A Packet Loss Concealment Algorithm Based on Time-Scale Modification for CELP-type Speech Coders

Moon-Keun Lee, Sung-Kyo Jung, Hong-Goo Kang, Young-Cheol Park and Dae-Hee Youn*

MCSP Lab., Dept. of Electrical & Electronic Eng., Yonsei University, Seoul, Korea
*Division of Information Technology, Yonsei University, Wonju-city, Korea

samalim@mcsp.yonsei.ac.kr

ABSTRACT

In this paper, we propose a packet loss concealment algorithm for a code-excited linear prediction (CELP) speech coder. We perform a time-scale modification (TSM) using a waveform similarity overlap-add (WSOLA) technique to reconstruct the excitation signal of the lost or dropped frames. In addition, when a lost frame is classified as a voiced, an adaptive codebook gain and a fixed codebook gain are estimated by a modified gain parameter re-estimation (GRE) technique. By applying these techniques, we can reduce quality degradation of the decoded speech and error propagation effect through the adaptive codebook memory. We apply the proposed scheme to the ITU-T G.729 standard speech coder to evaluate the performance of the proposed method. The perceptual evaluation of speech quality (PESQ) and AB preference tests under various packet loss conditions verify that the proposed algorithm is superior to the concealment algorithm embedded in the G.729.

1. INTRODUCTION

The quality of real-time voice communications in mobile links or packet-switched transmissions is degraded by frame erasures. For wireless communication networks, frame erasure is declared when the channel coder cannot protect the bit errors due to the channel impairments such as noise, co-channel interference and fading [1]. In voice transmission over IP networks, the packet loss is caused by the transmission impairments such as the excess of the transmission capacity and congestion. Since even a single missing packet may generate an audible artifact in the decoded speech signal, the receiver needs a packet loss concealment algorithm to minimize the quality degradation at the packet loss regions.

Most packet loss concealment algorithms embedded in the standard coders are based on an extrapolation method or a repetition method in which the speech coding parameters are extrapolated or repeated from the parameters of the last good frame. Since the loss of packets causes the corruption of the long-term prediction memory, an extra performance degradation may occur from the use of the incorrect memory even at the correctly received frames in the future. Recently, many error concealment schemes for the CELP-type coders were proposed in order to lessen the quality degradation and the error-propagation problem. Some of them tried to accurately estimate the excitation information of the missing packets using a voicing classification [2][3]. Others efficiently estimated the gain parameters of the lost and the subsequent frames [1][4]. In [1] an adaptive and a fixed codebook gains are re-estimated by a gain matching method (GRE). Though the above gain-parameter re-estimation method provided the good performance under a variety of frame erasure conditions, the method needed another memory for the adaptive codebook that should be updated continuously even for correctly received packets.

In this paper, we propose a new voicing-driven frame erasure concealment algorithm based on a TSM scheme using a WSOLA technique [5]. We apply the proposed algorithm to the ITU-T G.729 Conjugate-Structure Algebraic CELP(CS-ACELP) speech coder [6] that is widely used in Voice over IP (VoIP) applications. We compare the performance of the proposed algorithm with the embedded standard method by measuring the perceptual evaluation of speech quality (PESQ) [7] and by performing an AB preference listening test. The rest of this paper is organized as follows. In Section 2, we first briefly review the frame erasure concealment algorithm embedded in G.729. The proposed algorithm is introduced in Section 3. The performance evaluations are given in Section 4. Finally, we make a concluding remark in section 5.

2. G.729 FRAME ERASURE CONCEALMENT

In G.729 speech coding, an erased frame is reconstructed using the speech coding parameters of the previous good frame. Once frame erasure is detected, the new parameters are generated by analyzing the spectral parameters of the last good speech frame. The method replaces the missing excitation signal of the erased frames by taking one of the similar characteristics, while gradually decaying its energy.

