IMPROVED FRAME ERASURE CONCEALMENT FOR CELP-BASED CODERS

Juan Carlos De Martin[*], *Takahiro Unno and Vishu Viswanathan*

DSPS R&D, Texas Instruments
Dallas, Texas
E-mail: demartin@polito.it, [takahiro|vishu]@hc.ti.com

ABSTRACT

This paper describes new techniques for concealing frame erasures for CELP-based speech coders. Two main approaches were followed: interpolative, where both past and future information are used to reconstruct the missing data, and repetition-based, where no future information is required. Key features of the repetition-based approach include improved muting, pitch delay jittering, and LPC bandwidth expansion. The interpolative approach can be employed in Voice over IP scenarios at no extra cost in terms of delay. Applied to the ITU-T G.729 ACELP 8 kb/s speech coding standard, both interpolation- and repetition-based techniques outperform standard concealment in informal listening tests.

1. INTRODUCTION

Concealment of missing or corrupted frames (or *frame erasures*) is recommended whenever speech is transmitted over a noisy channel. Even a single corrupted frame, in fact, can generate an audible, potentially annoying, artifact in the decoded output speech.

Depending on the speech coder employed and on the error statistics, frame erasures can decrease average speech quality well below nominal level even for relatively low percentages of lost frames.

Important speech transmission scenarios are affected by frame erasures. Wireless links, both cellular and satellite, are noisy and can generate, at least locally, high rates of corrupted frames. Voice transmission over IP networks can also be severely affected by discarded or late packets.

When a frame is missing or corrupted, the corresponding portion of output speech, usually between 5 and 25 ms of signal, equivalent to 40–200 samples for narrowband speech, needs to be reconstructed. Jayant *et al.* studied odd-even sample interpolation in PCM and DPCM systems [1]. PCM systems were again analyzed by Goodman *et al.*, who proposed waveform substitution techniques such as pattern matching [2]. More recent work on the waveform substitution in PCM was presented by Erdol *et al.* [3]. Voice packet reconstruction for CELP-based coders was studied by Yong, who presented results on the performance of four methods in a simulated packet network environment [4]. Concealment of lost frames based on classification and spectral extrapolation, applied to the ITU G.728 standard, was proposed by Husain and Cuperman [5]. Leung *et al.* extended to CELP-based coders the concept of using future as well as past information to reconstruct missing frames [6].

Current state-of-the-art frame reconstruction for modern speech coders, whether vocoders or CELP-based systems, essentially consist in repeating the speech parameters contained in last correctly received frame. If two or more consecutive frames are lost, increasingly strong muting is applied. This approach, which has the advantage of not introducing any extra delay, has been followed by most recent speech coding standards.

We propose improvements over the standard practice both in the regular case of concealment based on past data only and in the case of concealment that uses some future information as well.

The paper is organized as follows. In Section 2, improvements to the standard repetition based techniques are presented for the specific case of ITU-T standard G.729 [7], one of the most widely used speech coder in Voice over packet networks applications. In Section 3, instead, interpolative concealment is presented and its case made for Voice over IP applications. Results of A/B listening tests comparing the standard method against the proposed techniques are presented in Section 4. Finally, conclusions are presented in Section 5.

2. REPETITION-BASED CONCEALMENT

2.1. Overview

Figure 1 shows the block diagram of the G.729 decoder containing our repetition-based frame erasure concealment method. There are three key features in this method. The first feature is a new muting algorithm which mutes the excitation signal directly with the muting factor $g_e^{(n)}$ to decay the signal gradually, instead of attenuating the codebook gains in the previous frame as is done in the G.729 standard

[*]Work done at Texas Instruments while on leave from IRITI-CNR, Politecnico di Torino, Italy.

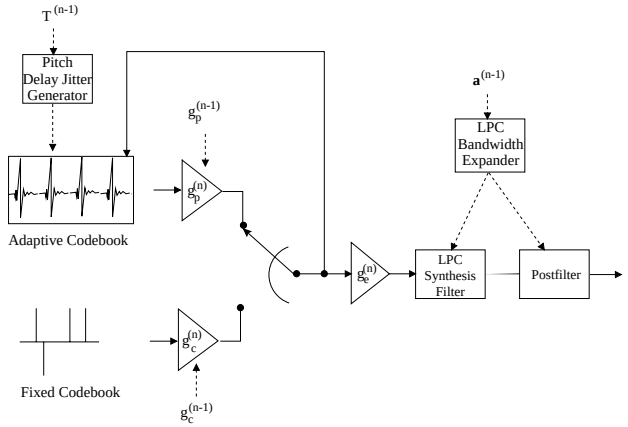

Figure 1: Block diagram of proposed frame erasure concealment in G.729 decoder.

