A LOW RESOLUTION PULSE POSITION CODING METHOD FOR IMPROVED EXCITATION MODELING OF SPEECH TRANSITION

Jongseo Sohn and Wonyong Sung

School of Electrical Engineering,
Seoul National University, Seoul, Korea.
e-mail: sohn, wysung@dsp.snu.ac.kr.

ABSTRACT

We propose a new excitation model for transitional speech to reduce the distortion due to the traditional two-excitation source, voiced and unvoiced, model. The proposed low resolution pulse position coding (LRPPC) algorithm detects the existence of pulses at frames of weak periodicity, which are determined as unvoiced, and transmits the approximate pulse positions. In the decoder, dispersed pulses that have a flat magnitude spectrum are synthesized at the decoded positions to form the excitation signal. A subjective quality test shows that the vocoder employing the LRPPC algorithm produces better quality of speech, and is very robust to mode decision errors.

1. INTRODUCTION

Traditional low bit rate vocoders assume that the excitation signal for the vocal tract filter consists of two, voiced and unvoiced, sources, and synthesize the excitation using a mixture of them. The existence of periodicity is the main feature that characterizes the two different sources, which validates the harmonic models employed in low bit-rate coders such as the sinusoidal transform coder (STC) and the waveform interpolation (WI) coder [1]. The noise-like characteristics of unvoiced speech enables many efficient coding methods that describe only the second order statistics [2].

Although this type of vocoder can compress speech with high intelligibility at a bit rate as low as 2.4 kbps, it fails to produce a high quality of speech even before the quantization of the relevant parameters [1]. The limited performance is mainly due to its simple excitation source model, especially for the transition segments such as onsets and irregular glottal pulses in voiced speech [3][4]. In [3] and [4], the analysis-by-synthesis based waveform coders are used for coding the transition segments at 4 kbps, but it can not be applied to lower bit-rate coding.

The limitation of the two source model in the coding of transition segments can be explained by observing the behavior of a simple binary voiced/unvoiced (V/UV) mode vocoder. In steady-state voiced speech, misclassification of a voiced frame as unvoiced one causes two kinds of distortion. First, the decoder fails to generate the pitch pulses in the misclassified frame and the overall periodicity is destroyed, which results in an artifact in the synthesized speech. The worse distortion is caused by the large amount of noise that corresponds to the total energy of the pitch pules. Transition segments usually show weak periodicity, and they are often classified as unvoiced. In this case, the noise generated by the decoder is the major source of distortion because there is little periodicity to be destroyed. On the contrary, when it is classified as a voiced frame and the excitation is synthesized as usual periodic voiced sounds, we found that there is no or only slight degradation of perceived quality as long as the estimated pitch is not extremely high or low.

A research to model the erratic glottal pulses by introducing a new excitation source has been reported in the development process of the mixed excitation linear predictive (MELP) vocoder [5]. In the MELP coder, the frame that shows a low correlation but a high peakiness value in the residual signal is set to the zittery voiced mode, where the aperiodic pulses are synthesized with a maximum of 25 % random offsets from their periodic positions. The exictation signal synthesized by the aperiodic pulses produces a natural quality of speech in transition segment, although it differs from the original one in the actual pulse positions and the number of pulses. This implies that whether the excitation is pulse-like or noise-like is perceptually important.

However, the analysis/synthesis method of the MELP coder for the transition segment is not directly related to the actual pulses in the origianl residual signal. Although strong pulses yield a high peakiness value, a high peakiness value does not necessarily indicate the existence of pulses. In addition, the synthesis method based on the estimated pitch is not robust, because the pitch is not defined and the estimated pitch is usually random in transition segments.

In this paper, we study an analysis method that directly determines the pulse existence and positions at frames of weak periodicity. Moreover, we propose an efficient coding scheme that quantizes the approximate pulse positions in a frame, and describes a method to synthesize the excitation signal using the decoded information.

