SPEECH- AND NETWORK-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
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{batusat, wah}@uiuc.edu

ABSTRACT

In this paper, we propose a layered CELP speech coding (LC) scheme that adapts dynamically to the characteristics of the speech encoded and the network loss conditions in real time transmissions of voice over IP. Based on the ITU G.729 CS-ACELP codec operating at 8 Kbps, we design a variable bit-rate codec that is robust to losses and delays in IP networks. To cope with bursty losses while maintaining an acceptable end-to-end delay, our scheme employs LC with redundant piggybacking of perceptually important parameters in the base layer, with a degree of redundancy adjusted according to feedbacks from receivers. Under various delay constraints, we study trade-offs between the additional bit rate required for redundant piggybacking and the protection of perceptually important parameters. Experimental results show that our scheme works well and has quality comparable to full replication.

1. INTRODUCTION

Background. In this paper, we study the loss concealment of ITU G.729 coded speech transmitted in real time by voice-over-IP (VoIP). These transmissions may suffer quality degradations when packets are lost or delayed because pervasive dependencies in low bit-rate speech coding may lead to sustained distortions over a number of consecutive frames. Recovering losses from information received is also difficult because most implicit redundancies have been removed by the encoder in order to achieve a high coding efficiency. As a result, the minimal protection provided by the built-in loss-concealment algorithm in G.729 does not perform well even under low-loss scenarios.

Our analysis on the Internet shows the following.

(a) The loss rate can be highly varying. Our measurements of transmissions to some international sites show a sustained loss rate of 50% or more and bursty losses of three or more consecutive packets. Moreover, the loss behavior is non-stationary and connection dependent. Such a loss behavior precludes the use of source coding methods.

(b) The end-to-end delay of an IP packet can be highly varying and can range from tens to hundreds of milliseconds within a short duration. Since ITU G.114 specifies the worst-case one-way delay in real-time speech to be 400 ms, loss concealment cannot be accomplished by retransmissions in TCP but must be done by embedding explicit redundancies in the UDP packets sent. Further, jitter buffers must be used at receivers to smooth out irregular arrivals.

(c) The loss rate may go up dramatically when UDP packets are transmitted at a very high rate (say 100 packets/sec). Since G.729 encodes speech frames of 10-ms duration each into a set of 10-byte parameters, multiple frames (called *group of frames* or GOF) will have to be placed in one packet in order to reduce the packet rate. The disadvantage of this approach is that multiple frames in close proximity will be lost when a single packet is lost.

(d) The loss behavior is not sensitive to the packet size as long as it is within the MTU.[1] Hence, one may use *piggybacking* to place duplicate copies of GOFs transmitted in the past in each packet in order to ensure that at least one copy of each frame will reach the receiver on time. In practice, only a small number of past GOFs will need to be duplicated because the prescribed end-to-end delay will dictate an upper bound on the delay a frame can tolerate.

Problem statement. The above observations lead to unique requirements on the design of speech codecs for VoIP applications. In our previous work, we have chosen a fixed 8-Kbps bit rate in the design of a speech-adaptive layered G.729 codec [1] for concealing losses at receivers.

RESEARCH SUPPORTED BY THE MOTOROLA CENTER FOR COMMUNICATIONS, UNIV. OF ILLINOIS, URBANA-CHAMPAIGN.
IEEE WORKSHOP ON MULTIMEDIA SIGNAL PROC., 2005.

[1] The MTU (maximum transfer unit) in IPv4 is 576 octets. This is the largest value to ensure that an IP packet will not be fragmented.

To conceal information lost in heavy-loss scenarios without increasing the bit rate, we extend the size of each frame in order to create additional bit space for carrying redundant information. The increased frame size leads to a dramatic degradation in the coded speech quality, even under no loss. Obviously, according to (d) above, a constant bit rate is unnecessary in designing a speech codec for VoIP applications in the Internet. To this end, we study in this paper the design of a speech codec with a relaxed bit-rate requirement, as well as a network-adaptive piggybacking scheme to combat the non-stationarity of packet losses and delays.

Without indiscriminately increasing the bit rate, we study two related issues in the design of a robust G.729 coder that allows loss concealments at receivers without retransmissions. Assuming redundant copies of GOFs transmitted in the past are piggybacked in each UDP packet transmitted, we first investigate trade-offs between the degree of piggybacking and the *unconcealable frame loss rate* (*UFLR* or fraction of frames whose loss cannot be recovered from the redundant frames piggybacked later) over a wide range of loss conditions and delays. Figure 1 illustrates the grouping of three G.729 frames into a GOF and the placement of redundant copies of two previous GOFs in the current UDP packet transmitted. Next, we study trade-offs between the importance of a parameter encoded by G.729 and the number of bits required to protect it by a redundant copy. We protect parameters that have the largest impact on perceptual quality at receivers, while saving the bit space by not protecting perceptually unimportant parameters. We measure perceptual quality by the ITU P.862 PESQ objective measure [2] designed for evaluating the perceptual quality of speech coded by low bit-rate coders.

