SPEECH-ADAPTIVE LAYERED G.729 CODER FOR LOSS CONCEALMENTS OF REAL-TIME VOICE OVER IP

Batu Sat and Benjamin W. Wah

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{batusat,wah}@uiuc.edu

ABSTRACT

In this paper, we propose a speech-adaptive layered-coding (LC) scheme for the loss concealments of real-time CELP-coded speech transmitted over IP networks. Based on the ITU G.729 CS-ACELP codec operating at 8 Kbps, we design a loss-robust speech-adaptive codec at the same bit rate. Our scheme employs LC with redundant packetization in order to conceal losses and adapt to dynamic loss conditions characterized by the loss rate and the degree of burstiness, while maintaining an acceptable end-to-end delay. By protecting only the most important excitation parameters of each frame according to its speech type, our approach enables more efficient use of the bit budget. Our scheme delivers good-quality speech with a level of protection similar to full replication under medium loss rates, provides speech quality similar to the standard G.729 under very low loss rates, and outperforms both for low-to-medium loss rates.

1. INTRODUCTION

Background. In this paper, we study the loss concealment of low bit-rate coded speech transmitted in real time by voice-over-IP (VoIP). These transmissions may suffer quality degradations when packets are lost because pervasive dependencies may lead to sustained distortions over a number of consecutive frames. These distortions cannot be overcome by source coding methods using a channel-loss model because losses in the Internet are non-stationary and connection dependent [1]. Measurements on the Internet show that losses to some international destinations can sustain a loss rate that can be as high as 50% and bursty losses of three or more consecutive packets. Moreover, the loss rate goes up dramatically when packets are transmitted at a very high rate (say 100 packets per second). Further, the end-to-end transit delay of an IP packet can be highly varying and can range from tens to hundreds of milliseconds within a short amount of time. In order to allow the perception of real-time interactive speech within an end-to-end delay of 400 ms or less (ITU G.114), the application must employ a combination of UDP as the transport protocol, jitter buffers at the receiver to smooth out irregular arrivals, an encoder that enables loss concealments, and a reconstruction algorithm that recovers lost information at receivers.

To allow a receiver to reconstruct lost information without retransmissions, there must be redundancies in the data stream received. Such redundancies are usually absent in low bit-rate coded speech because most redundancies have been removed by the encoder in order to achieve high coding efficiency. Hence, when frames are lost, the quality of a speech sequence is usually less than satisfactory.

As an example, consider the built-in loss-concealment algorithm of the ITU G.729 codec. When a frame is lost, the decoder reconstructs the LPC parameters of the lost frame from those of the last received frame. For the excitation parameters, it either reconstructs them from past excitation samples using the previous pitch period if the frame contains voiced samples, or randomly generates algebraic codebook (ACB) pulses and omits the contribution of previous excitation signals if the frame contains unvoiced samples. It further scales down the excitation gain in order to reduce the perception of losses. The scheme can only cope with infrequent and isolated frame losses because dependencies across frames may result in the propagation of the decoder-state error over multiple frames. Figure 1b illustrates the effect of a single lost frame that lasts over 15 frames.

An alternative to improve the reconstruction quality at receivers without increasing the bit rate is to reduce the amount of information in a coded speech sequence at senders in order to reserve the space needed for carrying redundant information. If a channel-loss model is available, the sender may employ FEC or unequal error protection to provide different levels of protection for different layers of information. These schemes, however, are not applicable when the loss behavior is non-stationary.

RESEARCH SUPPORTED BY THE MOTOROLA CENTER FOR COMMUNICATIONS, UNIVERSITY OF ILLINOIS, URBANA.
IEEE INT'L CONF. ON MULTIMEDIA & EXPO, 2005.

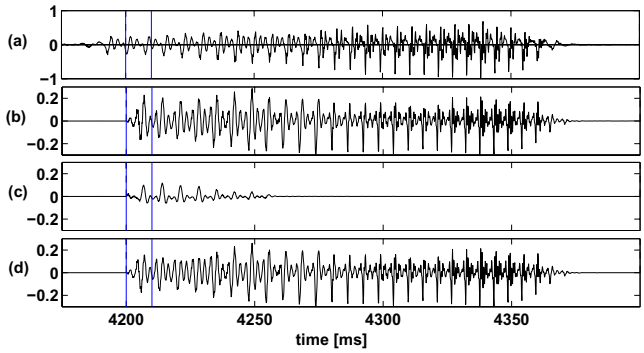

Figure 1: Effects of loss of a single ITU G.729 frame (between 4200 ms and 4210 ms): (a) original input speech; (b) distortions when the frame at 4200 ms is lost and built-in loss concealment is applied; (c) distortions when only the LPC parameters are lost; and (d) distortions when only the excitation parameters are lost.