2. LOW RESOLUTION PULSE POSITION CODING

The lack of periodicity in transitional speech may be due to the irregularity of the glottal pulse positions and shapes, or due to the analysis frame size that is too long to discriminate a few glottal pulses at the beginning or end of the frame from the remaining non-periodic signal. In either case, we can observe that there are usually a few dominant pulses regardless of the overall pitch contour. Moreover, the exact pulse position is not perceptually important unlike in the steady-state voiced case where a small offset in the pulse position destroys the periodicity of the overall signal. Therefore, we detect the existence of pulses at frames of weak pe-

riodicity, and transmit their approximate positions. In our implementation, the pulse detection and position analysis is performed for each 10 ms subframe, twice in a 20 ms frame, for the speech sampled at 8 kHz. We assume that there can be only two or smaller number of pulses in a subframe. In the following subsections, the analysis, quantization, and synthesis procedures are described.

2.1. Pulse Detection and Position Analysis

In voiced speech, the linear prediction (LP) residual signal has a pulse-like structure due to the slope discontinuity of the glottal pulse. A simple method to determine the existence and position of pulses is to inspect the instant where the short-term energy of LP residual is concentrated. Typical LP residual shows a dispersed waveform such as bipolar swings near the excitation instant due to the effect of the formant phase angle, and has a trend to tilt before the next epoch in the opposite direction of the next pulse [6]. This makes the peak of the short-term energy envelope less sharp and hinders the detection of pulse existence. This problem can be alleviated by using the HEWLPR (Hilbert envelope of windowed LP residual) proposed in [6]. To avoid the transformation to frequency domain, we use the half sine wave window instead of the hamming window in [6], as follows:

$$W(e^{j\omega}) = \sqrt{2}\sin(\omega). \tag{1}$$

The windowing is implemented in time domain using a linear phase band-pass filter, $H(z) = (1 - z^{-2})/\sqrt{2}$, which reduces the tilting trend of LP residual. The HEWLPR, $e_H(n) = \sqrt{r(n)^2 + r_H(n)^2}$, is obtained using the band-pass filtered LP residual $r(n)$ and its Hilbert transformed signal $r_H(n)$.

To examine the concentration of energy in a subframe, we first calculate a short-term energy contour given by

$$E_p(n) = (1/5) \sum_{k=-2}^{2} e_H^2(n-k). \tag{2}$$

Then two candidate pulse positions, n_1 and n_2, are selected as follows:

$$n_1 = \underset{n \in S_1}{\text{argmax}}\{E_p(n)\}, \quad \text{where} \quad S_1 = [0, N-1] \tag{3}$$

$$n_2 = \underset{n \in S_2}{\text{argmax}}\{E_p(n)\}, \quad \text{where} \quad S_2 = [0, N-1] \cap [n_1 - \delta_2, n_1 + \delta_2]^c \tag{4}$$

where $[a, b]$ represents the set of integers from a to b, A^c represents the complementary set of A, $N = 80$ is the subframe size, and δ_2 is introduced so that n_1 and n_2 are separated at least δ_2 samples apart. To identify two pulses, we check the following conditions:

$$E_p(n_1) > T_2 \cdot E_s \tag{5}$$

$$E_p(n_2) > T_2 \cdot E_s \tag{6}$$

where $E_s = \sum_{n=0}^{N-1} e_H^2(n)$ is the subframe energy, and T_2 is a value between zero and unity. Only one pulse is identified if the above test fails and if

$$E_p(n_1) > T_1 \cdot E_s \tag{7}$$

otherwise, no pulse is found. We found that the values of $T_2 = 0.3$ and $T_1 = 0.6$ are appropriate.

