PITCH PERIOD ESTIMATION USING MULTIPULSE MODEL AND WAVELET TRANSFORM[†]

Prasanta Kumar Ghosh, Antonio Ortega and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory &
Signal and Image Processing Institute,
Department of Electrical Engineering,
University of Southern California,
Los Angeles, CA 90089
prasantg@usc.edu, antonio.ortega@sipi.usc.edu, shri@sipi.usc.edu

Abstract

Wavelet transform-based pitch period estimation is well known in the literature. This approach to pitch estimation assumes that the glottis closures are correlated with the maxima in the adjacent scales of the wavelet transform and for pitch period estimation, one needs to detect these correlated maxima across these scales, which is often prone to error especially in the case of noisy signals. In this paper, we develop an optimization scheme in the wavelet framework using a multipulse excitation model for the speech signal and the pitch period is estimated as a result of this optimization. We report experiments on both clean and noisy conditions and show that the proposed optimization works better than wide used heuristic approach for maxima detection.

Index Terms: Pitch period estimation, multipulse excitation, dyadic wavelet transform.

1. Introduction

Pitch period estimation of speech and music signals is an essential component in various speech processing applications such as speech coding, speaker identification and verification, and speech segregation. The existing pitch detectors are broadly classified into either event detection pitch detectors or nonevent detection pitch detectors. Event detection pitch detectors estimate the pitch period by locating the instant at which the glottis closes (called an event) and then measuring the time interval between two such events. The non-event based pitch detectors estimate pitch period by a direct approach like autocorrelation or cepstrum method. The estimation of pitch using wavelet transform falls under the category of event detection pitch detectors. The pitch detection based on classical wavelet transform (CWT) in [1]-[3] estimates the pitch period by determining the glottal closure instant (GCI) and measuring the time period between such two events. In [10], the proposed algorithm reduces the computational load of those CWT based algorithms. They demonstrate the suitability of these approaches for a wide range of pitch periods and different speakers.

However, it should be noted that the wavelet transform is appropriate for pitch period estimation based on the assumption that the glottal closure causes sharp changes (discontinuities) in the derivative of the air flow in the glottis and transients in the speech signal [1], which results in maxima in the scales of the wavelet transform around the point of discontinuity [4]. In addition, one needs to detect the correlated maxima of the wavelet coefficients at successive scales by heuristic algorithms. In this paper, we demonstrate analytically, using multipulse model of excitation that the glottal pulse positions are preserved in successive scales of wavelet transform of speech signal and instead of any heuristic approaches, we formulate an optimization problem for finding pitch period under the wavelet framework. We model the speech signal using an excitation-filter model [5] as follows:

$$x[n] = e[n] \star h[n] \tag{1}$$

where $\star$ denotes the convolution. $h[n]$ is the vocal tract filter impulse response and $e[n]$ is the glottal excitation, which is modeled as a series of delta functions, spaced irregularly in general, i.e.,

$$e[n] = \sum_k \beta_k \delta[n - n_k] \tag{2}$$

where β_k are the amplitudes of the glottal excitation pulses and n_k are the pulse positions. This is the multipulse model of the excitation signal, first proposed by Atal et al [6]. Thus from (1) and (2),

$$x[n] = \sum_k \beta_k h[n - n_k] \tag{3}$$

With this model of the speech signal, the signal at successive scales of the wavelet transform can also be expressed in terms of the pulse amplitude and pulse positions. Assuming that the excitation pulse positions occur at an interval of the pitch period, we minimize a suitable cost function with respect to pulse positions. Minimization yields the optimum pulse positions and thus, the desired pitch period is obtained.