Layered coding (LC) generates multiple *layers* from a data stream, followed by the partitioning of each layer into UDP packets, each of MTU size.[1] Since the information in lower layers is more important than that in higher layers in determining the quality of the received speech, it is given more protection or transmitted at higher priorities. Without a channel-loss model and priority transmissions in the Internet, the extra protection is usually provided by full replications (characterized by the *degree of redundancy*). For instance, the example in Figure 1 indicates that the excitation parameters are more important and should be in Layer 0, whereas the LPC parameters can be in Layer 1.

Multiple description coding (MDC) has been applied in the loss concealments of FS-1016 CELP, ITU G.723, and FS MELP low bit-rate coded speech [1]. The approach identifies that the LPC parameters are correlated across frames and can be reconstructed from those in adjacent frames, whereas the excitation parameters are uncorrelated and cannot be reconstructed easily. It then interleaves the LPC parameters to different descriptions, while replicating the excitation parameters, in order to enable reconstruction when a packet is lost. To avoid increasing the bit rate, the space required for replicating the excitation parameters is created by increasing the sub-frame size of each frame.

Full replication. The above MDC scheme has to be modified when applied to the G.729 coder. In G.729, temporal redundancies in LPC parameters have been removed by predictive coding. Hence, both the LPC and the excitation parameters must be replicated across the multiple descriptions in order to allow reconstructions. To keep the bit rate constant at 8 Kbps while assuming two-way MDC, the original speech sequence must be coded at 4 Kbps and a frame size of 25 ms with the same code structure.

The replication of all the parameters in every description is over-conservative. Figures 1c-1d illustrate the degradations caused by the loss of either the LPC or the excitation parameters when losses are concealed by the built-in algorithm. They reveal that the magnitude and the duration of the distortions caused by the loss of the LPC alone are insignificant when compared to those of the excitations. Further, the distortions caused by the loss of the excitations are very similar to those of the built-in algorithm. Hence, it suffices to replicate only the excitation parameters in both descriptions in order to achieve good reconstructions.

[1] The MTU (maximum transfer unit) in the Internet has 576 octets. This is the largest value to ensure that an IP packet will not be fragmented.

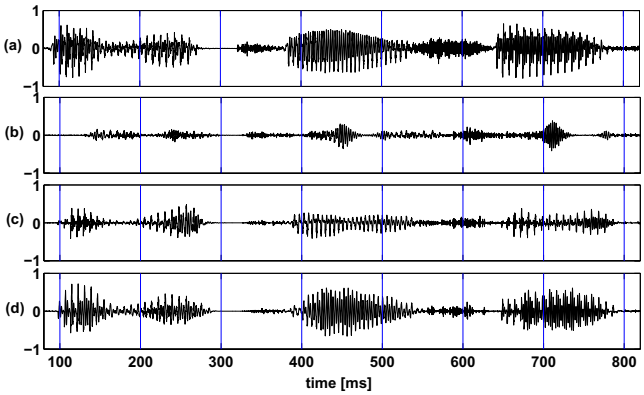

Figure 2: Synthesized speech under isolated and infrequent losses (every 100ms at vertical lines): (a) original input speech; (b) reconstructed speech when losses are concealed by the built-in G.729 algorithm; (c) reconstructed speech using the hybrid LC scheme; and (d) reconstructed speech using full replications.

Hybrid LC scheme. The above observation shows that the LPC parameters do not have to replicated but can be copied from adjacent frames when a frame is lost. This leads to a hybrid LC scheme that interleaves the LPC (Layer 1) and that replicates the excitations (Layer 0) during packetization. The coder uses a frame size of 20 ms and a bit rate of 4.4 Kbps, with an aggregate bandwidth of 8 Kbps.