We compare the quality of our integrated piggybacking and coding scheme with that of a reference scheme which places redundant copies of as many GOFs as possible in each packet, up to the MTU allowed. By adapting the degree of redundancy in each piggybacked packet using feedbacks from receivers and by protecting only perceptually important parameters, our scheme will have a perceptual quality very close to that of the reference scheme, while requiring a bit rate close to that of the original G.729.

2. LOSS-ADAPTIVE REDUNDANT PIGGYBACKING UNDER DELAY CONSTRAINTS

In this section, we present a loss-adaptive redundant piggybacking scheme that aims to reduce the UFLR to an acceptable level within a given delay constraint. Observations reveal that the distortion effects due to an isolated lost frame last for about ten frames [1] or until a voiced-unvoiced speech boundary is reached. Moreover, if a loss is unconcealed when the distortion effects still persist from previous frames, then the accumulated effect will degrade

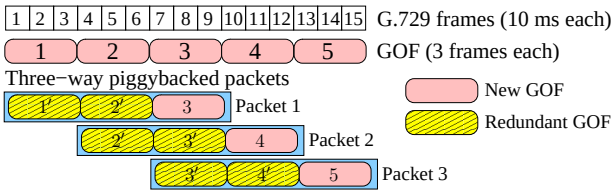

Figure 1: Piggybacking of three GOFs in each packet transmitted.

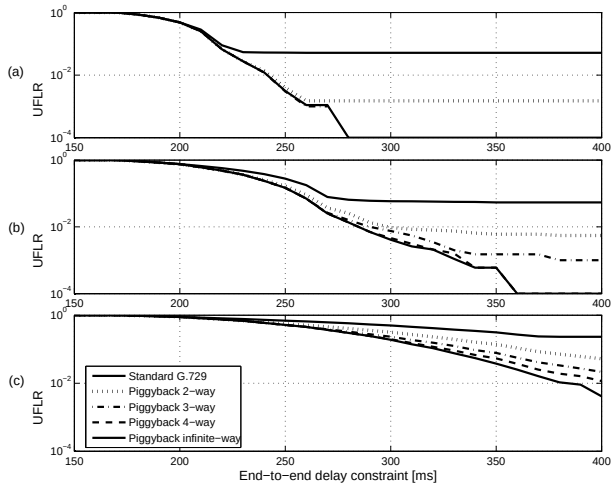

Figure 2: UFLR with respect to end-to-end delays for piggybacking under different degrees of redundancy between UIUC and a) Taiwan, b) Thailand, and c) Argentina. The traces were collected in April 2003.

quality further. As a result, we deem a 5% UFLR to be acceptable in order to limit the probability of convoluted distortions in multiple losses.

Figure 2 depicts the effect of redundant piggybacking of 2,000 UDP packets on the UFLR for various end-to-end delays between UIUC and Taiwan, Thailand, and Argentina (representing, respectively, low, medium, and high loss scenarios). For any delay constraint, a scheme with a higher degree of piggybacking always performs as good as or better than schemes with a lower degree. Under very tight delay constraints, piggybacking does not help in loss concealments because late arrivals dominate most losses. In that case, the performance of a non-redundant scheme and that of redundant piggybacking with an infinite degree are identical. In contrast, a higher piggybacking degree is beneficial for reducing the UFLR when the delay constraint is relaxed. Under very loose delay constraints, no packet is late, and only long bursty losses can cause unconcealable losses. The results also show that, for a given delay constraint, the degree of piggybacking required to keep the UFLR below 5% is connection dependent. Hence, it is important for receivers to feed this information back to senders and for senders to adapt its piggybacking degree dynamically.

Figure 3a depicts the variations in the UFLR for different piggybacking schemes and a medium-loss connection to Thailand with a 300-ms delay constraint. It shows that the

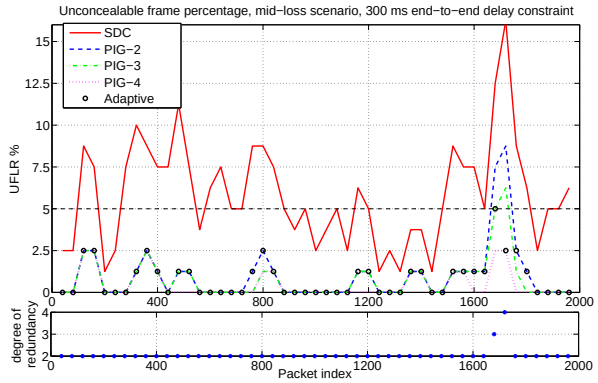

Figure 3: Performance of static and network-adaptive redundant piggybacking schemes: a) UFLR; b) degree of redundancy in network-adaptive piggybacking scheme (each point is an average over a sliding window of 80 packets, with 50% overlap).

