1998 International Computer Symposium
Workshop on Image Processing and Character Recognition
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

A PYRAMID-STRUCTURED WAVELET ALGORITHM FOR DETECTING PITCH PERIOD OF SPEECH SIGNAL

Shi-Huang Chen and Jhing-Fa Wang

Department of Electrical Engineering
National Cheng Kung University, Tainan, Taiwan 701, R.O.C.
TEL: +886-6-2747076, FAX: +886-6-2747076
Email: shchen@cad.ee.ncku.edu.tw & wangjf@server2.iie.ncku.edu.tw

ABSTRACT

A new time-spectral domain algorithm based on the pyramid-structured wavelet transform for detecting pitch period has been developed in this paper. The development of the pyramid-structured wavelet transform for pitch detection is motivated by the observation that the human speech signals can be modeled as quasi-periodic signals whose fundamental frequencies are located in 30 ~500 Hz. The pyramid - structured wavelet transform is able to zoom into dominant frequency (30~500 Hz) channel, and reserves the fundamental characteristics of speech signals. Therefore, the proposed algorithm can efficiently determine the pitch period in the dominant frequency channel. From the various experiments, they show that the proposed scheme is not only suitable for wide-range pitch period of speech signals but also is robust to noise. The performance of the proposed scheme is compared with that of the time domain and the spectral domain pitch detectors and other presented wavelet based pitch detectors.

1. INTRODUCTION

In the last years, wavelet transforms have been intensively applied to signal processing, numerical analysis, and mathematical modeling. Unlike the Fourier, cosine and sine transform which are not local and only appropriate for periodic signals, the wavelet analysis is well-suited to transient signals. The wavelet representation can offer the location in both time and frequency simultaneously. This localizing property of wavelets that allows the wavelet expansion of a transient event to be modeled with a small number of coefficients [4, 5]. This turns out to be very useful for the pitch detecting.

The pitch period is regarded as one of the important features for the analysis and synthesis of speech signals. Pitch detectors could be used in vocoders, speaker identification, verification systems, linguistic and phonetic knowledge acquisition and voice disease diagnostics [1, 2]. However, it is very difficult to estimate the pitch period in consideration of following reasons: (1) the human vocal tract is not a perfect train of periodic pulses and its characteristics vary from person to person, (2) the pitch period can vary from 1.25 ms to 40 ms, (3) the pitch period of the same speaker can vary depending upon the emotional state of the speaker and (4) the background ambient noise can also seriously affect the performance of the pitch detector [2, 3, 8].

In the traditional case, the pitch detection algorithms can be classified into two types, spectral-domain (nonevent) based and time-domain (event) based pitch detectors. The spectral based pitch detectors, such as the autocorrelation and the cepstrum methods, estimate the average pitch period over a segment of a speech signal that they obtain by using a window whose length is fixed. The time based pitch detectors estimate the pitch period by locating the glottal closure instant (called an event or GCI) and then measuring the time period between two such events. Generally, the time based pitch detectors give more accurate estimation of pitch period than the spectral based types. On the other hand, the spectral based pitch detectors are computationally simpler than the time based types.

Pitch period can, in some sense, be related to the edge detection problem in the image processing. In [7], Mallat has shown that the multiscale edge detection is equivalent to finding the local maximum of its wavelet representation. Several wavelet based pitch determination algorithms have been presented [2, 6]. These methodologies are based on the Mallat's work on image [7] essentially. In [2] and [6], the presented works use the assumption that the points of GCI occur in the original speech waveform will be the same as those in the several consecutive scales of the wavelet transform. However, the proposed method in [2] seems only suitable for synthesized speech signals, but not for real speech recordings [6]. In [6], the authors proposed a modified method embedded the conception of a single filtering function to improve the disadvantages in [2]. Nevertheless, it leaves an accuracy problem. This problem implies that the matching of GCI point between the original speech signal and the single filtering function representation of this signal is not very correct.

To overcome these problems cited above, this paper describes a new algorithm based on the pyramid-structured wavelet transform for detecting pitch period of speech signals under ideal and noisy conditions. The proposed new method is not only computationally attractive but also has excellent performance and is robust to noise. From the various experimental results, the performance of the proposed scheme is better than that of the classical spectral-domain and time-domain pitch detectors, and other presented pitch detectors with wavelet based methods.

