IMPROVING 16 KB/S G.728 LD-CELP SPEECH CODER FOR FRAME ERASURE CHANNELS

Craig R. Watkins† and Juin-Hwey Chen

Speech Coding Research Department
AT&T Bell Laboratories
Murray Hill, New Jersey, USA

ABSTRACT

We have improved G.728 output speech quality for frame erasure channels. Three cases are considered: (1) no change to G.728, (2) change only the G.728 decoder, and (3) change both the encoder and decoder. In case 1, we synthesize a bit-stream during erased frames so that the decoder decodes an excitation with low energy or with characteristics similar to the excitation of previous good frames. In case 2, the gain-scaled excitation and LPC coefficients are extrapolated, and vital operations of backward LPC and gain adaptations are continued. Case 3 adds spectral smoothing and increases bandwidth expansion for the LPC and gain predictors. These techniques are quite effective, as the speech quality degradation due to 1% frame erasures ranges from just slightly noticeable in case 1 to almost unnoticeable in case 3. For case 3, the output speech is still intelligible for frame erasure rates up to 10% or even 20%.

1. INTRODUCTION

Frame erasure is a common condition in wireless communication systems such as Personal Communications Systems (PCS) and Future Public Land Mobile Telecommunication Systems (FPLMTS). It refers to the situation when entire frames of bits are lost or unreliable at the receiver (and thus considered "erased"). The ITU-T (formerly CCITT) G.728 16 kb/s Low-Delay CELP speech coding standard [1] was originally designed to be robust to random bit errors. However, handling frame erasures was not a design consideration when it was created.

The ITU-R (formerly CCIR) is looking for a speech coder suitable for FPLMTS applications and has requested that the ITU-T standardize an extension of G.728 that is capable of handling frame erasures in radio channels. This extension of G.728 is now in the ITU-T's official agenda for standardization. The performance requirement is no more than 0.5 Mean Opinion Score (MOS) degradation from the MOS of clear-channel G.726 32 kb/s ADPCM, for both random and bursty frame erasures at a 3% erasure rate and a frame size of 10 ms. It is assumed that some form of error detection is provided externally to G.728, and it can reliably detect frame erasures. It is also assumed that each received frame is either totally good (no bit errors), or totally bad (no bits received or bit error rate approaching 50%).

To participate in this standardization effort, we developed several frame erasure concealment techniques for G.728. Since this work is about an extension of the existing G.728 standard, compatibility with G.728 is an important issue to consider. Therefore, we developed different techniques for each of the following three levels of compatibility.

(1) Strict compatibility: Neither the G.728 encoder nor the G.728 decoder is allowed to change. Output speech quality enhancement is achieved by using a decoder preprocessor to intercept the received bit-stream and synthesize a bit-stream when a frame is erased.

(2) Bit-stream compatibility: No change is allowed in the G.728 encoder, but the internal operations of the decoder can be modified. The modified decoder can still interoperate with any G.728 encoder.

(3) No compatibility: Both the G.728 encoder and the G.728 decoder are changed. The backward compatibility with G.728 is sacrificed in order to accomplish better robustness to frame erasures.

The preprocessor approach for the strict compatibility case (case 1 above) is useful in at least two situations. First, if G.728 becomes available on a low-cost Application Specific Integrated Circuit (ASIC) chip, it is not likely that a user of the chip can change the internal operations of the G.728 encoder and decoder inside the chip. Second, if a designer buys a commercial G.728 DSP object code to implement G.728 on a DSP chip, the operations of the G.728 encoder and decoder still cannot be changed, unless the designer has access to the DSP source code for G.728. In either case, the G.728 output speech quality can still be improved if we use a preprocessor to manipulate the received bit-stream to influence the G.728 decoder output in a desirable way.

In the three levels of compatibility listed above, the speech quality improves as we go from case 1 to case 3. Case 1 works well for a frame erasure rate of 1% or below, but the performance is not adequate at 3% frame erasures. Case 3 gives a very robust coder against frame erasures, but the compatibility with G.728 is completely lost. Case 2 appears to be the most appropriate solution. Therefore, we concentrated most of our effort on case 2 and produced a modified G.728 decoder that met the performance requirement for the ITU-T extension of G.728. This was achieved without increasing the decoder complexity or sacrificing the bit-stream compatibility with G.728. AT&T has submitted this modified G.728 decoder as a candidate for the ITU-T standard for FPLMTS extension of G.728 [2].

