THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.27, No.1E 2008. 3. pp. 11-16

Time-Domain Quantization and Interpolation of Pitch Cycle Waveform

Moo Young Kim*
*Dept. information and communications Eng., Sejong University
(Received November 14; Received January 21 2008; Accepted February 25, 2008)

Abstract

In this paper, a pitch cycle waveform (PCW) is extracted, quantized, and interpolated in a time domain to synthesize high-quality speech at low bit rates. The pre-alignment technique is proposed for the accurate and efficient PCW extraction, which predicts the current PCW position from the previous PCW position assuming that pitch periods evolve slowly. Since the pitch periods are different frame by frame, the original PCW is converted into the fixed-dimension PCW using the dimension-conversion method, and subsequently quantized by code-excited linear predictive (CELP) coding. The excitation signal for the linear predictive coding (LPC) synthesis filter is generated using the time-domain interpolation and interlink of the quantized PCW's. The coder operates at 4.2 kbit/s and 3.2 kbit/s depending on the pitch period. Informal listening test demonstrates the effectiveness of the proposed coding scheme.

Keywords: *Speech Coding, Quantization, Harmonic Coder, Waveform Interpolation*

I. Introduction

Code-excited linear predictive (CELP) coding can produce synthesized speech of toll quality around at 8 kbit/s such as ITU-T G.729 [1]. It can be also used for the text-to-speech (TTS) system to compress its speech database. However, the performance of CELP coding is seriously degraded at lower bit-rates as it requires lots of bits to represent the frame-based excitation signal for the linear predictive coding (LPC) filter.

On the other hand, parametric vocoders have been widely used to design low bit-rate speech coders and TTS systems: mixed-excitation linear predictive (MELP) coding, sinusoidal transform coding (STC), multi-band excitation (MBE) coding, waveform interpolation (WI), and harmonic plus noise model (HNM) [2-4]. Vocoder-based TTS system can be effectively extended to the speech modification system, as well as can compress its speech database at low bit-rates.

In this paper, we propose a parametric vocoder whose quantization is performed with CELP coding in time domain. A novel pitch-cycle waveform (PCW) extraction method is used based on the evolving characteristics of PCW. Time-domain interpolation and interlink schemes are utilized, which is advantageous over frequency-domain methods in terms of search time complexity. Finally, time-domain quantization of PCW is performed by using the dimension conversion and the CELP -based technique. The CELP coding was performed for each PCW instead of for each frame. The proposed time-domain approach can produce high -quality of speech over frequency-domain vocoders

Corresponding author: Moo Young Kim (mooyoung@sejong.ac.kr)
Dept. Information and Communications Eng.
Sejong University, Gunja-dong, Gwangjin-gu, Seoul 143-747, Korea

since the proposed method can represent residual waveform efficiently at low bit rate.

This paper is organized as follows. In section II, we describe the outline of the proposed coder. In section III, we propose the novel PCW extraction and alignment methods. Time-domain interpolation and interlink methods of PCW's are discussed in section IV. Dimension conversion and quantization of PCW are presented in section V. Experimental results and conclusions are reported in section VI and VII, respectively.

II. Outline of the Proposed Coding Method

The overall structure of the proposed coding scheme is presented in Fig. 1. Upper and lower parts of Fig. 1 describe the encoder and the decoder, respectively. For the efficient quantization, LPC coefficients are transformed into line spectrum pairs (LSP) and quantized using linked-split vector quantization [5]. Open-loop pitch of each frame is estimated using spectro-temporal autocorrelation method [3]. Then, extracted PCW is aligned and converted to the fixed dimension PCW for quantization. At the decoder, LSP, pitch, and PCW parameters are received. PCW around frame boundary is decoded and converted back into the original pitch length to reproduce excitation signal. PCW's of previous and current frame boundaries are interpolated and interlinked in time-domain to reconstruct the remaining excitation signal. The excitation signal is used as the input of LPC synthesis filter to obtain the synthesized speech.

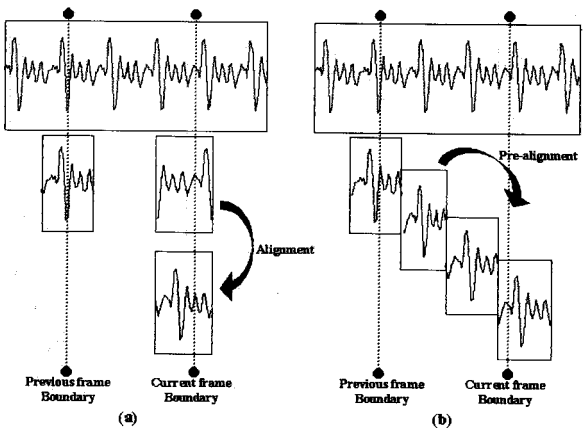

Figure 1. Block diagram of the proposed method.