If the *n*-th frame is detected as an erased frame, the G.729 considers the spectral parameters of the last good frame as those of the erased frame. In other words, the LPC parameters of the last good frame are repeated to an erased frame. In addition, an adaptive codebook gain and a fixed codebook gain are obtained by multiplying predefined attenuation factors by the gains of the previous frame. To avoid excessive periodicity a long-term prediction lag is increased by one to the value of the previous frame. The main reason that the speech coding parameters of the erased frame are basically assigned with slightly different or

This work was supported in part by Biometrics Engineering Research Center, (KOSEF).

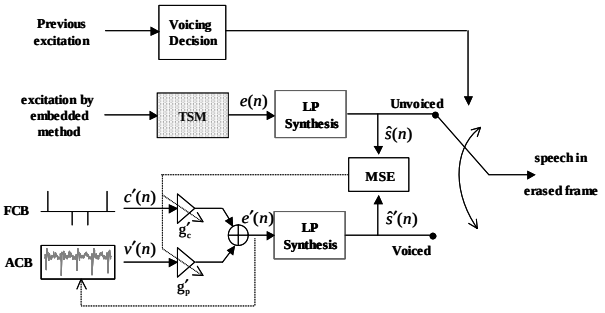

Figure 1 : Block diagram of the proposed algorithm

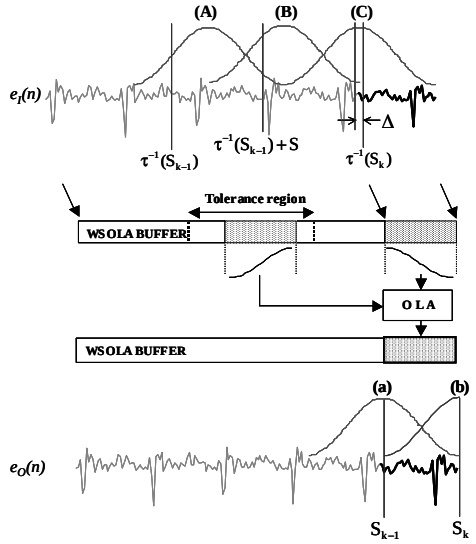

Figure 2 : WSOLA technique in the proposed algorithm

scaled-down values from the previous good frame is to prevent from generating a reverberant sound. However this simple scaling-down approach causes a fluctuation of an energy trajectory for the decoded speech and brings a more annoying effect to the listeners when longer frames are erased [1].

3. PROPOSED ALGORITHM

In this proposed algorithm, the key idea relates to the estimation procedure of the excitation signal. Comparing to the spectral parameter that is relatively slowly-varying, the excitation signal is important in the packet loss condition because it continuously affects the contents of the updated memory even after the erased frame. Figure 1 shows a block diagram for the proposed frame erasure concealment algorithm. As sketched in Figure 1, a new excitation generation module using a TSM and a modified gain parameter re-estimation module are the main functions of the proposed excitation recovery method. Once the frame erasure is detected, we classify the erased frame as voiced or unvoiced. If the erased frame is declared as unvoiced, we estimate the excitation information of the erased frame using the TSM method. If it is declared as voiced, we estimate the excitation information by combining the TSM and the modified GRE method.

3.1. Voicing Classification method

To classify the characteristics of the erased frame, we use a long-term prediction gain of the previous good frames. If the long-term prediction gain is higher than 3dB, the erased frame is classified as a voiced frame. Otherwise the frame is classified as an unvoiced frame [6].

3.2. Time-Scale Modification technique

For the case of an unvoiced frame, we reconstruct the new excitation using a TSM technique only. We use a WSOLA technique for the TSM. WSOLA specifies the timing tolerance of the input segments during an overlap-and-add (OLA) procedure [4]. Figure 2 shows the WSOLA technique applied in the proposed algorithm. This algorithm performs on a subframe basis. The synthesis equation for the proposed algorithm is

$$e_O(n) = \sum_k w_h(n - kS)\, e_I\left(n + \tau^{-1}(kS) - kS + \Delta_k\right) \qquad (1)$$

where S is a subframe length and $w_k(n)$ is a hamming window. $e_I(n)$ is an extended input excitation generated by an embedded method and $e_O(n)$ is a time-expanded output excitation. To generate (b) which is a time-scaled excitation for an erased subframe, maximally similarity to the excitation (C) reconstructed by the embedded method is searched in the tolerance region of WSOLA buffer. We construct the WSOLA buffer using the long-term prediction memory of G.729 coder. To apply WSOLA to the next erased subframe, we have to consider the time-instant of the excitation according to the time-warping function, $\tau^{-1}(kS)$. We solved the problem by using a dynamic buffer.