2. Wavelet Transform

The wavelet transform (WT) could be classified as either continuous wavelet transform or discrete wavelet transform (DWT). A continuous wavelet transform of a signal $x(t) \in L^2(R)$ results in:

$$WT_x(\lambda, \tau) = \frac{1}{\sqrt{\lambda}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t - \tau}{\lambda} \right) dt \quad \lambda > 0 \tag{4}$$

where the function $\psi(t)$ is usually referred to as mother wavelet, λ is the scaling factor, τ is the shift and * stands for com-

[†] Work supported in part by NIH and ONR-MURI.

plex conjugation. The DWT can be performed via the multiresolution analysis wavelet decomposition/reconstruction algorithm developed by Mallat. At the m^{th} level, the multiresolution space, V_m, is spanned by the basis functions $\left\{2^{m/2}\phi(2^m t - n);\ n \in Z\right\}$ and the space, W_m, orthogonal to V_m in V_{m-1} is spanned by $\left\{2^{m/2}\psi(2^m t - n);\ n \in Z\right\}$, where $\phi(t)$ is called the scaling function and $\psi(t)$ is called the wavelet function. Mallat's algorithm allows wavelet coefficients (also called the detailed version of the signal), $d_{m,n} = \langle x(t), \psi_{m,n}\rangle$ and scaling coefficients (also called the approximation version) $x_{m,n} = \langle x(t), \phi_{m,n}\rangle$ at the m^{th} scale to be calculated recursively from the representation of the signal, $x(t)$, at the preceding, finer scale, $x_{m-1,n}$, through the following filtering operation:

$$x_{m,n} = \sum_k a_0(k-2n)x_{m-1,k} \qquad (5)$$

$$d_{m,n} = \sum_k a_1(k-2n)x_{m-1,k} \qquad (6)$$

$$\begin{aligned} \text{where } \ a_0[n] &= \langle \phi_{1,0}, \phi_{0,n}\rangle \\ a_1[n] &= \langle \psi_{1,0}, \psi_{0,n}\rangle \end{aligned}$$

3. Proposed Optimization Technique

In this section, we show analytically how pulse positions in adjacent scales are preserved and formulate the optimization problem for pitch period estimation. Let us denote $x^m[n]$ the approximation version of the input signal $x[n]$ at the m^{th} scale. Thus using (5),

$$\begin{aligned}
x^1[n] &= (x[n] \star a_0[n]) \downarrow_2 \\
&\quad (\downarrow_2 \text{ denotes downsampling by a factor of 2}) \\
&= \left(\sum_l x[l]a_0[n-l]\right)\downarrow_2 \\
&= \left(\sum_l \left(\sum_k \beta_k h[l-n_k]\right) a_0[n-l]\right)\downarrow_2 \\
&\qquad\qquad\qquad\qquad [\text{using (3)}] \\
&= \left(\sum_k \beta_k \left\{\sum_l h[l-n_k]a_0[n-l]\right\}\right)\downarrow_2 \\
&= \sum_k \beta_k \left(\sum_l h[l]a_0[n-n_k-l]\right)\downarrow_2 \\
&\qquad\qquad\qquad\qquad (l \rightarrow l+n_k) \\
&= \sum_k \beta_k h^1[n-n_k] \qquad (7)
\end{aligned}$$

$$\text{where,} \qquad h^1[n] = (h[n] \star a_0[n]) \downarrow_2 \qquad (8)$$

Similarly[1],

$$x^2[n] = \sum_k \beta_k h^2[n-n_k] \qquad (9)$$

$$\text{where, } h^2[n] = \left(h^1[n] \star a_0[n]\right)\downarrow_2 \qquad (10)$$

In general, the model for $x^m[n]$ is

$$x^m[n] = \sum_k \beta_k h^m[n-n_k] \qquad (11)$$

$$\text{where, } h^m[n] = \left(h^{m-1}[n] \star a_0[n]\right)\downarrow_2 \qquad (12)$$

From (11), we see that the output at the m^{th} scale has a similar excitation pattern as in the original signal; only the filter impulse response has been changed from $h[n]$ to $h^m[n]$.

