A MULTIRESOLUTION APPROACH TO CLOSED GLOTTIS INTERVAL DETERMINATION

Atanas Gotchev[1], *Karen Egiazarian*[2], *Tapio Saramäki*[2]

[1]Tampere International Center for Signal Processing,
[2]Signal Processing Laboratory, Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, FINLAND
Tel: +358 3 365 2929; fax: + 358 3 365 3857
e-mail: {agotchev, karen, ts}@cs.tut.fi

ABSTRACT

We consider the closed glottis interval determination problem from the point of view of getting a multiresolution signal representation and searching for maxima in successive scales. Within this framework, an octave band decomposition, using half-band finite impulse response (FIR) filters is very promising, since those filters are less constrained than the orthogonal and bi-orthogonal wavelet filters. A number of experiments with synthesized signals argue the superior performance of the multiresolution approach in comparison with the traditional covariance approaches.

1. INTRODUCTION

The glottal source waveform is an important characteristic to determine in voice analysis, speaker emotional state identification, naturally sounding speech synthesis, etc. It is obtained most frequently through an inverse filtering. The basic stage in the inverse filtering is getting an adequate vocal tract model. The latter is usually of an autoregressive type. Its coefficients are used to define the inverse filter. An important part of vocal tract modeling is the proper choice of the length and localization of the analyzing frame. It is most commonly placed in the closed glottis interval (CGI). During the CGI, the vocal tract is free oscillating and resembles a pure autoregressive process. Most of the inverse filtering methods choose an analyzing frame being 35-40% of the pitch period long and locate it after the glottal closure instant (GCI). For the GCI detection, several measures can be applied, e.g. the residual prediction error [1], the minimized squared prediction error [3], the Frobenius measure of the voice signal covariance matrix [4], the wavelet decomposition maxima [5], etc. A drawback of these approaches is that only the GCI is determined without detecting the glottal opening instant (GOI) and, thus, without finding the CGI length. In an earlier work [6] the use of Battle-Lemarie wavelets for both GCI and GOI determination has been proposed, based on the wavelet property to detect signal transients.

In the present work we determine the actual CGI by over-complete multiresolution pyramidal decomposition. We collect a set of linear phase decomposition filters and study their performance for different synthesized sounds in different noisy conditions.

2. CGI DETERMINATION

The voice production channel can be modeled to consist of an excitation source or oscillator and of a resonating part, as shown in Fig. 1.

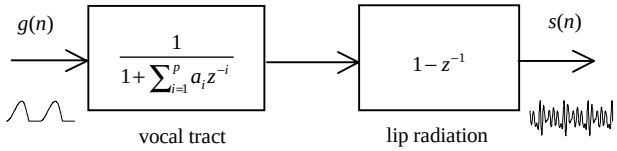

Figure 1. The "source-tract" model of the voice production channel.

From an anatomical point of view, the source is represented by the glottis (the area where the vocal folds are situated) while the resonating part is represented by the vocal tract. From our model point of view, the vocal tract is considered as an all-pole filter excited by the voice source while the lip radiation impedance is considered as a differencing filter [7]. The glottal wave generated by the voice source is quasi-periodic for the vocalized sounds (Fig. 2b). Every glottal wave's period consists of phases, related to vocal folds' abduction and adduction referred also as glottal closure and glottal opening. The glottal wave has more abrupt closing slope and more slanting opening slope.

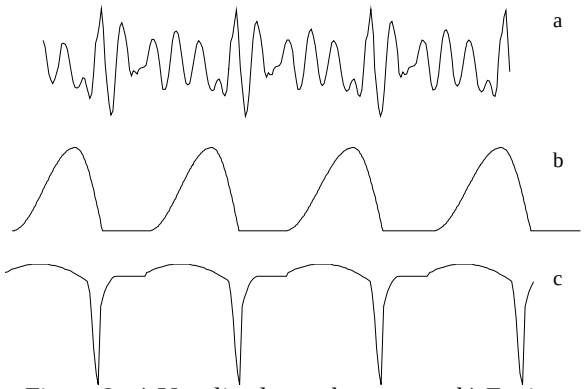

Figure 2. a) Vocalized speech segment; b) Excitation glottal wave; c) Differentiated glottal wave.