2. THE PYRAMID-STRUCTURED WAVELET TRANSFORM

One of the main approaches to wavelets is through the two-channel filter banks. The wavelet transform can be considered as filtering a signal by a pair of lowpass filter $h(n)$ and highpass filter $g(n)$. Then the lowpass and highpass filter outputs are downsampled by two, which removes the odd-numbered components after filtering, respectively [4, 5]. Consequently, the signal length of the lowpass or highpass filter output is only half of original input one. The block diagram of the two-channel filter bank is shown in Figure 1.

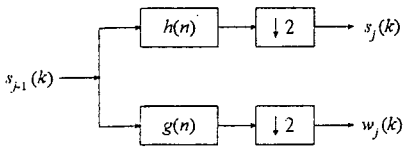

Figure 1. Two-channel filter bank.

The scaling coefficients, $s_j(k)$, shown in Figure 1 is given by

$$s_j(k) = \sum_m h(m-2k)s_{j-1}(m), \tag{1}$$

and the corresponding relationship for the wavelet coefficients, $w_j(k)$, shown in Figure 1 is

$$w_j(k) = \sum_m g(m-2k)s_{j-1}(m). \tag{2}$$

The filter coefficients $h(n)$ and $g(n)$ in (1) and (2), respectively, play a very crucial role in a given discrete wavelet transform and have to satisfy orthonormal and a certain degree of regularity. Several different sets of the orthogonal filter coefficients $h(n)$ and $g(n)$ can be found in [4, 5].

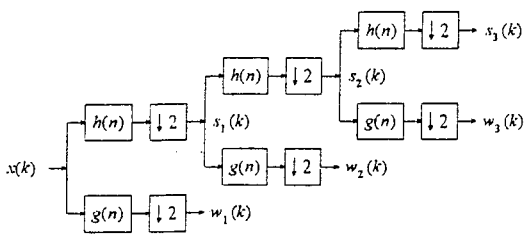

Figure 2. Three-stage pyramid-structured wavelet transform

The splitting, filtering, and decimation in the Figure 1 can be repeated on the scaling coefficients to give the idea of pyramid-structured wavelet transform. It provides a recursive algorithm for wavelet decomposition through $h(n)$ and $g(n)$ and a multiresolution filter-bank decomposition of a signal with narrower bandwidths in the lower frequency channels. Figure 2 is an example of three-stage pyramid -structured wavelet transform where $x(k)$ is the input signal.

The reasons that the pyramid-structured wavelet transform could be applied to detect the pitch period are summarized as follows. First, human speech generally spans only 10 octaves. Second, the pitch or fundamental frequency of voiced speech is located in the low frequency region (about 30 ~ 500Hz). Third, unvoiced signals are random in nature and contain high frequency information [2]. These phenomena imply that the pyramid-structured wavelet transform with limited decomposition stage can be used for determining the pitch period in the dominant frequency channel. This will lead to a lower computational complexity of detecting pitch period and a robustness of the high frequency noise.

3. PYRAMID-STRUCTURED WAVELET ALGORITHM FOR DETECTING PITCH PERIOD

Theoretically, the pyramid-structured wavelet transform will recursively decompose subsignals in the low frequency channels. However, since the most significant information of pitch will appear in the dominant frequency (30 ~ 500Hz) channel, excessive decomposition in the lower frequency region may not help much for the purpose of pitch detection. In order to construct a required dominant frequency channel in the pyramid-structured wavelet transform, the order of decomposition stage and the filter types must be determined concurrently.

The suitability of the orthogonal filter bank, $h(n)$ and $g(n)$, for pitch detection depends on its ability to separate speech information into several independent frequency channels. Therefore, the orthogonal filter bank, $h(n)$ and $g(n)$, in the proposed pyramid-structured wavelet transform be selected for corresponding to the halfband lowpass and halfband highpass filters, respectively. And the coefficients of the halfband lowpass filter $h(n)$ have the following relation [5]:

$$\sum_n h(n)h(n-2k) = \begin{cases} 1 \text{ if } k = 0 \\ 0 \text{ if } k \neq 0. \end{cases} \tag{3}$$

The orthogonal halfband lowpass filter coefficients $h(n)$ of Haar and 4-tap Daubechies wavelet transforms are listed in Table 1.