$h[n]$ is estimated from the given finite length speech signal $\{x[n]\}_{n=0}^{N-1}$ in the following way:

$$h[n] \overset{\mathcal{Z}}{\leftrightarrow} H(z) = \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}}$$

where $\{a_k\}_{k=1}^{p}$ are the optimum p-order linear prediction coefficients obtained from $x[n]$ [7]. Once $h[n]$ is obtained we can compute $h^m[n]$ from (8) and (12). Thus, the signal domain modeling error $e[n]$ and the modeling error $e^m[n]$ at the m^{th} scale decomposition are

$$e[n] = x[n] - \sum_{k=1}^{k=K} \beta_k h[n-n_k] \qquad (13)$$

$$e^m[n] = x^m[n] - \sum_{k=1}^{k=K} \beta_k h^m[n-n_k] \qquad (14)$$

where $m = 1, 2, ..., M$; M is the maximum level of wavelet decomposition and K is the number of pulses in the given speech segment.

We construct a cost function J which is the sum of the energy of the errors in the signal domain and in each scale of the wavelet decomposition. Thus,

$$\begin{aligned}
J(\beta_k, n_k, K) &= \sum_n (e[n])^2 + \sum_{m=1}^{m=M}\sum_n (e^m[n])^2 \\
&= \sum_n \left(x[n] - \sum_{k=1}^{k=K}\beta_k h[n-n_k]\right)^2 \\
&+ \sum_{m=1}^{m=M}\sum_n \left(x^m[n] - \sum_{k=1}^{k=K}\beta_k h^m[n-n_k]\right)^2
\end{aligned} \qquad (15)$$

Assuming constant pitch period over the given voiced segment, we rewrite n_k as follows

$$n_k = kN_p + N_0 \qquad (16)$$

where N_p is the constant pitch period over the signal segment and N_0 is the offset of the first pitch pulse in the segment. Hence, the cost function becomes

$$\begin{aligned}
J(\beta_k, N_p, N_0, K) &= \sum_n \left(x[n] - \sum_{k=1}^{k=K}\beta_k h[n-kN_p-N_0]\right)^2 \\
+ \sum_{m=1}^{m=M}\sum_n & \left(x^m[n] - \sum_{k=1}^{k=K}\beta_k h^m[n-kN_p-N_0]\right)^2
\end{aligned} \qquad (17)$$

We minimize $J(\beta_k, N_p, N_0, K)$ and obtain the optimum values of $\{\beta_k^{opt}\}_{k=1}^{K}$, N_p^{opt}, N_0^{opt}, K^{opt}.

4. Pitch period estimation by minimizing J

From (17), we see that the minimization of J with respect to $\{\beta_k\}$ is straightforward. Once we know the N_p, N_0, K, we can set the derivatives of J with respect to β_k, $1 \le k \le K$ equal to zero to obtain equations

$$\sum_{k=1}^{k=K} \beta_k c_{ki} = d_i, \ \ 1 \le i \le K \qquad (18)$$

[1] Note that $h^1[n-n_k] = \left(\sum_l h[l]a_0[n-n_k-l]\right)\downarrow_2$

where

$$\begin{aligned}
c_{ki} &= \sum_n h[n-kN_p-N_0]h[n-iN_p-N_0] \\
&+ \sum_{m=1}^{M}\sum_n h^m[n-kN_p-N_0]h^m[n-iN_p-N_0] \\
d_i &= \sum_n x[n]h[n-iN_p-N_0] \\
&+ \sum_{m=1}^{M}\sum_n x^m[n]h^m[n-iN_p-N_0]. \qquad (19)
\end{aligned}$$

The minimization of J with respect to the $N_p,\ N_0,\ K$ is a combinatorial problem and does not have a closed form solution. We exploit the advantage of wavelet decomposition here. From (3) and (11), we observe that the quasi-periodic structure of the signal is also maintained in adjacent scales of wavelet transform. We estimate an initial value of $N_p^{ini},\ K^{ini}$ from the 2nd or 3rd scale coefficients of wavelet decomposition and perform a combinational search for $N_p,\ N_0,\ K$ in the following ranges:

$$\begin{aligned}
N_p &\in \left[N_p^{ini}-20, N_p^{ini}+20\right] \\
N_0 &\in \left[1, 2N_p^{ini}\right] \\
K &\in \left[K^{ini}-2, K^{ini}+2\right]
\end{aligned}$$

