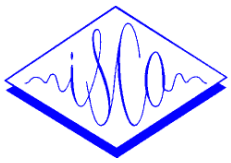


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A PITCH DETECTOR BASED ON EVENT DETECTION USING THE DYADIC WAVELET TRANFORM

Shubha Kadambe and G.F. Boudreaux-Bartels[1]

Dept. of Elect. Engg., University of Rhode Island, Kingston, RI-02881, USA

ABSTRACT

Several pitch detectors[1-9], which have been developed so far, are not always suitable for both low pitched and high pitched speakers and are not robust to noise. In this paper, we describe an event based pitch detector which overcomes the above mentioned problems. We estimate the pitch period by detecting the glottal closure, which we label here as an event, using a time-scale representation viz., the Discrete Wavelet Transform(DWT) and by measuring the time interval between two such events. We illustrate the applicability of this pitch detector to a wide range of pitch periods i.e. for both low pitched and high pitched speakers and its robustness to noise with various examples. We then highlight the merits and demerits of this method by comparing it with existing pitch detection methods. The results of this study indicate that the algorithm works well for noise free and noisy vowels as well as voiced consonants.

I. INTRODUCTION

The instant at which the glottis closes is defined here as an event. Pitch detectors which estimate the pitch period by locating the instant at which the glottis closes and then measuring the time interval between two such events are defined as event detection pitch detectors.

Only a few event based pitch detectors[6-9] were developed in the past several years. The instant at which the glottis closes is determined in [6] by locating the instant at which the determinant of the autocovariance matrix of a given signal is maximum. The advantage of this autocovariance method is: it can estimate the pitch period very accurately in the case of vowels that are produced by vigorous vocal cord vibrations with sharp glottal closure. However, the disadvantages of this autocovariance based method are: a) it is unsuitable for all vowels and non-stationary pitch periods and b) it is computationally complex. The pitch detection techniques in [7] and [8] use the occurrence of discontinuities in the derivatives of glottal airflow to detect the Glottal Closure Instant (GCI). These two methods can detect precisely the instant at which the vocal tract is excited. However, these two epoch extraction methods are applicable for only 'clean' data and certain vowels. Finally, the maximum likelihood epoch determination technique is applied in [9] to detect the GCI. This method gives accurate estimation of pitch period in the case of synthetic signals(all vowels), noisy signals(up to 0 dB Signal to Noise Ratio) and phase distorted signals. However, this method is not suitable for high pitched speakers since the data length available for the Linear Predictor could be very small. This method is also computationally intense.

Classical pitch detectors estimate the pitch period by a direct approach and hence they are referred to here as non-event based pitch detectors. Some of the non-event based pitch detectors developed so far [1-5] estimate the average pitch period by segmenting a speech signal using a window whose length is fixed and, by using the autocorrelation of the infinitely and centrally clipped signal in [1], by obtaining the cepstrum of a given segment of a signal in [2], by evaluating the autocorrelation of a inverse filtered signal in [3], by finding the average magnitude difference function of a given signal in [4] and by calculating the autocorrelation of a given signal and looking for the value of pitch period which maximizes the sum of the autocorrelation functions in [5]. These non-event based pitch detectors are computationally simple, however, they assume that the pitch period is stationary within each segment and each segment contains at least two full pitch periods. Hence the disadvantages of these pitch detectors are: they are a) insensitive to non-stationary variations in the pitch period over the segment length and b) unsuitable for both low pitched and high pitched speakers.

In this paper, we describe an event detection pitch detector which overcomes the above mentioned problems. We measure the pitch period by detecting the glottal closure(event) and by determining the time interval between two such events. During glottal closure, the vocal tract is strongly excited and there is an abrupt change(transient) in the speech signal. Hence this event can be detected by locating the occurrence of transients in a speech signal. We apply a time-scale representation viz., the Wavelet Transform (WT) for the task of locating these transients.

II. WAVELET TRANSFORM

In this section, we review the WT and list its properties that are useful for speech analysis.