This paper is organized as follows. In Section 2, we discuss the problems the G.728 decoder faces during and after frame erasures. Section 3 describes the modified G.728 decoder submitted by AT&T. Section 4 discusses other techniques we studied that are

† Work done at AT&T Bell Labs while on leave from Australian National University

0-7803-2431-5/95 $4.00 © 1995 IEEE

not part of our proposal to ITU-T: Kalman filtering, the decoder preprocessor, and the G.728 encoder changes. Finally, we report in Section 5 some simulation and MOS test results.

2. G.728 DURING AND AFTER FRAME ERASURES

The G.728 decoder faces two problems when a frame is erased. First, it loses all the bits corresponding to a large consecutive block of excitation signal samples. It needs to get a substitute excitation signal that would give as little speech quality degradation as possible. Second, during frame erasures, the G.728 decoder loses track of all internal variables in the backward adaptation of the LPC filter and gain predictor. When the next good frame comes, these variables differ from those of the encoder and may take a while to re-converge. Hence, re-convergence of the decoder backward adaptation variables after frame erasures is an important issue for G.728. (This is not a problem for forward-adaptive CELP, because the filter coefficients are immediately "refreshed" to the encoder values by decoding the received bits, and the filter memories usually re-converge within a short period of time.)

3. G.728 DECODER MODIFICATIONS

Our solution to the decoder problems mentioned above involves special techniques to handle the excitation, the LPC filter, and the backward adaptation schemes for LPC and gain predictors. To maintain the bit-stream compatibility with G.728, these techniques are used only in the decoder and only during erased frames.

3.1 Excitation Extrapolation

During erased frames, we extrapolate the gain-scaled excitation signal (rather than the excitation VQ codebook output before the gain scaling unit). We use the pitch predictor information of the G.728 postfilter to control such extrapolation. Let p and β be the pitch period and the optimal first-order pitch predictor tap of the last frame [1]. The voicing threshold V is chosen to be 0.6/1.4. If $\beta > V$ in the last frame, we periodically repeat a scaled-down version of the last p samples in the gain-scaled excitation buffer. The scaling factor is 0.8. Such periodic extrapolation continues until a good frame is received. The extrapolated excitation is used to update the gain-scaled excitation buffer and is treated as if it were the true excitation.

If $\beta \le V$ in the last frame, we do not perform periodic extrapolation. We first developed an extrapolation scheme based on waveform matching [3]. This was used in two MOS tests in 1993. Later, we developed a much simplified version which gave equal or better speech quality in informal A-B comparisons. In this version, each 5-dimensional excitation vector is extrapolated by randomly selecting 5 consecutive samples from the last 40 samples in the gain-scaled excitation buffer. Each extrapolated excitation vector is individually scaled such that it has the same average magnitude as the average magnitude of the last 40 samples of the excitation of the last frame. Again, the extrapolated excitation vectors are used in extrapolating future vectors.

3.2 LPC Filter

When a frame is erased, we use bandwidth expansion to "soften" the LPC coefficients in the last good frame. The bandwidth expansion factor is 0.97. Specifically, the i-th LPC coefficient a_i is replaced by $a_i' = (0.97)^i a_i$, $i = 1, 2, ..., 50$. This new set of $\{a_i'\}$ is used throughout the entire bad frame. Then, for each subsequent and consecutive erased frames, this set of $\{a_i'\}$ is further bandwidth expanded by a factor of 0.97. That is, in the k-th consecutive bad frame, the LPC coefficients are bandwidth expanded by a factor of $(0.97)^k$. This continues until the next good frame is received. When a bad frame is encountered again, the process start over again from a bandwidth expansion factor of 0.97. Note that no such bandwidth expansion of the gain predictor is necessary during erased frames. Since we directly extrapolated the gain-scaled excitation, there is no need to produce the backward-adapted gain or to update the gain predictor coefficients.

3.3 Backward Adaptation of LPC and Gain Predictors

In erased frames, we do not perform the usual backward LPC and gain adaptation. However, we still continue to update some internal variables of the backward adaptation schemes just to "keep them alive". If we kept all internal states "frozen", then they would be likely to have large discontinuities when the next good frame comes. This could cause slower convergence of decoder states after frame erasures. Therefore, it is advantageous to continue the vital operations of backward adaptation which would affect the backward adaptation in future frames.

