech using group delay function," *IEEE trans. on SAP*, vol. 3, no. 5, pp. 325–333, 1995.

[3] P.A. Naylor, A Kounoudes, J Gudnason, and M Brookes, "Estimation of the glottal closure instant using the dypsa algorithm," *IEEE Trans. on acoustics, speech and language processing*, vol. 15, pp. 34–46, 2007.

[4] E. Moulines and Di Francesco R., "Detection of the glottal closure by jumps in the statistical properties of the speech signal," *Speech Communication*, vol. 9, no. 5/6, pp. 401–418, 1990.

[5] Y. M. Cheng and O'Shaughnessy D., "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. on ASSP*, vol. 37, no. 12, pp. 1805–1814, 1989.

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

[7] S. Kadambe and G.F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE trans. on IT*, vol. 38, no. 2, pp. 917–924, 1992.

[8] A. Bouzid and N. Ellouze, "Open quotient measurements based on multiscale product of speech signal wavelet transform," *Research Letters in Signal Processing*, vol. 2007, 2007.

[9] Vu Ngoc Tuan and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Proceedings of the European Conference on Speech Technology, Eurospeech*, Budapest, Sep. 1999, pp. 2805–2808.

[10] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63(5), pp. 561–580, 1975.