frame erasure concealment. The second feature is a pitch delay jittering for a bursty frame erasure. The random jitter is added to the repeated pitch delay only when a consecutive frame erasure occurs. The third feature is LPC bandwidth expansion for bursty frame erasures. As is the case in the pitch delay jittering, the LPC bandwidth in the previous frame is expanded only when a consecutive frame erasure occurs. The proposed method is designed not only to reconstruct speech in bad frames but also to recover speech smoothly after the frame erasure.

2.2. Muting of Excitation Signal

In Figure 1, $g_p^{(n)}$ and $g_c^{(n)}$ are the adaptive and fixed codebook gains in the current frame respectively. In a bad frame, G.729 uses an attenuated version of the previous codebook gains $g_p^{(n-1)}$ and $g_c^{(n-1)}$ as the current codebook gains $g_p^{(n)}$ and $g_c^{(n)}$. Figure 2 (a) and (b) show the G.729 synthesized speech waveforms without and with a frame erasure. Figure 2 (d) shows the adaptive codebook gain at the decoder corresponding to the waveform in (b). The dashed vertical lines in Figure 2 (b) and (d) indicate the regions where the frame erasure occurs. The adaptive codebook gain in the last good frame is shown to be attenuated during the period of frame erasure. However, even after the frame erasure is over, the speech signal is further decayed in the subsequent frames. This is because the adaptive codebook is updated with the attenuated excitation signal so the attenuation propagates to the subsequent frames. To avoid such an excessive decay of the signal, our method applies the muting factor $g_e^{(n)}$ outside of the adaptive codebook feedback-loop to take better control of muting than attenuating codebook gains. In the proposed muting algorithm, the muting factor $g_e^{(n)}$ is decreased by 0.4 dB every subframe (5 ms) during the consecutive bad frames and subsequent few frames. A few frames after the last bad frame, $g_e^{(n)}$ is increased by 0.8 dB up to 1 every subframe for the smooth recovery at the end of frame erasures. $g_e^{(n)}$ is represented as follows:

$$\begin{aligned} g_e^{(n)} &= 0.95499 g_e^{(n-1)} && \text{if } C_m^{(n)} > 0 \\ g_e^{(n)} &= \min(1.09648 g_e^{(n-1)}, 1.0) && \text{Otherwise,} \end{aligned} \quad (1)$$

where C_m is a muting counter. C_m is set to four in consecutive bad frames and decremented by one down to zero only if $g_p^{(n)} < 1.0$ and the current frame is good. The codebook gains $g_p^{(n)}$ and $g_c^{(n)}$ are simply repeated in the bad frames instead of being attenuated. However, the upper bound g_{pmax} is set for $g_p^{(n)}$ to prevent the unpredicted surge of the excitation signal energy as follows:

$$g_{pmax} = \max(1.2 - 0.1(C_b - 1), 0.8), \quad (2)$$

where C_b is the consecutive number of bad frames. Figure 2 (c) shows the synthesized speech waveform with the proposed muting algorithm. As is shown in Figure 2 (c), the speech signal is not excessively decayed in and after the frame erasure. In the proposed method, no upper bound is set for the fixed codebook gain $g_c^{(n)}$, and the MA-prediction memory for the fixed codebook gain is updated in the same manner as the G.729 standard in bad frames.

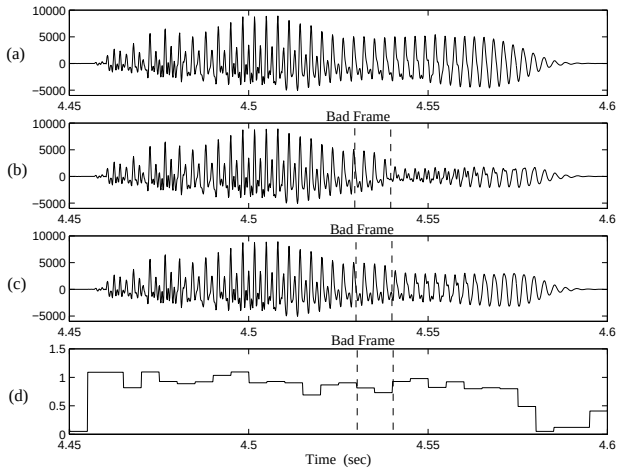

Figure 2: Synthesized speech and adaptive codebook gain. (a) G.729 synthesized speech without frame erasure. (b) G.729 synthesized speech with standard concealment. (c) G.729 synthesized speech with proposed concealment. (d) Adaptive codebook gain.

