A Reliable Pitch Determination Algorithm (PDA) Based on Dyadic Wavelet Transform (DyWT)

Namhoon Kim*, Yongsung Kang*, Hanseok Ko*

Abstract

This paper presents a time-based Pitch Determination Algorithm (PDA) for the reliable estimation of Pitch Period (PP) in speech signals. Based on the Dyadic Wavelet Transform (DyWT), the proposed PDA detects the presence of Glottal Closure Instants (GCI) and uses the information to determine the pitch period. We also examine the problem of conventional PDAs based on DyWT; their performance is compared with the proposition of this paper. The effectiveness of the proposed method is tested with real speech signals containing a transition between the voiced and the unvoiced interval where the energy of the voiced signal is unsteady. The result shows that the proposed method provides good performance in estimating both the unsteady GCI positions as well as the steady parts.

Keyword : GCI, Pitch period, U/V detection, Dyadic wavelet transform

1. Introduction

Accurate PP detection can improve the performance of many speech applications requiring signal processing, e.g speech recognition, speaker verification, and lip-synch since PP provides a discerning feature useful for classifying phonemes. In short, it can be used as an indicator for phoneme segmentation. However, reliable and accurate PP detection poses a formidable challenge for the following reasons. First of all, the human vocal tract is very flexible and as a result, the characteristics of vocal tract vary widely depending on the individual. Second, even though the information of PP comes from the same speaker, it can be changed by the emotional state of the speaker. Finally, PP can be affected by the individuals speaking style, e.g. accent or intonation.

This paper focus mainly on the event-based PDA, more specifically, that based on DyWT. The wavelet transform is a multi-scale analysis which has been shown to be

* Dept. of Electronics Engineering, Korea University.

very well suited for speech signal processing in that it decomposes speech signals into a series of band-pass components. It is similar to the way human ears process sound. Kadambe and Bordreaux-Bartels used the assumption that as GCI occurs in speech signals, maximums also occur in the adjacent scales of the wavelet transform. Therefore, the classification between the unvoiced and the voiced can not be done by only comparing only the maximum amplitude of the DyWT with some threshold, but by also checking whether the local maxima of the DyWT is identical across the two scales [1, 4]. However, although Kadambe's work shows relatively good performance for detecting steady GCI, unfortunately, it does not provide reliable results for estimating unsteady GCI positions such as the transition between voiced and unvoiced or the beginning and ending of the voiced where the energy is not steady. Wendt's method [2], another PDA based on DyWT, used a filtering function, which is linked to the bandwidth properties of the wavelet transform at different scales. It shows better performance in detecting unsteady GCIs in comparison to Kadambe's method, because it does not use a threshold to detect GCIs. But it also suffers from unwanted maxima, which incur PP detection errors. As a remedy to the problem discussed in this paper, we suggest new PDA. The proposed PDA explores the compensation of the above-mentioned two PDAs based on DyWT to be free from PP estimation errors.

This paper is organized as follows. A brief description of the adopted DyWT is presented in Section 2; Section 3 discusses detail problem of conventional PDAs based on DyWT; and Section 4 gives the proposed PDA. Then, Sections 5 and 6 present the representative results and conclusions respectively.

2. Background

The DyWT of signal x(t) is defined as, [1, 3]

$$DyWT_x(b,2^j) = \frac{1}{2^j}\int_{-\infty}^{\infty} x(t)g^*\left(\frac{t-b}{2^j}\right)dt = x(t) * g^*_{2^j}(t) \quad (1)$$

where $g_{2^j}(t)$ is the dilated and scaled version of the wavelet. Note that b and 2^j are limited to integer, and that $DyWT(b,2^j)$ are the Dyadic wavelet transform coefficients representing the wavelet transform at each 2^j scale.

The DyWT acts as a constant-Q filter bank and splits the signal into bandpass components. This is very useful in the analysis of characteristics of the signal, which are localized in frequency. In realization of Kadambe's method in this paper, the cubic

spline wavelet is used, and $DyWT(b, 2^j)$ at scale $j = 3, 4$ is used to estimate the GCI position. Mallat's algorithm is also utilized for fast implementation of DyWT [2]. On the other hand, haar wavelet function at scale $j = 3$, and scaling function at scale $j = 6$ are used to implement Wendt's PDA [2]. Finally, in the proposed method, both wavelet function and scaling function at scale $j = 5$ are adopted to construct a filtering function in a similar fashion as Wendt's.

