Fast Harmonic Estimation Using a Low Resolution Pitch for Low Bit Rate Harmonic Coding

Yong-Soo Choi[*]*, and Dae-Hee Youn*[+]

[*]Digital Network R&D Lab.,
LG Electronics Inc.,
Seoul 153-023, Korea
*cando@lgic.co.kr

[+]Center for Signal Processing Research,
Yonsei University,
Seoul 120-749, Korea
+dhyoun@yonsei.ac.kr

Abstract

This paper describes a fast harmonic estimation, referred to Delta Adjustment (DA), using a low resolution pitch. The presented DA method is based on modification of the Generalized Dual Excitation (GDE) technique [1] which was proposed to improve speech enhancement performance. We introduce the GDE technique and modify it to be suitable for low bit rate harmonic coding that uses only an integer pitch estimate. Unlike the GDE, the DA matches a frequency-warped version of the original spectrum that conforms to a fixed pitch at all harmonic bands. In addition, complexity and performance of the presented method are described in comparison with those of the conventional Fractional Pitch (FP) based harmonic estimation. Experimental results showed that the DA algorithm significantly reduces the complexity of the FP method while maintaining the performance.

1. Introduction

Harmonic coders that can be classified as parametric coders such as Sinusoidal Transform Coding (STC) [2], MultiBand Excitation (MBE) [3], and Waveform Interpolation (WI) [4] have proven useful for producing speech at a 2.4 kbps or lower rate. In recent years, most harmonic coders use the fundamental frequency (or pitch), linear predictive coefficients (LPC), and spectral harmonics as encoding parameters. Pitch and harmonic estimation are important in speech quality and complexity of the harmonic coder.

Improved MBE (IMBE) [5] and Harmonic Vector eXcited Coding (HVXC) [6] employ the frequency domain analysis-by-synthesis technique, referred to the fractional pitch (FP) method, for the harmonic estimation. The method has good performance for low bit rate harmonic coding, but requires considerable complexity since it includes an iterative error minimization process for fractional pitch candidates.

In speech enhancement area, Yoo and Lim proposed the Generalized Dual Excitation (GDE) [1] to improve performance of the DE model [7] that uses a similar method to the FP by taking pitch variation into consideration. The GDE allows a more complete decomposition of the speech into voiced and unvoiced components. Harmonics of the synthesized voiced component by the GDE are faithful to those of the original speech spectrum more accurately.

The objective of this paper is to present a fast harmonic estimation method using only an integer pitch estimate, while maintaining the FP performance. To achieve this goal, we partly adopt the speech decomposition algorithm of the GDE model and modify it to be suitable for low bit rate harmonic coding. We refer the presented method to Delta Adjustment (DA) in this paper. Unlike the GDE, the DA does not attempt to match the original speech spectrum exactly. Instead, the DA matches a frequency-warped version of the original spectrum that conforms to a fixed pitch at all harmonic bands. This concept is similar to the Relaxation Code Excited Linear Prediction (RCELP) [8].

Complexities and performances of the conventional FP and the DA harmonic estimation are described. From results of objective subjective tests, the DA showed performance comparable to that of the FP while reducing its complexity significantly.

The paper is organized as follows: First, we give a general description of the speech decomposition algorithm of the GDE. In Section 3, we present the DA algorithm focused on modification of the GDE technique. Section 4 states complexity and performance comparison of the presented DA and conventional FP methods. Finally, we make a conclusion in Section 5.

2. The GDE decomposition algorithm

The GDE algorithm [1] generalizes DE model [7] by taking pitch variation into consideration to improve speech enhancement performance. The algorithm allows a more complete decomposition of the speech into voiced and unvoiced components. The GDE characterizes speech in two ways: First, it allows for some irregularities in the periodicity. Second, some small inaccuracies in the pitch estimate can be compensated for. Overall by allowing small variations in the pitch, the harmonics of the synthesized voiced component match those of the speech spectrum more accurately. This effect can be significant for high frequency harmonics and transitional speech segments. Figure 1 conceptually shows the voiced component extraction of the GDE that consists of frequency adjustment and harmonic estimation processes.