3.3. Modified Gain parameter Re-Estimation method

In the GRE method [1], the adaptive and fixed codebook gains are estimated by using a minimum mean square error criterion. Actually, this process should be continued even after receiving the correct frames because the buffers for the re-estimation procedure should be updated to use them for a later stage. It is a redundant processing that increasing the average complexity of the decoding process. We modify the GRE method in order to re-estimate the gain parameters only for the loss frame. The block diagram of the proposed algorithm is depicted in Figure 1. It shows that the target signal for the modified GRE procedure is modified using WSOLA technique. Thus the new gains, g_p' and g_c' are re-estimated by using a following criterion

$$\arg\min_{g_p', g_c'} \sum_{n=0}^{L-1} \left(\hat{s}(n) - \hat{s}'(n)\right)^2 \qquad (2)$$

Considering the synthesis procedure, (2) can be rewritten as

$$\begin{aligned} &\arg\min_{g_p', g_c'} \sum_{n=0}^{L-1} \left[h(n) * \left(e(n) - e'(n)\right)\right]^2 \\ &= \arg\min_{g_p', g_c'} \sum_{n=0}^{L-1} \left[h(n) * \left(e(n) - \left(g_p' v'(n) + g_c' c'(n)\right)\right)\right]^2 \end{aligned} \qquad (3)$$

where, L is a subframe length, $h(n)$ is the impulse response of the LPC-synthesis filter, and $e'(n)$ is the recovered excitation by the embedded method. From partial derivation of (3), the re-estimated values of g_p' and g_c' are

$$g_c' = \frac{\sum_n Z_e(n) Z_{c'}(n) - g_p' \sum_n Z_{p'}(n) Z_{c'}(n)}{\sum_n Z_{c'}^2(n)},$$

$$where,\ g_p' = \frac{\sum_n Z_e(n) Z_{p'}(n) \cdot \sum_n Z_{c'}^2(n) - \sum_n Z_e(n) Z_{c'}(n) \cdot \sum_n Z_{p'}(n) Z_{c'}(n)}{\sum_n Z_{c'}^2(n) \cdot \sum_n Z_{p'}^2(n) - \left(\sum_n Z_{p'}(n) \cdot \sum_n Z_{c'}(n) \right)^2} \quad (4)$$

$$Z_{p'}(n) = h(n) * v'(n),\ Z_{c'}(n) = h(n) * c'(n),\ Z_e(n) = h(n) * e(n)$$

In CELP-type coders, the effect of the erased frame propagates continuously even after receiving good frames. This effect causes an extra quality degradation. When a good frame is received after at least two consecutive bad frames, we felt the quality degradation of the reconstructed speech signal. This quality degradation results from the discrepancy between pitch contour used in encoder and that used in decoder. The discrepancy between pitch contours results in the corruption of the long-term prediction memory in a decoding stage, consequently error propagates. In order to apply the modified GRE procedure for only the lost frame, we set a hangover frame called a pseudo-good frame. If the previous frame is a bad frame and the current frame is a good frame, this current frame is set to a pseudo-good frame. In order to lessen the disagreement of the long-term prediction memories between encoder and decoder, we applied an overlap-add and a pitch smoothing technique while updating the adaptive codebook of the pseudo-good frame. Based on the degree of pitch variation between the artificial pitch and the transmitted pitch, we assign a new pitch to the pseudo-good frame not to vary abruptly. The simple algorithm for this purpose is using a weighted sum of two pitches values.

4. PERFORMANCE EVALUATIONS

We briefly explained the proposed concealment algorithm in the previous section. In this section, we compare the performance of the proposed method with that of the embedded method in G.729.

