A SCALABLE CODING SCHEME BASED ON INTERFRAME DEPENDENCY LIMITATION

José L. Carmona, José L. Pérez-Córdoba, Antonio M. Peinado, Angel M. Gómez, José A. Gonzalez

Dpt. Teoría de la Señal, Telemática y Comunicaciones, University of Granada
maqueda@ugr.es

ABSTRACT

While VoIP (Voice over IP) is gaining importance in comparison with other types of telephony, packet loss remains as the main source of degradation in VoIP systems. Traditional speech codecs, such as those based on the CELP (Code Excited Linear Prediction) paradigm, can achieve low bit-rates at the cost of introducing interframe dependencies. As a result, the effect of a packet loss burst is propagated to the frames correctly received after the burst. iLBC (internet Low Bit-rate Codec) alleviates this problem by removing the interframe dependencies at the cost of a higher bit-rate. In this paper we propose a combination of iLBC with an ACELP (Algebraic CELP) codec in which a variable number of ACELP-coded frames is inserted between every two iLBC-coded frames. The experimental results show that the combined codec can achieve a performance close to that of iLBC at different loss conditions but with a smaller bit-rate. Also, scalability is achieved by modifying the number of inserted ACELP-coded frames.

Index Terms— Speech codec, iLBC codec, ACELP, packet loss.

1. INTRODUCTION

The fast growth of the communications over Internet and, particularly, VoIP, requires the development of new speech codecs adapted to the characteristics of the transmission channel and, in particular, robust against impairments such as packet losses. Avoiding interframe dependencies, as the iLBC codec does [1], is a possible solution. In this case, a packet loss does not affect the frames correctly received after the loss. iLBC presents two operation modes depending on the length of the frames, 20 ms (15.2 kbps) or 30 ms (13.3 kbps). The 20 ms mode has the inconvenience of a high bit-rate with an end-to-end delay of 25 ms. The 30 ms mode reduces the bit-rate, although it presents a longer delay (40 ms). CELP-based codecs allow a bit-rate reduction [2, 3], although they present a lower performance in presence of packet losses due to interframe dependencies.

To cope with these inconveniences several techniques have been proposed. A comparison between several approaches to improve the robustness of the speech codecs against packet losses can be found in [3]. For example, MDC (Multiple Description Coding) can be used to mitigate packet losses [4, 5]. In this technique several replicas of each packet are sent. Nevertheless, MDC increases notably the bit-rate and the end-to-end delay. Another possibility is to *reengineering* (as its authors name it) the parameters of iLBC [6, 7] in such a way that the gross bit-rate is lower. Other kinds of techniques work trying to minimize the effect of packet losses in CELP decoders through procedures which perform a *glotal pulse resynchronization* [8] or using received delayed packets that were discarded [9].

In this work, we propose a new coding scheme that is robust against packet losses and is based on the 20 ms mode of iLBC. Our scheme is a combination of the iLBC codec with an ACELP-based codec, where both codecs share the same kind of LP (Linear Prediction) synthesis filter differing only in the coding of the excitation signal. Between two iLBC frames, that act as key frames, is inserted a variable number of ACELP-coded frames, repeating this pattern periodically.

This combination leads to a lower bit-rate than iLBC, while maintaining the same delay. In addition, the propagation of the errors due to packet losses caused by the ACELP-based frames is limited because of the periodic presence of iLBC frames. Finally, the scalability in the bit-rate is achieved by introducing a variable number of ACELP-coded frames between each two consecutive iLBC frames.

This paper is organized as follows. In Section II, the structure of our proposed technique is described. The method to evaluate the quality of our proposal is presented in Section III, while the performance of the new codec and the comparison with other codecs are shown in Section IV. Finally, Section V summarizes this paper.