Thus, the main excitation occurs at the instant of glottal closure. The differentiation of the main excitation results in a very sharp pulse at the instant of the glottal closure (Fig. 2 c). It is the same if we change the places of the lip radiation system and vocal tract system and use the differentiated glottal wave (DGW) as an excitation (input) signal for the vocal tract while the output is the speech signal. Hence the greatest energy changes in the output speech signal will be provoked by the glottal closures and there will be also energy changes due to the glottal openings.

2.1. Covariance methods for GCI determination

There are several GCI detection criterions exploiting the covariance structure of the speech signal.

One of the first approaches was proposed by Strube [2]. He suggested that the maxima of the logarithm of the autocovariance signal matrix's determinant detect the transitions from opened glottis phase to closed one.

The next criterion was proposed by Wong *et al.* [3]. It is based on the usage of the overall minimized squared prediction error. For the current sliding analyzing segment *m* with length *N*, this error is given by

$$Cr_W = \min\left\{\alpha(m) = \sum_{j=m}^{m+N-p-1}\left[s(j)+\sum_{k=1}^{p} a_k s(j-k)\right]^2\right\}. \quad (1)$$

Here *p* is the linear prediction order.

Theoretically, if the analyzing frame is within the closed glottis interval then the sum of residual prediction errors must be equal to zero (precise prediction) and the total error α must be equal to zero too. If the analyzing frame is no longer within the range between glottal closure and glottal opening instants, then the assumption for pure autoregressive model of the vocal tract is not valid and α cannot be reduced to zero. Then, at the first sample with α=0, glottal closure is detected.

For actual speech data, Cr_w defines a squared error measure checking how adequate is the covariance prediction model during the closed phase. The actual squared error curve is characterized by flat top and bottom areas with many ripples. It cannot reach zero, and it is difficult to associate the absolute minimum with the glottal closure. Hence a threshold is used.

The next criterion is based on a singular value decomposition and was proposed by Ma *et al.* [4], as a generalization of the Strube's and Wong's methods. For the signal segment *m* in the matrix form one gets:

$$\mathbf{S}(m) = \begin{bmatrix} s_{m+p} & s_{m+p-1} & \cdots & s_m \\ s_{m+p+1} & s_{m+p} & \cdots & s_{m+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m+N-1} & s_{m+N-2} & \cdots & s_{m+N-1-p} \end{bmatrix} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T, \quad (2)$$

where $\mathbf{U}_{(n-p)\times(n-p)}$ and $\mathbf{V}_{(p+1)\times(p+1)}$ are square orthogonal matrices and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_0, \sigma_1, \ldots \sigma_p)$ contains the singular values of **S**. Hence the Strube's criterion becomes

$$Cr_S = \log\left[\det\left[\mathbf{S}^T\mathbf{S}\right]\right] = \log\left[\prod_{i=0}^{p}\sigma_i^2\right] = \sum_{i=0}^{p}\log(\sigma_i^2). \quad (3)$$

The Wong's criterion itself can be associated with the minimal singular value σ_0. The new criterion, proposed in [4], is just the Frobenius norm of the matrix **S**, which can be expressed in terms of the singular values as:

$$Cr_M = \|\mathbf{S}\|_F = \left(\sum_{i=0}^{p}\sigma_i^2\right)^{1/2}. \quad (4)$$

This norm is bounded by the Strube's criterion and the Wong's criterion [4]. These lower and upper bounds both increase in the open phase and both decrease in the closed phase that determines the same behavior of the Ma's criterion. Furthermore, the latter is very easy to compute [4].

2.2. Multiresolution methods for CGI determination

As we mentioned above, the amplitude changes in the speech signal can be related with the glottal wave's phases. It is natural to express these amplitude changes in successive scales and search for local maxima corresponding to both the GCI and the GOI. Assuming dyadic scales, this leads to an over-complete signal decomposition by octave band filter banks [8]. The idea has been inspired by the Mallat's research [9] and can be considered also as a wavelet frame decomposition if filters representing wavelet bases are used. From this point of view it is an extension of [5].

