A STUDY OF DESIGN COMPROMISES FOR SPEECH CODERS IN PACKET NETWORKS

Roch Lefebvre, Philippe Gournay*, Redwan Salami***

*University of Sherbrooke, Sherbrooke, Quebec, Canada
**VoiceAge Corporation, Montreal, Quebec, Canada

ABSTRACT

In this paper, we present an objective and subjective comparison of alternate methods for improving the robustness of speech coders in packet networks. Two approaches are considered: 1) adding redundancy in the packets to improve the robustness of a baseline encoder or 2) reducing (or eliminating) inter-frame dependencies at the encoder. It is shown that both approaches have to trade bit rate and/or delay for quality over lossy channels. Formal subjective tests clearly show that using relatively simple forward error correction methods, standard coders such as ITU-T recommendation G.729 can be made significantly more robust than "frame-independent" coders, at a lower or similar bit rate.

1. INTRODUCTION

Speech coders achieve bit rate reductions by expressing the speech signal redundancies into compact representations. Linear predictive (LP) coding is a widely used coding paradigm, where the speech signal is efficiently encoded as an excitation signal and a set of linear filters. Long-term (LTP) and short-term predictive filters are typically used. The prediction filter coefficients are transmitted once per frame or sub-frame. Significant gains in information transmission can be achieved since: 1) there are much fewer filter coefficients than there are speech samples per frame (or sub-frame), and 2) the excitation signal can be encoded very compactly, for example as a set of (few) non-zero pulses as in ACELP coders [1]. Predictive coding actually also applies to the prediction filter coefficient themselves: in the LSF domain [4], a quantization gain can be achieved by using predictive VQ instead of memoryless VQ.

Since all predictors have memory, the coding gain of LP models is somewhat offset by a lower resilience to channel errors. In particular, in the case of missing speech frames, the decoder becomes "desynchronized" from the encoder. This is because, when a missing frame is detected, the decoder applies concealment, which is essentially a set of extrapolation techniques to obtain the missing frame from past frames information. Furthermore, this desynchronization propagates to several frames after a lost frame due to a mismatch between the filter memories. The impact is most important for LTP filters, or equivalently for adaptive codebooks, whose content is the past excitation signal. When the LTP filter memories become desynchronized at the encoder and at the decoder, a difference of just a few samples can produce large errors in the synthesis signal and propagate over several frames after a lost frame.

Since lost frames will always occur with a certain probability, tailoring speech coders for packet networks requires proper attention. There are invariably some compromises to consider. The end-to-end delay has to be maintained below an acceptable level for real-time, duplex communications. The bit rate has to fit the particular system capacity. But at the same time, the subjective quality must remain above an acceptable level even when there are significant packet losses. In general, these quality requirements imply higher bit rates (for the use of error protection, or to allow "non-predictive" coders). And using longer delays improves the decoder's resilience to packet losses (better interpolation for concealment, multiplexed redundant frames to protect against consecutive frame losses as in [2]).

At a given "operating point", i.e. for a given set of bit rate, delay and quality constraints, there are several solutions to consider. We propose to divide these solutions in two classes: 1) solutions that add redundancy in the packets to improve the robustness of a baseline encoder or 2) solutions that reduce (or eliminate) inter-frame dependencies at the encoder. To illustrate the design compromises that can be made at similar bit rates and delays, we present in this paper several techniques to improve the robustness of recommendation G.729 in the presence of missing packets (solutions that fall in the first class). We compare these solutions to a recently proposed speech coder that falls into the second class of solutions: the Internet Low Bitrate Coder (iLBC) [3].

The paper is organized as follows. Section 2 briefly describes the attributes of the iLBC speech coder. In Section 3, we describe the proposed methods for improving the robustness of the G.729 speech coder, where the overall bit rate is lower or similar to the iLBC coder bit rate, and different delay constraints are considered. Sections 4 and 5 describe the formal listening test that was performed for comparison. A conclusion is then presented in section 6.

