GLOTTAL OPENING INSTANT DETECTION FROM SPEECH SIGNAL

Aïcha Bouzid and Noureddine Ellouze

National School of Engineers of Tunis
Le Belvédère , B. P. 37, 1002 Tunis, Tunisia (Africa)
phone: +216 71 874 700, fax: +216 71 872 729 , email: aicha.bouzid@enit.rnu.tn, N.Ellouze@enit.rnu.tn

ABSTRACT

Nowadays, new techniques of speech processing such as speech recognition and speech synthesis use the glottal closure and opening instants. Recognition techniques use them for the vocal folds description and for the classification of speaker's state or for speaker classification, and speech synthesis techniques use them for the speech timbre.
In an effort to develop techniques that enhance data-driven techniques in speaker characterisation for speech synthesis, this paper describes a new method for automatically determining the location of the closed phase delimited by the glottal closure and opening instants.
The proposed approach for detecting the glottal opening is based on multiscale products of wavelet transform of speech signal at different scales with enhancement of edge detection and estimation. It is shown that the method is effective and robust for speech singularity detection such as glottal opening instant as product is a processing which reinforces edge detection.

1. INTRODUCTION

This work is an important part of a current research in pitch synchronous speech analysis using an automatic and accurate vocal tract characterisation and vocal source parameter estimation such as glottal closure and opening instants.
Accurate vocal tract estimation is one of the key requirements of glottal wave analysis using the source-filter model of speech. In recent years, considerable progress has been achieved in estimating the glottal source characteristics from speech through single-channel speech analysis. An initial pitch-asynchronous speech deconvolution technique that uses the autoregressive Linear Prediction (LP) model builds on the assumption that the vocal tract parameters are slowly and smoothly varying and performs analysis over a number of pitch periods [20], [4].
However because fixed-frame analysis is performed during excitation and open phases of the glottal cycle, there are two adverse effects on the estimation of the vocal tract filter parameters when the glottis is open.
First, the vocal tract tube is no longer open at one end-invalidating the LP model. So when the glottis is open, coupling takes place with the subglottal cavity introducing subglottal resonances and antiresonances to the spectrum. These are superimposed on the supraglottal spectrum. The typical effects of this sub-glottal interference are to reduce formant frequencies while increasing formant bandwidths [12]. Thus, if the period of analysis is over both closed and open glottal phases, there will be a smearing or averaging of the parameters, and consequent loss of speaker characteristic information when we inverse filter with these parameters.
Second, the speech is no longer excitation free. LP autoregressive analysis techniques assume zero-mean input to the vocal tract filter. This assumption is no longer valid while the glottis is open. Thus, if the analysis is performed only during the closed phase, we can more accurately parameterise the vocal tract resonances [7]. That's why determining the glottal closure and opening instants from speech signal with accuracy are of major interest. This parameter can be implied in a wide range of applications. Among these are speech synthesis and transformation, voice quality enhancement, speaker identification, voice pathology classification, speech coding and transmission [3]. In particular, knowing these instants can improve the natural speech synthesis. In speech coding, they could improve the speech compression rates.
The main aim of this paper is to present a new method for determining glottal opening and closure instant from only speech signal. The proposed approach consists of computing the multiscale products of speech wavelet transform at different scales in order to enhance edge detection of speech signal.
The paper is structured as follows. First we outline some closed phase detection methods. In the next section, we present a multiscale method for GCI. Then we briefly review the principles of the multiscale products and its ability of peak detection and estimation as reported in [11]. We then step through the method for automatically locating glottal opening instant. We illustrate results for both male and female voices. Section 6 concludes this work.

2. CLOSED PHASE DETECTION METHODS

When a closed phase of the glottal cycle is assumed to exist, attempts have been made to locate the GOI in order to perform pitch synchronous processing of speech signal. These approaches can be classed as single channel analysis or dual channel analysis. It has been fairly common for studies and analyses to use a dual channel approach [18], [19], where the laryngograph is used to locate the closed phase by locating especially the glottal opening. However, this will not be appropriate for speech analysis outside laboratory conditions.

