EXTRACTION OF PITCH INFORMATION IN NOISY SPEECH USING WAVELET TRANSFORM WITH ALIASING COMPENSATION

Shi-Huang Chen and Jhing-Fa Wang

Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan 701, R.O.C.
Email: shchen@cad.ee.ncku.edu.tw & wangjf@server2.iie.ncku.edu.tw

ABSTRACT

Although many wavelet-based pitch detection methods have been proposed in the literatures, there still remains a need to investigate new wavelet-based methods for more accurate and more robust pitch determination. In this paper, an improved wavelet-based method is developed for extraction of pitch information in noisy speech. At each decomposition in the wavelet transform, an aliasing compensation algorithm is applied to approximate and detail signals, in which the distortion of aliasing due to downsampling and upsampling operations of the wavelet transform is eliminated. In addition, this paper utilizes the concept of spatial correlation function used in signal denoising to improve the performance of pitch detection in noisy environment. It is shown in various experimental results that this new type of method has a considerable performance improvement compared with other conventional methods and wavelet-based methods.

1. INTRODUCTION

The extraction of pitch information is one of the most essential tasks in many speech processing applications. Generally, the pitch information refers to both the pitch period and the instants of glottal closure (GCI) in voiced speech. A well-designed pitch detection algorithm can be used to improve performance in a variety of systems, including a low bit-rate speech coding system [1], a PSOLA based text-to-speech (TTS) system [2], and a speech recognition system [3]. It is therefore no wonder that there are multiform pitch detection algorithms have been proposed in the literatures. However, due to the non-stationarity and quasi-periodicity of the speech signal as well as the interaction between the glottal excitation and the vocal tract, the development of more accurate and more robust pitch determination algorithms still remains an open problem [4].

In recent years, the multiresolution analysis with wavelet transforms received considerable attention due to its tremendous potential for dealing with non-stationary signal, such as speech. An effective approach to pitch determination using wavelet transform was proposed by Kadambe *et al.* [5]. Most of the subsequent wavelet-based pitch detection algorithms are originally inspired by the work presented in [5]. Essentially, the basic procedure of wavelet-based pitch detection algorithms consists of

(a) use of a wavelet transform to decompose the input speech signal into certain subband signals called approximate or detail signals;

(b) an exhaustive search of maximum values (GCI's) from the approximate or detail signals obtained in (a);

(c) correction of the locations of GCI's detected in (b).

Even though it was shown that the wavelet-based methods are superior to the traditional pitch detection techniques [5]-[6], they still can not meet the requirements of robustness and accurateness. There are two important issues which need to be improved in the classic wavelet-based pitch detection algorithms. First, the cost of a direct search of GCI's from the approximate or detail signals is too high, and its performance is usually degraded by background ambient noise. Second, there are some unwanted aliasing components in the approximate and detail signals, and they may affect the results of pitch detection.

The pitch detection algorithm presented in this paper overcomes the above two problems. Although the basic framework of the proposed algorithm is similar to that of the classic wavelet-based pitch detection algorithms, there are some significant differences in comparison with them. Unlike the classic wavelet-based method, the proposed algorithm utilizes a spatial correlation function [7] to sharpen and enhance GCI's while suppressing noise and other small sharp features. With this improved technique, the robustness of the proposed pitch detection method is increased substantially from previous approaches. This paper also applies an aliasing compensation algorithm [8] for eliminating the distortion of aliasing due to downsampling and upsampling operations of the wavelet transform from the approximate and detail signals. Without the interference from the aliasing, the accuracy of the proposed method can be further improved. To illustrate this, the proposed method is applied to both the synthetic and natural speech signals under additive non-stationary noises. It yields a considerable performance improvement compared with other conventional methods and wavelet-based methods.

The remainder of this paper is organized as follows. In Section 2, the wavelet transform with aliasing compensation will be described briefly. Then, the detailed description of the proposed wavelet-based pitch detection algorithm will be given in Section 3. In Section 4, the various experimental results are illustrated. Finally, Section 5 concludes the paper.

2. WAVELET TRANSFORM WITH ALIASING COMPENSATION

The wavelet transform discussed in this paper is implemented via filter banks structure. A fast discrete algorithm [9] is shown in Fig.