2. ILBC: A RECENT EXAMPLE OF "FRAME-INDEPENDENT" SPEECH CODER

One of the major characteristics of iLBC is the fact that there is no inter-frame dependencies [3]. Each 20 (or 30) ms frame is encoded separately from adjacent frames. Prediction is used within a frame to achieve good coding gains, but the predictors do not extend across frame boundaries. This means that the error caused at the decoder by the concealment algorithm in case of missing packets does not propagate, in principle, over the following frames, provided these frames are received properly. This is however not completely accurate, since iLBC uses an enhancement module to process the excitation signal at the

decoder, using a memory of up to 6 frames. The increased robustness of iLBC against packet loss comes at the expense of increased bit rate, since its performance in error-free conditions is equivalent to CELP-type codes at half the bit rate. The iLBC encoder can operate with a frame length of either 20 or 30 ms. The (fixed) bit rate depends on the frame length chosen. With 20 ms frames, the bit rate is 15.2 kbps, and with 30 ms frames the bit rate is 13.33 kbps. The decoder also requires an additional 5 ms to 10 ms lookahead, which makes the overall delay equal to 25 or 40 ms.

In the subjective test described in sections 4 and 5, we used the 20 ms version of iLBC, at 15.2 kbps.

3. PROPOSED METHODS FOR IMPROVING THE ROBUSTNESS OF G.729

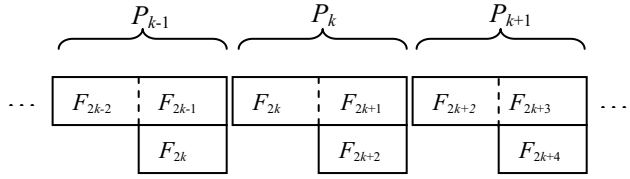
The G.729 speech coder [4] was standardized by ITU-T in 1995. It has been selected as one of the default speech coders for Voice over Frame Relay and in several VoIP applications. The main G.729 Recommendation operates at a bit rate of 8kbps, with a frame length of 10 ms and a lookahead of 5 ms at the encoder, resulting in an overall delay of 15 ms. With G.729 Annexes the codec can operate also at extended bit rates of 6.4 and 11.8 kbps. It can also operate in VAD/DTX/CNG mode for saving the average bit rate by reducing the bit rate of inactive speech periods. We consider only the 8 kbps mode here. The LP and LTP filters used in G.729 operate continuously on the speech signal, i.e. across the frame boundaries. In case of missing frames at the receiver, the decoder applies concealment and then continues decoding normally when later frames are correctly received. The concealment error can then propagate across several successive frames, as was discussed in the introduction. The G.729 concealment algorithm was designed to be optimal for 10 ms frames, i.e. for a relatively low probability of having two or more successive 10 ms frames lost.

In IP packets, several bytes are used for control and routing information (packet header). A 10 ms frame of speech encoded with G.729 at 8 kbps takes only 10 bytes. The ratio between control information and payload (here, the compressed speech) then becomes important, which increases the channel load. Hence, when using low rate coders, larger delays are generally considered, but still low enough to meet the end-to-end delay constraint. Packets of 20 ms are typically used. In what follows, we consider 20 ms packets. This is one of the possible frame lengths in the iLBC coder. To create packets of 20 ms with G.729, two consecutive 10 ms frames are sent in each packet. A direct impact is that each lost packet can generate two consecutive missing frames of 10 ms in duration.

To improve the robustness of G.729 at a total bit rate equivalent or lower than that of the iLBC coder, we propose four different approaches. We use the following notation: F_i will denote the i^{th} 10 ms frame at the G.729 encoder, F'_i will denote the same i^{th} 10 ms frame but without the 18 LSF bits, and P_k will denote the k^{th} 20 ms packet in the transmitted stream. Each 20 ms packet P_k will contain frames F_{2k} and F_{2k+1}, plus redundant information which depends on the approach used.

Approach 0: To serve as a benchmark, we first consider G.729 without any added redundancy. The payload of each 20 ms packet P_k will then contain only frames F_{2k} and F_{2k+1} produced by the G.729 encoder. When a missing packet is detected at the receiver, the decoder applies the G.729 concealment algorithm for the two consecutive missing frames. The bit rate is thus 8 kbps, and the total delay is 25 ms (two frames plus 5 ms look ahead).

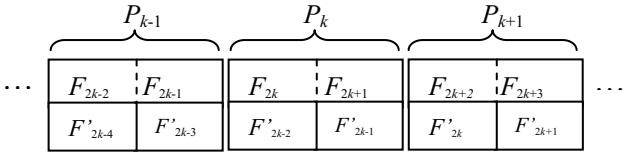
Approach 1: We extend the nominal 8 kbps bit stream of G.729 by repeating in each 20 ms packet P_k the first 10 ms frame of the next packet P_{k+1}. The payload in the packet stream has the following structure:



This requires an additional 10 ms delay at the encoder (to encode, for example, F_{2k} and add it as redundant information in packet P_{k-1}). There is no additional delay required at the decoder. The total delay is thus 35 ms, i.e. 3 frames of 10 ms plus 5 ms of look ahead. Since each packet carries half the next packet as redundancy, the total bit rate is 50% greater than the baseline bit rate. For G.729, this translates into 12 kbps.

