RESYNCHRONIZATION OF THE ADAPTIVE CODEBOOK IN A CONSTRAINED CELP CODEC AFTER A FRAME ERASURE

Mohamed Chibani, Roch Lefebvre and Philippe Gournay

University of Sherbrooke, Sherbrooke, Quebec, Canada
Mohamed.Chibani@USherbrooke.ca

ABSTRACT

The adaptive codebook used in CELP (Code Excited Linear Prediction) codecs to model the pitch excitation allows the attainement of a high quality of synthesized speech but introduces a strong inter-frame dependency and consequently causes error propagation in case of frame erasure. In a previous work we showed that the error propagation can be greatly reduced by constraining, at the encoder side, the innovative codebook to partially model the pitch excitation. In this paper we extend this work by exploiting, at the decoder side, the pitch-related information present in the innovative excitation to speed up the recovery of the decoder. The method consists in adequately shifting the last pitch pulse present within the corrupted adaptive codebook memory so that it is resynchronized with the excitation parameters of the frame that follows the erased one.

1. INTRODUCTION

In packet-switched telephony applications, such as voice over IP (VOIP), some packets may be lost due to network congestion. The missing frames have to be regenerated at the decoder side using packet loss concealment techniques. For CELP codecs, the errors not only affect the concealed frame but may propagate over several frames. Although all the codebooks that make use of prediction are potential sources of error propagation, the adaptive codebook (ACB) is undoubtedly the main cause.

Using frame-independent coding as in [1] or redundant transmission of the encoded parameters of a frame-dependant codec as in [2] prevents the error propagation but at a cost of an increased bit rate. In a previous work [3], we showed that the robustness of a CELP codec can be significantly improved by constraining the search for the excitation parameters. Modifying only the encoder, the method is based on forcing the innovative codebook (ICB) to partially model the pitch excitation. Thus, the decoder will recover faster after a frame erasure due to a lower dependency on the ACB.

It was observed that the pitch-related information present in the ICB excitation can be further exploited, at the decoder, to speed up the recovery. If a frame erasure occurs in a highly voiced speech signal, the last pitch pulse present within the (erroneous) ACB memory after the frame erasure will be different from the correct one, both in waveform and position. Whereas mismatch in pitch pulse waveform may cause distortions, differences in pitch pulse positions are much more critical and may yield a complete asynchrony between the encoder and the decoder.

In this paper, we propose a method to resynchronize the ACB memory after a frame erasure. The method is based on adequately shifting the last pitch pulse within the ACB memory in order to resynchronize it with the excitation parameters of the frame that follows the concealment. The combined methods (constraint and resynchronization) may be applied to virtually any CELP codec; however, in our research, we focused on their application to the AMR-WB standard [4].

The paper is organized as follows. In section 2, we briefly recall the constrained search of the excitation parameters. The resynchronization algorithm is described in section 3. In section 4, some experimental results are presented and section 5 gives some conclusions.

2. CONSTRAINED SEARCH OF THE EXCITATION PARAMETERS

In the AMR-WB speech codec [4], the speech signal is first segmented into 20-ms frames. The coefficients of a 16^{th} order LP filter are estimated and converted to the LSF (Line Spectral Frequencies) representation to be quantized. To search for the excitation parameters, the frame is further divided into 4 subframes. For each subframe, the speech signal is weighted using a perceptual weighting filter to form a new signal usually referred to as the target signal. This signal is noted $x_1(n)$.

The excitation signal is composed of the ACB excitation, which models the pitch-related (or periodic) component of the speech signal, and the ICB excitation. First, the ACB parameters, corresponding to the delay and the gain for each subframe, are searched, then the ICB excitation parameters are calculated. The details of the excitation parameter search can be found in [4].

In the standard encoder, the contribution of the ACB is removed from the target signal $x_1(n)$ to form the new target $x_2(n)$ needed for the ICB search:

$$x_2(n) = x_1(n) - g_0 y_0(n), \quad (1)$$

where $y_0(n)$ is the non-scaled ACB contribution at the optimal delay T_0, and g_0 is the optimal ACB gain calculated using equation (2):

$$g_0 = \frac{\sum_n x_1(n) y_0(n)}{\sum_n y_0^2(n)}. \quad (2)$$

For highly voiced speech, most of the excitation signal is modeled by the ACB which, as explained above, renders the decoder vulnerable in case of frame erasure. We proposed in [3] to modify the value of the gain to force the ICB to partially model the already modeled pitch-related excitation. The ACB gain is modified if the ratio R, given in (3), between the energy of the ACB contribution and the energy of the target $x_1(n)$ is greater than the predefined threshold R_{th} (set to 0.55).