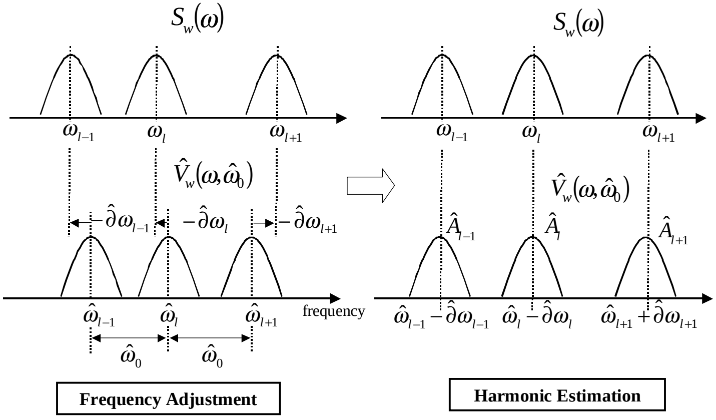


Figure 1. Basic concept of the voiced component extraction of the GDE.

In the GDE, the estimate of the windowed voiced component is obtained by minimizing the following error criterion,

$$\varepsilon = \frac{1}{\pi}\int_0^{\pi} \left| S_w(\omega) - \hat{V}_w(\omega) \right|^2 d\omega, \tag{1}$$

where $S_w(\omega)$ and $\hat{V}_w(\omega)$ are spectra of windowed input and synthetic speech signals. In (1), the synthetic voiced spectrum $\hat{V}_w(\omega)$ is modeled as L windowed harmonics and is given by

$$\hat{V}_w(\omega) = \sum_{l=1}^{L} \hat{A}_l W\left(\omega - l\hat{\omega}_0 + \hat{\partial}\omega_l\right), \tag{2}$$

where $W(\omega)$, $\hat{\omega}_0$, $\hat{A}_l$, $\hat{\partial}\omega_l$ and L are the spectrum of the window function, the estimated fundamental frequency, the l-th estimated harmonic amplitude, and the estimated harmonic frequency variation, the number of harmonics, respectively. $\hat{A}_l$ and $\hat{\partial}\omega_l$ are obtained by differentiating (1) and setting to zero as

$$\hat{A}_l = \frac{\int_{a_l}^{b_l} S_w(\omega) W^*\left(\omega - l\hat{\omega}_0 - \hat{\partial}\omega_l\right) d\omega}{\int_{a_l}^{b_l} \left| W\left(\omega - l\hat{\omega}_0 - \hat{\partial}\omega_l\right) \right|^2 d\omega}, \tag{3}$$

$$\hat{\partial}\omega_l = \arg\max_{\partial\omega_l} \left| \int_{a_l}^{b_l} S_w(\omega) W^*\left(\omega - l\hat{\omega}_0 - \partial\omega_0\right) d\omega \right|^2, \tag{4}$$

where $a_l = (l - 0.5)\hat{\omega}_0$, $b_l = (l + 0.5)\hat{\omega}_0$, $|\partial\omega_l| \le C\frac{l}{L}\hat{\omega}_0$, $0 \le C \le 0.5$, and * denotes complex conjugate.

3. Fast harmonic estimation

As a matter of convenience, first, we briefly describe the conventional Fractional Pitch (FP) method used in IMBE [5] and HVXC [6] as the reference. Then we give a detail description of a fast harmonic estimation process, referred to Delta Adjustment (DA) method.

The FP utilizes the frequency domain analysis-by-synthesis technique and has good performance for low bit rate coding. However, the FP requires high complexity since it includes an iterative error minimization process, i.e. $\hat{\partial}\omega_l = 0$ in (1) and (2), between the original and synthetic spectra for M fractional pitch candidates. Figure 2 shows a block diagram of the conventional FP harmonic estimation method. In Figure 2, N and N_1 are the frame length and the Discrete Fourier Transform (DFT) size of the input spectrum, respectively.