When a missing packet occurs at the decoder, the first 10 ms frame of the missing packet can be recovered from the previous received packet. For example, if packet P_k is missing and P_{k-1} and P_{k+1} are received, then concealment only has to be applied to frame F_{2k+1}. This means that in case of single packet losses, there are only 10 ms frame losses even though the packets are 20 ms in duration. In fact, in this approach, G.729 reverts to its normal operation since its concealment was optimized with 10 ms frames (equivalent to using 10 ms packets). Nevertheless, the bit rate increase from 8 to 12 kbps is still more efficient than using 10 ms packets since it results in higher bit rate increase overall if we consider the packet header information.

Approach 2: An alternative approach consists in repeating partial information for all frames instead of all information for a single frame. Specifically, the packet structure now looks as follows (note that each packet repeats information from the previous packet and not the next packet as in approach 1 – this will allow different design compromises by playing on the delay at the decoder):

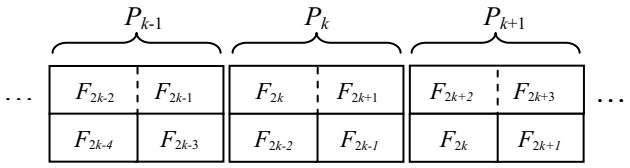


Here, F'_k is the same as F_k but with some information removed. As defined above, we choose F'_k to denote frame F_k but without the 18 LSF bits and pitch parity bit. Hence, each 20 ms packet contains 80+80+61+61 bits, which translates into a total bit rate of 14.1 kbps. The total delay will depend on the compromises made at the decoder. The shortest delay is 25 ms: encoding a 20 ms packet, plus 5 ms overhead. With no additional delay at the decoder (unlike in approach 3 described below), the redundant information cannot be used during a missing packet. So, for example, if packet P_k is missing at the

decoder, concealment has to be applied on the whole 20 ms packet since the redundant information of packet P_k is actually transmitted in the next packet, P_{k+1}. However, before decoding packet P_{k+1} (provided it is properly received), packet P_k can be "re-decoded" using F'_{2k} and F'_{2k+1} to allow the decoder to properly update its filter memories (adaptive codebook memory) before decoding packet P_{k+1}. This memory update process was recently described in [5], for use in another context (using late arrival frames at the decoder to limit error propagation after a concealed frame).

Approach 3: Here, encoding is identical to approach 3, but the decoder uses an additional 20 ms delay. Now the decoder waits for the arrival of packet P_{k+1} before decoding and synthesizing packet P_k. The bit rate and packet format are exactly the same as in approach 2. However, with the additional 20 ms delay, the decoder can properly decode a missing packet since it is re-transmitted in the next packet (except for the LP filter).

Approach 4: In this last approach, we completely duplicate each packet in the next packet as redundant information, and use the same delay as in approach 3. The packet structure is now as follows:



With this approach, the decoder can completely eliminate single packet losses.

Table 1 summarizes the attributes of the approaches described above, and gives their corresponding bit rate and total delay. G.729-0 refers to approach 0 in Section 3, G.729-1 corresponds to approach 1, and so on.

Method	Bit rate (kbps)	Delay (ms)
G729-0	8	25
G729-1	12	35
G.729-2	14.1	25
G.729-3	14.1	45
G.729-4	16	45
iLBC	15.2	25

Table 1. Attributes of the proposed approaches

Figure 1 shows signal examples for all the coding methods and redundancy approaches presented above. We see both the effect of concealment and error propagation across frames. The error signals (curves (b) to (g)) refer to the difference between the synthesis signals produced at the decoder with and without packet losses. All curves are shown at the same scale. In the figure, only the 3rd packet was lost. We see that for G.729 (curve (b)), the error due to concealing 20 ms propagates over several frames. Curve (c) shows how this can be reduced by approach 1, where 20 ms packet losses actually are transformed into 10 ms frame losses. Curves (d) and (e) show the effect of the delay tradeoff: in curve (d), which uses a 25 ms delay, concealment is applied but a memory update mechanism allows fast convergence after the packet loss, while in curve (e), which uses a 45 ms delay, the missing packet can actually be decoded properly (except for the LPC filter) because the packet was repeated in the next packet. With approach 4 (curve (f)), a single packet loss has no impact, which translates into a null synthesis error. Finally, the synthesis error for the iLBC coder is shown in curve (g). Note that the error also propagates over more than the missing packet. In terms of bit rate, recall that curve (b) requires 8 kbps, curve (c) requires 12 kbps, curves (d) and (e) require 14.1 kbps, curve (f) 16 kbps and curve (g) 15.2 kbps.