TABLE 1
Halfband Filter Coefficients of Wavelet Transform

	Haar	4-tap Daubechies
$h(0)$	0.7071067811865	0.48296291314453
$h(1)$	0.7071067811865	0.83651630373708
$h(2)$	—	0.22414386804201
$h(3)$	—	-0.12940952255126

The pyramid-structured wavelet transform with the orthogonal halfband filter bank will cut the input signal frequency band in half. The lower half of the band goes through the lowpass filter, and the upper half of the band goes through the highpass filter. Therefore, the order of the

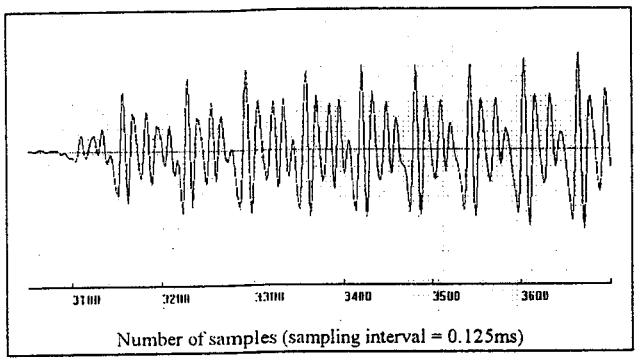

Figure 3(a). A segment of the phoneme /o/ spoken by a Chinese male speaker.

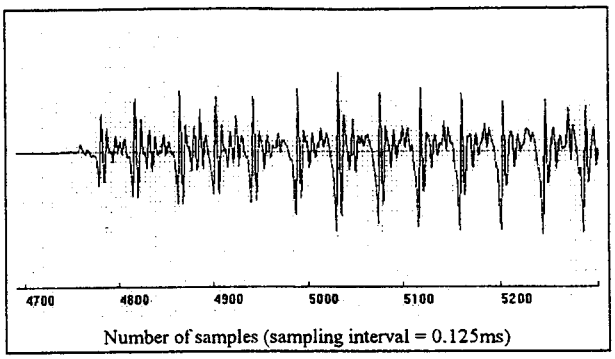

Figure 4(a). A segment of the phoneme /i/ spoken by a Chinese female speaker.

decomposition stage can be decided as follows: Given an input signal with bandwidth A Hz, one can determine the decomposition stage 'N' corresponding to the required dominant frequency bandwidth B Hz using the following equation:

$$N = \left\lfloor \log_2\left(\frac{A}{B}\right) \right\rfloor. \quad (4)$$

The input speech signals in the proposed scheme are sampled at 8000 Hz, so that the bandwidth of the input speech signals is 4000 Hz, and the required dominant frequency bandwidth is 500 Hz. By the equation (4), the order of decomposition stage can be computed to equate 3. Hence, the proposed method can determine the pitch period at only one-eighth ($1/2^3$) signal data instead of full one; this has the advantage of significantly reducing the computational complexity of detecting pitch period.

The algorithm of pitch detection with pyramid-structured wavelet transform is given below.

1) Decompose the input speech signal $x(k)$ with three - stage pyramid - structured wavelet transform and preserve its 3rd scale coefficients $s_3(k)$.
2) Set window length = L sampling points, and load L samplings from the $s_3(k)$ once. And then compute the maximum value called MG within this window. In the proposed algorithm, the window length L = 40ms × 8000 Hz × $1/2^3$ = 40 where 40ms is the maximum pitch period, 8000 Hz is the sample rate of input signal, and the $1/2^3$ is the remnant data length after three-stage pyramid-structured wavelet transform.
3) Locate the positions of the $s_3(k)$ whose amplitude exceeds 80 percentage of MG. Then, mark the corresponding $x(k)$ of these located $s_3(k)$ whose amplitude exceed or equate these of located $s_3(k)$.
4) Reject one of the contiguous marked sample points of $x(k)$ whose distance is less than 25 sampling points and keep the largest one.
5) Shift the window segment by L / 2 = 20 sampling points and go to step 2) until end of $s_3(k)$.
6) Measure the time period between two marked sampling points of $x(k)$ and output the pitch period estimate.