To present a fast harmonic estimation method while maintaining the FP performance, we adopt the GDE technique and modify it to be suitable for low bit rate speech coding. Unlike the GDE, the DA does not attempt to match the original spectrum exactly. Instead of attempting to match the original spectrum, the DA matches a frequency-warped version of the original spectrum that conforms to a fixed pitch at all harmonic bands. This concept is similar to the Relaxation Code Excited Linear Prediction (RCELP) [8], though its effect is less meaningful than that of the RCELP. A basic concept of the DA method is shown Figure 3.

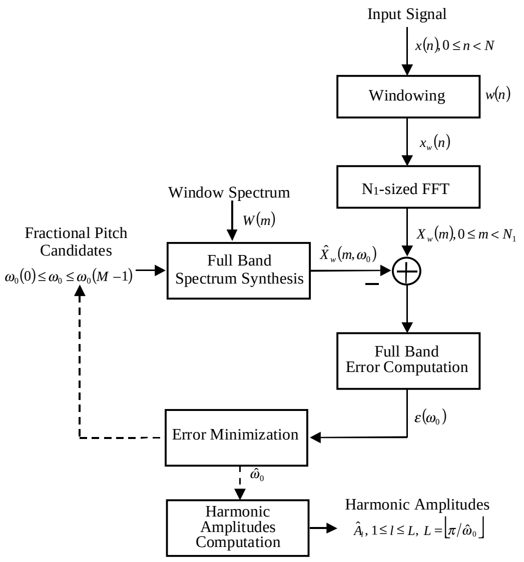


Figure 2. Block diagram of the conventional FP.

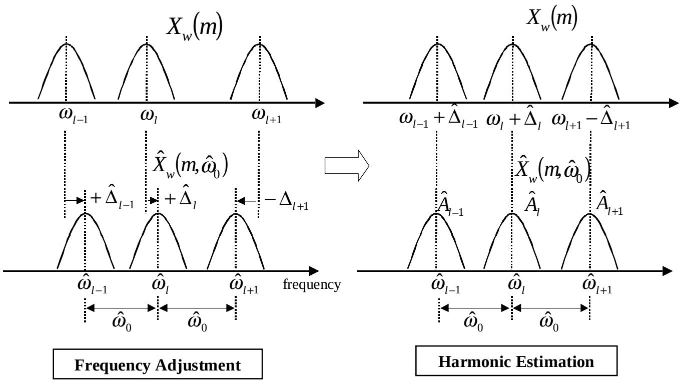


Figure 3. Basic concept of the presented DA.

In the DA, original harmonic frequencies with some variations are shifted to fixed synthetic harmonic frequencies at each harmonic band. The l-th harmonic amplitude A_l is independently obtained by minimizing $\varepsilon_l(\Delta_l)$ between the windowed input spectrum $X_w(m)$ and synthetic spectrum $\hat{X}_w(m, \hat{\omega}_0)$ at the l-th harmonic band in the DFT domain as follows

$$\varepsilon_l(\Delta_l) = \sum_{m=\lfloor a_l + 0.5 \rfloor}^{\lfloor b_l + 0.5 \rfloor} \left[\left| X_w(m + \Delta_l) \right| - \left| \hat{X}_w(m, \hat{\omega}_0) \right| \right]^2, \tag{5}$$

where $\hat{\omega}_0$ is determined using an integer pitch estimate that is obtained by means of the time-domain autocorrelation pitch search algorithm, $\lfloor \cdot \rfloor$ denotes the smallest integer less than or equal to the argument and

$$\hat{X}_w(m, \hat{\omega}_0) = A_l \left| W\left(\left\lfloor \frac{N_2}{N_1} m - \frac{N_2}{2\pi} \hat{\omega}_0 l + 0.5 \right\rfloor \right) \right|. \tag{6}$$

In (5), N_2 is DFT size of the window spectrum, the frequency variation is $|\Delta_l| \le d_l$, and as the frequency increases, d_l increases to reflect characteristics of frequency errors [1] as

$$d_l = \left\lfloor \frac{\alpha_2 - \alpha_1}{L-1}\hat{\omega}_0 (l-1) + \alpha_2 \hat{\omega}_0 \right\rfloor, \tag{7}$$

where α_1 and α_2 are constants which is determined through experiments.