4.1. Waveform and Spectral difference

Figure 3 shows an example of speech quality degradation when frame erasure is occurred. Two burst frames are erased in each region. The proposed algorithm gives an improved waveform shape compared to the embedded method in G.729 coder. Since the embedded method decreases the gain of the adaptive codebook and fixed codebook in the erased frames and the pitch interval is increased by one to that of the previous frame, it shows a high level of magnitude distortion to the decoded speech for the erased frame. However, as shown in Figure 3 (e), the proposed concealment algorithm reduces the distortion in a waveform shape aspect. Figure 4 shows the speech spectrum obtained by the embedded method and the pro-posed method. Each spectrum is obtained by applying 256-pint FFT to the corresponding segment of 20 ms duration, where the spectrum without any frame errors is denoted by a dotted line. Figure 4 (a) and (b) give

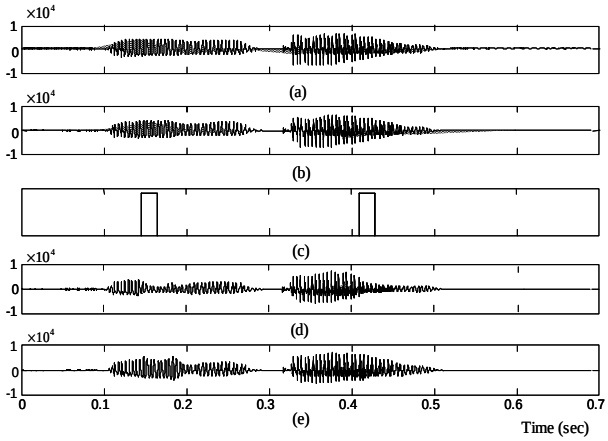

Figure 3. Example of speech quality degradation due to frame erasure: (a) original speech, (b) G.729 decoded speech without any frame error, (c) frame erasure patterns, each of which occurs burst errors of 2 frames (20msec), (d) decoded speech by the Embedded method in G.729, (e) decoded speech by the proposed concealment method

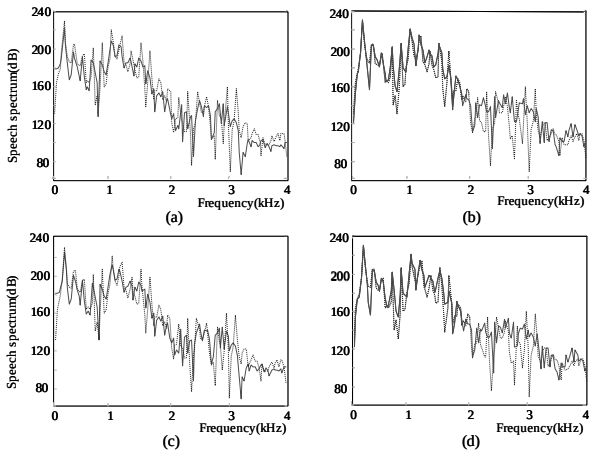

Figure 4. Comparison of decoded speech spectra (solid line) obtained by the embedded method and the proposed method. (a) the embedded method during the erased frames, (b) the proposed method during the erased frames, (c) the embedded method after a good frame, (d) the proposed method after a good frame

the spectrum matching performance at the packet loss frame and Figure 4 (c) and (d) give the performance just after a good frame is received. The proposed method matches the spectrum more accurately at the packet loss frame, especially in low frequency regions, than the embedded method. It also recovers the error-free spectrum more quickly than the embedded method.

4.2. Objective quality measure

We use PESQ for an objective quality measure. Two male and female Korean speakers recorded 8 seconds long sentences, and the speech was down-sampled to 8 kHz. We evaluated the performance by varying the frame error rate (FER) from 1% to 10%.