3. The Conventional PDA based on DyWT

3.1 S. Kadambe's Method

Kadambe's method [1, 4], which compares the maximum amplitude of the DyWT with a certain threshold level T in addition to checking whether the local maxima of the DyWT, correlates across two scales. However, we have found through extensive experimentation that the fixed threshold is too strong in some low energy voiced segments as in the transition between voiced and the unvoiced or the beginning and ending of the voiced where the energy is not steady. As shown in Fig. 1, we failed as the accurate estimation of GCI position using their method.

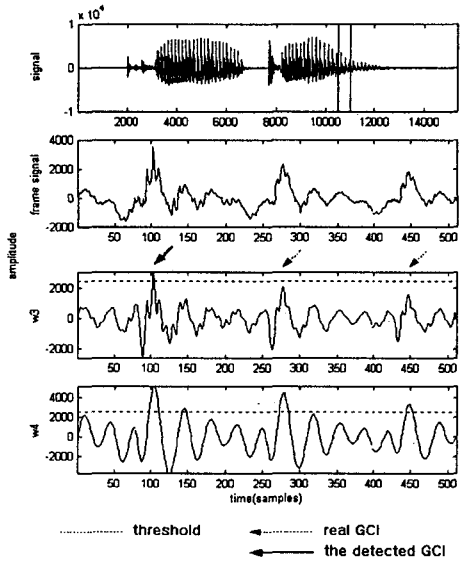

Fig.1 Detection of GCI's with Kadambe and Bordreaux-Bartels's method

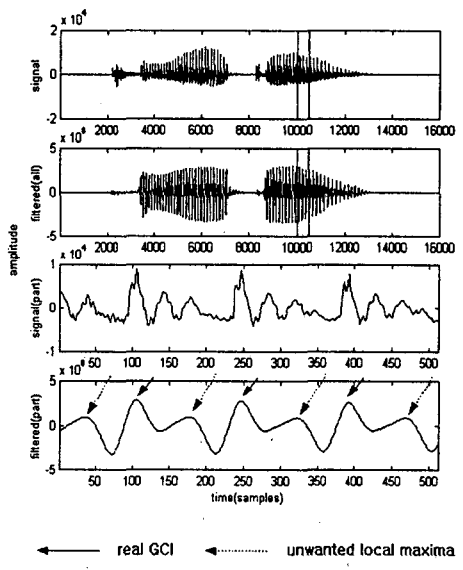

Fig.2 Detection of GCI's with Wendt and A. Petropulu's method

3.2 C. Wendt's Method

In Wendt's method, the idea is to use a wavelet with derivative properties, described by Mallat [3], that also combine the bandwidth properties of the wavelet transform at different scales. Since the frequency range of voiced speech is between *30-500 Hz*, a filtering function is constructed to cover a similar range by using both lowpass scaling function and highpass wavelet function. The filtering function $\rho(t)$ is obtained as below,

$$\rho(t) = \psi_{K_a}(t) * \phi_{K_b}(t) \qquad (2)$$
$$s'(t) = s(t) * \rho(t)$$

where $\psi_{K_a}(t)$ and $\phi_{K_b}(t)$ are wavelet and scaling functions, K_a and K_b are corresponding upper- and lower- bound scales, $s'(t)$, $s(t)$, and $\rho(t)$ are the filtered signal, speech signal and derivative filtering function, respectively [2].

However, as shown in Fig. 2, even though the speech signal is filtered out in the range of about *30-500 Hz*, the unwanted local maxima still remain to cause GCI detection errors. Furthermore, although it is possible to get rid of unwanted local maxima as increasing scaling index, in this case, there exists a significant difference between GCIs obtained and real GCIs.

4. The Proposed Method

The proposed method also follows the C. Wendt and A. Petropulu's method [2], which constructs filtering function with both lowpass scaling function and highpass wavelet function to cover the range of the voiced, that is, about *30-500 Hz*. But, unlike Wendt's method, we utilize the filtering function, which is made of cubic spline wavelets at scale $j = 5$; it also covers the frequency range between *30-500 Hz* approximately. The proposed method in the following subsections consist of two important stages; the variant threshold stage and the GCI detection stage.