[1]This work was supported in part by ONR grant # N00014-89-J-1812

The Continuous Wavelet Transform(CWT), a time-scale representation, was implemented recently by J. Morlet for the analysis of seismic data[12]. The CWT of a signal $x(t)$ is defined as,

$$CWT_x(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) g^*(\frac{t-b}{a}) dt \quad (1)$$

where the wavelet function(analysis window) $g(t)$, satisfies the conditions mentioned in [11]. The CWT is the convolution of a signal with an analysis window $g(t)$ shifted in time by a translation parameter 'b' and dilated by a scale parameter ' a'. The wavelet $g(t)$ is either compressed or expanded depending on the choice of 'a'. Hence, the CWT can extract both the local and global variations in the signal $x(t)$. Some of the properties of the CWT which makes it an useful tool for the analysis of speech signals are:

- the CWT is linear i.e. $CWT_{x_1+x_2}(b,a) = CWT_{x_1}(b,a) + CWT_{x_2}(b,a)$,
- the CWT is shift invariant i.e. if the signal $x(t)$ is shifted in time by t_0, its CWT is also shifted in time by t_0 [$CWT_{x(t-t_0)}(b,a) = CWT_x(b-t_0,a)$],
- the CWT is scale invariant i.e. if the signal $x(t)$ is scaled in time by λ, then its CWT is also scaled by the same amount $[CWT_{\sqrt{\lambda}x(\lambda t)}(b,a) = CWT_x(\lambda b, \lambda a)]$, and
- if the signal $x(t)$ or its derivatives have discontinuities then the modulus of the CWT of $x(t)$, $|CWT_x(b,a)|$, exhibits local maxima around the point of discontinuities and the lines of constant phase of the $CWT_x(b,a)$ converge toward the point of discontinuities.

The CWT is defined as the Discrete Wavelet Transform(DWT) if both the scale parameter 'a' and the translation parameter 'b' in equation(1) are discretized. The uniform discretization in the translation parameter 'b' makes the DWT translation variant. That is, if the signal $x(t)$ is shifted in time then its DWT is not shifted in time by the same amount necessarily. Hence, the DWT is not useful for speech as well as pattern analysis. However, if we discretize only the scale parameter 'a' in equation(1) then such a DWT retains the translation invariance property of the CWT. Therefore, in this paper we make use of such a DWT and we discretize the scale parameter along the dyadic sequence 2^j where $j = 1, 2, \cdots$. The DWT which uses the scale parameter that is discretized along the dyadic sequence is called the Dyadic Wavelet Transform(D_yWT) and is defined as[10]

$$\begin{aligned} D_yWT_x(b,2^j) &= \frac{1}{2^j} \int_{\infty}^{\infty} x(t) g^*(\frac{t-b}{2^j}) dt \\ &= x(t) \star g^*_{2^j}(t) \end{aligned} \quad (2)$$

where $g_{2^j}(t) = \frac{1}{2^j} g(\frac{t}{2^j})$ and $\star$ represents the convolution operator.

In [10] Mallat has shown that if we choose a wavelet function which is the first derivative of a smoothing function(a smoothing function is a function whose Fourier transform has energy concentrated in the low frequency region) $\theta(t)$, then the local maxima of the D_yWT indicate the sharp variations in the signal while local minima indicate the slow variations. In Figure 1, we plot the signal $x(t)$, $x(t)$ convolved with the smoothing function $\theta_{2^j}(t)$ where $\theta_{2^j}(t) = \frac{1}{2^j}\theta(\frac{t}{2^j})$, D_yWT computed using a wavelet function which is the first derivative of $\theta(t)$ ($D_yWT^1_x$) and D_yWT computed using a wavelet function which is the second derivative of $\theta(t)$ ($D_yWT^2_x$), respectively, from top to bottom. From Figure 1, we can see that in the case of $D_yWT^1_x$, the local maxima correspond to the sharp variations in the signal $x(t)$ while local minima correspond to the slow variations. However, in the case of $D_yWT^2_x$, the local minima correspond to both sharp and slow variations in the signal. Therefore, if we chose a wavelet which is the first derivative of a smoothing function, then the local maxima of $D_yWT^1_x$ locate the sharp variations in the signal $x(t)$.

III. DESCRIPTION OF THE EVENT BASED PITCH DETECTOR

In this section, we describe the event based pitch detector using the D_yWT which overcomes the problems mentioned in section I.

The pitch period can be estimated by locating the instant at which the glottis closes and by measuring the time interval between two such glottal closures. During glottal closure, the vocal tract is strongly excited which causes an abrupt change in a speech signal. In the previous section we showed that if we use a wavelet function which is a first derivative of a smoothing function, then the local maxima of the D_yWT using such a wavelet function indicate the sharp variations in the signal. We apply this technique to detect the instant at which the glottis closes. We use a cubic spline wavelet function which is a first derivative of a smoothing function shown in Figure 2. We detect the instant at which the glottis closes by locating the local maxima of the D_yWT which are $\geq \alpha\times$ (global maximum) where α is the threshold level. In this work, α is chosen to be equal to 0.8. We then estimate the pitch period by measuring the time interval between two such local maxima.