III. PCW Extraction and Alignment

The extraction and alignment of a PCW are performed in residual domain utilizing estimated open-loop pitch. In [2], a PCW is extracted at the end of the current frame to minimize the pitch-boundary energy, and then aligned with a PCW at the end of the previous frame, as shown in Fig. 2 (a). This method may arise two problems.

· The alignment procedure should be performed in the range of pitch period to find the maximum correlation between previous and current frames.

· The alignment result may not be adequate in case that the pitch pulse shape of a PCW is not impulse-like, especially in nasal or highly-pitched sounds.

Thus, we propose a pre-alignment method based on the evolving characteristics of PCW's. The position of a current PCW can be estimated from the position of a previous PCW because we assume that pitch period evolves slowly as shown in Fig. 2 (b). After that, conventional alignment can be performed

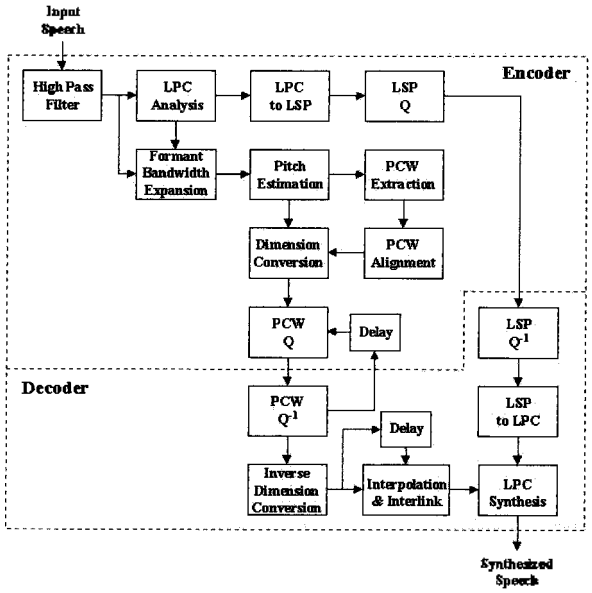

Figure 2. Pitch cycle waveform extraction and alignment methods: (a) frame-based and (b) pitch-based methods.

for fine tuning. We found that the proposed pre-alignment scheme has following advantages over conventional methods:

- search range of the alignment procedure can be considerably reduced,
- and alignment error is alleviated even though the pitch pulse shape is not impulse-like.

In order to find more accurate PCW position, input signal is passed through the formant bandwidth expansion filter of the form,

$$w(z)=\frac{A(z)}{A(z/\gamma)},\ 0\le\gamma\le 1 \tag{1}$$

where $A(z)$ is a LPC filter. The filter output is the interpolated signal between speech and residual domain in accordance with the weighting factor γ. As γ increases to one, speech domain signal is emphasized compared with residual domain signal, and vice versa as γ decreases to zero. The value of γ is adjusted on the idea that speech-domain signal is more useful for highly-pitched signal, whereas residual-domain signal is preferred for lower-pitched signal. Hence, based on the experiments to reduce the pitch error, we set γ=0.8.

IV. Time-Domain Interpolation and Interlink Methods

In WI coder [2], each PCW is represented as Fourier series:

$$u(t,\phi)=\sum_{k=1}^{K}\left[C_k(t)\cos(k\phi)+D_k(t)\sin(k\phi)\right] \tag{2}$$

where K is the number of harmonics, and $C_k(t)$ and $D_k(t)$ are the Fourier series coefficients. If the previous and current frames are denoted as $t=T_{m-1}$ and $t=T_m$, respectively, the residual signal between $T_{m-1}\le t\le T_m$ can be interpolated by:

$$\begin{aligned}e(t)&=\sum_{k-1}^{K}[\{(1-\alpha(t))C_k(T_{m-1})+\alpha(t)C_k(T_m)\}\cos\phi_k(t)\\&\quad+\{(1-\alpha(t))D_k(T_{m-1})+\alpha(t)D_k(T_m)\}\sin\phi_k(t)]\end{aligned} \tag{3}$$