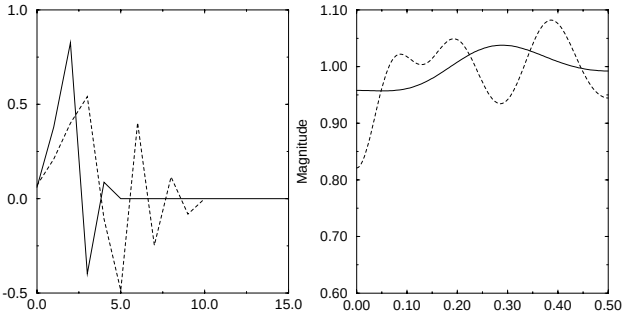

Figure 1: Characteristics of the prototype pulses. Solid line is for the 5-tap pulse and dotted for the 10-tap pulse. (a) Impulse response. (b) Frequency response.

2.2. Pulse Position Coding and Synthesis

The positions of the identified pulses are coarsely quantized. When two pulses are identified, the subframe of 80 samples is divided into 13 disjoint subsets. Each subset except the first and the last one consists of consecutive $\delta_2 = 6$ samples. Two indices of the subsets to which the pulses belong are coded instead of the exact pulse positions. When there is only one pulse, the frame is divided into 11 subsets of $\delta_1 = 7$ samples (except the first and the last one). Then the total number of cases is

$$\binom{13}{2} + \binom{11}{1} + 1 = 90, \tag{8}$$

where the last 1 is for the case of no pulse. Since the maximum entropy of the 90 cases is 6.49 bits per subframe, 13 bits are allocated for each 20 ms frame. Regardless of the number of pulses, the overall subframe gain is transmitted once for each subframe.

In the decoder part, the number of pulses and the subset indices are decoded and the fine pulse position in the decoded subset is randomly selected. Since the pulse analysis procedure guarantees that most of the subframe energy is concentrated on the pulses, the pulse gains are determined from the decoded subframe gain so that the energy of the pulses is 90% of the subframe energy. The remaining energy is used to add the noise component.

For the naturalness of the synthesized speech, we used a 5-tap dispersed pulse instead of one-tap impulse. The pulse is obtained by truncating the impulse response of a signal having a flat amplitude spectrum and a phase spectrum which is a quadratic function of frequency [7]. Figure 1 shows the waveform shape and the magnitude spectrum of the prototype pulse.

Although the pulse has a flat magnitude spectrum, two adjacent pulses can yield frequency nulls of large bandwidth, which may cancel the poles of the synthesis filter. In the actual LP residual signal, the dominant pulses are separated far enough, thus two adjacent pulses are in fact one dispersed pulse. Therefore, two pulses in contiguous subsets are merged into a 10-tap dispersed pulse as shown in Fig. 1. This merged pulse is useful for modeling the excitation of plosive sounds.

3. OVERALL CODER DESCRIPTION

A 2.4 kbps vocoder employing the proposed low resolution pulse position coding (LRPPC) algorithm is developed. Basically, it is a binary mode vocoder that employs a harmonic model for steady-state voiced speech and the LRPPC for unvoiced and transitional speech. Note that the two modes are harmonic/non-harmonic (H/NH) rather than the V/UV modes. Though the LRPPC is a model mainly for voiced transitional speech, it is applied to both unvoiced and transitional speech. This increases the robustness against mode decision errors and maintains the number of bits for coding the steady-state voiced speech which has relatively large perceptual information.

The vocoder operates on a 20 ms frame basis and requires a 7.5 ms look-ahead region for spectral and pitch analysis. The pitch is estimated by analyzing the correlations of low pass filtered LP residual and speech signals. The H/NH mode decision is made based on some parameters such as the zero crossing rate, first cepstrum coefficient, and first autocorrelation coefficient [8] as well as the correlation value.

3.1. Analysis/Synthesis of Harmonic Speech

The analysis/synthesis procedure for a harmonic frame is similar to that of the WI coder [1, ch. 5]. The pitch cycle waveforms (PCW's) in a frame are extracted from the LP residual and the magnitude spectrum of the prototype PCW is estimated by averaging the squared magnitudes of the Fourier series coefficients of the estimated PCW's with an appropriate weighting. This estimated magnitude spectrum is quantized and transmitted to the decoder. The PCW's are almost critically sampled according to the pitch period to reduce the computational complexity.

