ATLANTIS PRESS

Advances in Computer Science Research, (ACSR), volume 52

2016 International Conference on Computer Engineering and Information Systems (CEIS-16)

Fast Wavelet-based Pitch Period Detector for Speech Signals

Y.H. Goh, Y.H. Ko , Y.K. Lee
Department of Mechanical Engineering
Tunku Abdul Rahman University College
53300 Kuala Lumpur, Malaysia
Email: gohyh@acd.tarc.edu.my

Y.L. Goh
Department of Mathematical and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Science
Sungai Long Campus, Malaysia

***Abstract*-A fast algorithm which can extract pitch period information from speech is described. The solution is developed by processing speech signals using Haar wavelet function. We have shown mathematically that the pitch period of a speech signals can be computed using global minimum of the total energy level of the Haar processed signal. The accuracy of the proposed method is examined by observing the differences in pitch period obtained between the proposed method and the ETSI pitch estimation algorithm. By reducing the computational complexity, proposed pitch detector is 8 times faster than the conventional ETSI pitch estimation algorithm. Besides, the proposed method has been implemented in a energy-based harmonic-features speech recognition system, recognition results have been compared and discussed.**

Keyword-pitch period; haar wavelet function

I. INTRODUCTION

The extraction of pitch period is an essential component in the analysis and synthesis of speech signals. A well-designed pitch detector can be used to improve the performance in a variety of systems. Recently, researchers have used pitch information to enhance different types of speech processing systems such as speech recognition system developed in [1], and make uses of pitch-synchronous averaging to make their speech features robust. Pitch period information has been utilized in [2] to extract the Energy-based Harmonic Features from the speech signals and make the speech recognition system robust. Computationally efficient implementations for the ITU-T G.729 speech codec described in [3] first estimates the pitch period of the speech frame being coded. Picth information can be further used to improve the perceptual quality in noisy environment in [4]. A pitch detection is required in the cosine-modulated filter bank with TV number of channels in [5]. Method to extract pitch synchronous cepstrum (PSC) for robust speaker recognition over telephone channels proposed by [6] required analysis of consecutive pitch periods to compensate for spectrum distortion.

Various methods have been proposed to detect the pitch period. Classical or non-event based pitch detectors estimate the average pitch period by computing the autocorrelation function (ACF) [7], average magnitude difference function (AMDF) [8], cepstrum [9-10], inverse filtering based on linear prediction (LPC) [11] and others over a fixed windowed segment of speech. The computed spectrum consists of a series of impulses at the fundamental frequency and at the harmonics. Hence, the pitch period can be detected based on the detected spectrum. Though they are computationally simple to obtain, they suffer from two defects : (1) they are insensitive to non-stationary variations in the pitch period and (2) they are unsuitable for both low pitched and high pitched speakers [12].

Unlike nonevent-based pitch detectors that use a direct approach to estimate the pitch interval, event- based pitch detectors locate the glottal closure instants (GCI) where the time interval between two consecutive GCI is considered as the pitch interval [12]. Event based pitch detectors using wavelet have been extensively used. The first wavelet-based pitch determination of speech was proposed by [12] This method is superior to traditional pitch detection techniques in dealing with non-stationary signals in at least two ways: (1) it estimates the pitch period accurately and (2) it is suitable for a wide range of pitch periods. Other wavelet-based pitch detection algorithms inspired by [12] are presented in [13-15].

In this paper, we proposed a new wavelet-based pitch detection method which reduces the complexity and the average computation time needed compared to the Telecommunications Standards Institute (ETSI) pitch estimation algorithm [16]. Besides, the proposed method shows similar pitch period detection results with the pitch period results obtained by conventional ETSI pitch estimation algorithm. The reduction of time consuming makes the proposed pitch detection method suitable to be used in practical speech processing system. Besides, the proposed method has been implemented into a speech recognition system. The pitch period extracted using the proposed method has been utilised to extract energy-based harmonic features for speech recognition purpose. Recognition results have been compared and discussed.