2. STRUCTURE OF THE PROPOSED CODEC

iLBC is a speech codec specially conceived for packet networks, such as Internet, since it was designed to combat packet losses. To achieve this goal, iLBC does not exploit the correlation between adjacent frames in the excitation encoding. Thus, iLBC removes the interframe dependencies at the cost of a higher bit-rate than other coding techniques.

On the other hand, ACELP codecs do exploit the cor-

Work supported by project MEC/FEDER TEC/2007-66600.

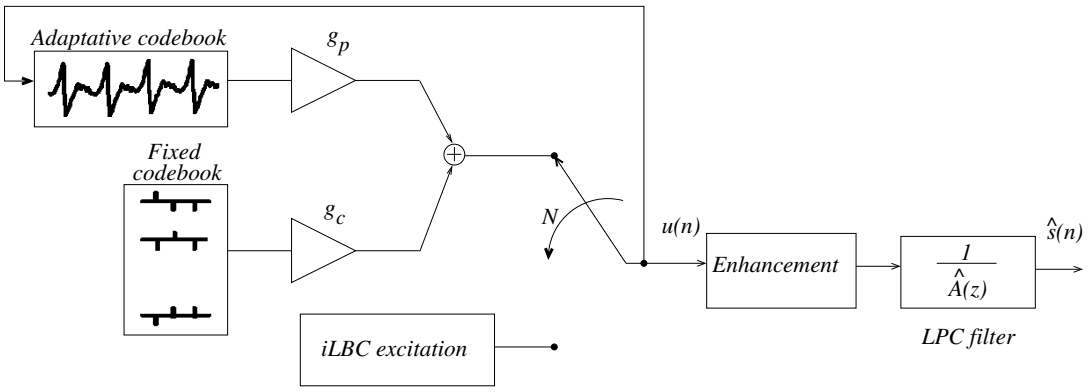

Fig. 1. Structure of the proposed decoder.

relation between consecutive frames to reduce the bit-rate. ACELP is based on the *analysis by synthesis* paradigm, which consists of choosing an excitation signal which minimizes the error between the synthesized signal and the target signal. The excitation is produced by summing the contributions from an adaptive codebook and a fixed codebook. The fixed codebook contains a number of innovation sequences, while the entries of the adaptive codebook consist of delayed versions of the excitation. Although the adaptive codebook makes possible to efficiently code quasi-periodic signals, such as voiced segments, propagates the errors forward in case of packet losses.

We propose combining both coding schemes, iLBC (15.2 kbps) and ACELP (10.1 kbps), in order to obtain a robust performance against packet losses while reducing the bit-rate of iLBC. The idea is based on using iLBC and ACELP frames, as shown in Fig. 2. Thus, in case of packet losses, the error propagation of ACELP frames is limited by the iLBC frames (key frames), which act as firewalls. Also, the insertion of ACELP frames reduces the average bit-rate.

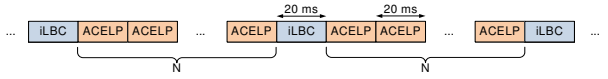

Fig. 2. Combination of different types of frames.

The reduction of the total bit-rate is controlled by the number of ACELP frames (N) inserted between two adjacent iLBC frames. Thus, a trade-off between robustness against packet loss and bit-rate is achieved. As the distance between iLBC frames is increased, the robustness against packet losses decreases, since a larger separation of iLBC frames allows a longer propagation of errors. Regarding the delay, our proposal does not increase it, since the size of all frames is 20 ms and it is not necessary a lookahead in the encoding process.

Fig. 1 shows the structure of the decoder. It can be seen that the iLBC and ACELP sections share the same LP filter. Furthermore, both types of frames use the enhancement and packet loss concealment blocks defined in [10] that works on the excitation signal and introduces an additional delay of 5 ms. The total delay of our proposal is 25 ms, just like the 20 ms mode of iLBC.