Single channel analysis uses only the speech signal to locate the GOI. The most methods that rely on using the speech signal alone have proved unreliable in locating the closed phase. Because of the difficulty in locating the glottal opening instant, many of these techniques, e. g. [16], [17], rely on simply estimating the GCI and assuming that an ad-hoc choice of post-GCI interval length will lie within the closed phase. These lengths are generally chosen to be either: a fixed constant length e. g. 2ms; or a percentage of the pitch period, e. g. 30%. Others methods, like that of Wong and al. rely on appropriate thresholds being applied [21]. Work presented in [5], [7] outlined a method for automatic closed phase location by excluding the intervals that are not within the closed phase; the indicator used is the log determinant of the Kalman filter (KF) estimate error covariance matrix.
Recently work has been reported and it was suggested that the signal representing acoustic input power at the glottis can be used to determine the instants of glottal closure and opening [9].

3. MULTISCALE GCI DETECTION

Glottal closure instants are often points of sharp variations or singularities in speech signal, [15], [1]. According to Mallat [9], the wavelet transforms demonstrated excellent capabilities for detection of singularities in signals. Furthermore, in the last years, wavelet transforms have been intensively applied in different pitch and GCIs detection algorithms [8], [13], [15]. Most of those algorithms are based on the dyadic wavelet transform. In [8], Vu Ngoc and al. proposes speech representation in the time-scale domain by wavelet transform and a filterbank implementation. The main idea presented is that all dyadic scales are used for speech analysis. As a result, not only high frequency features are analysed with accuracy but also smooth singularities in the signal can be detected. The work presented in [6], explores similar concept and proposes a robust strategy for glottal closure instants detection. This strategy uses significant minima and maxima time localization of the filterbank outputs; it takes decision from different scale minima giving the best estimation of the GCIs. Figure 1 shows the strategy of this algorithm. GCIs are located inside the meantime defined by minima and maxima of the scale 6 filter. The minima and maxima positions converge to the reference GCIs for the channels where these extrema are detected and satisfy the condition of being included in the alternation minimum and maximum at scale 6. Thus the GCI is estimated as the position of the minimum given by the lowest scale which satisfies the inclusion condition. In the worst case, the mid instant of the minimum and maximum interval of scale 6 is chosen. Figure 1 illustrates an example where GCI detection fails at scale 0, but gives the best estimation at scale 1.

4. MULTISCALE PRODUCTS

We consider a multiscale analysis by forming the product of the wavelet transform of a function f(n) at some dyadic scales

$$p(n) = \prod_{j=j_0}^{j=j_L} w_{2^j} f(n). \quad (1)$$

This is distinctly a non linear function of the input time series f(n). The function p(n) will show peaks at speech signal edges, and will have relatively small values elsewhere [10].

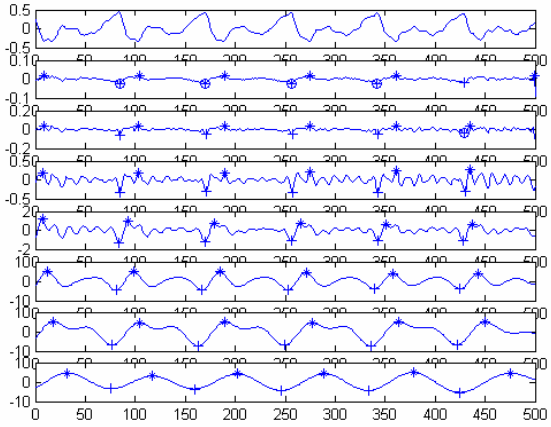

Figure 1: GCI detection for a voiced speech signal (female voice). From up to bottom: speech signal-speech wavelet transforms (WT) from scale 2^0 to scale 2^6. *Symbols*: + minimum, * maximum, o GCI

Singularities produce cross-scale peaks in wavelet transform coefficients, which are reinforced in p(n). Although particular smoothing levels may not be optimal, the non linear combination tends to reinforce the peaks while suppressing spurious noise peaks. The signal peaks will align across scale for the first few scales, but not for all scales because increasing the amount of smoothing will spread the response and cause singularities separated in time to interact. Thus choosing it too large will result in misaligned peaks in p(n). An odd number of terms in p(n) preserves the sign of the edge [11]. Choosing three levels of wavelet decomposition is an optimal solution of multiscale product method for detecting small peaks.