$$R = \frac{g_0^2 \sum_n y_0^2(n)}{\sum_n x_1^2(n)}. \quad (3)$$

The new gain is then given by [3]:

$$g_0' = \begin{cases} \sqrt{R_{th} \dfrac{\sum_n y_a^2(n)}{\sum_n x_1^2(n)}}, & \text{if } R > R_{th}, \\ g_0, & \text{otherwise.} \end{cases} \quad (4)$$

When the contribution of the ACB is high, using the gain g_0' calculated in (4) guarantees that a fraction, proportional to 1-R_{th}, of the already modeled ACB contribution is left in the new target signal $x_2(n)$. Therefore, the pitch-related excitation will be modeled partially by the ICB.

It has been shown in [3] that, without modifying the decoder, the constraint already improves the recovery after a frame erasure. In the next section we present a method that exploits the residual information of the pitch excitation present in the ICB excitation to further speed up the recovery of the decoder.

3. DESCRIPTION OF THE RESYNCHRONIZATION ALGORITHM

For stable voiced speech, the pitch pulse features change slowly over a frame. After a frame erasure, the pulse present within the concealed ACB memory will be fairly similar to the correct one. A small difference in the pitch pulse waveform has moderate but tolerable effects on the synthesized speech, as will be shown in next paragraphs. However, difference in the pitch pulse position is much more critical. Even a slight drift in the pitch pulse position is generally sufficient to completely desynchronize the ACB. In the proposed method, the last pitch pulse in the erroneous ACB memory is shifted in such a way as to be resynchronized with the excitation parameters of the frame that follows the concealment.

The resynchronization algorithm consists in two main operations. First, a potential shift δ_0 is estimated using the position P(-1) of the pulse in the corrupted ACB memory and an estimation of its correct position P(0) (i.e., the pulse position if the frame had been received and properly decoded). Then an optimal shift is searched in a closed-loop around δ_0. The algorithm is summarized as a block diagram in Fig. 1.

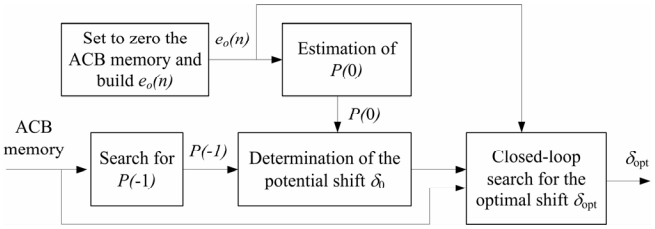

Fig.1. Block diagram of the resynchronization algorithm

3.1. Determination of the potential shift

The potential shift δ_0 is the difference between the estimated position, noted P(0), of the last pitch pulse and the position P(-1) of the pitch pulse present within the concealed ACB memory:

$$\delta_0 = P(0) - P(-1). \quad (5)$$

$P(-1)$ is simply the position of the sample that has maximum absolute value in the concealed ACB memory. The search range is limited to the last ACB delay of the previous frame.

Estimating the correct position P(0) is more complex, however, and is done as follows. An excitation signal noted $e_0(n)$ is built after setting the ACB memory to zero . Thus, the resulting excitation will be cleared from the corrupted ACB memory. The constraint applied at the encoder makes it possible to build a good approximation of the excitation signal using only the contribution of the current frame. The ICB contributions for every subframe combine with the long-term prediction within the frame to progressively shape the pitch pulses, so that the last pitch cycle in the excitation signal $e_0(n)$ is the most similar to the corresponding pitch cycle in the correct excitation. This is illustrated in Fig. 2. The position, noted P(1), of the last

pitch pulse in $e_0(n)$ together with the ACB delays of the current frame are used to predict $P(0)$; starting from $P(1)$, the distances between successive pulses are accumulated in backward direction. We assume that these distances, or equivalently, the pitch cycle lengths, are equal to the ACB delays of the corresponding subframes.

3.2. Closed-loop search for the optimal shift

The assumption made when using the ACB delays as distances between successive pitch pulses may not always hold. For instance, for high-pitched speech signals, such as those produced by female speakers, the pitch contour may evolve within a subframe. Therefore, the potential shift δ_0 calculated in (5) has to be refined in a closed-loop fashion by testing a set of values around the estimated shift δ_0. The search range is limited to 5 shift candidates, that is $\delta \in [\delta_0 - 2, \delta_0 + 2]$.