2.3. Jittering of Pitch Delay

In the G.729 standard concealment, the pitch delay in a bad frame is the previous pitch delay increased by one to mimic the pitch evolution in natural speech [8]. It can avoid reconstructing an excessively periodic signal in a bursty frame erasure, but it may accumulate the estimation error

of the pitch delay in consecutive bad frames. In the proposed frame erasure concealment method, the pitch delay is repeated in the bad frames, but random jittering of 3% is added to the repeated pitch delay in the consecutive bad frames to avoid reconstructing an excessively periodic signal without accumulating the estimation error.

2.4. LPC Bandwidth Expansion

In the G.729 decoder, the LPC parameters in the last good frame are repeated in bad frames. However, it may result in a synthetic speech quality if the LPC spectrum in the last good frame contains a sharp formant peak. To avoid this problem, the proposed concealment method progressively expands the LPC bandwidth in the consecutive bad frames only if the minimum Line Spectral Frequency (LSF) bandwidth in the last good frame is less than 100 Hz. The bandwidth expansion factor $\gamma^{(n)}$ is updated as follows:

$$\gamma^{(n)} = \max(0.95\gamma^{(n-1)}, 0.8), \quad (3)$$

where $\gamma^{(n)}$ and $\gamma^{(n-1)}$ are the current and previous LPC bandwidth expansion factors. The factor $\gamma^{(n)}$ is applied to the LPC parameter in the last good frame. When the decoder receives the good frame after a bursty frame erasure, $\gamma^{(n)}$ is progressively increased for the smooth recovery from frame erasures:

$$\gamma^{(n)} = \min(1.05\gamma^{(n-1)}, 1.0). \quad (4)$$

In this case, the factor $\gamma^{(n)}$ is applied to the LPC parameter in the received good frame.

3. INTERPOLATIVE CONCEALMENT

3.1. Overview

If future speech data is, or can be made, available, then an interpolative approach to frame erasure concealment becomes possible. This should intuitively produce better concealment than the simpler repetition-based approach, at the expense of extra delay.

Interpolation-based concealment for CELP coders has hardly been investigated. The reason for such relative neglect is probably the extra delay entailed by the approach, not acceptable in applications, like wireless, where delay is tightly controlled.

The emergence of a new, important application, however, *voice over IP networks*, makes interpolative concealment attractive. In VoIP systems, in fact, one or more future frames are, at least most of the time, available at the decoder, stored in the so-called *playout buffer*. Such buffer, introduced to smooth out the effects of delay jitter, is an essential component of all VoIP receivers. Interpolative concealment can exploit the delay introduced by the playout buffer to improve performance under frame erasures at no extra cost in terms of delay.

A block diagram illustrating interpolative concealment within a typical VoIP receiver is shown in Figure 3.

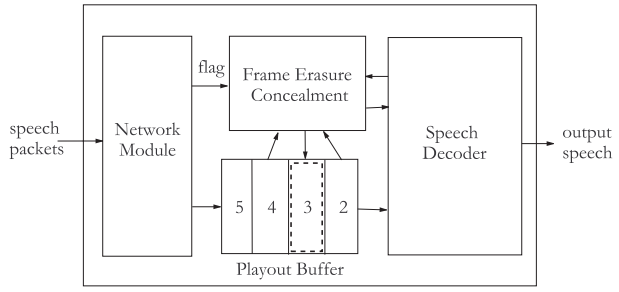

Figure 3: Typical VoIP receiver: Interpolative concealment

Packets arriving from the network are first processed by the network module. Statistics are collected, packets ordered and transferred to the playout buffer. If near the time of playback the packet has not yet arrived, it is declared lost and the frame erasure concealment module reconstruct it using both past and future frames. In the figure, packet 3, missing, is reconstructed by interpolating the previous (2) and following (4) packet.

3.2. Interpolative Concealment for G.729

A frame erasure concealment scheme based on interpolative reconstruction was implemented in the ITU-T G.729 8 kb/s speech coding standard. The speech decoder was modified so that if a frame erasure was detected, and if the next frame was not erased as well, interpolation based concealment was applied instead of the method defined by the standard.