The organization of the paper is as the following. In section 2, we show mathematically that the Haar transformed of harmonic model speech signal is suitable to be used in the detection of the pitch period. Recurrence relationship of the Haar transform for fast computation is also included in this section. Section 3 carries out the experimental study on the proposed method in pitch period detection using isolated and continuous speech signals contained inside TIDIGIT train subset. The obtained pitch period results and the average computation time of the proposed method are compared to the pitch period results obtained from ETSI pitch estimation algorithm [16]. Besides, the extraction of energy-based harmonic features for speech recognition purpose using pitch period extracted by the proposed method has been included also. Finally, the

conclusions of the study are summarized and discussed in Section 4.

II. FAST COMPUTATION OF PITCH PERIOD DETECTION

The harmonic model of a voiced frame, V(t) can be represented as

$$V(t) = b_{c0} + \sum_{n=1}^{N}\left(b_{cn}\cos\left(\frac{2n\pi}{T}t\right) + b_{sn}\sin\left(\frac{2n\pi}{T}t\right)\right) \quad (1)$$

where bc0 is the mean of the voiced speech signal, coefficients bcn and bsn carry the information of the intensity and phase of the nth harmonic component. N denotes the number of harmonics and T is the pitch period of the signal [17]. Let $b_{cn} = A_{n,t}\sin\frac{2n\pi}{T}(k_{n,t})$, $b_{sn} = A_{n,t}\cos\frac{2n\pi}{T}(k_{n,t})$, bc0 is zero, and assume that speech signals become stationary signals in short interval. Equation (1) can be represented as following:

$$V(t) = \sum_{n=1}^{N}\left[A_n\, sin\left[\frac{2n\pi}{T}(t + k_n)\right]\right] \quad (2)$$

A. Pitch Period Detection

The Haar function $\Psi(t_H)$ with bandwidth BT can be defined as

$$\Psi(t_H) = \begin{cases} 1 & 0 \le t_H < \frac{BT}{1} \\ -1 & \frac{BT}{2} \le t_H < BT \\ 0 & otherwise \end{cases} \quad (3)$$

$$\Psi_{j_H,k_H}(t_H) = \Psi(2^{j_H}t_H - k_H) \quad (4)$$

Integer j_H determines the dilation, while k_H specifies the translation. In this proposed method, both j_H and k_H are set to zero. By setting $t=t_0+t_H$, where t_0 is the starting time of the convolution process, the Haar wavelet series expansion of voiced speech signal function at time t relative to the wavelet $\psi(t_H)$ is shown below.

$$VH(t_0, t_H) = \int_0^{BT}\Psi(t_H)\left[\sum_{n=1}^{N}A_n\sin\left[\frac{2n\pi}{T}(t_0 + t_H + k_n)\right] + N(t_0 + t_H)\right]dt_H \quad (5)$$

The frequency content of noise N(t0+tH) in (5) is much higher than the range of fundamental frequencies of human speaker, hence, the expected value E(N(t0+tH)) is relatively low in amplitude after the convolution process. By integrating and further processing of (5), noise N(t0+tH) gets filtered out and we get

$$VH(t_0, BT) = -\sum_{n=1}^{N}\left[\frac{2A_nT}{n\pi}\sin^2\frac{n\pi BT}{2T}\cos\frac{n\pi(2t_0+2k_n+BT)}{2T}\right] \quad (6)$$

Discrete-time speech signals are processed using (7)

$$VH(t_0, t_H) = \sum_{t_H=0}^{BT}[\Psi(t_H)V(t_0 + t_H)] \quad (7)$$

The total energy level of VH(k,BT) over a fix segment C for different bandwidths, BT is defined as

$$EVH(BT) = \sum_{i=0}^{C}\left(VH(t_0 + i, BT)\right)^2 \quad (8)$$

The total energy level EVH is the summation of square values. Hence, it must be greater or equal to zero. When BT=2mT, where m is any positive integer number, then VH=0 and brings EVH=0 as well. However, since speech signals are non-stationary signals, slightly changes happen from time to time, the differences in the speech signals make the calculated EVH value using (8) deviated from the expected value. Obtained EVH at BT=2mT will be higher than 0, however, global minimum will happen at BT=2T where the changes in speech is the minimum in each speech frame. By locating the global minimum in each speech frame, pitch period T is obtained.