where $\alpha(t)$ and $\phi_k(t)$ are the linear interpolation coefficient and phase function, respectively. When the pitch period is not changed from previous to current frames, the interpolation of Fourier coefficients yields good performance as the number of harmonics is fixed. However, in general, pitch period is varied frame by frame, and direct interpolation of Fourier coefficients may generate undesired frequency components as shown in Fig. 3 (a). To alleviate this problem, zero padding method in frequency-domain [2] can be applied as depicted in Fig. 3 (b). This frequency-domain normalization may make important peak pattern in time-domain be changed. Thus, the proposed interpolation and interlink methods are performed in time domain by utilizing the characteristics of the PCW whose signal energy around the both ends is relatively small. Thus, as shown in Fig. 3 (c), we proposed to use the time-domain zero padding at the PCW boundary. The zero-padded PCW is used only for the interpolation purpose, thus

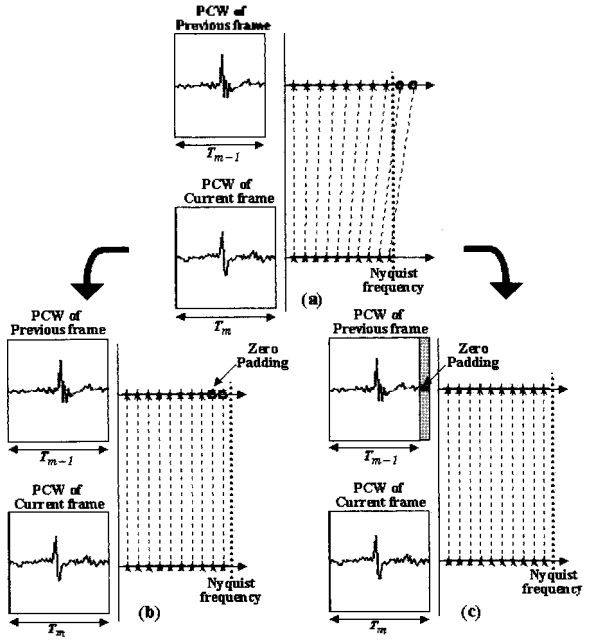

Figure 3. Comparison of PCW interpolation methods: (a) sampling frequency problem, (b) the conventional method, and (c) the proposed method.

the original-size PCW is used at the frame boundaries.

To reduce computational complexity, we also propose the pitch-period dependent interpolation and interlink methods in time domain. The number of required PCW's for interpolation can be calculated based on the frame size and the estimated pitch -period. The interpolation between the previous and current PCW's can be performed in time-domain as PCW's have the same length by using time-domain zero padding method.

To generated the concatenated excitation signal for LP filter, interpolated PCW's are interlinked in time-domain. For example, if the number of interpolated PCW's is 2 as shown in Fig. 4, then the interpolated PCW's, PCW_1 and PCW_2, can be estimated by

$$\begin{aligned} PCW_0 &= u(T_{m-1},\phi) \\ PCW_1 &= \frac{2}{3}u(T_{m-1},\phi)+\frac{1}{3}u(T_m,\phi) \\ PCW_2 &= \frac{1}{3}u(T_{m-1},\phi)+\frac{2}{3}u(T_m,\phi) \\ PCW_3 &= u(T_m,\phi) \end{aligned} \tag{4}$$

where PCW_0 and PCW_3 are the extracted PCW's at the frame boundaries. The proposed interpolation and interlink methods save computational complexity because all the procedures can be performed in time-domain.

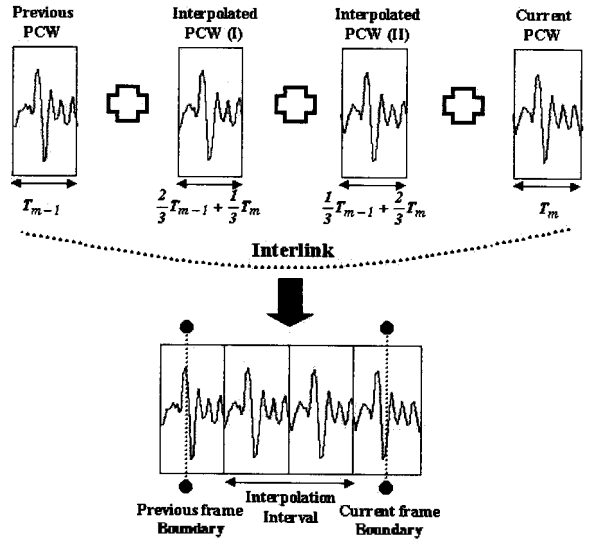

Figure 4. Example of interpolation and interlink methods.