5. GOI DETECTION BY MULTISCALE PRODUCT

The Keele university database has been used to experiment the proposed method for estimating GOIs in voiced speech signal. Keele database includes acoustic speech signals and laryngograph signals. Five adult female and five adult male speakers were recorded in low ambient noise using a sound proof room. Each utterance consisted of the same phonetically balanced English text. In each case, the acoustic and laryngograph signals are time-synchronised and share the same sampling rate value of 20 kHz [14].
We have shown in [2] that opening is more regular than closure instant on the EGG signal. Thus the glottal flow presents approximately the same behaviour in the neighbourhood of these instants. These singularities are smoothed due to the effect of the vocal tract and then the speech signal presents smoothed singularities at these instants. It is then

more difficult to detect them especially at GOI which is characterised by a more regular behaviour.
In an effort to circumvent these problems, it is argued that if we use a non linear combination of wavelet transforms of speech signal at different scales, we can give more accurate estimation of the GOI. The wavelet given by the equation (2) is used in this work and the multiresolution product analysis is operated for three scales $s_1=2^2$, $s_2=2^3$ and $s_3=2^4$.

$$g(t) = -\cos(2\pi f_0 t).\exp(-t^2/2\tau^2) \qquad (2)$$

With $\tau = 1/2f_0$, f_0= Fe /2 and Fe=20 kHz.

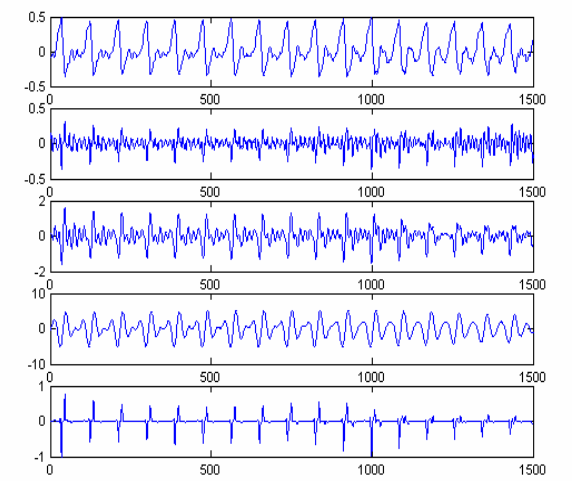

Figure 2: From up to bottom: speech signal (female voice f1) – speech wavelet transforms from scale 2^2 to scale 2^4- normalized multiscale product of the 3 scale speech WT.

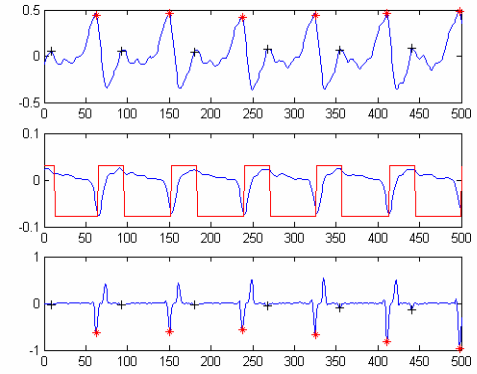

Figure 3: GOI and GCI detection for a voiced speech signal (female voice /o/). From up to bottom: speech signal - EGG- normalized product of the 3 scale speech WT. *Symbols*: * GCI, + GOI.