1 where $h(n)$ and $\widetilde{h}(n)$ are low-pass filters, whereas $g(n)$ and $\widetilde{g}(n)$ are high-pass filters. Also, the symbols ↓2 and ↑2 shown in Fig. 1 denote the downsampling by 2 and the upsampling by 2, respectively.

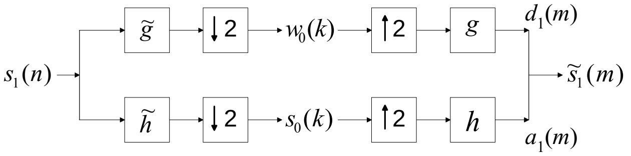


Fig. 1. Two-channel orthogonal wavelet filter banks.

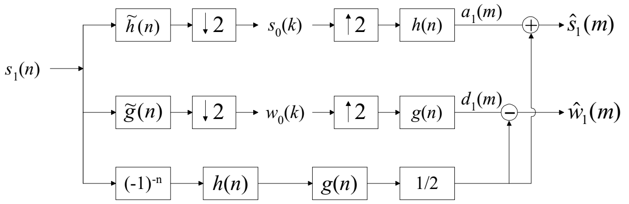


Fig. 2. Two-channel orthogonal wavelet filter banks with aliasing compensation.

Let $s_1(n)$ be the input signal, then the output of the analysis filter bank are

$$\begin{cases} s_0(k) = \sum_n \widetilde{h}(2k-n)s_1(n) \\ w_0(k) = \sum_n \widetilde{g}(2k-n)s_1(n) \end{cases} \tag{1}$$

where $s_0(k)$ and $w_0(k)$ are called the approximation coefficients and the detail coefficients, respectively, of the first level wavelet decomposition of $s_1(n)$. The output of the synthesis filter bank shown in Fig. 1 is

$$\widetilde{s}_1(m) = a_1(m) + d_1(m) \tag{2}$$

where $a_1(m)$ and $d_1(m)$ are called the approximate signal and the detail signal, respectively, of the first level wavelet transform of $s_1(n)$. The definitions of $a_1(m)$ and $d_1(m)$ are

$$a_1(m) = \sum_{k\in Z} s_0(k)h(m-2k)\,, \tag{3}$$

$$d_1(m) = \sum_{k\in Z} w_0(k)g(m-2k)\,. \tag{4}$$

Since the aliases will arise from the downsampling and upsampling operations, the filter banks discussed above are designed to cancel these undesired aliasing components, and to ensure that $s_1(n) = \widetilde{s}_1(m)$. Under this requirement, these four filters have to be related as [10]

$$\widetilde{g}(n) = (-1)^n h(1-n)\,, \quad g(n) = (-1)^n \widetilde{h}(1-n)\,. \tag{5}$$

However, these aliasing terms still can not completely remove from each subband of the synthesis filter bank (i.e., $a_1(m)$ and $d_1(m)$) due to the imperfect magnitude response of filters [8]. Directly utilization of these approximate and detail signals, regardless of aliasing effects, for pitch detecting may cause some unexpected errors. Therefore, this paper utilizes an aliasing compensation algorithm proposed in [8] to further eliminate these unwanted aliasing components from the approximate and detail signals. An example of one-level two-channel orthogonal wavelet transform embedded the aliasing compensation algorithm is given in Fig. 2 where $\hat{s}_1(m)$ and $\hat{w}_1(m)$ are the aliasing compensated approximate and detail signals, respectively. It is worth to note that the aliasing compensation algorithm still remain the perfect reconstruction property of wavelet transform (i.e., $\hat{s}_1(m) + \hat{w}_1(m)$ $= a_1(m) + d_1(m) = \widetilde{s}_1(m) = s_1(n)$) [8]. In addition, this aliasing compensation algorithm can be easily extended to higher level wavelet decomposition with a pyramid structure [8]. Fig. 3 gives an illustrative example of two-level orthogonal wavelet transform with the aliasing compensation.