B. Fast Computation Using Recursive function

After getting the first $VH(t_0,t_H)$ values, other VH can be fast computed using the following two recurrence relationships.

$$VH(t_0 + 1, t_H) = VH(t_0, t_H) - V(t_0) + 2V\left(\frac{t_H}{2} + t_0\right) - V(t_H + t_0) \quad (9)$$

$$VH(t_0, t_{H+2}) = VH(t_0, t_H) + 2V\left(\frac{t_H}{2} + t_0\right) - V(t_H + t_0) - V(t_H + t_0 + 1) \quad (10)$$

Fig. 1 shows the flow chart of the proposed algorithm to detect the pitch period. There are 4 types of voicing class: 1)non-speech, 2) unvoiced, 3) mixed-voiced and 4) fully-voiced, only mixed-voiced and fully-voiced speech frames have pitch period. All types of speech frame will be processed by the same proposed algorithm. To distinguish between the voiced speech frame and unvoiced speech frame, threshold values for the global maximum, global minimum and average value of the computed EVH was obtained through experimental study. A speech frame will be classified as voiced speech frame if only its global maximum and average EVH are higher than their threshold values respectively and the EVH global minimum of the speech frame must be lower than the threshold value. For unvoiced speech frame, the pitch period will equal to 0. The threshold values used in this proposed algorithm are different with the threshold values used in the ETSI pitch estimation algorithm. For voiced speech frame, by locating the global minimum of the obtained EVH from the signals, pitch period can be found using BT=2T where BT is the pitch period and T is the corresponding x value of the global minimum.

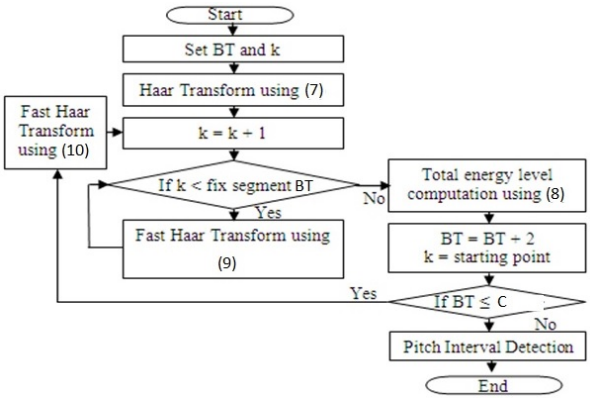

Figure 1. Proposed pitch period detection flow chart.

Fig. 2 shows a EVH plot of an isolated speech 'one' spoken by a female speaker. From the figure, there are many local minimum, this agrees with the earlier statement that when BT=2mT, EVH value is close to zero. Global minimum happens at 8.6ms, detected pitch period is equal to 4.3ms.

III. PERFORMANCE EVALUATION AND EXPERIMENTAL RESULTS

The proposed Haar wavelet-based pitch detector algorithm has been tested on all the isolated and continuous speech signals contained inside TIDIGIT train subset. All the speech signals have been down sampled to 8 kHz before the pitch extraction process. Total of 8598 speech signals have been processed and a total of 756863 voiced speech frames have been detected. Pitch period of each voiced speech frame has been extracted.

A. Accuracy

To show the accuracy of the proposed method, obtained pitch period was compared to the pitch period result of the same speech frame extracted using ETSI pitch estimation algorithm. Fig. 3 shows the percentage of difference in millisecond for the pitch period detected using the proposed algorithm and the ETSI pitch estimation algorithm. Results show that around 80% of the detected pitch periods using proposed algorithm are very close to the results obtained using ETSI pitch estimation algorithm. The differences are smaller than 0.25 ms, in other words smaller than 2 data points in 8 kHz discrete signals. Overall, ETSI algorithm shows more accurate detection on pitch interval than the proposed method. This is because ETSI algorithm is more complicated and it extracts more detail information such as Voice Activity Detection, Spectrum and Energy Computation, Mel Filtering, Low-Band Noise Detection and others. All these detail information make the ETSI pitch estimation algorithm more accurate but time consuming.