Such vital operations include updating the filter memory of the gain predictor and the internal state variables of hybrid windows [1], such as the recursive components of autocorrelation coefficients, and the synthesized speech and logarithmic gains stored in the buffers. These updates are performed by using the extrapolated excitation as if it were the true excitation. Many computationally intensive operations for backward adaptation (such as Durbin's recursion) need not be done during erased frames. We can use the saved computation (or DSP processor time) to perform excitation extrapolation and LPC bandwidth expansion, both of which take very little computation. This is why our G.728 decoder modifications do not increase the computational complexity.

3.4 Postfilter

During frame erasures, we do not "freeze" or bandwidth expand the postfilter coefficients. Instead, we perform the functions of the postfilter exactly the same way as in good frames. This means that the postfilter coefficients are updated based on the speech synthesized from the extrapolated excitation. This avoids the potential cancellation or de-emphasis of the spectral peaks of the synthesized speech by the spectral valleys in the postfilter frequency response due to a mismatch of the two.

4. OTHER TECHNIQUES STUDIED

4.1 Kalman Filtering

We tried to replace the LPC filter and gain predictor by Kalman filters to provide some smoothing to the extrapolated excitation and the corresponding log-gain sequence. Such Kalman filtering techniques generally improved the SNR of the decoded speech. Unfortunately, such an SNR improvement did not translate into statistically significant improvement in the MOS scores. Therefore, we did not include Kalman filtering in our proposal.

4.2 Decoder Preprocessor Techniques

Suppose we directly use G.728 without any modification, then a frame erasure can be assumed to correspond to random bits being at the G.728 decoder input. The corresponding decoder output speech often has very severe distortions, characterized by many short duration "explosions." This is because the random bits are often decoded into a high-magnitude excitation signal. Such a "do nothing" situation gives us the worst case scenario. An obvious and easy way to improve this is to modify the G.728 decoder and set the excitation signal to zero during erased frames. This indeed improves the speech quality significantly and eliminates the "explosions." However, it requires a change to the internal operation of the G.728 decoder, and it does not work as well as the techniques described in Section 3 above.

An effect similar to such a zero excitation can be achieved by a decoder preprocessor without the need to change the G.728 decoder. The preprocessor intercepts the received bit-stream to process it before passing it on to the G.728 decoder. In good frames, the bit-stream is passed on unmodified. During erased frames, the preprocessor receives random bits, and it "masks out" the two magnitude bits of each 10-bit G.728 excitation channel index codeword so that the two magnitude bits are zero. With such a modified bit-stream, the G.728 decoder will always use the smallest magnitude from the 2-bit magnitude codebook. After a few excitation vectors, the backward gain adaptation will bring the gain-scaled excitation signal to almost zero. Thus, such a preprocessor achieves speech quality improvement similar to (actually slightly better than) the zero excitation approach.

In fact, we can further improve the speech quality by generalizing this preprocessor idea. For example, during good frames, while passing on the bit-stream unchanged, the preprocessor can locally decode and store the excitation signal (without scaling by the backward-adapted gain). When a bad frame comes, the preprocessor can extrapolate the stored excitation signal of previous good frames into the current frame, using extrapolation techniques similar to those described in Section 3.1. Next, the preprocessor can directly perform vector quantization of the extrapolated excitation signal using the G.728 excitation codebook and a simple mean-square error (MSE) distortion measure. The resulting codebook indices form the bit-stream that is passed to the G.728 decoder. The decoder will decode the bit-stream into the quantized version of the extrapolated excitation signal. This approach requires a higher preprocessor complexity, but it will produce better speech quality. This generalized preprocessor idea can be extended to other internal coder variables or even to other speech coders under similar compatibility constraints.

4.3 G.728 Encoder Modifications

By relaxing the compatibility constraint and allowing modifications in the encoder, we can further enhance the robustness against frame erasures, especially when frame erasures occur at the beginning of a talk spurt. When this happens, the decoder has nothing useful to extrapolate with, and the synthesized speech has a large distortion. A much more serious consequence is that even after the frame erasure is over, the decoder internal states may take a long time to re-converge to the encoder states. We modified the backward adaptation algorithms in the encoder as well as the decoder to increase the inherent convergence speed of G.728. We achieved this by using increased bandwidth expansion and the Spectral Smoothing Technique (SST) [4] for both the LPC and the gain predictors. Such changes gave a noticeable improvement in speech quality when the frame erasure rate was high (e.g. 10%). However, at 3% erasure rate, no meaningful improvement was observed. Thus, we could not justify these changes and did not include them in our ITU-T proposal.