4. EXPERIMENTAL RESULTS

Five voiced phonemes, /a/, /e/, /i/, /o/ and /u/, spoken by male and female Chinese are the test signals in this paper. Each test signal is sampled at 8000 Hz and saved with 8-bit resolution. All of these speech signals are recorded by a dynamic coil microphone.

A. Performance of Pitch Detection with Pyramid - Structured Wavelet Transform

In this subsection, the performance of the pitch detection with pyramid - structured wavelet transform will be demonstrated. As far as different wavelet functions are concerned previously, the Haar wavelet function gives the quite well results while the other wavelet functions achieve similar performance. By reason of the Haar wavelet function has the easiest computational complexity of wavelet transform, the proposed scheme will apply the Haar wavelet function.

Figure 3(a) and 4(a) show a segment of voiced speech signal, /o/ and /i/, which spoken by a male and female Chinese, respectively. It is clear that the pitch period of the female speaker is shorter than that of the male speaker. Firstly, the results of the pitch detection by using proposed method with Haar wavelet function were shown in Figure 3(b) and 4(b). It can be found that the proposed scheme with 4-tap Daubechies wavelet function provides the similar performance in Figure 3(c) and 4(c). From these experimental results, one can indicate the proposed algorithm is able to detect the pitch location very accurate.

B. Comparison with Classical Spectral-domain Pitch Detection Methods

With data reported in this and the following subsections, it will perform the experiments of the same speech signals, /o/ and /i/, described in the previous subsection. It firstly compares the performance of the proposed pitch detector

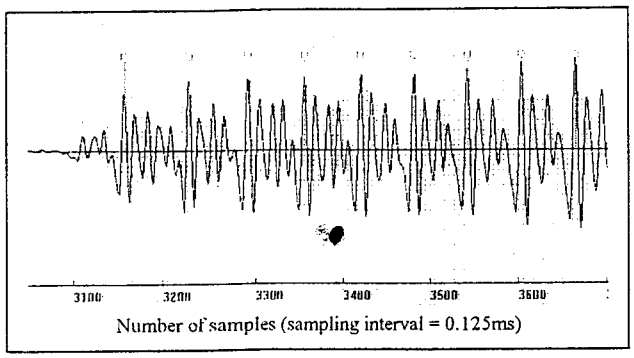

Figure 3(b). Pitch detection by using proposed scheme with Haar wavelet function. (The vertical lines indicate the pitch locations.)

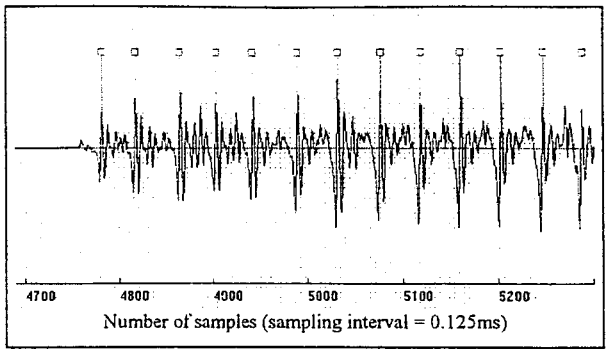

Figure 4(b). Pitch detection by using proposed scheme with Haar wavelet function. (The vertical lines indicate the pitch locations.)

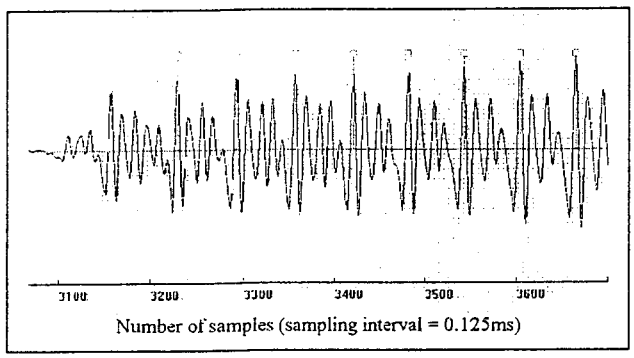

Figure 3(c). Pitch detection by using proposed scheme with 4-tap Daubechies wavelet function. (The vertical lines indicate the pitch locations.)