V. Dimension Conversion and Quantization of PCW

The aligned PCW is quantized using the CELP coding with the proposed dimension-conversion method.

For the CELP-type codebook search, each PCW should have the same fixed-dimensionality, N, for every frame. However, pitch period of a PCW, T, is changed frame by frame. Thus, a variable length PCW is converted to a fixed-length PCW with a proposed dimension-conversion method. To convert from the variable T-dimensionality to the fixed N-dimensionality, T-dimension PCW is transformed to frequency domain using T-point DFT. Then, zero padding into higher frequency bins or redundant frequency removal is performed to convert into the N-point frequency bins. If N is greater than T, zero padding for higher frequency bins is applied as shown in Fig. 5 (a). Otherwise, high frequency removal is performed. Finally, N-point inverse Fourier transform is performed to generate the time-domain excitation signal for the codebook search. The dimensionality of the impulse response of the synthesis filter is also converted by $T:N$.

A current PCW is firstly quantized using a previous PCW in a similar way to adaptive codebook search of CELP [1]. The remaining residual signal is required to be quantized with a fixed codebook.

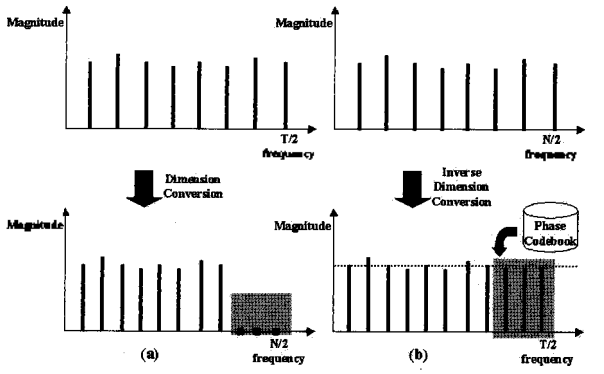

Figure 5. Fixed-dimension conversion for time-domain quantization: (a) dimension conversion and (b) inverse dimension conversion.

Table 1. Algebraic codebook pulse positions for N=40 and N=80.

Track #	Pulse Positions
0	0,5,10,15,20,25,30,35, 40,45,50,55,60,65,70,75
1	1,6,11,16,21,26,31,36, 41,46,51,56,61,66,71,76
2	2,7,12,17,22,27,32,37, 42,47,52,57,62,67,72,77
3	3,8,13,18,23,28,33,38, 43,48,53,58,63,68,73,78
4	4,9,14,19,24,29,34,39, 44,49,54,59,64,69,74,79

Track #	Pulse Positions
0	0,5,10,15,20,25,30,35
1	1,6,11,16,21,26,31,36
2	2,7,12,17,22,27,32,37
3	3,8,13,18,23,28,33,38, 4,9,14,19,24,29,34,39

For the fixed codebook search, algebraic codebook is used. If T<50, N is set to 40. Otherwise, N is set to 80. The algebraic codebook has two types according to N as shown in Table 1.

After the codebook search, the N-point PCW should be converted back into the T-point PCW. We call this process the inverse dimension-conversion. The best candidate codevector, which is a time-domain waveform, is transformed to frequency domain. Since N may be different from T, we need to remove or regenerate frequency components. If N is greater than T, high frequency bins are removed to generate T-point Fourier coefficients. Otherwise, a simple bandwidth extension techniques is applied to regenerate high-frequency Fourier coefficients. The average magnitude of $N/2$-point Fourier coefficients is used for these high-frequency bins as shown in Fig. 5 (b) under the assumption that LP residual has flat spectral magnitude. Phase information for these regenerated frequency bins is modeled with a stored male speech signal. Final excitation signal can be obtained using inverse Fourier transform.

VI. Experimental Results

The proposed method is implemented with 20 ms frame size and 15 ms lookahead.

Table 2. Bit allocation for the proposed method.

Parameters		double subframes		single subframe
LSP		24		24
Adaptive Codebook	Index	3	7	7
	Gain	3	3	3
Fixed Codebook	Index	17	17	25
	Gain	5	5	5
Total		84		64

Table 3. Preference test results.

Condition	Conven.	Same	Proposed
Clean	25.00%	39.58%	35.42%
Car	18.75%	33.33%	47.92%
Total	21.88%	36.46%	41.67%

The bit allocation is summarized in Table 2. A 10-th order LPC is transformed to LSP and quantized with linked-split vector quantization with 24 bits [5]. The pitch lag at the frame boundary is encoded with 7 bits.