In order to preserve the initial signal resolution in all octaves, we have performed a pyramidal decomposition using undecimated octave band filter banks [8]. Suppose the filter bank is formed by a low-pass filter *H*(*z*) and a high-pass filter *G*(*z*). Then, the decomposition signals in successive octaves can be represented as follows:

$$\begin{aligned} S_1(z) &= G(z)S(z); \\ S_2(z) &= G(z^2)H(z)S(z) \\ S_3(z) &= G(z^4)H(z^2)H(z)S(z), \text{etc.} \end{aligned} \quad (5)$$

Here *S*(*z*) is the initial speech segment in the *z*-domain. The decomposition scheme is shown in Fig. 3.

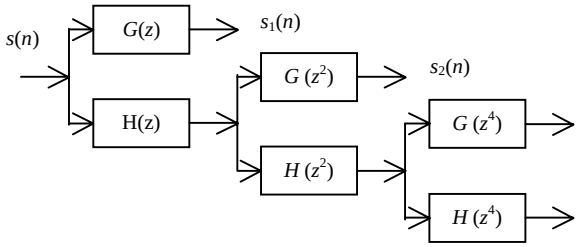

Figure 3. Octave-band pyramidal decomposition.

As we need to detect time phenomena, the appropriate filters for our task should have a linear phase and certain time-frequency localization properties. We have collected a set of filters as follows: filter banks corresponding to Battle-Lemarié wavelets of different orders [10], [11]; two-channel filter banks corresponding to biorthogonal spline wavelets [12], [13]. While they had been designed with some constraints required for the perfect reconstruction, we do not need any synthesis part in our decomposition scheme. Hence, we can exploit those more degrees of freedom in designing shorter FIR filters with a better frequency selectivity. This better frequency selectivity would provide us with registering the more slanting glottal opening fronts. Within this assumption we have constructed half-band filters of different orders optimizing them in the least-mean-square sense. For the properties of half-band FIR filters and for the details on how to implement effectively a lowpass – highpass complementary half-band filterpair, see, e.g. [14].

Using octave-band decompositions, we have observed that there are multiscale maxima related with the GCI, as reported in [5]. The maxima can be retraced through the octaves starting from the octave containing the pitch frequency. In addition, there are also maxima related with the GOI, which are clearly detectible in the octave preceding the pitch octave.

3. EXPERIMENTAL RESULTS

To verify the applicability of the octave-band decomposition mentioned above we have carried out a number of experiments with synthesized vocalized voice signals. Each synthesized vowel has been obtained through exciting a linear prediction vocal tract model with a differentiated glottal waveform of Fujisaki-Ljungqvist model [15]. By using this model one can tune three amplitude and three timing parameters, giving the desired excitation source. We have detected the GCI and the GOI examining clean (noise-free) and noisy signals as well. We have also tried to detect the same phenomena in signals with fast pitch changes (highly nonstationary signals).

When detecting the GCI, all octave band decompositions showed superior performance in comparison with the covariance methods. It is especially true for the highly nonstationary signals and noisy signals with a very low SNR. For short-pitched signals the Wong's criterion is strongly influenced by the number of data in the analyzing frame and the order of the prediction filter. Its minimum often occurs within the closed phase, instead in the beginning. For noisy signals the Cr_w curve has no meaningful transitions from maximum to minimum, related with the GCI. Up to 10 dB, the Ma's criterion showed good results. It failed for very heavy noisy conditions (SNR=0 dB).

The octave-band decompositions offer one more possibility to detect also the GOIs, and hence the real glottis closure interval.

Between the filter sets, the case of half-band filters is quite promising because it assures the possibility to use shorter filters. As an example filter we have used a half-band filter of order ten. It has been optimized by minimizing the stopband energy in the interval $[0.75\pi, \pi]$. The corresponding frequency domain octave-band separation is shown in Fig. 4.