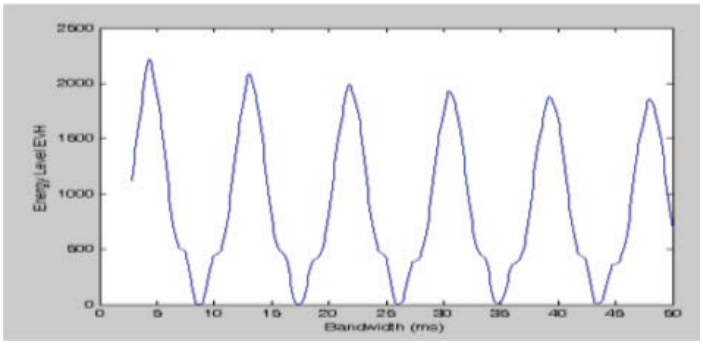

Figure 2. Total energy level, EVH of an isolated speech 'one' spoken by a female speaker.

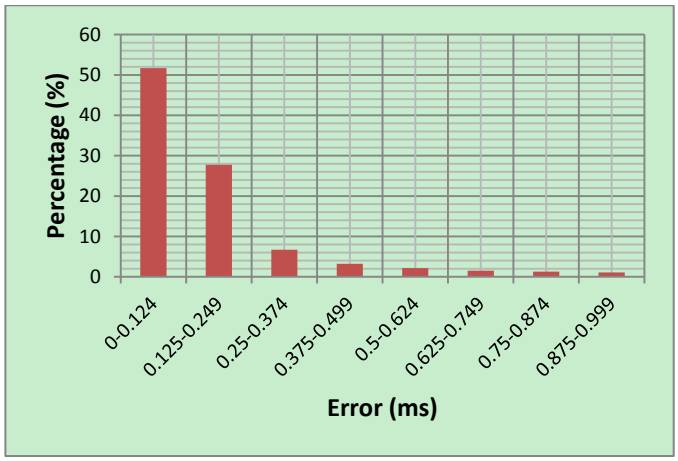

Figure 3. Percentage of difference in millisecond between the detected pitch period using proposed algorithm and ETSI pitch estimation algorithm.

B. Computation Time

Using a computer with Intel(R) Core(TM) i5 3.1 GHz processor, running Matlab R2015a under Windows 8.1 operating system, to classify all the speech signals contained inside the TIDIGIT train subset into voiced and unvoiced speech frames and then extract the pitch period information from each speech frame, proposed Haar wavelet-based pitch detector algorithm is 8 times faster than the ETSI pitch estimation algorithm. This is caused by the simplicity of the proposed algorithm and the used of the recurrence relation in (9) and (10) for fast processing.

C. Speech Recognition Using Harmonic Features

The proposed wavelet-based pitch detector algorithm was further tested in a speech recognition system which utilised both MFCC features and Energy-based harmonic features proposed in [2]. The recognition test was carried out using TIDIGIT database. Each speech signal is first down sampled to 8 kHz. Then, direct current (DC) offset of the input speech signal is removed. The offset-free input signal is then divided into overlapping frames. The frame

length is 25 ms and the frame shift interval is 10 ms. A pre-emphasis filter is applied to the framed offset-free input signal. This follows by applying a Hamming window of length 25 ms to the output of pre-emphasis block. Pitch information of each frame is estimated using the proposed Harr wavelet-based pitch detector algorithm and the ETSI pitch estimation algorithm. By using the extracted pitch periods, dynamic harmonic features of each speech frame were extracted using the equations provided in [2]. The extracted dynamic harmonic features were further concatenated with the MFCC features. 12 MFCC features, 1 energy level, 12 delta MFCC, 1 delta energy level and 8 delta harmonic features were used in training and testing the speech recognition system. HTK-toolkit was used in building Hidden Markov models (HMM).

Table I shows the recognition rates of the speech recognition system using the proposed algorithm and the ETSI pitch estimation algorithm. Results show that the recognition rate of the harmonic-based speech recognition system using pitch period information extracted by the proposed algorithm is slightly lower than the recognition rate of the system using pitch period information extracted by the ETSI pitch estimation algorithm. According to [2], accuracy in pitch period detection affects the speech recognition rate for the speech recognition system using Energy-based Harmonic Features. This may be the cause that the speech recognition system using ETSI pitch estimator as pitch period extrator performs better than the speech recognition system using the proposed algorithm as pitch period extractor. However, the processing time needed for the proposed algorithm is shorter than the ETSI algorithm.