2.1. Linear Predictive Filter

The LP analysis coincides with that of iLBC for all frames. The LP coefficients are calculated once per frame using an asymmetric window of 30 ms centered in the third subframe of 5 ms. The resulting LP coefficients (a_k) are finally transformed into LSF (Line Spectrum Frequencies) parameters prior to quantization. However, before being transformed, the LP coefficients are modified as,

$$\widetilde{a}_k = \gamma_1^k \cdot a_k \quad k = 1, 2, \ldots, M$$

where $\gamma_1 = 0.9025$ and $M = 10$. This operation introduces a distortion (bandwidth expansion) in the LPC spectrum, although it improves the stability of the filter.

Another advantage of the bandwidth expansion technique is the shortening of the impulse response length, which improves the robustness against channel errors. This is because the excitation signal distorted by channel errors is filtered by the synthesis filter, and a shorter impulse response reduces the propagation of the error.

2.2. ACELP Frames

A weighting filter ($W(z) = 1/A(z/\gamma_2)$) is used in the analysis-by-synthesis process in order to obtain the target signal. During this process, the entries of an algebraic codebook ($c(n)$), based on the AMR 10.2 kbps codebook, and the adaptive codebook ($v(n)$) are determined, obtaining the final excitation signal through the following expression,

$$u(n) = \hat{g}_p v(n) + \hat{g}_c c(n) \quad (1)$$

where $\hat{g}_p$ and $\hat{g}_c$ are the decoded pitch and code gains, respectively.

Parameter	1st Subf	2nd Subf	3rd Subf	4th Subf	Total
LSP set					20
Pitch delay	8	5	8	5	26
Algebraic code	31	31	31	31	124
Gains	8	8	8	8	32
Total					202

Table 1. Bit allocation of the ACELP coding algorithm for 20 ms frames.

To reduce the impact of packet losses in ACELP frames, all kind of predictive techniques are avoided in the coding process. Thus, both gains are quantized using a Vector Quantizer (VQ) codebook of 8 bits, that obtains the pair $(\hat{g}_p, \log(\hat{g}_c))$ which minimizes the error between the synthesized speech and the target vector. This process is carried out four times per frame corresponding to 4 subframes of 5 ms. Table 1 shows the number of bits used for coding ACELP frames.

This ACELP codec has a bit-rate of 10.1 kbps, so, the bit-rates (B_N) of our scalable proposal are determined by the following expression.

$$B_N = \frac{15.2 + N \cdot 10.1}{N+1} \; kbps \qquad (2)$$

3. PERFORMANCE EVALUATION

To evaluate the quality of the proposed method an objective test based on the PESQ algorithm was carried out. The reason to use PESQ was the number of conditions to be tested, which makes a subjective MOS test impractical.

We have used the clean utterances from set A of the Aurora-2 database in order to evaluate the performance of each codec. This database is uttered by a balanced number of male and female speakers. However, original test utterances were concatenated into groups of seven, resulting in a total of 572 sentences. The reason for this is that PESQ algorithm has not been designed to evaluate short sentences [11]. Lengths between 8 and 20 s are recommended, but Aurora-2 utterances have a mean duration of only 1.5 s. Through this grouping, the mean duration is extended to 12 s (approx.), with minimum and maximum values of 7.5 s and 20 s respectively. To obtain an overall score for the tested condition, the score of each sentence is weighted by its length. Although the Aurora-2 database was designed for automatic speech recognition, it is appropriate to evaluate the intelligibility of a codec in presence of packet loss, since the PESQ scores are corroborated by informal listening tests and comparable with the results obtained in [6, 7] (for the TIMIT database), so that they provide a good indication of actual MOS scores.

In order to emulate the behavior of an IP channel, a two-state Markov model [12] is used. The model parameters can be set in accordance with an average burst length (L_{burst}) and a loss rate (P_{loss}).