Since the excitation $e_0(n)$ is greatly similar to the correct excitation, it is a natural candidate to be used as reference when searching for the optimal shift δ_{opt}. A correlation function is calculated between the excitation $e_0(n)$ and the excitation $e_\delta(n)$ built after correcting the ACB memory for every shift candidate:

$$C(\delta) = \frac{\sum_{n=L\_FRM-L}^{L\_FRM-1} e_0(n) e_\delta(n)}{\sqrt{\sum_{n=L\_FRM-L}^{L\_FRM-1} e_0^2(n)}}. \quad (6)$$

where L_FRM=256 and L_SBFR=64 are the frame and the subframe lengths respectively, and $T(3)$ is the ACB delay of the 4th subframe. The length of the correlation L is given by $L=max(2*L\_SBFR, T(3))$.

The calculation of the correlation in equation (6) is limited to the two last subframes or to one pitch cycle if the last ACB delay is greater than the length of two subframes. This choice was motivated by the fact that the similarity is higher at the end of the frame, as mentioned previously (see Fig. 2). The optimal shift is then the shift value that maximizes the correlation function:

$$\delta_{opt} = \arg\max_\delta (C(\delta)) \quad (7)$$

Note that the correlation in equation (6) is normalized using only the energy of $e_0(n)$, which does not depend on the shift. This choice was driven by the observation that, for the optimal shift, not only the resemblance of the excitation $e_\delta(n)$ with $e_0(n)$ increases but also its energy. Not normalizing the correlation with the energy of the excitation $e_\delta(n)$ favors shifts that increase both waveform similarity and energy.

To increase the reliability of the algorithm, only shift values that yield high correlation are considered, otherwise the ACB memory is not modified. The decision is made by comparing the correlation to a threshold (determined experimentally and set to 1.2). In general, for low voiced speech, the pulse within the ACB memory may be quite different in shape from the correct one, resulting in a lower correlation, and correcting those frames will not improve the recovery and increases the risks of introducing artifacts. Moreover, during weakly voiced speech, the decoder automatically recovers relatively fast without the resynchronization because of the higher ICB contribution. Fig. 2 gives an example of a resynchronized excitation for the frame that follows a concealed one.

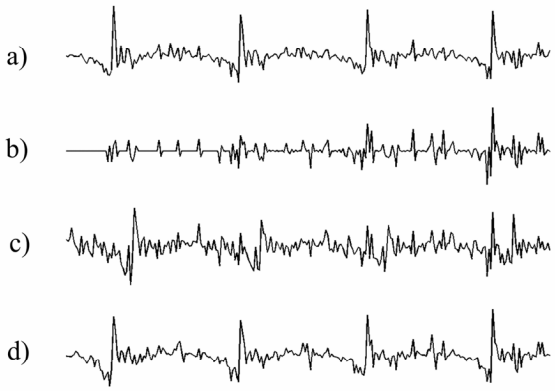

Fig. 2. Example of a resynchronized excitation: (a) the correct excitation, (b) the excitation $e_0(n)$, (c) the excitation signal built using the erroneous ACB memory and (d) the excitation signal built after correcting the ACB memory.

Correcting the position of the pitch pulse in the erroneous ACB memory resynchronizes the ACB but causes an abrupt change in the pitch contour at the frame boundary (one pitch cycle is either too short or too long). This may result in an audible artifact, especially if the applied shift is large. In the specific case of VOIP application, the abrupt change in the pitch contour may be handled by making the playout buffer process the optimal shift as a change in the playout delay; as a result, the frame will be shorter or longer than the regular frame length. In this work, a different approach that consists in modifying the pitch contour of the resynchronized frame is considered. The modification is achieved by segmenting and moving the pitch pulse cycles individually in such a way that the pitch contour evolves smoothly.

This approach has the advantage of confining the resynchronization process to the frame following the concealment; the shift is not accumulated with the other delays. Experiments showed that the distortion due to the pitch contour modification is not significant compared to the overall distortion caused by the frame erasure.