The initial estimate of the above parameters from the wavelet decomposition is effective specially in case of noisy condition, when the estimates in the signal domain is very poor. In the successive levels, the noise is reduced due to filtering, while the periodic structure is maintained. We use the autocorrelation and average magnitude difference function (AMDF) [11], to estimate N_p^{ini}. We follow an approach similar to what was taken by Shimamura et al in [8], where the peak in the AMDF weighted autocorrelation function is detected to estimate N_p^{ini}. From N_p^{ini} the K^{ini} is obtained by finding the possible number of period of length N_p^{ini} over the given signal of length N, i.e. $K^{ini} = \left\lceil \frac{N}{N_p^{ini}} \right\rceil$[2]. For each combination of N_p, N_0 and K, the $\{\beta_k\}_{k=1}^{K}$ are solved using (18) and the value of J is noted. The optimum parameters are chosen for the smallest value of J among all these combinations. The N_p^{opt}, obtained after minimizing J is declared to be the pitch period of the given speech segment. $f_p^{opt} = \frac{f_s}{N_p^{opt}}$ is called the pitch frequency, where f_s is the sampling frequency of the speech signal.

As a result of the optimization, we also get the pitch pulse positions $n_k^{opt} = kN_p^{opt} + N_0^{opt}$ and their amplitudes β_k^{opt}.

5. Evaluation

Given a speech segment, $h[n]$ is obtained by linear prediction of order p =10. We have used the Haar wavelet in all our experiments for decomposing the input segment in successive levels. Use of other wavelets did not change the result drastically. 3 levels of decomposition is used for all experiments, i.e. M=3.

The experiment is performed on voiced data from the TIMIT database [9]. The test set contains 10 utterances from each of the eight dialect regions of TIMIT (total of 80 utterances), half spoken by male and half by female speakers. The voiced segments for each utterance are detected based on short-time energy using frame based analysis. A frame length of 20 msec is used for this and if the short-time energy of a frame is more than 60% of the maximum short-time energy of an utterance, that frame is decided to be a voiced frame. These voiced frames are used in our optimization and pitch values are obtained for all voiced frames.

Table 1: *Coefficient of similarity*

Method	Praat	Optimization
ρ	.872	.878

To evaluate the performance of the optimization based pitch period estimation for clean speech, we define an objective measure which avoids the knowledge of the 'true' underlying pitch. The optimization results in K^{opt} number of pitch periods in a given speech segment. Let $\underline{S}_k$ denote the signal vector of the k^{th} period; we define the following coefficient of similarity over all periods,

$$\rho = \frac{1}{K^{opt}-2}\sum_{k=2}^{K^{opt}-1}\frac{\text{Cov}\left[\underline{S}_k, \frac{\underline{S}_{k-1}+\underline{S}_{k+1}}{2}\right]}{\sqrt{\text{Var}[\underline{S}_k]\text{Var}\left[\frac{\underline{S}_{k-1}+\underline{S}_{k+1}}{2}\right]}} \qquad (20)$$

where Cov[$\cdot$] and Var[$\cdot$] denote sample covariance and variance, respectively. The closer ρ is to 1, the more the estimated periods are correlated i.e. better is the pitch estimation. ρ is determined for each frame and the resulting values are averaged to obtain an overall value. As a reference method to our approach, we chose the pitch marking algorithm of Praat [12], a state-of-the-art event-based pitch estimation method. Table 1 shows ρ values for Praat and our optimization algorithm. It is seen that in terms of such an objective measure, the wavelet based optimization method works almost as efficiently as that of fine-tuned modified autocorrelation based approach, used in Praat.