4.1 Variant threshold stage

The threshold is imposed according to the detected local maxima. As shown in Fig. 3, once the local maxima have been detected, the proposed PDA takes its threshold from the lowest local maxima value to the highest. And, simultaneously, thresholded local maxima are compared with this variant threshold. The adopted threshold is given below.

$$threshold = s'(n) \qquad 1 \le n \le M \tag{3}$$

where $s'(n)$ is the frame signal and M is the number of local maxima within the frame.

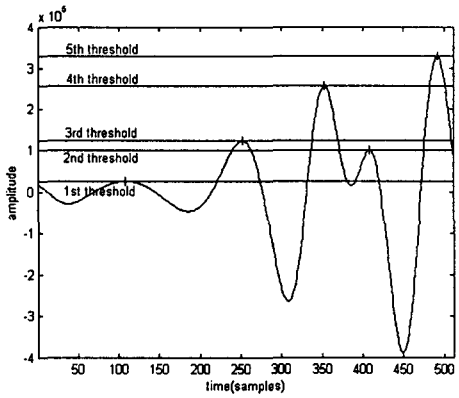

Fig.3 Variant threshold

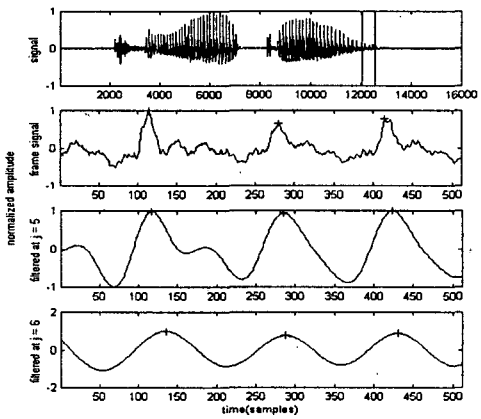

Fig.4 Comparison of GCI position in the filtered signal at scale j = 5 and 6

4.2 GCI detection stage

With given thresholded local maxima, it is affirmed whether or not that they are GCIs. If it is assumed that local maxima above or equal to each variant threshold are n_k, then the distance between them and the mean of p_k are defined as below,

$$\begin{cases} p_k = n_{k+1} - n_k \\ p_m = \frac{1}{N} \sum_{k=1}^{N-1} p_k \end{cases} \tag{4}$$

$$|p_m - p_k| \le \epsilon \quad \text{where k = 1, ···, N-1} \tag{5}$$

where N is the number of local maxima within the frame, p_m and are the mean of p_k and tolerance, respectively.

If p_k in the analyzed frame is quasi-periodic, each of p_k is almost the same as p_m within a relatively small tolerance, ϵ. It is adopted 0.125ms throughout this paper. However, if p_k is aperiodic due to one or two n_k, it needs another treatment. As noted in (6), as scale k increases, the bandwidth of filtering function, which is composed of both $\psi_k(n)$ and $\phi_k(n)$, decreases as much as 2^{-k} in the frequency domain. In other words, the signal is filtered at the higher scale, the periodicity

information is higher in that quasi-periodic factors in the signal get filtered out, on the contrary, the gap between real GCI and local maxima increases,

$$\begin{aligned} s'(n) &= s(n) * \rho(n) \\ &= s(n) * \psi_k(n) * \phi_k(n) \\ S'(w) &= 2^{-k} S(w) \cdot \Psi(2^{-k}w) \cdot \Phi(2^{-k}w) \end{aligned} \tag{6}$$

where $s'(n)$, $s(n)$, $\rho(n)$, $\psi_k(n)$, and $\phi_k(n)$ are filtered signal, analyzed signal, filtering function, wavelet, and scaling function, respectively. Also, $S'(w)$, $S(w)$, $\Psi(2^{-k}w)$, and $\Phi(2^{-k}w)$ are Fourier transform pairs of $s'(n)$, $s(n)$, $\rho(n)$, $\psi_k(n)$, and $\phi_k(n)$, respectively.

Accordingly, we use the periodicity information to confirm whether or not it is GCI at the higher scale filtered signal, that is, the filtered signal at scale $j = 6$. However, in practice, we do not separate the two cases, but test all GCIs with the above-mentioned method. The Fig. 4 shows a comparison of the GCI position in the filtered signal at scales $j = 5$ and 6.

5. Results and Discussion

We provide a performance comparison of the three PDAs on real speech signals; the proposed method, Kadambe's method, and the Wendt's method. The speech database used is composed of four sentences, also uttered by four male speakers, and sampled at 16kHz. In all the PDAs, we use the following measurements : GGDE(Gross GCI Determination Error, say, errors greater than 1ms); FGDE(Fine GCI Determination Error, say, errors below than 1ms); and VDE(Voicing Determination Error, over 30ms). Table 1 show the overall results.