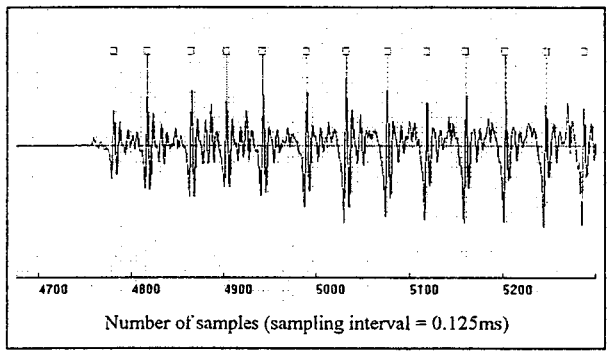

Figure 4(c). Pitch detection by using proposed scheme with 4-tap Daubechies wavelet function. (The vertical lines indicate the pitch locations.)

with classic spectral based techniques such as the autocorrelation and the cepstrum method.

The algorithm embedded the autocorrelation function of a speech signal to estimate the pitch period can be found in [1]. The autocorrelation function $\phi(k)$ of a discrete-time signal $x(n)$ is defined as

$$\phi(k) = \sum_{n=-\infty}^{\infty} x(n)x(n+k) \cdot \qquad (5)$$

The autocorrelation function of a period signal also exhibits periodicity equal to that of the signal [1]. The cepstrum $c(\tau)$ of a discrete-time signal $x(n)$ is defined as

$$c(\tau) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)|e^{j\frac{2\pi kn}{N}}, \quad 0 \le n \le N-1 \qquad (6)$$

where $X(k)$ is the discrete Fourier transform (DFT) of $x(n)$ and N is the data length of $x(n)$. The cepstrum of a speech signal can be applied to estimate the pitch period, since the cepstrum of a periodic signal exhibits the same periodicity as the signal under consideration. And the algorithm of pitch detection using the cepstrum can also be found in [1].

These spectral based pitch detectors assume that the pitch period is stationary within each segment, and each segment contains at least two full pitch periods. It is well known that the real voiced speech is not perfectly periodic, therefore, the spectral based pitch detectors are unsuitable for both low pitched and high pitched speakers. Figure 5 and 7 are the experimental results of detecting the pitch of the phoneme /o/ and /i/ by using the autocorrelation function, respectively. And Figure 6 and 8 show the corresponding experimental results but using cepstrum method. Obviously, the performance of the proposed scheme is better than those of the spectral based pitch detectors.

C. Comparison with Time-domain Based Pitch Detection Methods

The performance of the proposed method in comparison with the time-domain based pitch detector will be demonstrated in this subsection. Generally, the time-domain based method will estimate the pitch period more accurately and is computationally complex. Only a few time-domain based pitch detectors have been developed recently [9, 10]. The method proposed in [9] uses the maximum likelihood epoch determination technique to detect the pitch period. Although this method has excellent performance, it is not suitable for high pitched speakers since the data length available for the linear predictor

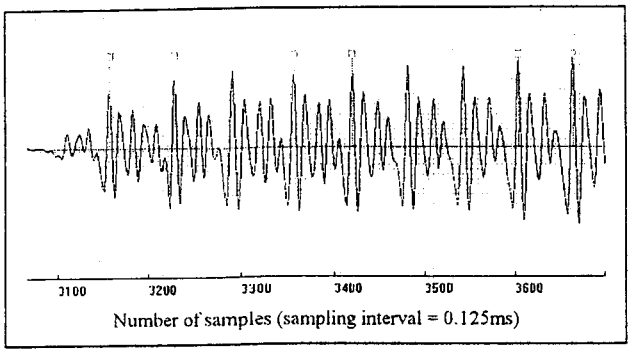

Figure 5. Pitch detection with autocorrelation method on the phoneme /o/ spoken by a Chinese male speaker. (The vertical lines indicate the pitch locations.)