We now choose a critical speech segment to demonstrate how heuristic approach for maxima detection fails whereas the optimization scheme works well. Fig. 1(a) shows the original speech segment and its wavelet (Haar) coefficients in 1st, 2nd and 3rd scales are shown in (b)-(d). These are repeated in the right hand column, Fig. 1(e)-(h), where the maxima are detected using algorithm[3], reported in [1, 10]. For this speech segment it can seen that the maxima detected across scales does not match always, which leads to a wrong estimate of the pitch period. In contrast, the optimum pulse positions obtained through optimization are shown in the left hand column Fig. 1(a) on the original signal segment.

To evaluate the performance under noisy conditions, we performed experiments on the same data set using additive white Gaussian noise at 10 dB, 5 dB and 0 dB. The pitch value obtained from Praat is taken as reference. The % relative error (= (reference pitch - estimated pitch)/(reference pitch)X100) averaged over all frames is shown in Table 2 for both optimization based and maxima detection based pitch estimation. The optimization based approach shows better performance in all noisy conditions.

6. Conclusions

Through multipulse excitation based signal modeling in the wavelet framework, we analytically show that the pseudo-periodic structure of speech signal is maintained in adjacent scales. The proposed optimization scheme for pitch estimation is based on this analytical result and is shown to be robust by experimental evaluation.

[2] $\lceil x \rceil$ is the smallest integer greater than x

[3] The algorithm described in [1] picks the maxima which are above the 0.8 of the global maxima in a scale. Two dash-dotted lines in Fig. 1(f)-(h) indicate the range between global maxima and its 80% value.