After selecting $\hat{\Delta}_l$, finally the *l*-th harmonic amplitude $\hat{A}_l$ is calculated by the following equation used in IMBE [5] as

$$\hat{A}_l = \left[\frac{\sum_{m=\lfloor a_l+0.5 \rfloor}^{\lfloor b_l+0.5 \rfloor} \left| X_w\left(m+\hat{\Delta}_l\right) \right|^2}{\sum_{m=\lfloor a_l+0.5 \rfloor}^{\lfloor b_l+0.5 \rfloor_l} \left| W\left(\left\lfloor \frac{N_2}{N_1} m - \frac{N_2}{2\pi}\hat{\omega}_0 l + 0.5 \right\rfloor \right) \right|^2} \right]^{0.5}. \tag{8}$$

With (3), leakage of the harmonic energy is observed in the week voiced or unvoiced components, but (8) reduces this leakage of harmonic energy. Figure 4 shows a block diagram of the DA harmonic estimation in detail. In the technique, the error minimization process shown in Figure 4 is performed independently at each harmonic band, while the FP performs the process at the full frequency band.

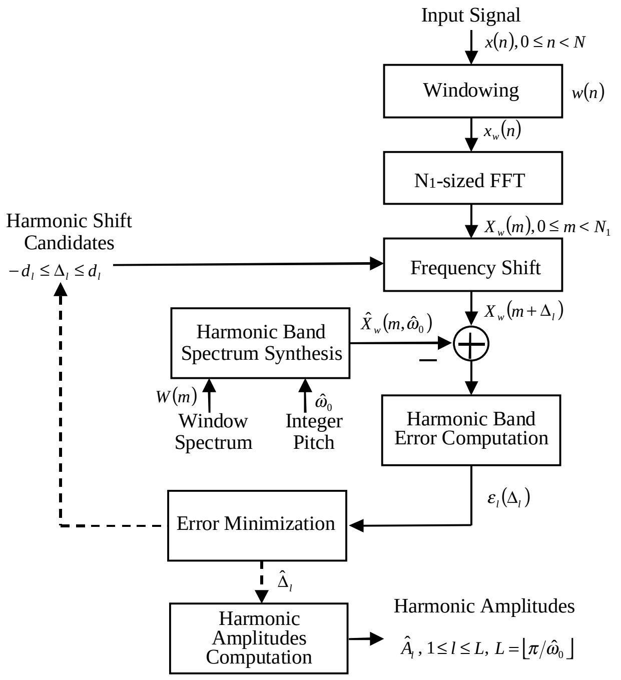

Input Signal
$x(n), 0 \le n < N$
Windowing
$w(n)$
$x_w(n)$
N1-sized FFT
$X_w(m), 0 \le m < N_1$
Harmonic Shift Candidates
$-d_l \le \Delta_l \le d_l$
Frequency Shift
$X_w(m+\Delta_l)$
Harmonic Band Spectrum Synthesis
$\hat{X}_w(m, \hat{\omega}_0)$
−
$W(m)$
Window Spectrum
$\hat{\omega}_0$
Integer Pitch
Harmonic Band Error Computation
$\varepsilon_l(\Delta_l)$
Error Minimization
$\hat{\Delta}_l$
Harmonic Amplitudes Computation
Harmonic Amplitudes
$\hat{A}_l, 1 \le l \le L, L = \lfloor \pi / \hat{\omega}_0 \rfloor$

Figure 4. Block diagram of the DA method.

In addition, it has been reported in [9] that pitch contour is more important than pitch resolution in speech quality. This means that an integer pitch faithful to the pitch contour is sufficient for producing good quality in the sinusoidal decoder. To utilize this fact, the DA uses only an integer pitch estimate. Figure 5 shows an example of spectral harmonic estimations by three methods when the input signal is the linear predictive residual signal by a female speaker: the FP, the DA and the case that an integer pitch is used without frequency adjustment. In Figure 5, we can see that the DA matches the FP result while a harmonic estimation using an integer pitch without frequency adjustment does not.