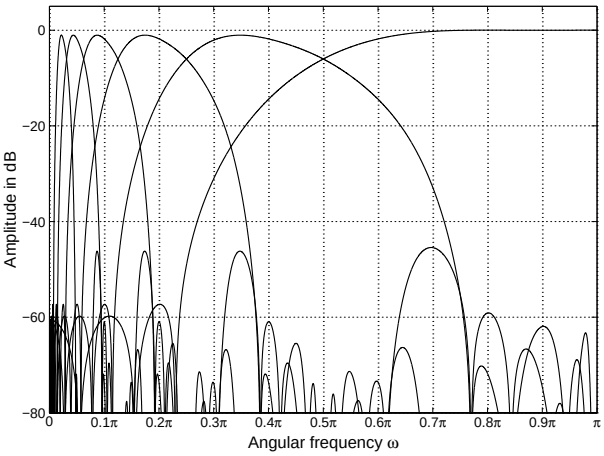

Figure 4. The amplitude response of the example half-band filters in the octave-band decomposition structure.

Table 1 shows some results obtained by applying this filter bank, for four different vowels in different noisy conditions. We have compared the relative mean and max deviation of CGI according to the pitch period.

Figure 5 illustrates the obtained results for a synthesized signal /a/ contaminated by white noise with SNR=0 dB.

The lower octave signal maxima in the octave 5 (this is the octave preceding the octave containing the pitch frequency for the concrete signal) are related with GOI and the maxima in octave 6 (the octave of pitch frequency) are related with GCI, respectively.

In the same figure (1e) the Frobenius measure of covariance signal matrix is presented [4]. It can be seen that the Frobenius measure fails completely in heavy noisy conditions while the multiresolution detector remains stable.

4. CONCLUSIONS

We have performed a multiresolution signal decomposition aimed at the CGI detection. We have chosen linear phase octave band filters with sufficient time-frequency localization, most of them representing wavelet bases, but also some half-band filters have been considered. We have compared their performance with the traditional covariance methods. It seems the multiresolution structure is the most promising in this task since the half-band filters (which are not wavelet filters at all) showed the quite acceptable performance. Moreover, we are not constrained with perfect reconstruction conditions and we can improve the frequency selectivity. This

is especially important in the GOI detection. Furthermore, we can also use shorter filters.

Vowel	Deviations, [%]	SNR, [dB]			
		90	20	10	0
/a/	Δ_{mean}	4.23	4.63	4.42	5.78
	Δ_{max}	5.45	5.63	5.63	6.53
/e/	Δ_{mean}	3.12	3.78	3.47	3.85
	Δ_{max}	3.71	6.67	3.71	4.17
/i/	Δ_{mean}	1.41	2.01	3.56	3.62
	Δ_{max}	2.14	2.82	4.22	4.25
/u/	Δ_{mean}	3.16	2.60	2.73	3.13
	Δ_{max}	4.58	3.12	3.12	3.44

Table 1. Deviations in determining the CGI for four vowels, using a half-band filter bank. The deviations are taken in percentages with respect to the current pitch period. The signals are mixed with white noise with different variances.

5. REFERENCES

[1] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval", IEEE Trans. ASSP, vol. 27, pp.309-319, 1979.

[2] H. Strube, “Determination of the instant of glottal closures from the speech wave” J. Acoust. Soc. Am., vol. 56, No. 5, pp. 1625-1629, 1974.

[3] D. Wong, J. Markel, and A. Gray, "Least squares inverse filtering from the acoustic speech waveform" IEEE Trans. ASSP, vol. 27, No.4, pp.353-362 1979.

[4] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal", IEEE Trans. Speech and Audio Proc., vol. 2, No. 2, pp. 258-264, 1994.

[5] S. Kadambe and G.F. Boudreaux-Bartels, “Application of the wavelet transform for pitch detection of speech signals”, IEE Trans. Inf. Theory, vol. 38, No. 2, pp. 917-924, 1992.