If the lag is less than 50 samples, a frame is divided into two subframes (double-subframe structure). Otherwise, a single frame is not divided into subframes (single-subframe structure). In the double-subframe structure, the differential pitch lag for the first subframe is also encoded with 3 bit. For each subframe, algebraic codebook index is encoded with 13 position bits and 4 sign bits. Gains for adaptive and fixed codebooks are encoded with 3 and 5 bits, respectively. For a single-subframe structure, a major difference in the bit allocation comes from the algebraic codebook structure as shown in Table 1. In this case, five pulses are selected and each pulse is encoded with 4 bits for position and 1 bit for sign. Total bit rates for the proposed coder are 4.2 kbit/s (for pitch lag < 50) and 3.2 kbit/s (for pitch lag ≥ 50), respectively.

At the decoder, the PCW at each frame boundary is reconstructed using the algebraic and fixed codebooks, and converted back into the original pitch length using the inverse dimension-conversion. PCW's of previous and current frame boundaries are interpolated and interlinked in time-domain to reproduce the excitation signal. The excitation signal is used as an input for the LPC synthesis filter to obtain the synthesized speech.

Since the interpolation and interlink schemes are utilized in time domain, the proposed method is advantageous over frequency-domain methods [2-4] in terms of complexity. Time-domain quantization of PCW is also more efficient than conventional CELP coding in terms of codebook search complexity because the alignment information between the previous and the current PCW's can be utilized instead of finding the closed-loop delay.

We compare the search time complexity between our method and the conventional harmonic coding method [4]. Most of coding blocks, including LPC and pitch estimation, are shared between two coders for fair comparison. With 3 minutes Korean speech data, the conventional and the proposed methods require 33.8 and 28.3 seconds to process both encoder and decoder, respectively, using the pentium 2.0 GHz machine. Thus, we can obtain 16.3% improvement in complexity.

As shown in Table 3, subjective speech quality of the proposed coding scheme (at 3.2 or 4.2 kbit/s depending on the pitch period) is compared with that of the conventional harmonic coding method (at 4.0 kbit/s) [4]. For the preference listening test, 24 Korean speech sentences (12 female and 12 male speakers were used) processed by two coders were played in a random order to 12 listeners. The listeners decide which coder is preferable. The listening test results show that the proposed method produces higher speech quality than the conventional method not only for clean speech, but also for noisy speech (15 dB car noise). The proposed time-domain approach can be preferred over the frequency-domain vocoder, since the proposed coder can represent not only the residual magnitude but also the phase information (shape of the waveform in time domain) more efficiently at low bit rate. Hence, the energy localization in time domain can be more effectively performed with the propose method.

VII. Conclusions

In this paper, the methods for the extraction, quantization, and interpolation of a pitch cycle waveform (PCW) are presented. For the extraction of a PCW, the pre-alignment technique is proposed. To quantize a variable-dimension PCW in time- domain, the dimension-conversion scheme with CELP coding is applied. At the decoder, the received PCW's are interpolated and interlinked to synthesize the speech signal. The proposed coder obtains higher preference than the conventional vocoder not only for speech quality, but also for search time complexity.

Acknowledgement

This work was supported by the faculty research fund of Sejong University in 2007.

References

1. R. Salami, C. Laflamme, J. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," IEEE Trans. Speech Audio Processing, 6(2), 116-130, Mar. 1998.
2. W. B. Kleijn and J. Haagen, *Speech Coding and Synthesis, Amsterdam*, (The Netherlands: Elsevier, 1995).
3. T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practices, Upper Saddle River, NJ: Prentice Hall, 2002.
4. Y. D. Cho, M. Y. Kim, and S. R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameters determination," in Proc. IEEE ICASSP, 601-604, Seattle, WA, USA, 1998.
5. M. Y. Kim, N. K. Ha, and S. R. Kim, "Linked Split-Vector Quantizer of LPC Parameters," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Atlanta, GA, 741-744, May 1996.

[Profile]

• **Moo Young Kim**

1989.2-1993.3: B.Sc. in EE, Yonsei Univ., Korea
1995.2: M.Sc. in EE, Yonsei Univ., Korea
2000.12: Member of Research Staff, Samsung Advanced Institute of Technology, Korea
2004.11: Ph.D. in EE, KTH, Sweden
2006.8: Senior Research Engineer, Ericsson Research, Sweden
Present: Assistant Professor, Dept. Info. Comm., Sejong Univ., Korea
* Interested Area: Multimedia Signal Processing and Coding, Information Theory, Pattern Recognition