In summary, the protection needed is loss dependent. The built-in loss-concealment algorithm provides minimal protection and does not perform well under low-loss (Figure 2b) as well as medium-loss scenarios. In contrast, the full-replication and the hybrid LC schemes provide good protection under high-loss scenarios but are over-conservative in low-loss (Figure 2c-2d) and medium-loss scenarios. The degradations for full replications (*resp.* hybrid LC scheme) under low-loss scenarios are due to the reduced efficiency of pitch extraction when the sub-frame size is extended from 5 ms to 12.5 (*resp.* 10) ms.

Problem statement. In this paper, we design a modified G.729 coder that achieves quality similar to full replication under medium loss rates, that provides quality similar to the standard G.729 under very low loss rates, and that outperforms both for low-to-medium loss rates. We assume the same 8-Kbps rate as the original G.729, a packet period of 30 ms, periodic (but delayed) feedbacks of loss behavior from receivers, and without an accurate channel-loss model.

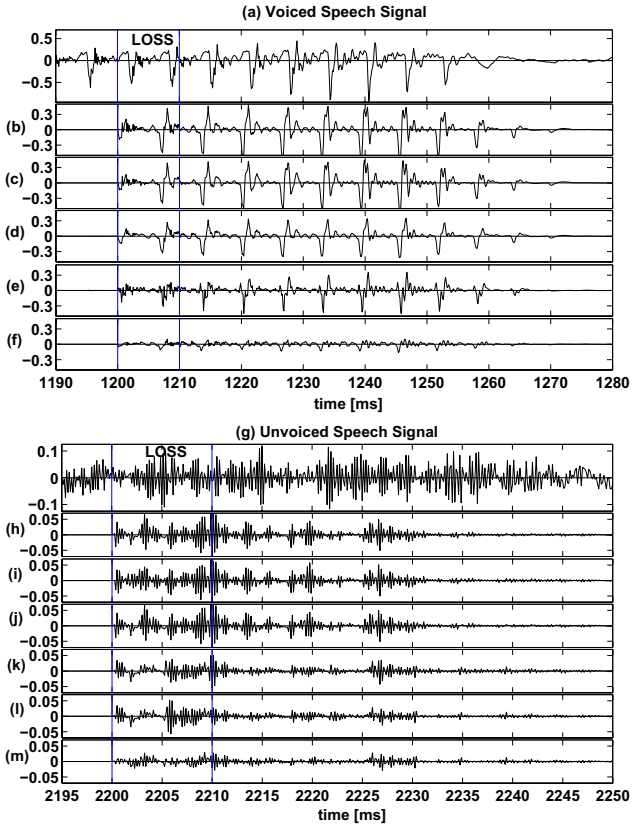

Figure 3: Distortions in the loss of a single 10-ms frame in the voiced (b-f) and unvoiced regions (h-m); (a,g) original speech sequence in the voiced and unvoiced regions. The protection scheme for each component of the G.729 coder is shown in Table 1.

2. SPEECH-ADAPTIVE LC SCHEME

Our modified G.729 coder is designed by reserving less space in the original speech sequence for carrying redundant information, as compared to that of full replication, in order to achieve an 8-Kbps rate and a quality similar to that of the original G.729 under low-loss scenarios. Using the space reserved, our coder protects only a subset of the parameters that are vital in the reconstruction of excitations in order to achieve a quality similar to that of full replication under medium-loss scenarios.

We have observed that the vital Layer-0 excitation parameters that need protection change with the voicing characteristics of a speech sequence. For example, the structure of CELP coders rely heavily on previously decoded excitation signals in order to reconstruct the current excitations in voiced regions. In contrast, algebraic (G.729) or stochastic (FS-CELP) codebooks that are used to represent aperiodic pulses are more important in unvoiced regions. This dependence is especially vulnerable to error propagations when the pitch gain is large (in voiced regions).

Figure 3 illustrates the above observations for different protection schemes listed in Table 1. Figures 3b-3e show

Table 1: Various schemes for protecting excitation parameters in voiced and unvoiced regions in Figure 3. (PP: pitch period; PG: pitch gain; AP: ACB pulses; AG: ACB gain; Bits: protection space required; S: built-in loss concealment; R: replicated; I: reconstruction by interpolations; 2,3,4: number of ACB pulses protected).