[6] Z. Nikolov and A. Gotchev, “Inverse filtering method for glottal waveform analysis using wavelet transform to determine the analyzing frame”, *in Proc. Med. Conf. Electronics and Automation*, Grenoble, France, Sept. 1995, pp. 201-206.

[7] J. Markel and A. Gray, "Linear prediction of speech", Berlin: Springer-Verlag, 1976.

[8] M. Veterli and H. Kovacevic “Wavelets and subband coding”, N.Y.: Prentice-Hall, 1995.

[9] S. Mallat and W. Hwang, “Singularity detection and processing with wavelets”, IEEE Trans. Inf. Theory, vol. 38, 1992, pp. 617-643.

[10] P. Lemarié, “Ondelettes à localisation exponentielles”, J. Math. Pures et appl., vol. 67, No. 3, pp.227-236, 1988.

[11] G. Battle, “A block spin construction of ondelettes. Part I: Lemarie functions” Commun. Math. Phys., vol. 110, pp. 601-615, 1987.

[12] A. Cohen, I. Daubechies and J. Feauveau, “Bi-orthogonal bases of compactly supported wavelets”, Comm. Pure Appl. Math., vol. 45, pp. 485-560, 1992.

[13] M. Unser, A. Aldroubi and M. Eden, “A family of polynomial spline wavelet transforms”, Signal Processing, vol. 30, No.2, pp. 141-162, 1993.

[14] T. Saramäki, “Finite impulse response filter design”, in Handbook for Digital Signal Processing, edited by S. Mitra and J. Kaiser, N.Y.: John Wiley and Sons, 1993, pp. 155-277.

[15] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of Models for the Glottal Source Waveform" *Proc Int. Conf. ASSP*, Tokyo, Japan, April 1986, pp. 1605-1608.

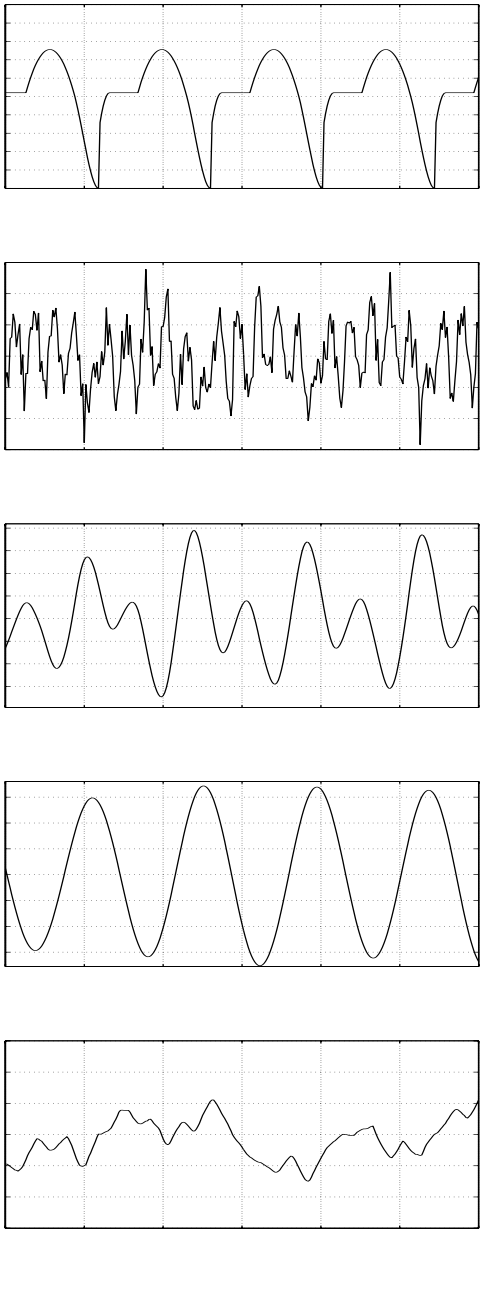

Figure 5. a) Excitation source signal (DGW); b) Voice signal contaminated by white noise with SNR=10 dB; c) The octave signal in the 5^{th} octave; d) The octave signal in the 6^{th} octave; e) The Frobenius measure for the signal in b).