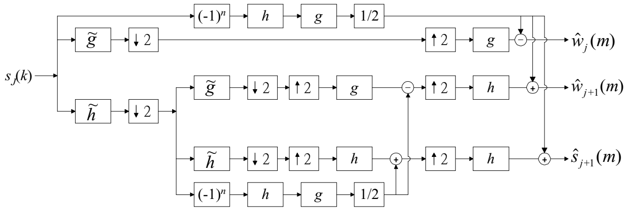


Fig. 3. An illustrative example of two-level orthogonal wavelet transform with aliasing compensation.

3. THE PROPOSED PITCH DETECTION ALGORITHM

Since the GCI is marked by a sharp discontinuity in the speech signal, it can in some sense be related to the edge detection problem. Based on the multiresolution analysis with wavelet transform, the GCI will appear at the same point in the approximation or detail signals over several successive decomposition levels [5]-[6]. Although this property is very useful for pitch detection, the cost of a direct search of GCI's from the approximate or detail signals is too high. Furthermore, its performance is usually degraded by background ambient noise. To overcome this problem, this paper applies a spatially selective noise filtration (SSNF) technique proposed in [7] to pitch detection. Cooperating with the aliasing compensation wavelet transform described in the previous section, a modified spatial correlation $Corr_l(k,m)$ is defined to sharpen and enhance GCI's while suppressing noise and small sharp features

$$Corr_l(k,m) = \prod_{i=0}^{l-1} \hat{s}_{k+i}(m),\ m = 1,2,\cdots,N \tag{6}$$

where $\hat{s}_k(m)$ denotes the approximate signals with aliasing compensation, k is the index of decomposition level, m is the translation index, N is the length of input signal, $l < M-k+1$, and M is the total number of decomposition levels. Usually, $l = 2$.

The first step of using spatial correlation function to pitch detection is to decide the decomposition index k. Due to the first formant frequencies of human are generally below 1 kHz [5], the decomposition index k is determined by following equation

$$k = \left\lfloor \log_2\left(Fs\,/\,1\text{ kHz}\right)\right\rfloor \tag{7}$$

where *Fs* is the bandwidth of the input speech signal. Thus, the aliasing compensated approximate signal $\hat{s}_k(m)$ will contain most of pitch information. Then, the power of the $\left\{Corr_2(k,m)\right\}$ is rescaled to that of the $\left\{\hat{s}_k(m)\right\}$ and get $\left\{NewCorr_2(k,m)\right\}$. Above procedure can be expressed as follows.

$$NewCorr_2(k,m) = Corr_2(k,m)\sqrt{Ps(k)\ /\ PCorr(k)} \qquad (8)$$

where

$$Ps(k) = \sum_m \hat{s}_k(m)^2 \text{ , and} \qquad (9)$$

$$PCorr(k) = \sum_m Corr_2(k,m)^2 . \qquad (10)$$

The GCIs' are identified in $\hat{s}_k(m)$ and $NewCorr_2(k,m)$ by comparing the absolute values of $\hat{s}_k(m)$ and $NewCorr_2(k,m)$. If $\left|NewCorr_2(m)\right| \geq \left|\hat{s}_k(m)\right|$, the position *m* will be accepted as a GCI and saved in the vector *G*(*m*).

Finally, the accurate locations of GCI's and the pitch period of input speech signal are obtained by a pitch correction algorithm as follows.

(1) Calculate the average distance *Da* between two adjacent elements in *G*(*m*).

(2) Eliminate the GCI whose distance between its adjacent GCI's is shorter than 0.5*Da* or longer than 2*Da* from *G*(*m*).

(3) Repeat (1) and (2) until no unsuitable GCI is available.

(4) Locate the final existed GCIs' positions and calculate the average distance between two adjacent GCIs' as the estimated pitch period.

4. EXPERIMENTAL RESULTS

In this section, it will first discuss the performance of the proposed method on synthetic and natural speech data. The robustness of proposed method is then examined for additive white Gaussian noises. In all the illustrations to follow, the speech signals were sampled at 8 kHz with 8-bit resolution, and Daubechies' length-8 wavelet [10] is used.