Table 1: Comparison of PESQ for the G.729 decoded speech with the embedded and the proposed frame erasure concealment under random frame erasure conditions. (A: the embedded method, B: the proposed method)

Random Error (%)	A	B
No error	3.887	3.887
1	3.749	3.772
3	3.436	3.513
5	3.177	3.297
7	3.050	3.181
10	2.832	2.973

Table 2: Comparison of PESQ for the G.729 decoded speech with the embedded and the proposed frame erasure concealment for the burst length under each FER rates. (A: the embedded method, B: the proposed method)

Burst length	FER = 1%		FER = 3%		FER = 5%	
	A	B	A	B	A	B
1	3.749	3.772	3.436	3.513	3.177	3.297
2	3.459	3.514	3.012	3.217	2.751	2.904
3	3.359	3.487	2.548	2.935	2.191	2.793

Table 3: AB preference test results at both random error and burst error. (A: the embedded method, B: the proposed method, C: no preference)

FER(%)	Random Error			Burst Error		
	A(%)	B(%)	C(%)	A(%)	B(%)	C(%)
1	8	58	34	8	67	25
3	7	60	33	17	60	23
5	20	43	37	10	65	25
7	13	50	37	7	60	33
10	3	64	33	5	70	25

Table 1 shows comparison results. As FER increases, the PESQ scores of the two algorithms decrease. However, the proposed algorithm has higher scores than the embedded algorithm for all FERs. Table 2 shows the PESQ scores according to the burst length of each random error. In this measure, we set the range of random error from 1% to 5%, and increase the burst length in each random error condition. This means that the FER varies from 1% to 15%. As shown in Table 2, the PESQ scores of the proposed method are also always higher than those of the embedded method.

4.3. Subjective quality measure

We also performed an AB preference test to evaluate subjective quality. 30 listeners attended in each condition. As shown in Table 3, they preferred the proposed method. Especially, as the burst errors increase, the preference ratio of the proposed algorithm also increases. It means that the proposed algorithm is more robust to the burst error than the embedded algorithm. While analyzing the test results we also found that the quality improvement for female speakers was superior to that of the male speakers. One reason might be the addition procedure of the pitch interval in the embedded method, which changes the characteristics of the updated memory more rapidly for female.

5. CONCLUSION

In this paper, we proposed a frame erasure concealment algorithm for CELP-coders and compared its performance with the conventional algorithm embedded in G.729. The proposed algorithm generated a new excitation using a time-scale modification technique for the erased frame. In addition, if the erased frame was classified as a voiced frame, the contribution of gain parameters to the new excitation was re-estimated using a minimum mean square error criterion. From the PESQ measurement and an AB preference test under a variety of frame erasure conditions, we found that the proposed algorithm significantly improved the speech quality compared to the embedded extrapolation method.

The proposed algorithm is a separate module that can be inserted before the conventional speech decoding procedure. Since this internal operation does not need any additional information, it can be easily adopted to other CELP-type speech coders.

REFERENCES

[1] Hong Kook Kim and Hong-Goo Kang, "A Frame Erasure Concealment Algorithm Based on Gain Parameter Re-estimation for CELP coders," in *IEEE Signal Processing Letters*, vol. 8, pp.252-256, Sept 2001.

[2] A. Husain and V. Cuperman, "Reconstruction of missing packets for CELP-based speech coders," in *Proc. ICASSP-95*, vol. 1, 1995, pp.245-248.

[3] Jhing-Fa Wang, Jia-Ching Wang, Jar-Ferr Yang and Jian-Jia Wang, "A voicing-driven packet loss recovery algorithm for analysis-by-synthesis predictive speech coders over Internet," in *Multimedia, IEEE Transaction* , vol. 3, pp.98-107, March 2001.

[4] de Martin, J.C, Unno, T. and Viswanathan, V, "Improved frame erasure concealment for CELP-based coders," in *Proc. ICASSP'00*, vol. 3, pp.1483-1486.

[5] Verhelst, W and Roelands, M, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP-93*, pp.554-557, 1993.

[6] ITU-T Draft Rec G.729 "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic Code-Excited Linear-Prediction (CS-ACELP)." Feb. 1996.

[7] ITU-T Draft Rec P.862 "Perceptual evaluation of speech quality (PESQ), an objective method of end-to-end speech quality assessment of narrowband telephone networks and speech codecs," May. 2000.