UFLR for SDC (single description coding with no redundancy) is unpredictable and wildly varying, leading to unstable and inferior speech quality at receivers. On the other hand, redundant piggybacking with a fixed degree results in smaller and infrequent changes in the UFLR, although the loss rate cannot be guaranteed to be always below 5%.

To ensure a stable UFLR, we propose a network adaptive scheme that uses the smallest degree of piggybacking to achieve an UFLR of 5%. The scheme is updated periodically (every 40 packets) using information in the past 80 packets. We only consider two-way, three-way and four-way piggybacking because consecutive losses of four or more packets are rarely encountered. Figure 3b shows the degree of piggybacking used in our network-adaptive scheme. The resulting UFLR experienced is further indicated by circles in Figure 3a. Using infrequent feedbacks to senders, our network-adaptive scheme allows the UFLR to be bounded to below 5%, except for infrequent cases with four or more consecutive lost packets.

3. SPEECH-ADAPTIVE LAYERED CODING

Based on redundant piggybacking, we present in this section some protection schemes that exploit the trade-offs between bit rate and perceptual quality. We identify G.729 parameters that have the largest impact on perceptual quality and protect them by redundant piggybacking at senders.

Layered coding (LC) divides an information stream into individually decodable units for the purpose of applying different protection techniques, depending on their importance. The most important information is placed in the base layer, whereas other information is placed in enhancement layers. Our approach is to classify G.729 parameters in each frame into layers such that the base layer has the most perceptually important parameters, while enhancement layers carry less perceptually important parameters.

Table 1: Various schemes for protecting the parameters in voiced and unvoiced frames. (PP&PG: pitch period & gain; AP&AG: ACB pulses & gain; S: built-in loss concealment; R: replicated;).

Scheme	Voiced				Unvoiced				Onset		Silence
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
LPC	S	R	S	R	S	S	R	R	S	R	S
PP&PG	R	R	R	R	S	R	S	R	S	R	S
AP&AG	S	S	R	R	R	R	R	R	R	R	S

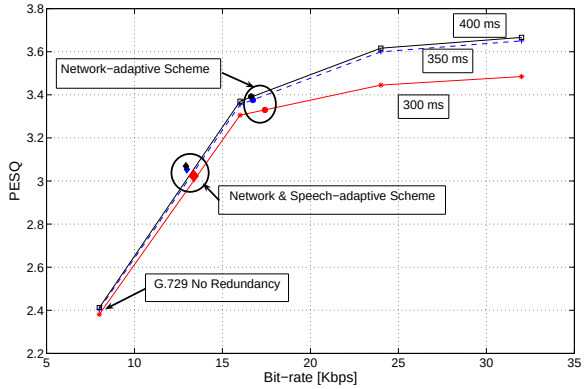

Figure 4: Bit rate-quality trade-offs for piggybacking, network-adaptive and network- and speech-adaptive schemes for a medium-loss scenario in Figure 2b with, respectively, 300, 350, and 400 ms end-to-end delays.

We have shown in our previous work [1] that, in low to medium loss scenarios, excitation parameters in G.729 are more important than LPC in terms of perceived speech quality. This observation leads to a **hybrid-LC scheme** that places LPC in the enhancement layer and interleaves them during packetization, while placing the excitation parameters in the base layer and replicating them in multiple packets according to the degree of interleaving.

We have also found that the most important parameter for voiced (*resp.*, unvoiced and onset) frames is the pitch (*resp.*, ACB and ACB) information. (An onset frame is the first voiced frame following a series of unvoiced frames.) Hence, in a **speech-adaptive LC scheme** [1], it first classifies a speech frame into voiced, unvoiced and onset, using the energy and the number of zero crossings in the frame. It then protects the pitch (*resp.*, ACB, ACB) information for voiced (*resp.*, unvoiced, onset) frames. This scheme is represented in Table 1 for the four classes of frames.