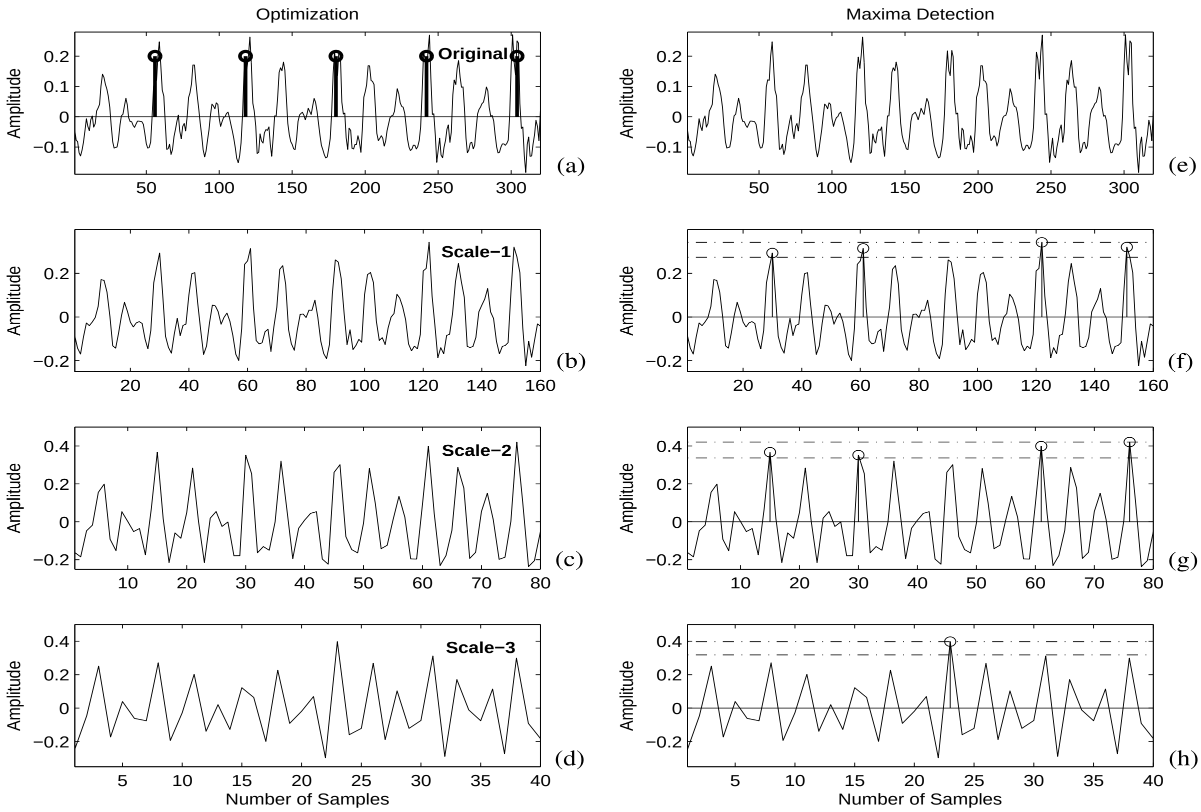

Figure 1: *Demonstration of the effectiveness of the optimization (left column) over maxima detection (right column). (a), (e) original signal; (b)-(d), (f)-(h) are its wavelet coefficients in 1-3 scale. Optimum pulses obtained by optimization are shown in (a). The maxima detected across scales are shown in (f)-(h). Two dash-dotted lines in each scale indicate the global maxima value and its 80% value.*

Table 2: *% relative error*

Noise Level	Maxima Detc.	Optimization
clean	0.01	0.004
10 dB	0.519	0.367
5 dB	1.012	0.958
0 dB	1.897	1.673

In formulating the cost function (17), one can also use different weight factors for different scales based on the estimated noise level in each scale. As a by-product of this optimization the pulse amplitudes $\{\beta_k\}$ are also estimated, which further could be useful for determining speaker characteristics. Validation of such ideas are part of future work.

7. References

[1] Kadambe S, Faye Boudreaux-Bartels G, “Application of the wavelets transform for pitch detection of speech signals”, IEEE Trans. on Information Theory 1992; 38(2), pp. 917-924.

[2] Obaidat MS, Lee C., Sadoun B., Neslon D., “Estimation of pitch period of speech signal using a new dyadic wavelet transform”, Journal of Information Sciences 1999; 119; pp. 21-39.

[3] Obaidat M.S., Bradzik A., Sadoun B., “A performance evaluation study of four wavelet algorithms for the pitch period estimation of speech signals”, Journal of Information Sciences 1998; 112; pp. 213-221.

[4] Mallat S.G., and Zhong S., “Complete signal representation with multiscale edges”, tech. rep. RRT-483-RR-219, Courant Inst. of Math. Sci., Dec. 1989.

[5] Singhal, S. and Atal, B.S., “Amplitude optimization and pitch prediction in multipulse coders”, IEEE Transactions on Acoustics, Speech, and Signal Proc., Volume 37, Issue 3, March 1989, pp. 317 - 327.

[6] Atal B.S. and Remde J., “A new model of LPC excitation for producing natural-sounding speech at low bit rates”, Proc. Int. Conf. Acoust. Speech. Signal Processing, Paris, France, 1980, pp. - 614-617.

[7] Makhoul, J., “Linear prediction: A tutorial review”, Proceedings of the IEEE, Volume 63, Issue 4, April 1975, pp. 561 - 580.

[8] Shimamura, T., Kobayashi, H., “Weighted autocorrelation for pitch extraction of noisy speech”, IEEE Transactions on Speech and Audio Proc., Volume 9, Issue 7, Oct. 2001, pp. 727 - 730.

[9] “DARPA-TIMIT”, Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1, 1990.

[10] Ergun Ercelebi, “Second generation wavelet transform-based pitch period estimation and voiced/unvoiced decision for speech signals”, Applied Acoustics 2003; 64, pp. 25-41.

[11] Lawrence R. Rabiner and Ronald W. Schafer, “Digital Processing of Speech Signals”, Prentice-Hall Series in Signal Processing, 1978.

[12] Paul Boersma and David Weenink, “Praat: doing phonetics by computer (Version 4.3.14)”, http://www.praat.org/.