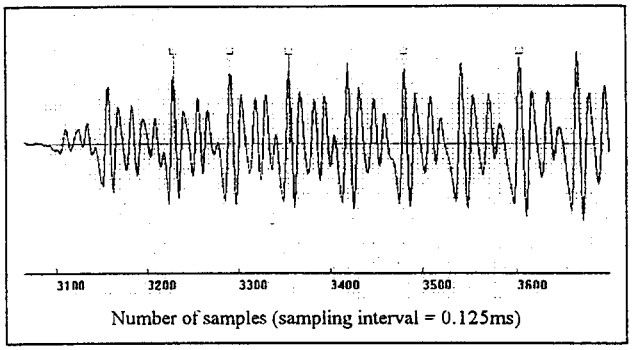

Figure 6. Pitch detection with cepstrum method on the phoneme /o/ spoken by a Chinese male speaker. (The vertical lines indicate the pitch locations.)

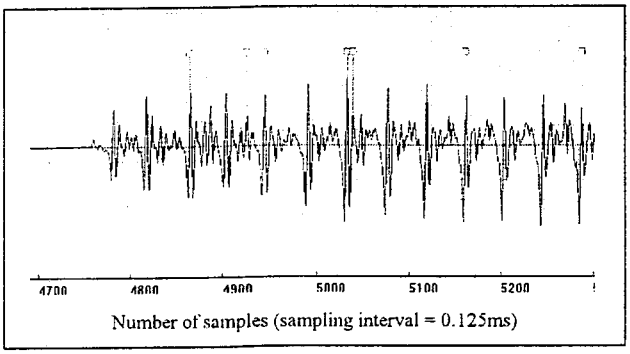

Figure 7. Pitch detection with autocorrelation method on the phoneme /i/ spoken by a Chinese female speaker. (The vertical lines indicate the pitch locations.)

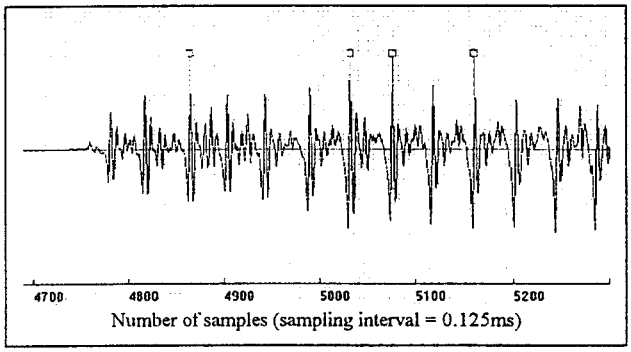

Figure 8. Pitch detection with cepstrum method on the phoneme /i/ spoken by a Chinese female speaker. (The vertical lines indicate the pitch locations.)

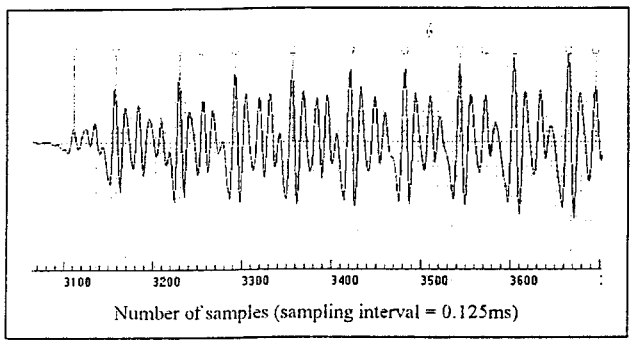

Figure 9. Pitch detection with the method described in [10] on the phoneme /o/ spoken by a Chinese male speaker. (The vertical lines indicate the pitch locations.)

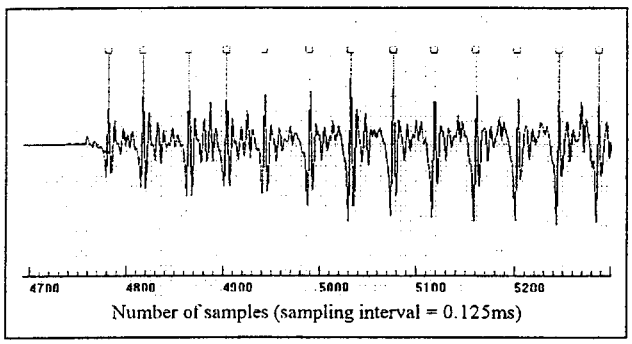

Figure 10. Pitch detection with the method described in [10] on the phoneme /i/ spoken by a Chinese female speaker. (The vertical lines indicate the pitch locations.)

could be very small [2].