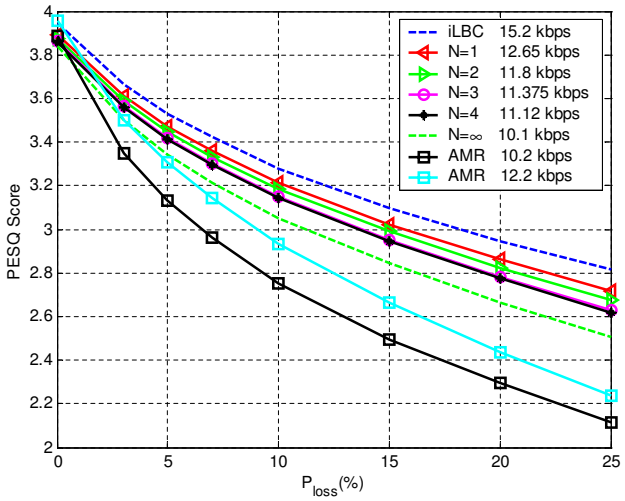

Fig. 3. PESQ scores for channels with $L_{burst} = 1$.

4. RESULTS

Fig. 3 and Fig. 4 show the PESQ performances obtained under different P_{loss} conditions with $L_{burst} = 1$ and $L_{burst} = 2$, respectively. Different values for the number of inserted ACELP-coded frames (N) are tested. The best performance is obtained for iLBC, which corresponds to the trivial case of $N = 0$ in our proposal, i.e. inserting zero ACELP frames between two adjacent iLBC frames. The case $N = \infty$ obtains the worst result of our proposal, and it corresponds to the proposed ACELP codec with a bit-rate of 10.1 kbps (no iLBC frames). These cases limit the performance of our scalable proposal (dashed lines in Fig. 3 and Fig. 4). Four values of N have been selected with bit-rates in the range of 12.65 kbps to 11.12 kbps. Furthermore, the AMR modes of 12.2 and 10.2 kbps have been included in this figure because they present similar bit-rates and delay to some configurations of our proposal, being adequate for a comparison [3].

Particularly, the configurations with $N = 1$ (12.65 kbps) and $N = 2$ (11.8 kbps) have bit-rates close to that of AMR 12.2 kbps. Without packet loss, AMR 12.2 kbps presents a better performance (PESQ score of 3.96) than our proposal. Otherwise, the performance of AMR 12.2 is worse than any configuration of our proposal.

Even for the case of $N = \infty$, the robustness against packet losses is higher in our proposal than in the AMR modes. Although both codecs, $N = \infty$ and AMR 10.2 kbps, use the same ACELP architecture, AMR uses predictive techniques to quantize more efficiently the codec parameters (the excitation gains are quantized using a predictor filter). This explains why the results obtained by AMR 10.2 without packet loss are slightly better (PESQ score of 3.89) than our

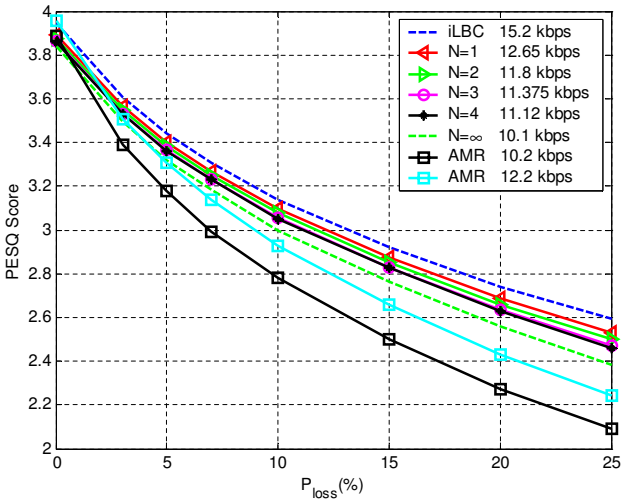

Fig. 4. PESQ scores for channels with $L_{burst} = 2$.