4. SUBJECTIVE EXPERIMENT

To compare the subjective quality of the proposed methods, a formal listening test was performed following the guidelines of ITU-T recommendation P.800 [6]. The test was realized in an isolated listening room, using binaural headphones. The test material included 4 male and 4 female speakers, with 6 sentence pairs for each speaker. All sentence pairs were formed of clean speech samples, with a sampling frequency of 8000 Hz. The samples were pre-processed using a modified IRS filter [7]. A total of 32 listeners were used. There were 36 conditions in all, including 6 MNRU reference conditions for calibration. The Frame Erasure Rate (FER) profiles were exactly the same for all coders. This means that, when looking at a specific FER value, the missing frames occurred at exactly the same speech samples for iLBC and all approaches involving G.729.

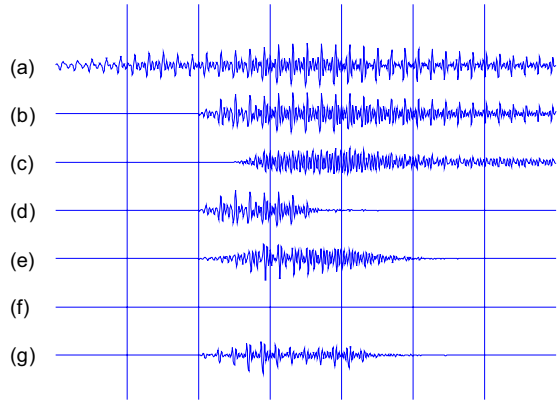


Figure 1. Examples of error signals for a single lost packet (3rd packet lost). Vertical lines indicate 20 ms packet boundaries. (a) G.729 synthesis at 8 kbps without packet loss; and the synthesis error signals: (b) G.729-0, (c) G.729-1, (d) G.729-2, (e) G.729-3, (f) G.729-4 and (g) iLBC.

5. SUBJECTIVE TEST RESULTS

Figure 2 summarizes the subjective test results in graphical form. The curve labels are as in Table 1. Mean Opinion Scores (MOS) are given as a function of Frame Erasure Rate (FER). The FER is varied between 0 and 20%. Note again that all frames were 20 ms in duration. Here, we use the terminology "frame" instead of "packet" to be consistent with the FER notation. The top two curves (original and G.729E) are the MOS

obtained for, respectively, the original signal and 11.8 kbps G.729 Annex E at 0% FER (given as a reference condition). For clarity, these are shown as horizontal lines and not single points at 0% FER. It is evident that G.729 Annex E is 11.8 kbps is significantly better than iLBC at 15.2 kbps.

Several observations can be made from Figure 2. First, the different approaches proposed in Section 3 for improving the robustness of G.729 produce a very broad range of subjective quality, especially at high FER. This shows the impact of the different delay/bit rate tradeoffs.

The lowest curve in Figure 2 corresponds to sending, without any redundancy, 2 consecutive G.729 frames in each 20 ms packet (G.729-0 in Table 1). The relatively poor performance of G.729 in this context can be explained by the fact that the concealment mechanism was optimized for 10 ms frames. The highest curve corresponds to sending each 10 ms frame twice, once in the current packet and once in the next packet as redundancy (G.729-4 in Table 1). This approach can sustain very high FER; in fact, the MOS score for this curve drops by only about 0.2 points when the FER increases from 0 to 10%. This is a significant gain in quality. The cost is twice the bit rate of G.729 (16 kbps) and three times its delay (45 ms). This is a significant increase compared to G.729, but this bit rate is actually very close to the bit rate of iLBC when used at 20 ms frames (15.2 kbps). Moreover, Figure 2 shows that the quality curve for iLBC is much closer to the curve for Approach 1 described in Section 3, which has a total bit rate of 12 kbps and a delay of 35 ms. We note also that the quality of iLBC in clear channel condition (0% FER) is almost equivalent to the quality of G.729 at 8 kbps. In other words, the additional bit rate (15.2 compared to 8 kbps) does not allow iLBC to encode speech better than a coder at half the bit rate. Rather, the higher bit rate in iLBC is required to encode each 20 ms frame sufficiently well while keeping the frames independent from each other to limit error propagation in case of missing frames.