5. SIMULATION AND MOS TEST RESULTS

Figures 1 through 5 show the speech waveforms obtained by our simulations of different conditions. Figure 1 shows 250 ms of clear-channel G.728 decoder output speech waveform. Figure 2 shows the same segment of waveform decoded with random bits used during the three erased frames around 30, 120, and 190 ms. The frame erasure rate is around 10%. In figures 3 through 5, the erasures occur at the same frames as in Fig. 2, but the waveforms are for the three levels of G.728 compatibility mentioned earlier: low-magnitude excitation (case 1), proposed decoder changes (case 2), and encoder and decoder changes (case 3). The waveform match becomes progressively better as we go from Fig. 2 to Fig. 5. It is quite remarkable that even at a very high frame erasure rate of 10%, the encoder and decoder changes (case 3) still preserve most characteristics of the speech waveform.

Our techniques were first tested in two MOS tests in October and November of 1993. The frame size was 10 ms. The frame erasure patterns were somewhat bursty, with typical bursts of two or three bad frames in a row, and the longest burst was 6 frames. Table 1 summarizes the resulting MOS scores. Each test used two kinds of source speech: Intermediate Reference System weighted (IRS) and unweighted (flat). The short-hand notations in Table 1 are: "Zero Exc" for the zero-excitation approach mentioned in Section 4.2, "D" for proposed decoder changes, "E+D" for encoder and decoder changes, and "D+KF" and "E+D+KF" are the same as "D" and "E+D", respectively, except that Kalman filtering is used. "D" is our proposal to ITU-T and is printed in bold-face along with the target of clear-channel G.726 ADPCM. Our modified G.728 decoder gave no MOS degradation at a frame erasure rate of 1%. It gave only very little MOS degradation (0.13 and 0.19) at a 3% erasure rate when compared with clear-channel G.726. This exceeds the ITU-R's target of 0.5 degradation. The encoder changes (SST and increased bandwidth expansion) improved the MOS significantly (by nearly 0.5) at an erasure rate of 10%, but they did not help at 3%. At 3% and 10% erasures, Kalman filtering did not give a statistically significant improvement of the MOS. Also note that the decoder changes alone easily beat a simple-minded zero-excitation strategy by a very large MOS margin (more than 1.3 for IRS-weighted speech).

Recently, another MOS test was conducted in September 1994. Both random and bursty frame erasures were tested. The bursty erasures have a much longer average burst length, with a large percentage of the bursts lasting more than 5 frames, and the maximum length was 10 frames (100 ms). The MOS results are shown in Table 2. Surprisingly, the MOS degradation relative to clear channel G.728 was much larger than the 1993 MOS results in Table 1. We do not know the real reasons for this. We can think of two possible reasons. First, the frame erasure bursts were much