Scheme	(b)	(c)	(d)	(e)	(f)	(h)	(i)	(j)	(k)	(l)	(m)
PP	S	S	R	I	R	S	R	S	S	S	S
PG	S	S	S	R	R	S	S	S	S	S	S
AP	S	R	S	S	S	S	S	4	2	3	4
AG	S	R	S	S	S	S	S	S	R	R	R
Bits	0	56	13	14	27	0	27	42	34	44	56

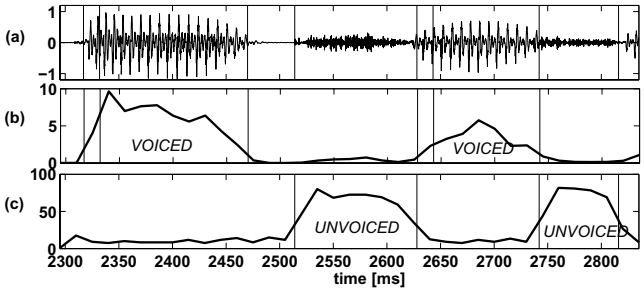

Figure 4: Heuristic for identifying voiced and unvoiced regions in a speech sequence. (a) original speech sequence, (b) energy and voiced regions identified, (c) zero-crossing percentage and unvoiced regions identified.

that, for the loss of a single frame in the voiced region (Figure 3a), the distortions with respect to the original speech sequence are large when the information in the lost frame is recovered by, respectively, the built-in loss concealment scheme, the replicated ACB parameters, the replicated pitch period, and the replicated pitch gain. Only when both the pitch period and the pitch gain are replicated using 27 extra bits per voiced frame will the distortions be low (Figure 3f).

Similarly, Figures 3h-3j show that, for the loss of a single frame in the unvoiced region (Figure 3g), the distortions with respect to the original speech sequence are large when the information in the lost frame is recovered by, respectively, the built-in loss concealment scheme, the replicated pitch period, and the replicated ACB pulses. Figures 3k-3m illustrate the trade-offs in protecting, respectively, two, three, and four ACB pulses as well as the ACB gain. It is clear that the protection of two ACB pulses and the ACB gain using 34 extra bits leads to acceptable distortions when a frame is lost in the unvoiced region.

Note that the onset of voiced regions cannot benefit from pitch extraction due to a lack of periodic signals in previous excitations and, thus, behave similar to unvoiced regions in terms of protection needs.

Based on the information on energy and zero crossings in the original speech sequence, we have designed an efficient heuristic to identify voiced and unvoiced regions and decide on the parameters to be protected on a frame-by-frame basis (illustrated in Figure 4).

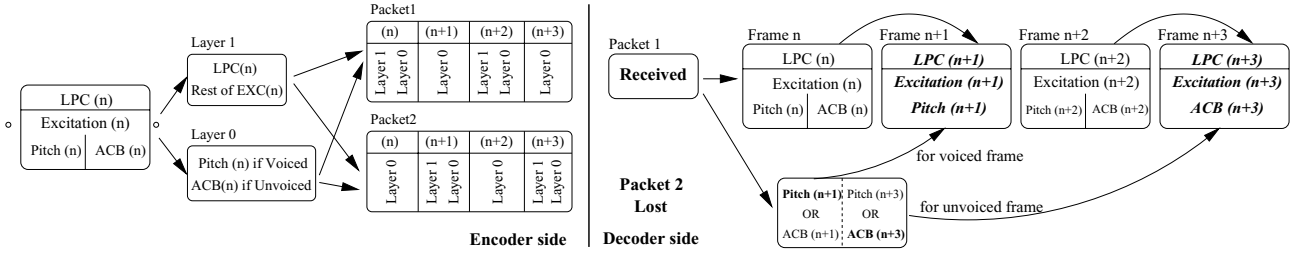

Figure 5: The reconstruction of a lost packet when the Layer-0 information in each frame is replicated two ways and the four frames in each packet are interleaved. The reconstruction of Frames $n+1$ and $n+3$ is done by using, respectively, Schemes f and k in Table 1.