TABLE 1. WORD ACCURACY (WACC) OF HARMONIC-BASED SPEECH RECOGNTION SYSTEM USING DIFFERENT PITCH ESTIMATION ALGORITHMS

Type of pitch estimator	Clean	20dB	15dB	10dB
Wavelet-based pitch estimator	99.0	92.0	90.6	82.1
ETSI pitch estimator	99.0	92.7	91.2	84.6

IV. CONCLUSION

A new method to detect the pitch period of speech frame has been proposed in this paper. We showed the mathematical reasons as to why the Haar transform can be used in the detection of the pitch period. The validity of the proposed method was tested on speech signals contained in the TIDIGIT train subset. Encouraging, 80% of the detected pitch periods using the proposed algorithm are very close to the detected pitch periods result obtained using ETSI pitch estimation algorithm. Due to the simplicity and the used of recurrence functions for fast processing, computation time needed for the proposed algorithm is 8 times faster than the ETSI pitch estimation algorithm. Furthermore, detected pitch period using the proposed algorithm has been tested on a Harmonic-based speech recognition system. The obtained recognition results are slightly lower than recognition results using pitch period detected by the ETSI pitch estimator algorithm.

REFERENCES

[1] J. Morales-Cordovilla, A. Peinado, V. Sanchez, and J. Gonzalez, "Feature extraction based on pitch-synchronous averaging for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 640–651, 2011.

[2] Y.H. Goh, P. Raveendran and S.S. Jamuar, "Robust speech recognition using harmonic features," *IET Signal Processing*, vol. 8, no. 2, pp. 167-175, 2014.

[3] S. Netto, "Efficient search in the adaptive codebook for itu-tg. 729 codec," *Signal Processing Letters, IEEE*, vol. 16, no. 10, pp. 881–884, 2009.

[4] H. Park, J. Yoon, J. Kim, and E. Oh, "Improving perceptual quality of speech in a noisy environment by enhancing temporal envelope and pitch," *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 489–492, 2010.

[5] P. Polotti, "A pitch-synchronous extension of fractal additive synthesis via time-varying cosine modulated filter banks," *Signal Processing Letters, IEEE*, vol. 15, pp. 433–436, 2008.

[6] Y. Kim and J. Chung, "Pitch synchronous cepstrum for robust speaker recognition over telephone channels," *Electronic Letters*, vol. 40, no. 3, pp. 207–209, 2004.

[7] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 25(1), pp. 24–33, 1977.

[8] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, pp. 353–362, 1974.

[9] A. Noll., "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, vol. 41(2), pp. 293–309, 1967.

[10] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithms," *IEEE transactions on speech and audio processing*, vol. 7, pp. 333–338, 1999.

[11] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 309–319, 1979.

[12] S. Kadambe and G. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Transactions on Information Theory*, vol. 38, pp. 917–924, 1992.

[13] C. Wendt and A.P.Petropulu, "Pitch determination and speech segmentation using the discrete wavelet transform," in *IEEE International Symposium on Circuits and Systems*, vol. 2, 1996, pp. 45–48.

[14] D.-Y. Huang and W. Lin, "Application of analytical optimum fir compaction filters for tracking pitch of musical signals," in *5th World Multi-Conference Systemics, Cybernetics and Informatics (SCI2001) proceedings*, vol. VI. Orlando, Florida, U.S.A., 2001, pp. 81–85.

[15] N. Sturmel, C. dAlessandro, and F. Rigaud, "Glottal closure instant detection using lines of maximum amplitudes (loma) of the wavelet transform," in *ICASSP*, 2009, pp. 4517–4520.

[16] Speech processing, transmission and quality aspects (stq); distributed speech recognition; extended front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm, ETSI ES 202 211 v1.1.1, 2003

[17] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 1, pp. 76–87, 2004.