Table.1 GCI detection results

	GGDE(%)	FGDE(%)	VDE(%)
the Proposed	0.91	2.20	0.65
the Kadambe's	9.16	3.31	3.69
the Wendt's	0.94	55.80	0.65

Table.1 and Fig.5 demonstrates that the proposed method shows the best performance in comparison of other two. It is also noted that in the case of the Kadambe's method, Table.1 shows the biggest GGDE due to the fixed threshold; Wendt's method also shows the biggest FGDE due to unwanted local maxima.

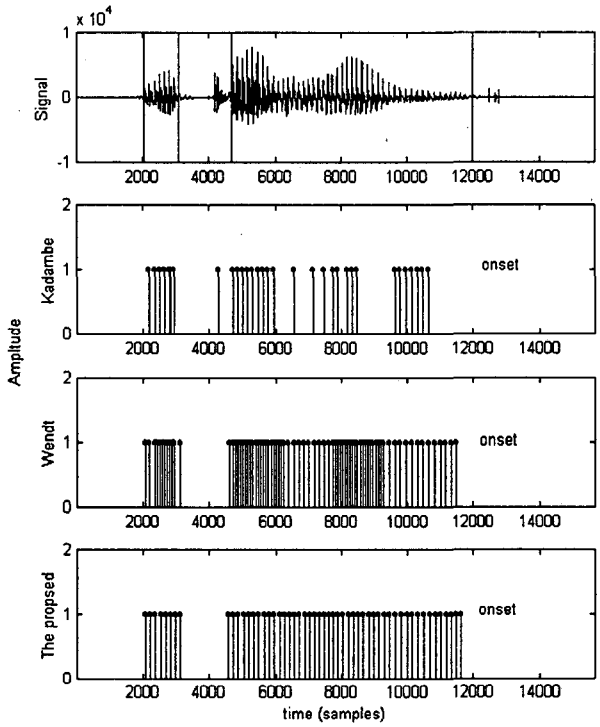

Fig.5 Speech signal and its GCI detection results

6. Conclusions

We presented an effective PDA, which is based on DyWT. By comparing with other prominent methods based on DyWT, we have shown that the proposed method is better in terms of performance especially in those speech segments where a transition between voiced and unvoiced take place, and where the beginning or ending of the voiced exists in the analyzed speech segment.

Acknowledgment : This work has been supported by University Basic Research grants from the Ministry of Information & Communication. The authors gratefully acknowledge the Ministry's support.

Reference

[1] Shubha Kadambe and G. Faye Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," IEEE Trans. on Information Theory, Vol. 38, No. 2, pp. 917-924, March 1992.

[2] Christher Wendt and Athina P. Petropulu, "Pitch Determination and Speech Segmentation Using the Discrete Wavelet Transform," Proc. IEEE International Symposium on Circuit and System, 1996, vol. 2, pp. 45-48

[3] Stephane Mallat and Sifen Zhong, Characterization of Signals from Multiscale Edges, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 14, No. 7, pp. 710-732, July 1992.

[4] Shubha Kadambe and G.F. Boudreaux-Bartels, A Comparison of a Wavelet Functions for Pitch Detection of Speech Signals, IEEE, pp. 449-452, 1991.

Received : October 26, 2000.
Accepted : November 24, 2000.

▲ Namhoon Kim
5ka-1 Anam-Dong sungbuk-ku Seoul 136-701, Korea
Dept. of Electronics Engineering, Korea University
Tel: +82-2-926-2909
Fax: +82-2-3291-2450
E-mail: nhkim@ispl.korea.ac.kr

▲ Yongsung Kang
5ka-1 Anam-Dong sungbuk-ku Seoul 136-701, Korea
Dept. of Electronics Engineering, Korea University
Tel: +82-2-926-2909
Fax: +82-2-3291-2450
E-mail: yskang@ispl.korea.ac.kr

▲ Hanseok Ko
5ka-1 Anam-Dong sungbuk-ku Seoul 136-701, Korea
Dept. of Electronics Engineering, Korea University
Tel: +82-2-3290-3239
Fax: +82-2-3291-2450
E-mail: hsko@korea.ac.kr