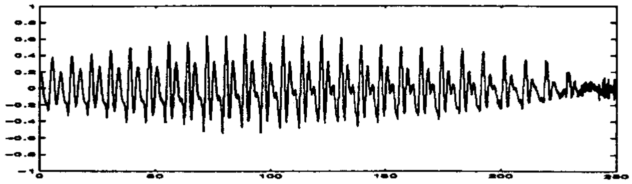

Fig. 1 Clear-channel G.728 decoder output waveform

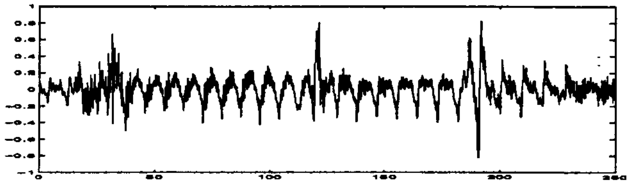

Fig. 2 Waveform produced by random channel index

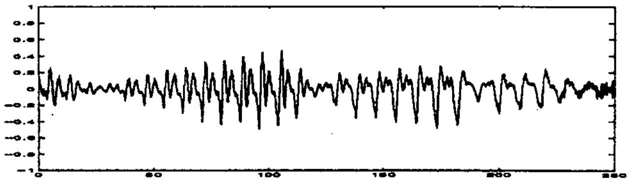

Fig. 3 Waveform produced by low-level random excitation

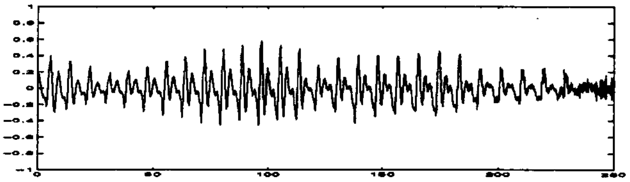

Fig. 4 Waveform produced by proposed decoder changes

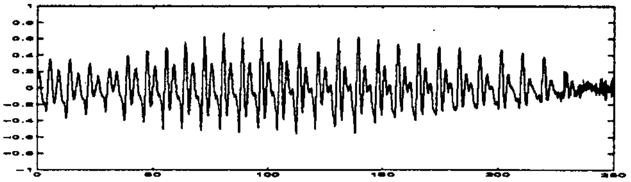

Fig. 5 Waveform produced by encoder and decoder changes

longer than in the 1993 tests; hence, the effects on the decoded speech are much more difficult to conceal. Second, the 1993 tests and the 1994 test had very different mixtures of coding conditions. The 1993 tests had many low-bit-rate coders (down to the 2.4 kb/s LPC vocoder) and had frame erasure rates as high as 10%. In contrast, the 1994 test contained mostly toll-quality speech coders and the maximum frame erasure rate was only 5%. When most stimuli sound good, the listeners may be more inclined to penalize the quality degradation more heavily.

A statistical analysis showed that the critical difference in MOS (for 95% confidence that the difference is "real") is about 0.5 for this test. G.726 has an MOS of 3.69; thus, any MOS greater than 2.69 is considered to be "no worse than 0.5 MOS lower than G.726" in a statistical sense. In any case, the ITU-T interpreted the MOS results as meeting the performance requirement.

Coder Condition	FER	10-93 MOS		11-93 MOS	
		IRS	Flat	IRS	Flat
D	**10%**	**2.27**	**2.55**	-	-
D+KF	10%	2.31	2.57	-	-
E+D+KF	10%	2.79	2.99	-	-
Zero Exc	3%	2.26	2.65	-	-
D	**3%**	-	-	**3.77**	**3.59**
D+KF	3%	3.58	3.51	3.70	3.61
E+D	3%	-	-	3.75	3.61
D	**1%**	-	-	**3.95**	**3.82**
G.728	0%	3.88	3.77	3.90	3.79
G.726	**0%**	**3.77**	**3.70**	**3.90**	**3.78**

Table 1 MOS Results for short bursts of frame erasures

Coder Condition	FER	Erasure	9-94 MOS
Decoder changes	3%	random	3.19
Decoder changes	3%	bursty	2.94
Decoder changes	1%	random	3.54
Decoder changes	1%	bursty	3.48
Low-magnitude excitation	1%	random	3.36
Low-magnitude excitation	1%	bursty	3.50
G.728	0%	-	3.78
G.726	0%	-	3.69

Table 2 MOS for random and long bursts of frame erasures

6. CONCLUSION

We have described several techniques for enhancing the G.728 coder's robustness against frame erasures. Different levels of G.728 compatibility require different techniques. Our modified G.728 decoder is now a candidate for ITU-T's FPLMTS extension of the G.728 standard. It meets ITU-T's performance requirement without sacrificing the bit-stream compatibility with G.728 and without increasing the overall computational complexity.

References

1. J.-H. Chen, R. V. Cox, Y.-C. Lin, N. S. Jayant, and M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE J. Selected Areas Communications*, pp. 830-849 (June, 1992).

2. AT&T, *High-level description of G.728 decoder modifications for frame erasure concealment*, Contribution to ITU-T SG15/Q5 (October 1994).

3. D.J. Goodman, G.B. Lockhart, O.J. Wasem, and W-C. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communication," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34, 6, pp. 1440-1448 (December 1986).

4. Y. Tohkura, F. Itakura, and S. Hashimoto, "Spectral smoothing technique in PARCOR speech analysis-synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-26, pp. 587-596 (December 1978).