Generally, the D_yWT is computed at scales $a = 2^j$ for, theoretically, all j. However, since the pitch period is low frequency information (20 Hz - 500 Hz), we have observed experimentally that we need to compute the D_yWT only at scales of 2^4 or 2^5 in order to determine the pitch period accurately. Hence, in this work we compute the D_yWT only at the two above mentioned scales; this has the advantage of significantly reducing the computational complexity of the D_yWT.

IV. RESULTS AND DISCUSSION

In this section we illustrate the applicability of the event based pitch detector described in the previous section, for a wide range of pitch periods and its robustness to noise, using representative examples.

Case 1: First we consider a simple synthesized signal /u/(see Figure 3) with pitch period equal to 20 ms. Although, from Figure 3, it can be seen that the onset of each pitch period can be easily located even by a simple amplitude thresholding method, we give this simple example to demonstrate the behavior of the D_yWT when computed at different scale parameters. In Figure 4, we plot the D_yWT of the signal /u/ computed at dilation scale $a = 2^1, 2^2, \cdots, 2^5$. It can be seen that the D_yWT exhibits local maxima across all these scale parameters, at the instant of onset of each pitch period. However, from Figure 4, we can see that the high frequency information is filtered out as we increase the scale parameter. Hence, in order to estimate the pitch period accurately, we need to compute the D_yWT at scale 2^4 or 2^5. In this case we obtain 100% accuracy of the pitch period estimate.

Case 2: Second we consider a synthesized voiced fricative /v/ (see Figure 5) with pitch period equal to 10 ms. From Figure 5, it is clear that the onset of each pitch period can not be located very easily. In Figure 6, we plot the D_yWT computed at $a = 2^5$ and we can see that the D_yWT exhibits local maxima corresponding to the onset of true pitch period. In this case we obtain a minimum of 98% accuracy of the pitch period estimate.

Case 3: Next we consider a synthesized nasal /n/(see Figure 7). In this example we deliberately vary the pitch period (11.5 ms to 21.7 ms) from period to period to make the signal non-stationary. In Figure 8, we plot the D_yWT computed at $a = 2^4$ and we can see that there exists local maxima corresponding to the onset of each true pitch period. In this case also we obtain 100% accuracy of pitch period estimate. This example illustrates that the proposed event based pitch detector is capable of detecting non-stationary variations from period to period, unlike some pitch detection methods in [1-6].

Case 4: Now we apply the proposed pitch detector to real speech signal spoken by a male and a female speaker(Figures 9 and 11). We plot the D_yWT of these two signals computed at $a = 2^4$ in Figures 10 and 12, respectively, and we can see that the local maxima of the D_yWTs correspond to the beginning of pitch period in both cases. This example illustrates that the proposed pitch detector is suitable for both male and female speakers.

Case 5: Finally, in order to illustrate the robustness of the proposed pitch detector to noise we add white Gaussian noise to a synthesized signal /i/ at various Signal to Noise Ratio(SNR) down to -18 dB. In Figure 13, we plot the relative percentage of error((true pitch period - estimated pitch period)/ (true pitch period)) versus SNR. We can see that the relative percentage of error increases as the SNR decreases. However, we can still estimate the pitch period with a minimum of 99% accuracy down to -18 dB SNR.

V. ADVANTAGES AND CONSIDERATIONS

To summarize, in this section we list the advantages of the proposed event based pitch detector and some of the points to be noted while developing this pitch detector.

5.1 Advantages

The main advantages of this method in comparison with the existing pitch detectors are: this method

- does not assume stationarity or quasi-stationarity within the analysis window,
- estimates the pitch period very accurately(minimum of 99% accuracy for SNR $\geq$ -18 dB),
- is suitable for a wide range of pitch period(3 ms - 40 ms),
- can detect the beginning of a pitch period and the number of pitch periods present in a given segment of a speech signal and, hence, can be used for pitch or event synchronous modeling application,
- is computationally simple since we need to compute the DWT only at scale 2^4 or 2^5 and
- exhibits superior performance[14] as compared to the pitch detectors described in [1-9].

5.2 Considerations

- The accuracy of pitch period estimation depends upon the choice of the wavelet function[13].
- There is need for a slightly more complicated algorithm for voiced fricatives. In this case, the accuracy obtained was up to 98%[13].

VI. CONCLUSION

In this paper we have described an event based pitch detector using the DWT. We have illustrated its applicability to both low pitched and high pitched speakers and its robustness to noise with various examples. We have also shown that this method is computationally simple. Hence, this method alleviates the problems associated with the existing pitch detectors. The results of this study also indicate that this method works well for both noise free and noisy vowels and voiced consonants.