Figure 2 and 4 depict respectively frame of vowel /o/ extracted from the word /north/ and uttered by a female and male speakers. Each of the two figures depicts the speech signal followed by its three wavelet transforms and at the bottom the cross-scale normalized product p(n). Firstly, we note in the cross scale product two types of minimum peaks; those corresponding to GCI are the most distinguishable. We can see clearly the effect of the product in suppressing the additional noise peaks and consequently the best detection of GOI.

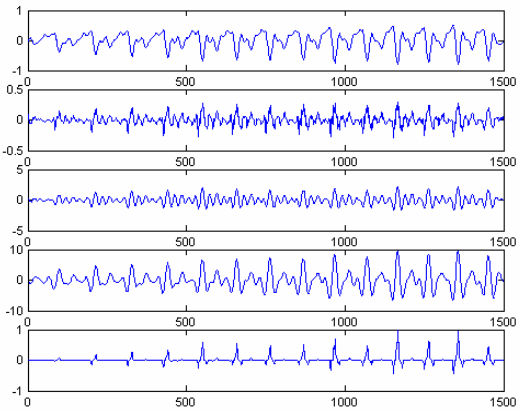

Figure 4: From up to bottom: speech signal (male voice) – speech wavelet transforms from scale 2^2 to scale2^4 - normalized multiscale product of the 3 scale speech WT.

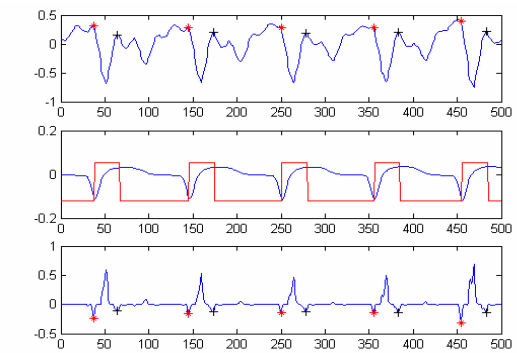

Figure 5: GOI and GCI detection for a voiced speech signal (male voice /o/). From up to bottom: speech signal - EGG- normalized multiscale product of the 3 scale speech WT. *Symbols*: * GCI, + GOI.

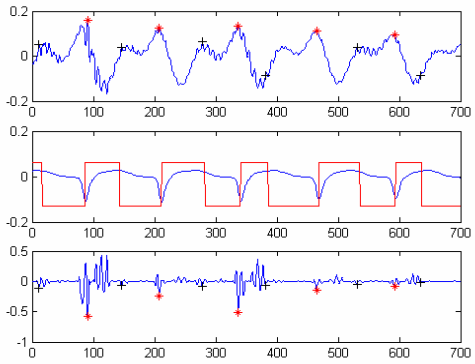

Figure 6: GOI and GCI detection for a voiced speech signal (male voice /i/). From up to bottom: speech signal - EGG- normalized multiscale product of the 3 scale speech WT. *Symbols*: * GCI, + GOI.

With reference to the GCIs and GOIs detected from the laryngograph signal, a robust strategy of GCI and GOI localization was deduced. Note that the derivative of EGG signals represented in figures 3, 5 and 6 are smoothed by a zero phase filter to show the GOI that is less obvious and can not be detected in some cases from the EGG signal [2].
The great minimum peak reinforced by the multiscale products corresponds to the sharpest variation in speech signal i. e. the GCI, however the smaller minimum peak situated between two successive GCIs is linked to the GOI. Figure 3 and 5 shows the efficiency of the proposed strategy for the two utterances of vowel /o/ comparing to the GOI given by the maximum of the filtered EGG. Figure 6 is an illustration of the efficiency of the proposed method for the vowel /i/ pronounced by a male speaker.

6. CONCLUSION

Detection and estimation of speech edges generated at GOI are considered. Characteristic points of speech signal are determined by a multiscale product method. This method consists of computing the wavelet transform of acoustic speech signal at various scales, forming the product of the wavelet transform coefficients and then looking for minimums to localize GOI.
Odd number of scales guarantees the sign edge preservation. The non linear combination permits to reinforce the cross-scale peaks produced at GOI and reduces spurious noise. Efficiency of this method is presented on three examples of speech utterance signal.