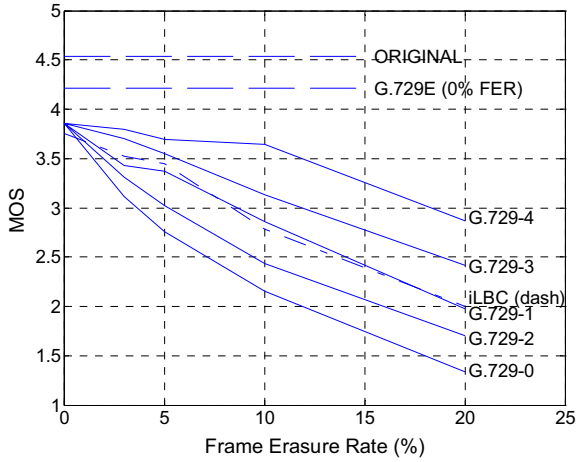


Figure 2. Subjective test results

The two curves above and below the iLBC curve in Figure 2 correspond respectively to approaches 2 and 3. The packet stream in these two approaches is actually identical (see Section 3). The only difference is the total delay: it is 20 ms longer for approach 3 (G.729-3) compared to approach 2 (G.729-2). The additional delay is taken at the decoder in approach 3 to replace concealment by actual decoding of the packet since it is transmitted (save for the LPC filter) in the next packet – thus the additional 20 ms required at the decoder.

6. CONCLUSION

This paper has presented an objective and subjective comparison of different approaches for ensuring or improving the robustness of speech coders in packet networks. Two main approaches were considered, either improving the robustness of a baseline coder with channel redundancy (using G.729 as the baseline coder), or using a coder designed to minimize inter-frame dependencies (as exemplified by the iLBC speech coder). Both approaches require tradeoffs between bit rate, delay and resilience to packet losses. A formal subjective test has shown that, in clean channel conditions, the 15.2 kbps mode of the iLBC coder has a performance equivalent to ITU-T Recommendation G.729 at 8 kbps. In Frame Erasure conditions, the quality of iLBC at 15.2 kbps was shown to be equivalent to the quality of G.729 using a total rate of 12 kbps (i.e. G.729 plus 50% frame repetition). At bit rates similar to iLBC (i.e. between 14 and 16 kbps), simple redundancy and memory update mechanisms allowed a baseline coder such as G.729 to perform significantly better than iLBC in Frame Erasures, provided the delay constraint can be relaxed (up to 45 ms compared to 25 ms in iLBC).

A general conclusion of this work is that selecting a particular speech coder for an application requires looking at the coder as more than a "black box". Simple extensions, such as those presented in this paper, can make a baseline coder such as G.729 very robust to channel errors while keeping the coder interoperable and allowing the use of all the coder functionalities (multiple bit rates, low-complexity extensions, VAD/DTX, etc.). Another conclusion is that when comparing different coders, to have a fair comparison the design parameters should be made similar. For example, if the goal is to compare the packet error robustness then the other parameters such as bit rate and delay should be made similar.

7. REFERENCES

[1] B. Bessette, et al. "The Adaptive Multi-Rate Wideband Speech Codec (AMR-WB)". *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 620-636, Nov. 2002.

[2] I. Johansson, et al., "Bandwidth efficient AMR operation for VoIP", *Proceedings of the IEEE Speech Coding Workshop*, pp. 150-152, 6-9 October 2002.

[3] S.V. Andersen, et al. "ILBC - A linear predictive coder with robustness to packet losses," In *Proc. 2002 IEEE Speech Coding Workshop*, pp. 23-25, Tsukuba, JAPAN, 6-9 October 2002

[4] ITU-T Recommendation G.729 (03/96), "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)".

[5] P. Gournay, et al. "Improved packet loss recovery using late frames for prediction-based speech coders", In Proc. ICASSP-2003, pp. 108-111, April 6-10 2003.

[6] ITU-T Recommendation P.800 (08/96). "Methods for subjective determination of transmission quality".

[7] Annex D of the ITU-T Recommendation P.830. "Subjective performance assessment of telephone-band and wideband digital codecs".