Another method proposed in [10] applies a progressive clipping algorithm to detect the pitch period and some complicated and unnecessary computations used in the traditional time based approaches are omitted. Therefore, this method can be implemented to operate very quickly. However, it is not suitable for some low pitched conditions. Figure 9 and 10 are the experimental results of detecting the pitch of the two phonemes, /o/ and /i/, by using the method described in [10], respectively. It can be found that the performance of the proposed scheme is still better than that of the method proposed in [10].

D. Comparison with Other Wavelet-Based Methods

In this subsection, it will compare the performance of the proposed pitch detector with other wavelet-based pitch detecting techniques described in [2] and [6]. The method proposed in [2], while using a derivative function as a

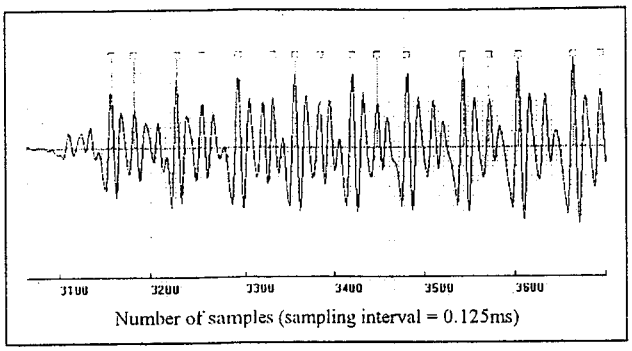

Figure 11. Pitch detection with the method described in [2] on the phoneme /o/ spoken by a Chinese male speaker. (The vertical lines indicate the pitch locations.)

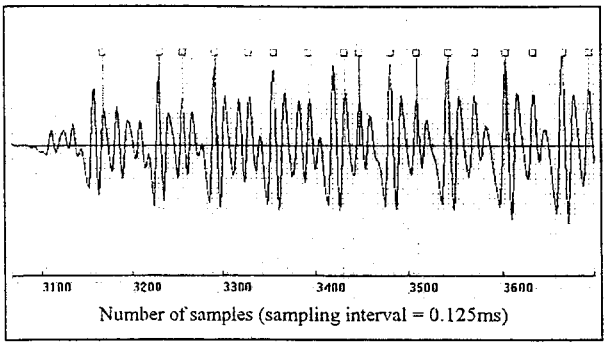

Figure 12. Pitch detection with the method described in [6] on the phoneme /o/ spoken by a Chinese male speaker. (The vertical lines indicate the pitch locations.)

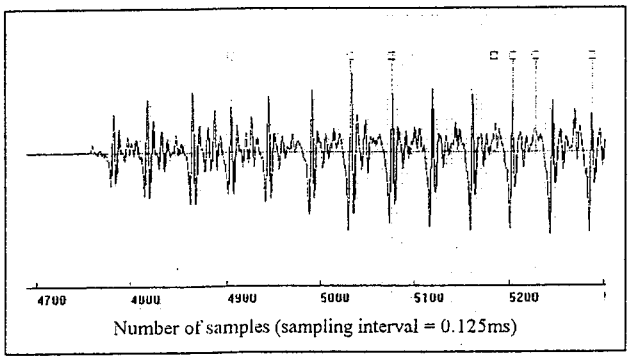

Figure 13. Pitch detection with the method described in [2] on the phoneme /i/ spoken by a Chinese female speaker. (The vertical lines indicate the pitch locations.)

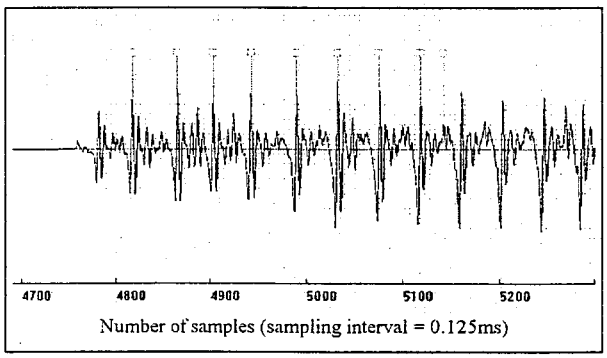

Figure 14. Pitch detection with the method described in [6] on the phoneme /i/ spoken by a Chinese female speaker. (The vertical lines indicate the pitch locations.)

mother wavelet, and applying the multiple scales in its analysis. Consecutive scale coefficients are searched for maximums occurring at or around the same positions. Figure 11 and 13 are the experimental results of detecting the pitch of the phoneme /o/ and /i/ by using the method described in [2], respectively.