REFERENCES

[1] A. Bouzid and N. Ellouze, "Caractérisation des singularités du signal de parole," in Proc. GRETSI 2003, Paris, Sep. 2003, pp. 273-276.

[2] A. Bouzid and N. Ellouze, "Local Regularity Analysis at Glottal Opening and Closure Instants in Electroglottogram Signal Using Wavelet Transform Modulus Maxima," in Proc. EUROSPEECH 2003, Geneva, Sep. 2003, pp. 2837-2840.

[3] C. Sturt, S. Villette and A. M. Kondoz, "LSF quantization for pitch synchronous speech coders," in Proc. ICASSP, Korea, 2003.

[4] O. O. Akande and P. J. Murphy, "Improved glottal wave estimation using split band inverse filtering," in Proc. AQL, 6th International Conference Advances in Quantitative Laryngology, Voice and Speech Research, N° 25, Apr. 2003.

[5] M. Tooher, "Machine learning of speaker characteristic speech dynamics and interactions," Transfer to Ph. D. Talk, Technical report, Feb. 2003.

[6] A. Ben Slimane, A. Bouzid and N. Ellouze, "Wavelet decomposition of voiced speech and mathematical morphology analysis for glottal closure instants detection," in Proc. EUSIPCO 2002, Toulouse, Sep. 2002, pp. 81-84.

[7] J. G. McKenna, "Automatic glottal closed-phase location and analysis by Kalman filtering," in Proc. ISCA 2001, Tutorial and Research Workshop on Speech Synthesis, Pittlochrie, Scotland, 2001.

[8] T. Vu Ngoc and C. d'Alessandro, "Robust glottal closure detection using wavelet transform," in Proc. EUROSPEECH 1999, Budapest, Sep. 1999, pp. 805-808.

[9] D. M. Brookes and H. P. Loke, "Modelling energy flow in the vocal tract with applications to glottal closure and opening detection," in Proc. ICASSP 1999, May 1999, pp. 213-216.

[10] S. Mallat, A wavelet tour of signal processing. Second Edition, Adress: Academic Press, 1999.

[11] B. M. Sadler, T. Pham, and L. C. Sadler, "Optimal and wavelet-based shock wave detection and estimation," Journal of the Acoustical Society of America, vol. 104 (2), pp. 955-963, Aug. 1998.

[12] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," IEEE Transactions on Speech and Audio Processing, vol. 6, pp. 313-327, Jul. 1998.

[13] L. Janer, J. J. Bonet and E. L. Lleida-Solano, "Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms," in Proc. ICSLP'96, Philadelphia, PA, USA, Oct. 1996.

[14] F. Plante, G.F. Meyer and W.A. Ainsworth, "A pitch extrac tion reference database," in Proc. EUROSPEECH 1995, Madrid, Sep. 1995, pp. 837-840.

[15] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signal," IEEE transactions on Information Theory, vol. 38, N° 2, pp. 917-924, 1992.

[16] D. Childers and C. K. Lee, "Vocal quality factors: analysis, synthesis and perception," Journal of the Acoustical Society of America, vol. 90, pp. 2394-2410, Nov. 1991.

[17] D. H. Detring, "Pitch-synchronous linear prediction," Cambridge papers in Phonetics and Experimental Linguistics, vol. 5, pp. 1-13, 1986.

[18] A. K. Krishnamurthy and D. G. Childers, " Two-channel speech analysis," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 34, no. 4, pp. 730-743, 1986.

[19] D. E. Veeneman and S. L. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 33, pp. 369-377, Apr. 1985.

[20] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis linear prediction of the speech waveform," Journal of the Acoustical Society of America, vol. 50, pp. 637-655, 1971.

[21] D. Y. Wong, J. D. Markel and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, pp. 350-355, Aug. 1970.