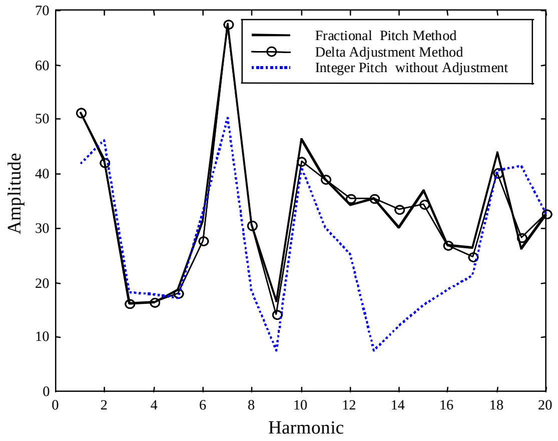


Figure 5. Example of harmonic estimations by the three methods.

4. Complexity and performance

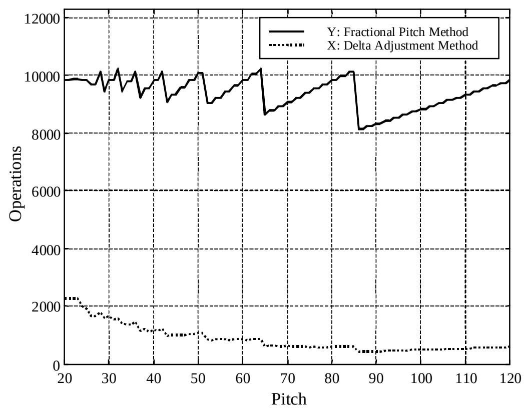
Complexities of the FP and DA in terms of approximated operations are summarized in Table 1. In this case, add, multiply and multiply-accumulation (MAC) are assumed to be one cycle operation.

Table 1: Approximated operations of the FP and DA.

Method	Approximated Operations
FP	$Y = M\left(5\left\lfloor \frac{N_1}{P_0} \right\rfloor L + 3\frac{N_1}{2}\right)$
DA	$X = \sum_{l=1}^{L} 5\left\lfloor \frac{N_1}{P_0} \right\rfloor (2\Delta_l + 1)$

In Table 1, note that *M*: the number of fractional pitch candidates, N_1 : FFT size, P_0 : pitch in sample, Δ_l : adjustment parameter, α_1 and α_2 : constants related to Δ_l.

Typically at 8 kHz sampling rate, *M*=10 [4], *N*$_1$=256 [4], and $20 \le P_0 \le 120$. In addition, α_1 and α_2 are set to 0.3 and 0, respectively through experiments. In this situation, as shown in Figure 6, complexity ratio *Y/X* ranges on the average 13.01 from 4.32 to 18.93. From Figure 6, we can see that the DA has considerable computational gains.



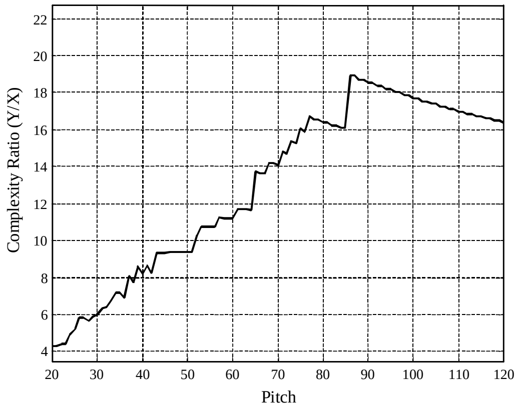


Figure 6. Complexities of the FP and DA with regard to pitch.