Under a fixed bit rate of 8 Kbps, we have further shown in our previous work [1] that the speech-adaptive LC scheme outperforms the built-in loss concealment of G.729, the hybrid-LC scheme, and the full-replication scheme in low- and medium-loss scenarios, but fails to do so under heavy losses. As the bit rate is fixed in these schemes, they must extend their frames and subframes in order to create additional space for carrying redundant information. The speech-adaptive scheme has better quality because it selectively protects critical parameters depending on the voice

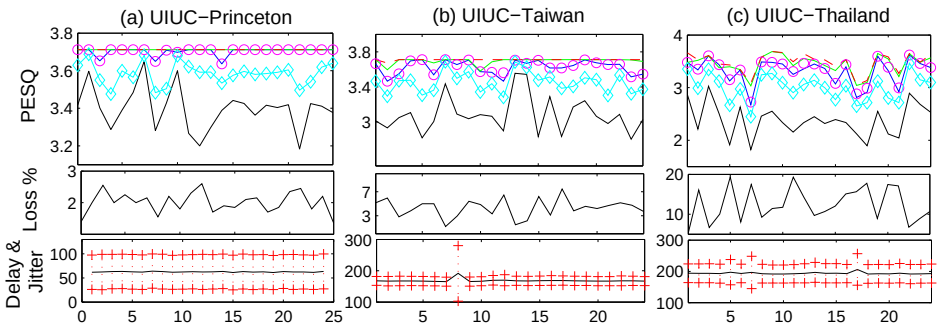

Figure 5: PESQ of received sequences [3] between UIUC and three destinations evaluated at each hour over a 24-hour period, using traces collected in April 2003 and a 300-ms delay constraint. In each PESQ plot in the top row, solid black: ITU G729 (8 Kbps); blue: two-way piggybacking (16 Kbps); green: three-way piggybacking (24 Kbps); red: four-way piggybacking (32 Kbps); circles: our network-adaptive piggybacking with full replication (average bit rate a) 16K, b) 16K, c) 16.2K); diamonds: our speech- and network-adaptive LC (average bit rate a) 12.3K, b) 12.3K, c) 12.6K). Second row: packet loss rates. Third row: average and standard deviation of network delays.

characteristics. As a result, it requires a smaller space for redundancy and, consequently, less increase in the frame and subframe sizes in order keep the bit rate fixed at 8 Kbps. However, because of increased frames and subframes, it performs worse than SDC under no loss. Moreover, it does not perform well in heavy-loss scenarios because the amount of protection is inadequate under the fixed bit rate.

To address these issues, we relax the requirement on bit rate and use piggybacking to carry redundant information of previous GOFs. Table 1 shows the trade-offs between the bit rate and perceptual quality for the various schemes.

Our study has shown that LPC can lead to better perceptual quality under any loss scenario with a minimal increase in bit rate. Hence, in addition to the information in the speech-adaptive LC scheme, we include LPC in Schemes (b) and (h) for, respectively, voiced and unvoiced frames.

Our study has also shown that onset frames are associated with high ACB gain and large changes in LPC parameters and pitch gain. As they cannot be predicted accurately by the standard loss-concealment algorithm, we always replicate them in piggybacking (Scheme (j)). Further, our scheme identifies silence frames and applies no replication on those frames (Scheme (k)).

4. EXPERIMENTAL RESULTS

Figure 4 shows the trade-offs between the bit rate and the perceptual quality among the various schemes, using UDP packet traces between UIUC and Thailand, for speech sequences in [3] and three end-to-end delay constraints.

The figure shows the rate-quality trade-offs among the six schemes under a 300-ms delay constraint: four-way piggybacking (32 Kbps, PESQ of 3.48), three-way piggybacking (24 Kbps, PESQ of 3.44), full-replication with network adaptation (17.4 Kbps and PESQ of 3.33), two-way piggybacking (16 Kbps, PESQ of 3.3), our proposed scheme with speech and network adaptation (13.3 Kbps and PESQ of 3.03), and the original SDC with built-in loss concealment (8 Kbps and PESQ of 2.39). It demonstrates that our proposed scheme requires 24% less bit rate than full replication, yet has only 9% degradation in PESQ.

Using UDP packet traces between UIUC and Princeton, Taiwan, and Thailand and a 300-ms delay constraint, Figure 5 further details the perceptual quality of our proposed network and speech adaptive LC scheme, the network-adaptive full-replication scheme, the G.729 with built-in loss concealment, and various piggybacking schemes.

For connections with a very low loss rate and a low average delay, Figures 5a and 5b show that our network-adaptive scheme achieves quality similar to that of four-way piggybacking with half the bit-rate, and that our speech-adaptive scheme achieves robust and acceptable quality with about 47% further reduction in redundant information.

For the connection with medium-loss rate and substantially higher delays and jitters, Figure 5c shows that, with an end-to-end delay constraint of 300 ms, some packets are late and are not useful for play-out, and even four-way piggybacking cannot achieve the maximum quality. Our network-adaptive scheme however, performs close to the four-way scheme with about half the bit-rate, while using a higher degree of redundancy only when needed. Our speech-adaptive scheme achieves similar and acceptable performance with 44% further reduction in redundant information.

5. REFERENCES

[1] B. Sat and B. W. Wah, "Speech-adaptive layered G.729 coder for loss concealments of real-time voice over IP," in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, July 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.

[3] 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.