4. EXPERIMENTAL RESULTS

A subjective evaluation of the resynchronization algorithm was conducted using the MUSHRA methodology [5]. The AMR-WB encoder was used in mode 2 (12.65 kbps). Ten listeners participated in the experiment. The test material was 14 clean-speech pairs of sentences in English: 8 uttered by female speakers and 6 uttered by male speakers. The listeners had to assess the quality of the decoded speech of the standard codec and the modified codec compared to the original speech signal (uncompressed signal). The latter should be scored at full scale corresponding to 100. The two codecs were evaluated for three conditions: clear channel (0% frame erasure rate), random frame erasure at a rate of 5% and random frame erasure at a rate of 10%. To evaluate the performance of the resynchronization algorithm over using the constraint only, the test was conducted with the resynchronization disabled then enabled. The results are presented in Fig. 3.

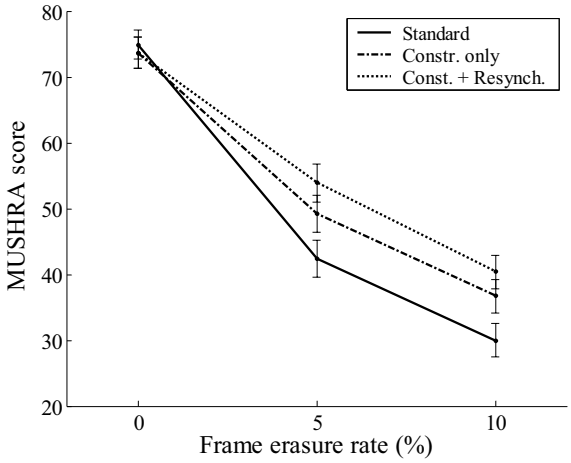

Fig. 3. Subjective test results (the error bars correspond to the confidence interval of 95%).

In the clear channel condition (no frame erasures) the standard codec scores 74.88, thus only slightly outperforms the constrained codec which scores 73.67. However, in the cases of frame erasures, the constrained codec clearly outperforms the standard codec. At a frame erasure rate of 5%, the constrained codec scores 49.18 with the resynchronization disabled and 53.91 with the resynchronization enabled, compared to 42.36 for the standard. The improvement also holds for the 10% frame erasure rate, where the constrained codec scores 36.70 and 40.40 with the resynchronization disabled then enabled respectively, while the standard codec scores 29.99.

As mentioned in section 3.2, the resynchronization is performed only if the value of the correlation indicates that it is likely to succeed. For the material used in the experiments, it is performed for only 10 to 15% of the frames that follow a concealment. Even with this low percentage, performing the resynchronization at the decoder brings a non-negligible improvement over only applying the constraint at the encoder. This is not completely unexpected since the resynchronization is effective in highly voiced segments of the speech signal, where, in general, the signal attains its highest perceptual relevance.

5. CONCLUSION

We showed in a previous work [3] that constraining the excitation search at the encoder to lower the ACB contribution improves the recovery after a frame erasure. In this work, the constraint is further exploited at the decoder side to speed up the recovery. More specifically, we used the redundant pitch-related information present in the ICB excitation to resynchronize the adaptive codebook. We showed that the drift of the ACB caused by the concealment can be corrected by adequately shifting the pitch pulses within the concealed ACB memory, which speeds up the resynchronization of the decoder. Although the waveform of the pitch pulse present within the erroneous ACB memory is slightly different from the waveform of the correct one, correcting the position of this pulse significantly improves the overall quality of the decoded speech in case of frame erasure. The modified codec (constraint and resynchronization) requires neither extra data nor additional delay and is completely interoperable with the standard.

6. REFERENCES

[1] S.V. Andersen, et al. “ILBC - A Linear Predictive Coder with Robustness to Packet Losses,” In Proc. *IEEE Speech Coding Workshop*, Tsukuba, Japan, pp. 23-25, Oct. 2002.

[2] R. Lefebvre, P. Gournay, and R. Salami,“A Study of Design Compromises for Speech Coders in Packet Networks,” In Proc. *International Conference on Acoustics, Speech and Signal Processing*, Montreal, pp. 265-268, May 2004.

[3] M. Chibani et al. “Increasing the Robustness of CELP-Based Speech Coders by Constrained Optimization,” In Proc. *International Conference on Acoustics, Speech and Signal processing*, Philadelphia, pp. 785-788, Mar. 2005.

[4] B. Bessette et al. “The Adaptive Multirate Wideband Speech Codec,” *Transactions on Speech and Audio Processing*, Vol. 10, No. 8, pp. 620-636, Nov. 2002.

[5] Recommendation ITU-R BS.1534-1, “Method for Subjective Assessment of Intermediate Quality Level of Coding Systems,” *International Telecommunication Union (ITU)*, Jan. 2003.