N	0	1	2	3	4	∞
Bit-rate(kbps)	15.2	12.65	11.8	11.375	11.12	10.1
PESQ score	3.94	3.89	3.87	3.86	3.86	3.84

Table 2. PESQ scores for our proposal without packet losses.

proposal for $N = \infty$ (PESQ score of 3.84). However, in presence of packet losses, these predictive techniques are not suitable. This, along with the use of bandwidth-expanded LP coefficients, is the reason why our proposed codec is more robust against packet losses.

Although our proposal for $N \geq 1$ presents lower bit-rates than iLBC, the behavior against packet loss is close to iLBC. As more ACELP frames are inserted between consecutive iLBC frames this robustness goes down. Nevertheless, our work provides an easy method to make the iLBC codec scalable with a small PESQ performance degradation in absence of packet loss, as shown in Table 2. Furthermore, in comparison with a scalable ACELP coding scheme such as AMR, the performance of our proposal is clearly higher in presence of packet loss. More degree of scalability could be reached using an ACELP codec with lower bit-rate than 10.1 kbps.

5. SUMMARY

In this paper we have proposed a new technique of speech coding based on the combination of iLBC with an ACELP-based codec. Thus, we combine the robustness of iLBC against packet losses with the lower bit-rates provided by ACELP coding. Furthermore, this scheme allows to control easily the trade-off between robustness and bit-rate by modifying the number of ACELP frames inserted between two consecutive iLBC frames. In addition, the experimental results without packet loss show that the combined codec achieves results slightly lower than the AMR modes with similar bit-rates and delay, while its performance against packet losses is close to iLBC and clearly higher than AMR.

6. REFERENCES

[1] S.V. Andersen, W.B. Kleijn, R. Hagen, J. Linden, M.N. Murthi, and J. Skoglund, "iLBC - a linear predictive coder with robustness to packet losses," in *Proceedings of Speech Coding Workshop 2002*. IEEE, pp. 23–25.

[2] 3GPP TS 26.090, "Adaptive multi-rate (AMR) speech codec," .

[3] R. Lefebvre, P. Gournay, and R. Salami, "A study of design compromises for speech coders in packet networks," in *Proceedings of ICASSP 2004*. IEEE, vol. I, pp. 265–268.

[4] E. Orozco, S. Villete, and A.M. Kondoz, "Multiple description coding for voice over IP using sinusoidal speech coding," in *Proceedings of ICASSP 2006*. IEEE, vol. I, pp. 9–12.

[5] H. Dong, A. Gersho, J.D. Gibson, and V. Cuperman, "A multiple description speech coder based on AMR-WB for mobile ad hoc networks," in *Proceedings of ICASSP 2004*. IEEE, vol. I, pp. 277–280.

[6] C.M. Garrido, M.N. Murthi, and S.V. Andersen, "Towards iLBC speech coging at lower rates through a new formulation of the start state search," in *Proceedings of ICASSP 2005*. IEEE, vol. I, pp. 769 – 772.

[7] C.M. Garrido, M.N. Murthi, and S.V. Andersen, "On variable rate frame independent predictive speech coding: Re-engineering iLBC," in *Proceedings of ICASSP 2006*. IEEE, vol. I, pp. 717–720.

[8] T. Vaillancourt, M. Jelinek, R. Salami, and R. Lefebvre, "Efficient frame erasure concealment in predictive speech codecs using glottal pulse resynchronisation," in *Proceedings of ICASSP 2007*. IEEE, vol. IV, pp. 1113–1116.

[9] P. Gournuy, F. Rousseau, and R. Lefebvre, "Improved packet loss recovery using late frames for prediction-based speech coders," in *Proceedings of ICASSP 2003*. IEEE, vol. I, pp. 108–111.

[10] RFC 3951, "Internet low bit-rate codec (iLBC)," 2004.

[11] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ)," .

[12] W. Jiang and H. Schulzrinne, "Modeling of packet loss and delay and their effect on real-time multimedia service quality," in *Proceedings of NOSSDAV 2000*.