REFERENCES

1. M. M. Sondhi, *New methods of pitch extraction*, IEEE Trans. Audio Electroacoust., Vol. AU-16, pp. 262-266, June 1968.
2. A. M. Noll, *Cepstrum pitch determination*, J. Acoust. Soc. Amer., Vol 41, no. 2, pp. 293-309, 1970.
3. J. D. Markel, *The SIFT algorithm for fundamental frequency estimation*, IEEE Trans. Audio Electroacoust., Vol. AU-20, pp. 367-377, December 1972.
4. M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley, *Average magnitude difference function pitch extractor*, IEEE Trans. on ASSP, Vol. ASSP-22, pp. 353-362, 1974.
5. J. D. Wise, J. R. Caprio and T. W. Parks, *Maximum likelihood pitch estimation*, IEEE Trans. on ASSP, Vol. ASSP-24, pp. 418-423, 1976.

6. H. W. Strube, *Determination of the instant of glottal closure from the speech wave*, J. Acoust. Soc. Amer., Vol 56, no. 5, pp 1625-1629, 1974.

7. T. V. Ananthapadmanabha and B. Yegnanarayana, *Epoch extraction of voiced speech*, IEEE Trans. on ASSP, Vol. ASSP-23, pp. 562-570, 1975.

8. T. V. Ananthapadmanabha and B. Yegnanarayana, *Epoch extraction from linear prediction residual for identification and closed glottis interval*, IEEE Trans. on ASSP, Vol. ASSP-27, pp. 309-319, 1979.

9. Y. M. Cheng and D. O'Shaughnessy, *Automatic and reliable estimation of glottal closure instant and period*, IEEE Trans. on ASSP, Vol. ASSP-37, December 1989.

10. S. G. Mallat and S. Zhong, *Complete signal representation with multiscale edges*, Robotics Research Technical report no. 483, Robotics report no. 219, New York University Courant Institute of Mathematical Sciences, December 1989.

11. R. Kronland-Martinet, J. Morlet and A. Grossmann, *Analysis of sound patterns through wavelet transforms*, Intnl. Journal of Pattern Recognition and Artificial Intelligence, Vol. 1, no. 2, pp 273-302, 1987.

12. J. Morlet, G. Arens, I. Fongeau, and P. Giard, *Wave Propagation and Sampling Theory*, Geophysics, 47, pp. 203-236, 1982.

13. S. Kadambe, *The application of Time-Frequency and Time-Scale Representations for Speech Analysis*, Ph.D dissertation, Dept. of Elect. Engg., University of Rhode Island, Kingston, Rhode Island, in preparation.

14. S. Kadambe and G. F. Boudreaux-Bartels, *A comparison of a Wavelet Transform event detection pitch detector with classical pitch detectors*, To be presented at the Twenty fourth annual Asilomar conference on Signals, Systems and Computers at Pacific Grove, California, Nov 5-7, 1990.

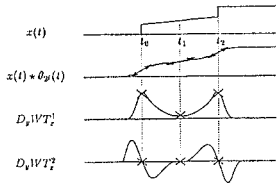


Figure 1: $x(t)$, $x(t) * \theta_\psi(t)$, $D_\psi WT_x^1$ and $D_\psi WT_x^2$, respectively(from top to bottom).

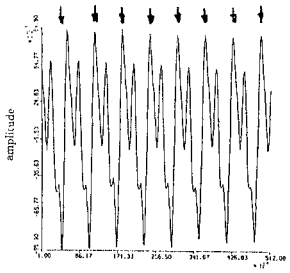


Figure 11: A real speech signal(letter 'o') spoken by a female speaker.

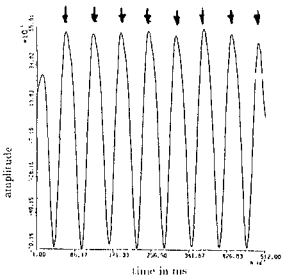


Figure 12: The $D_\psi WT$ of 'o' computed at $a = 2^4$.

* The arrows in all the above figures indicate the onset of true pitch period.

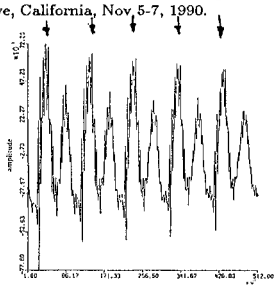


Figure 9: A real speech signal(letter 'e') spoken by a male speaker.