In performance evaluation, Spectral Distance (*SD*) and segmental Signal-to-Noise Ratio (*segSNR*) are used. *SD* is defined as

$$SD = \frac{1}{K}\sum_{k=1}^{K} SD_k \quad (dB), \tag{9}$$

$$SD_k = \left[\frac{1}{L_k}\sum_{l=1}^{L_k}\left[20\log_{10}\left(A_f(l)\right) - 20\log_{10}\left(A_d(l)\right)\right]^2\right]^{0.5}, \tag{10}$$

where $A_f(l)$ and $A_d(l)$ are the *l*-th harmonic amplitude estimates by the FP and DA methods, respectively, L_k is the number of harmonics in the *k*-th frame and *K* is the number of total speech frames. Also, *segSNR* is defined as

$$segSNR = \frac{1}{K}\sum_{k=1}^{K} segSNR_k \quad (dB), \tag{11}$$

$$segSNR_k = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1}\hat{s}_f^2(n)}{\sum_{n=0}^{N-1}\left(\hat{s}_f(n) - \hat{s}_d(n)\right)^2}\right), \tag{12}$$

where $\hat{s}_f(n)$ and $\hat{s}_d(n)$ are synthesized speech signals by the sinusodial synthesis [2] using $A_f(l)$ and $A_d(l)$, respectively, and *N* is the frame length.

In evaluating *SD* and *segSNR*, only voiced speech frames are taken into consideration. Results of these measures are summarized in Table 2 and show that the DA method has more efficient for female than male speeches.

Table 2: Results of the objective *SD* and *segSNR* tests.

	Female	Male	Average
SD (dB)	0.39	0.75	0.57
segSNR (dB)	35.55	29.53	32.54

In addition, the subjective prefernece test was performed informally and any degradation was not percievable as shown in Table 3. Through the results, we can state that the DA is comparable to the FP with low complexity.

Table 3: Results of the subjective preference test.

	FP	DA	No Preference
Preference (%)	15	13	72

5. Conclusions

We have presented the fast harmonic estimation, so called Delta Adjustment (DA). The method is based on modification of the GDE model. We have introduced the voiced component extraction technique of the GDE method and modified it to be suitable for low bit rate harmonic coding that uses only an integer pitch estimate. Unlike the GDE, the DA matches a frequency-warped version of the original spectrum that conforms to a fixed pitch at all harmonic bands. Then, we have given a description of complexity and performance of the presented method in comparison with those of the conventional FP one. From the experimental results, the DA algorithm showed good performance with very low complexity.

6. References

[1] C. D. Yoo, and J. S. Lim, "Speech Enhancement Based on The Generalized Dual Excitation Model with Adaptive Analysis Window," *IEEE Proc. Int. Conf. Acoust. Speech and Signal Proc.*, pp. 832-835, 1995.

[2] R. J. McAulay, and T. F. Quatieri, "Sinusoidal Coding," *Speech Coding and Synthesis by W. B. Kleijn and K. K. Paliwal,* Elsevier, Chapter 4, pp. 121-173, 1995.

[3] D. F. Griffin, and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. ASSP*, Vol. 36, No. 8, pp. 1223-1235, 1988.

[4] W. B. Kleijn, "Waveform Interpolation for Coding and Synthesis," *Speech Coding and Synthesis by W. B. Kleijn and K. K. Paliwal,* Elsevier, Chapter 5, pp. 175-207, 1995.

[5] *APCO Project 25 Vocoder Description Version 1.3,* Digital Voice Systems, Inc., 1993.

[6] *SO/IEC FCD 0.1 Subpart 2*, "Information Technology-Very Low Bit Rate Audio-Visual Coding," 1998.

[7] J. Hardwick, C. D. Yoo, and J. S. Lim, "Speech Enhancement Using The Dual Excitation Model," *IEEE Proc. Int. Conf. Acoust. Speech and Signal Proc.*, pp. 137-145, 1993.

[8] W. B. Kleijn, P. Kroon, and D. Nahumi, "The RCELP Speech-Coding Algorithm," *European Transactions on Telecommunications*, Vol. 5, No. 5, pp. 573-582, 1994.

[9] K. S. Lee and R. V. Cox, "TTS Based Very Low Bit Rate Speech Coder," *IEEE Proc. Int. Conf. Acoust. Speech and Signal Proc.*, pp. 181-184, 1999.