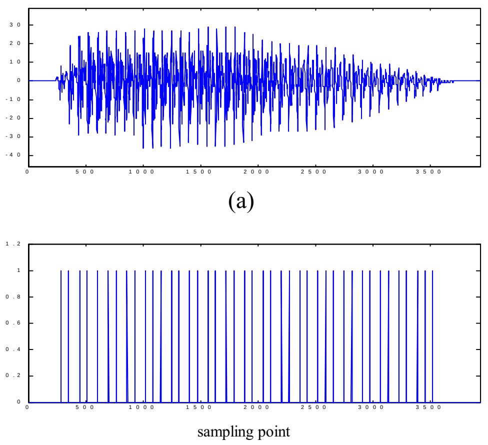
4.1 Illustration of the Method for Synthetic Speech Data

The synthetic speech data considered in this paper consist of five kinds of voiced phonemes, namely, /a/, /e/, /i/, /o/ and /u/. Fig. 4(a) shows a synthetic speech signal /a/ whose pitch period is constant and is equal to 10ms. The locations of GCIs' are illustrated in Fig. 4(b). The estimated pitch period is 9.92ms. Table 1 gives the experimental results of this new method on other synthetic speech data.

The error rate *Err* used in Table 1 is defined as

$$Err = \left(\left|\hat{T} - T\right| / T\right) \times 100\% \qquad (11)$$

where $\hat{T}$ is the estimated pitch period and the *T* is the true pitch period.



(a)

sampling point

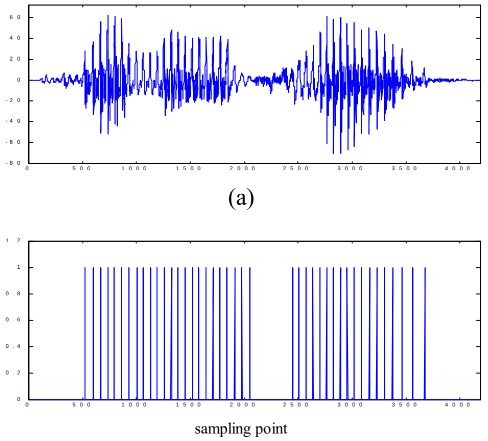
(b)

Fig. 4. (a) Clean synthetic speech signal /a/, (b) the detected locations of GCIs'.

Table 1. The experimental results of the synthetic speech data

Pitch Period / Speech Data	5 ms	10 ms	15 ms	20 ms	25 ms
	Estimated pitch period (ms)				
/a/	4.94	9.92	14.95	19.85	24.90
/e/	4.89	9.86	14.91	19.90	24.92
/i/	4.93	9.91	14.89	19.88	24.88
/o/	5.00	9.95	14.82	19.81	24.84
/u/	4.91	9.92	14.92	19.84	24.92
Average *Err*	1.27 %	0.88 %	0.68%	0.72%	0.43%

4.2 Illustration of the Method for Natural Speech Data



(a)

sampling point

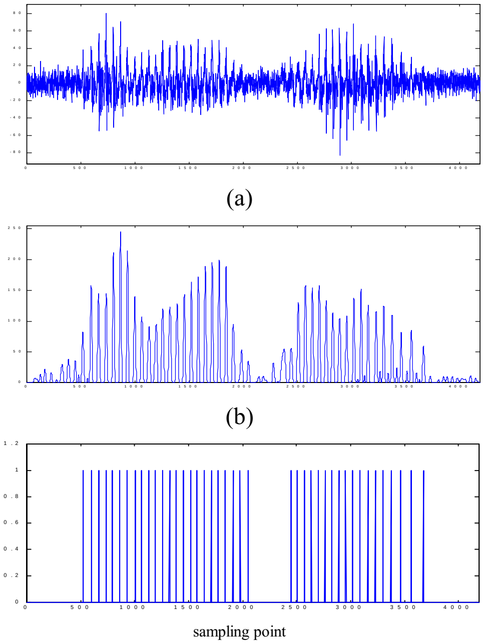
(b)

Fig. 5. Illustration of the method for a continuous speech "Any time…": (a) Speech waveform; (b) the estimated locations of GCI's.

Fig. 5 illustrates the experimental result of proposed method on the initial part of the utterance "Any time …." spoken by a male voice. Figs. 5(a) and (b) show the speech waveform and the estimated locations of GCIs', respectively.