A synthetic phase spectrum similar to that used to make the prototype pulse is generated by the decoder for the reconstruction of the prototype PCW. We used the phase spectrum where a constant phase is added to the quadratic phase function in [7], as follows:

$$\phi_n(k) = \begin{cases} 0 & \text{if } k = 0 \text{ or } L_n/2 \\ -\pi/2 - 3\pi(2k/L_n)^2 & \text{otherwise} \end{cases} \quad (9)$$

where k is the frequency index and L_n is the nth PCW length. The addition of the constant $\pi/2$ corresponds to the Hilbert transformation and models the tilting trend of LP residual pulses.

From the decoded magnitude spectrum and the synthesized phase spectrum of the prototype PCW's, the missing PCW's are interpolated. Direct linear interpoation of the complex spectrum of the PCW may cause distortions in the magnitude spectra of interpolated PCW's, when the phase spectra of the current and previous prototype PCW's are very different [9]. To prevent this distortion, the magnitude and phase spectra are interpolated separately. The magnitude spectra of the interpolated PCW's are scaled according to the voicing strength, and the noise component is injected. The noise component is extracted from a pseudo-noise sequence in the same manner as for the PCW extraction. The excitation signal is generated using these PCW's and the LPC synthesis filtering is applied to produce the output speech.

3.2. Interoperation between the LRPPC and the Harmonic Coding

In the harmonic mode, the reconstructed signal is usually asynchronous with the original one since the employed harmonic coder does not transmit the linear phase information, while the LRPPC operates time-synchronously. This may cause signal discontinuity at the frame boundary when there is a switching between the harmonic and non-harmonic modes. To solve this problem, we adopted a synchronization scheme similar to that proposed in [3]. In our case, the synchronization is easier since the pulse positions are explicitly given by the LRPPC. When switching from a non-harmonic frame to harmonic one, the phase spectra of the PCW's are adjusted so that the first pulse in the harmonic frame is located pitch period apart from the last pulse of the previous non-harmonic frame. In the other cases, the same synchronization methods described in [3] are applied, where the stationary unvoiced segment in [3] corresponds to the non-harmonic frame that has no pulse, and the transition segment in [3] corresponds to the non-harmonic frame having pulses.

3.3. Parameter Quantization

The bit allocation among the model parameters of the developed vocoder with the 20 ms frame length is given in Table 1. The LPC coefficients are transformed to the line spectral frequencies (LSF), which are quantized using a 24 bit split vector quantizer (VQ) [10]. A variable dimensional VQ [11] is used for encoding the magnitude spectrum below 1 kHz of the prototype PCW. To discriminate the voiced fricatives containing a large amount of noise due to the vocal tract constriction from the other voiced sounds, one of two voicing strength patterns is selected according to the normalized correlation. At harmonic frames, the second subframe gain is quantized using 5 bits, and the first subframe gain is predicted from the current and the previous second subframe gains and the prediction residual is quantized using 3 bits.

Table 1: Bit Allocation for the proposed 2.4 kbps vocoder

Parameters	Harmonic	Non-harmonic
Mode bit	1 bit	1 bit
LSF's	24 bit	24 bit
PCW Magnitude	7 bit	
Pitch	7 bit	
Gain	3+5 bit	5+5 bit
Pulse position		13 bit
Voicing strength	1 bit	
Total bits/frame	48 bit	48 bit

4. SUBJECTIVE TEST RESULTS AND CONCLUSIONS

To verify the perceptual effectiveness of the LRPPC scheme, we carefully listened to the speech segments where the coded pulses play a dominant role. During the periods containing irregular glottal pulses, we could perceive little distortion when we listened to the coded speech alone, although sometimes they are much different in timbre from the original. By the comparison of the two vocoder outputs with and without the LRPPC scheme, we could

find that the LRPPC scheme much reduces the noisy characteristics of the output speech. This is because the LRPPC scheme concentrates the frame energy on the pulses during the onsets, as shown in Fig. 2. In addition, the intelligiblity of the synthesized speech is increased, because the proposed excitation source models plosives better.