Another wavelet-based pitch detecting method proposed in [6] utilizes a single filtering function. And this single filtering function, $\rho(n)$, is obtained as

$$\rho(n) = \psi_{ka}(n) * \varphi_{kb}(n) \qquad (7)$$

where * is linear convolution and the highpass function, $\psi_{ka}(n)$, and the lowpass function, $\varphi_{kb}(n)$, are given in [6] respectively. This single filtering function is chosen to have both derivative characteristics and a bandwidth defined by voiced speech. Figure 12 and 14 are its experimental results of detecting the pitch of the phoneme /o/ and /i/, respectively. Although the method described in [6] provides a better performance than that proposed in [2]. However, for the low pitched situation, the accuracy of the method described in [6] is not very well. Through these experiments, the method proposed in this paper has the best performance.

E. Sensitivity to Noisy Signal

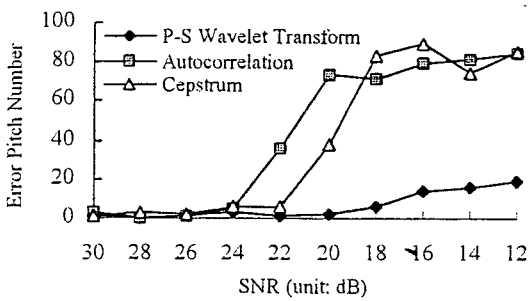

Figure 15. Error pitch number of the proposed scheme, the autocorrelation, and the cepstrum - based pitch detectors with noisy data.

The sensitivity of the proposed algorithm to noisy signal was simulated by adding white Gaussian noise to the test speech signals before pitch detection. The error pitch number is given to evaluate the sensitivity to noisy signal.

and it is defined as

Error pitch number = (Pitch number of noisy signal) – (Pitch number of noiseless signal). (8)

The simulated result is shown in Figure 15. The performance of the proposed scheme is almost unaffected when the SNR is higher than 20dB and is still acceptable for lower SNR which approaches 12dB. It is apparent that the proposed scheme is generally the most robust to noise as compared to the autocorrelation and the cepstrum based pitch detectors.

5. CONCLUSION

In this paper, the pyramid-structured wavelet transform is shown to provide an excellent tool for pitch detecting. From various experiments demonstrated, this paper has compared its performance with classical spectral based, time based and other wavelet based pitch detectors, and shown that it exhibits superior performance. It is worthwhile to point out several advantages of the proposed scheme in comparison with the other existing pitch detectors. First, the proposed method estimates the pitch period very accurately for both low pitched and high pitched speakers. Second, the computational complexity of the proposed scheme is quite simple. Third, the proposed scheme is robust to noise. The future work will focuse on algorithmic development and experimental justification with low bit rate hybrid voice codec and other applications.

6. REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[2] S. 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.

[3] W. Hess. *Pitch determination of speech signals: algorithms and* devices. Berlin: Springer Verlag, 1983.

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

[5] Gilbert Strang and Truong Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press, 1996.

[6] Christopher Wendt and Athina P. Petropulu, "Pitch Determination and Speech Segmentation Using the Discrete Wavelet Transform," ISCAS 96, Atlanta, GA, USA, pp. 45-48, 1996.

[7] S. G. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. of Patt. Analy. and Mach.* Intell., vol. 14, pp. 710-732, July 1992.

[8] A. M. Kondoz, Digital *Speech Coding for Low Bit Rate Communications Systems*, West Sussex, England, John Wiley & Sons Ltd, 1994.

[9] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 562-570, Dec. 1989.

[10] 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. on Signal Processing*, vol. 39, No. 9, pp. 2141-2146, September 1991.