4.3 Robustness of the Method

Finally, the robustness of the proposed method is evaluated under additive noisy conditions. A white Gaussian noise was added to the clean speech, and the performance is evaluated at SNR's 30, 25, 20, 15, 10, 5, and 0dB, respectively. To illustrate this, Fig. 6 shows the experimental results of the natural speech signal of Fig. 5(a) at a SNR of 5dB.



(c)

Fig. 6. (a) Natural speech of Fig. 5(a) with 5dB Gaussian white noise, (b) the waveform of $\left\{ NewCorr_2(k,m) \right\}$, (c) the estimated locations of GCIs'.

The robustness of the proposed method is also compared against the other present pitch detection methods including in the spectral, time and wavelet domains, and the results are given in Table 2. The definition of *Err* used in Table 2 is given in (11) where the true pitch period of natural speech is determined via time-domain method [3] with manual correction under clear condition. From these experimental results, one can find that the pitch information can be accurately and effortlessly extracted by using the proposed method.

Table 2. Performance of the proposed method at different SNR's

SNR (dB)	30	25	20	15	10	5	0
Methods	Average *Err* (%)						
Wavelet-based in [5]	1.2	1.4	2.7	4.9	7.5	12.4	23.6
Wavelet-based in [6]	1.1	1.3	2.4	4.2	6.9	8.6	17.8
Cepstrum-based in [4]	1.4	1.6	3.2	5.7	8.6	16.8	32.2
Time-domain in [3]	0.7	1.2	2.6	5.3	10.4	28.7	49.6
Proposed method	1.0	1.1	2.0	3.1	4.4	6.2	10.1

5. CONCLUSIONS

An improved wavelet-based pitch detection algorithm was presented. In this paper, the aliasing compensation wavelet transform and the spatially selective noise filtration technique were proposed to improve the accuracy and the robustness of the conventional wavelet-based pitch detection algorithm. The performance of the proposed algorithm was evaluated on synthetic and natural speech signals. Compared to other time, spectral and wavelet domain pitch detection methods, it was shown that the proposed method has the better performance under noisy conditions.

6. ACKNOWLEDGMENT

This work was supported by the National Science Council (NSC), Taiwan, R.O.C., under Grant NSC89-2218-E-006-045.

REFERENCES

[1] M. Oshikiri, M. Akanine, "A 2.4 kbps variable bit rate ADP-CELP speech coder," ICASSP 98, vol. 1, pp. 517-520, 1998.

[2] C. Hamon, E. Moulines, and F. Charpentier, "A diphone synthesis system based on time domain prosodic modifications of speech," ICASSP 98, vol. 1, pp. 238-241, 1998.

[3] J. F. Wang, C. H. Wu, S. H. Chang, and J. Y. Lee, "A hierarchical neural network model based on a C/V segmentation algorithm for isolated mandarin speech recognition," *IEEE Trans. Signal Processing*, vol. 39, No. 9, pp. 2141-2146, September 1991.

[4] Sassan Ahmadi, Andreas S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech and Audio Processing*, vol. 7, No. 3, pp. 333-338, May 1999.

[5] S. Kadambe and G. Faye Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Information Theory*, vol. 38, no. 2, pp. 917-924, March 1992.

[6] Shi-Huang Chen and Jhing-Fa Wang, "A Pyramid-Structured Wavelet Algorithm for Detecting Pitch Period of Speech Signal," 1998 International Computer Symposium (ICS), pp. 50-56, Dec. 1998.

[7] Y. Xu et al., "Wavelet transform domain filters: a spatially selective noise filtration technique," *IEEE Trans. Image Processing*, vol. 3, pp. 747-758, Nov. 1994.

[8] S. H. Chen, J. F. Wang," An aliasing compensation algorithm for (bi-)orthogonal wavelet transform," submitted to *IEEE Signal Processing Letter*, Nov. 2000.

[9] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.* Vol. 11(7), pp. 674-693, 1989.

[10] C. Sidney Burrus, Ramesh A. Gopinath and Haitao Guo, *Introduction to Wavelets and Wavelet Transforms*. Upper Saddle River, NJ: Prentice-Hall, 1998.