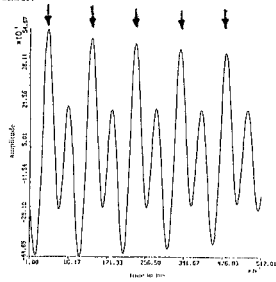


Figure 10: The $D_\psi WT$ of 'e' computed at $a = 2^4$.

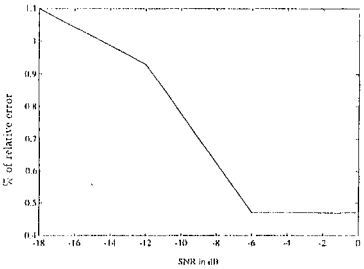


Figure 13: Percent of relative error in pitch period estimation of a speech signal(/i/) embedded in white Gaussian noise.

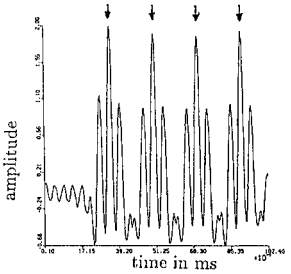


Figure 3: A synthesized signal /u/.

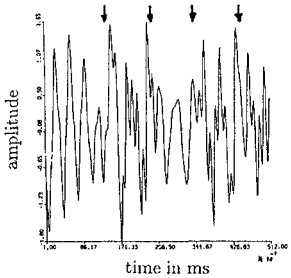


Figure 5: A synthesized voiced fricative /v/.

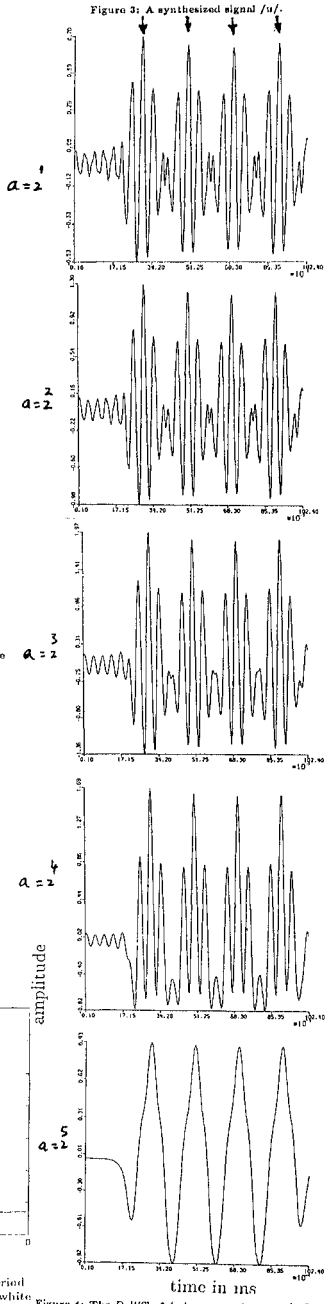


Figure 4: The $D_\psi WT$ of /u/ computed at $a = 2^1, 2^2, \ldots, 2^5$, respectively(from top to bottom).

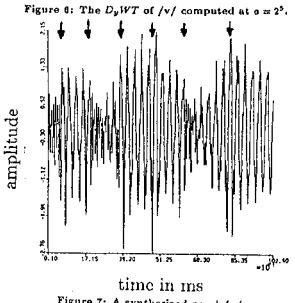


Figure 6: The $D_\psi WT$ of /v/ computed at $a = 2^5$.

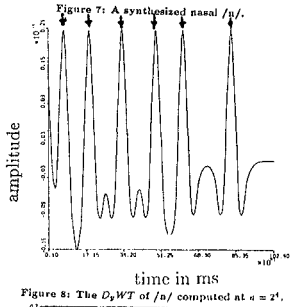


Figure 7: A synthesized nasal /n/.

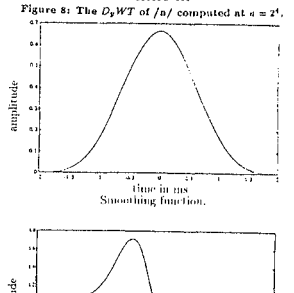
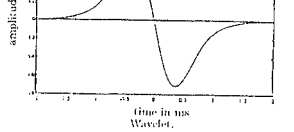


Figure 8: The $D_\psi WT$ of /n/ computed at $a = 2^4$.



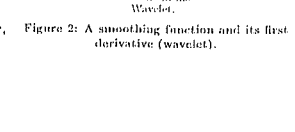


Figure 2: A smoothing function and its first derivative (wavelet).