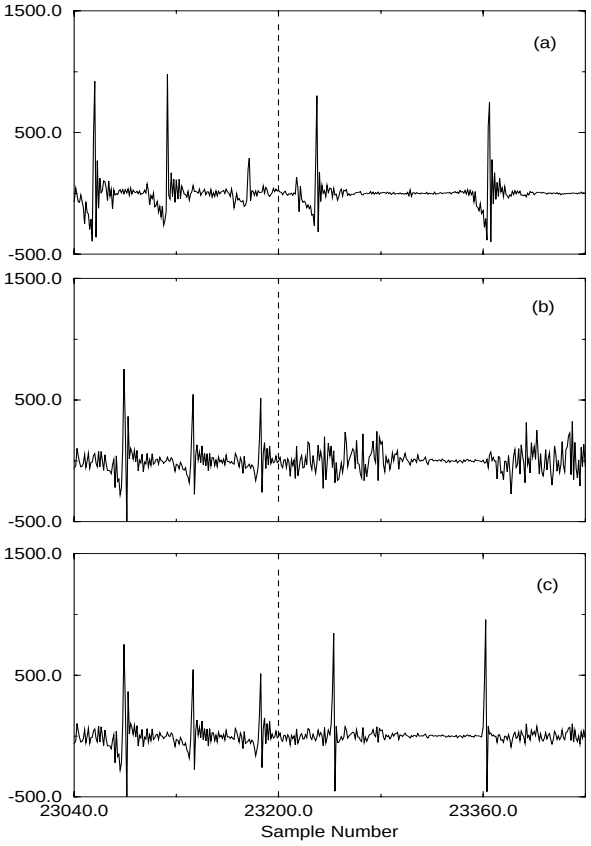

Figure 2: Residual signals of (a) original (b) harmonic+noise model (c) harmonic + LRPPC model. Major ticks of x-axis represent the frame boundaries. The frames at the lefthand side of the vertical dashed line are voiced frames and the frames at the righthand side are unvoiced or non-harmonic frames.

For a quantitative evaluation of the proposed vocoder, we conducted an informal degradation category rating (DCR) test. The MELP coder, the new U.S. Federal Standard at 2.4 kbps, was included in the test for a reference. Forty sentence (20 from female and 20 from male talkers) pairs of clean and processed speech are presented to ten non-expert listeners. The degradation mean opinion scores (DMOS) of the two coders are summarized in Table 2, which shows that the proposed vocoder performs better than the MELP coder, especially for female speech.

Table 2: DMOS of the proposed and the MELP vocoders.

	MELP	Proposed
Female	3.06	3.27
Male	3.68	3.79
Average	3.37	3.53

5. REFERENCES

[1] W. Kleijn and K. Paliwal, Eds., *Speech Coding and Synthesis*, Elsevier Science Publishers, Amsterdam, 1995.

[2] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," in *Proc. IEEE Workshop on Speech Coding for Telecomm.*, 1993, pp. 35–36.

[3] E. Shlomot, V. Cuperman, and A. Gersho, "Combined harmonic and waveform coding of speech at low bit rates," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1998, pp. 585–588.

[4] C. Li and V. Cuperman, "Enhanced harmonic coding of speech with frequency domain transition modeling," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1998, pp. 581–584.

[5] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.

[6] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.

[7] G. S. Kang and S. S. Everett, "Improvement of the excitation source in the narrow-band linear prediction vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 377–386, April 1985.

[8] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 3, pp. 203–212, June 1976.

[9] Y. Shoham, "High-quality speech coding at 2.4 to 4.0 kbps based on time-frequency interpolation," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1993, vol. 2, pp. 167–170.

[10] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[11] A. Das, A. V. Rao, and A. Gersho, "Variable-dimension vector quantization," *IEEE Signal Process. Letters*, vol. 3, no. 7, pp. 200–202, 1996.