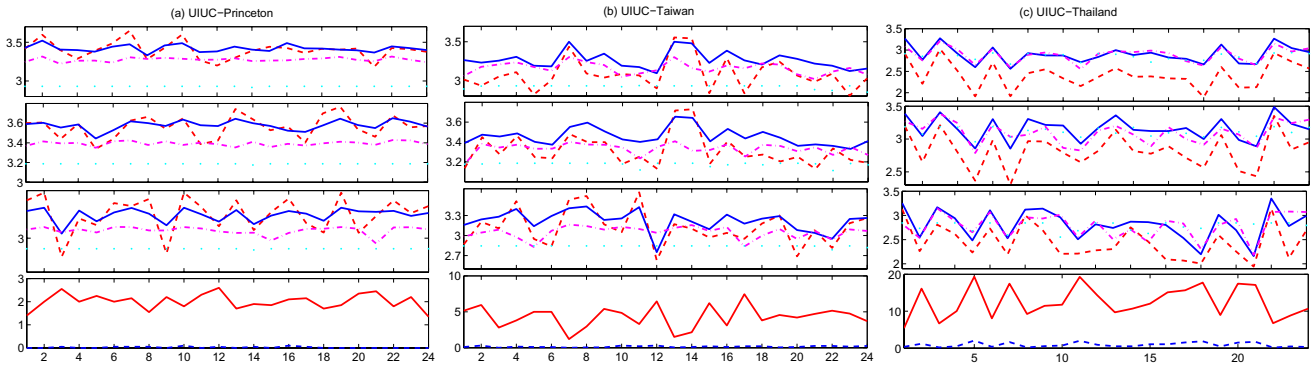

Figure 6: PESQ of three received speech sequences [1] (top three rows) to three destinations evaluated at each hour over a 24-hour period, based on traces collected in April 2003. In each PESQ plot, dashed: ITU G.729; dotted: LC with full replication; dot-dash: hybrid LC; solid: proposed speech-adaptive LC. In the plot on loss rates (bottom row), solid: single isolated loss in %; dashed: double losses in %.)

Figure 5 shows two packets that contain the two-way replication of the Layer-0 information in four frames and the interleaving of these frames. Using the redundant information in a packet, a receiver can conceal all single-packet losses as well as double-packet losses that span across two interleaving sets. To keep the bit rate at 8 Kbps and to create the space for holding duplicate information, the frame size has to be increased from 10 ms to 15 ms (*resp.* 20 ms) for two-way (*resp.* three-way) redundancy. Because the subframe size is only increased from 5 ms to 7.5 (*resp.* 10 ms), the scheme performs better than full replications and the hybrid LC scheme under low-loss scenarios. The additional interleaving and buffering delays incurred at both ends in two-way (*resp.* three-way) redundancy are 90 ms (*resp.* 150 ms). Since two-way redundancy incurs less end-to-end delays, the system should operate in this mode as much as possible and switch to three-way redundancy only when feedbacks from receivers indicate a high probability of bursty losses of two or more packets.

3. EXPERIMENTAL RESULTS

Figure 6 compares the quality of our speech-adaptive LC scheme with those of the built-in G729, the hybrid LC scheme, and the full-replication scheme. The experiments were done using trace-driven simulations, based on UDP packet traces to three destinations (UIUC-Princeton, UIUC-Taiwan, and UIUC-Thailand) [1]. Due to the non-stationary nature of loss characteristics, the encoder adapts its degree of redundancy between two ways and three ways every second based on periodic feedbacks from receivers. We have measured quality using the ITU P.862 PESQ metric [2], which was designed for evaluating the perceptual quality of speech coded by low bit-rate CELP coders.

Figure 6a depicts the speech quality for the UIUC-Princeton connection with very low loss rates. In this case, our scheme achieves a quality similar to that of G.729 on average, and is better than other schemes for all three speech sequences. Further, its level of quality is more stable as compared to that of G.729. Figure 6b depicts the speech quality for the UIUC-Taiwan connection with low loss rates. Our scheme outperforms all other schemes for over 90% of the cases. Last, Figure 6c shows the quality of the UIUC-Thailand connection with medium-loss rates. For this connection, unconcealed losses are frequent enough to warrant the occasional use of three-way redundancy. Our scheme also performs better than other schemes in most cases.

4. REFERENCES

[1] D. Lin and B. W. Wah, "LSP-based multiple-description coding for real-time low bit-rate voice over IP," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 167–178, Feb. 2005.

[2] A. W. Rix, J. G. Beerends, M. P.Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a. new method for speech quality assesment of telephone networks and codecs," *J. of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, Oct. 2002.