Interpolative reconstruction was first applied to the adaptive codebook parameters, index and gain. The adaptive codebook gain was linearly interpolated between past and future values. Median smoothing, instead, was applied to the adaptive codebook index. Voicing classification was used in the same manner as in G.729.

Line Spectral Frequencies and fixed-codebook gain, predictively quantized with a moving-average fourth-order predictor, proved difficult to interpolate. Since it seems reasonable to expect improvements from the use of future information even in the case of predictive quantization, more work seems possible in this direction. In our experiments the LSF's and the fixed-codebook gain were replaced according to the standard method. The fixed-codebook index was also generated as in G.729.

The resulting method entails virtually no added complexity with respect to the standard approach. Listening tests results on interpolative concealment are reported in the following Section.

Preference :		*Strong*	*Slight*	*No*	*Slight*	*Strong*
FER	ms	New Method			Standard Method	
3 %	10	0	20	21	7	0
3 %	20	2	26	14	6	0
8 %	20	1	26	16	5	0
8 %	40	2	16	18	12	0

Table 1: A/B listening test results for repetition-based concealment.

Preference :	*Strong*	*Slight*	*No*	*Slight*	*Strong*
FER	New Method			Standard Method	
3 %	0	24	18	6	0
5 %	1	22	18	7	0

Table 2: A/B listening test results for interpolative concealment. Packet size is 10 ms.

4. RESULTS

We conducted an A/B listening test to compare the performance of the proposed method to that of the G.729 standard frame erasure concealment. Four female and four male sentence pairs were presented to six listeners in each condition. In the test, frames are randomly erased with the specified frame erasure ratio. For the repetition-based concealment we tested on three packet sizes 10 ms, 20 ms and 40 ms to evaluate the robustness to a bursty frame erasure. For example, two frames are put into one packet (i.e. two frames are consecutively erased at least) in the packet size of 20 ms. As shown in Table 1, the proposed method is clearly preferred over the standard concealment method in the packet sizes of 10 ms and 20 ms. In the condition of 40 ms packet size, the proposed method is only slightly preferred because both of the repetition-based methods do not perform well for such a large packet size. In the test, it was also found that the quality improvement was clearer for female speakers than male speakers. This is because each subframe contains a pitch peak signal for female more likely than male speakers so the excitation signal may be more excessively decayed for female speakers in the G.729 standard concealment.

The same set-up was used to evaluate the interpolation-based approach. Random frame erasures at 3% and 5% were applied to G.729 bitstreams. Isolated frame erasures were reconstructed applying the interpolative concealment method described in Section 3.2. Six listeners compared the output of new and standard techniques in a blind A/B listening test. As shown in Table 2, the proposed technique was clearly preferred over the standard method in both conditions.

5. CONCLUSIONS

We have presented in this paper new, low-complexity techniques of concealing frame erasures in CELP coders. We have shown how our repetition-based approach improves on current state-of-the-art practices and presented the case for interpolative concealment in voice over IP applications. Both schemes were clearly preferred in A/B listening tests over the standard approach.

6. REFERENCES

[1] N. Jayant and S. Christensen, "Effects of Packet Losses in Waveform Coded Speech and Improvements Due to an Odd-Even Sample-Interpolation Procedure," *IEEE Transactions on Communications*, vol. COM-29, February 1981.

[2] D.J. Goodman et al., "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, December 1986.

[3] N. Erdol, C. Castelluccia, and A. Zilouchian, "Recovery of Missing Speech Packets Using the Short-Time Energy and Zero Crossing Measurements," *IEEE Transactions on Speech and Audio Processing*, vol. 1, July 1993.

[4] M. Yong, "Study of Voice Packet Reconstruction Methods Applied to CELP Speech Coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–125–128, 1992.

[5] A. Husain and V. Cuperman, "Reconstruction of Missing Packets for CELP-Based Speech Coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 245–248, 1995.

[6] T. Leung, W. LeBlanc, and S. Mahmoud, "Speech Coding over Frame Relay Networks," in *Proceedings IEEE Workshop on Speech Coding for Telecommunications*, (Quebéc, Canada), pp. 75–76, October 1993.

[7] R. Salami et al., "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 116–130, March 1998.

[8] P. Kroon and Y. Shoham, "Performance of the Proposed ITU-T 8 kb/s Speech Coding Standard for a Rayleigh Fading Channel," in *Proceedings IEEE Workshop on Speech Coding for Telecommunications*, (Annapolis, Maryland), pp. 11–12, September 1995.