Improved Frame Loss Recovery Using Closed-Loop Estimation of Very Low Bit Rate Side Information

Philippe Gournay
Speech and Audio Research Group, University of Sherbrooke, Sherbrooke (QC) J1K 2R1 CANADA
Philippe.Gournay@USherbrooke.ca

1. Abstract
In CELP coders, the past excitation signal used to build the adaptive codebook is known to be the main source of error propagation when a frame is lost. This paper presents a novel resynchronization technique using very low bit rate side information to correct the past excitation signal after a frame erasure, the novelty being that the correction is computed in a closed loop fashion, based on the actual error introduced by the concealment. Subjective test results show that this approach is a promising area for future research on frame loss recovery.

Index Terms: speech codecs, linear predictive coding, error correction, robustness

2. Introduction
In CELP coders, accurate long term (or pitch) prediction (LTP) [1] is a key requirement to attain a high quality of decoded speech at low bit rates. It is well known however that LTP is extremely sensitive to frame erasures [2]. This is because it relies on the past linear predictive (LP) excitation signal to build the adaptive codebook. When a frame is lost, the decoder uses a concealment procedure instead of normal decoding to regenerate the excitation signal. Concealment resembles normal decoding except that it extrapolates the pitch period (for the adaptive-codebook part of the excitation signal) and randomizes the innovation (for its fixed-codebook part). Since concealment strives to provide a close approximation to the missing actual input the resulting excitation signal is erroneous. When subsequent frames are decoded, concealment-induced errors propagate from one frame to the next through the adaptive codebook.

The extent to which these types of errors propagate depends on several factors with the primary ones being the type of signal that was lost and the magnitude of the adaptive codebook gain. When the loss strikes an onset frame for example, the error can propagate for dozens of frames [3]. Even during purely voiced segments, the decoder can totally lose synchronization a few frames after the frame loss because of small pitch variations that were not reproduced by the concealment (for an example signal, see Figure 9 in [4]).

Over the past few years, several solutions have been proposed to limit error propagation caused by long term prediction, while preserving the advantages of the adaptive codebook. In the 3GPP2 VMR-WB (Variable Multi-Rate Wideband) and ITU-T G.729.1 standards, specific frame erasure concealment information (voicing classification, energy, and pitch phase) is sent within the bitstream [5, 6]. This information is used by the decoder to improve both concealment (using the classification and energy) and recovery (using mainly the phase of the signal). In fact, sending side-information to improve concealment is known to be helpful also for non-CELP coders [7]. Another solution that does not require any side information is to introduce some constraints in the encoder to limit inter-frame dependencies [2]. These constraints can be further exploited at the decoder to speed up recovery [8]. This approach potentially requires a slightly higher bit rate to reach the same quality as the unconstrained codec. In other contexts (for example voice over packet networks) late frames can also be used to limit error propagation [9, 10].

This paper introduces a novel resynchronization technique using very low bit rate side information to correct the past excitation signal after a frame erasure. The correction is computed at the encoder in a closed loop fashion, based on the actual error introduced by the concealment.

The principle of the technique is detailed in section 3. To illustrate the concept, a very simple correction example which uses almost the same side information as VMR-WB is then presented. The extraction of the correction information, its quantization, and its use at the decoder to resynchronize the adaptive codebook are described in sections 4 and 5. Subjective test results presented in section 6 show that this approach is a promising area for future research on frame loss recovery. Finally, conclusions and perspectives are drawn in section 7.

3. The Basic Principle
A CELP encoder block diagram which includes the proposed technique for computing the side information is depicted in Figure 1. The white elements represent the items usually found in any CELP coder whereas the gray ones comprise the additions required for computing the side information.
A frame loss is simulated at the encoder (concealment) in order to determine the correction that should be applied to the past excitation signal.

The internal state of the encoder is a set of “static” variables (in the C-language sense) that are updated every time an audio frame is encoded. It contains, among other things, the past excitation signal required to compute the inter-frame contribution of LTP.

An encoder fed with a new audio frame generates a new bitstream frame. This involves: computing the LP residual signal corresponding to the input audio frame, retrieving the past excitation from the internal state to build an adaptive codebook (using LTP), removing the contribution of this adaptive codebook from the LP residual to build a target signal, encoding this target signal using the innovative codebook, combining the contributions of the adaptive and innovative codebooks to build the encoded excitation signal, and finally updating the past excitation signal.

Normally (i.e. when no frame is lost) most of the decoder’s internal state (specifically, all parameter values that are used to generate the encoded excitation signal) matches perfectly the encoder’s internal state. When a frame is lost the decoder’s internal state updated by the concealment procedure no longer matches the encoder’s internal state. To compute the correction that would render the “new bad past excitation signal” generated by the concealment as close as possible to the “new good past excitation signal” (should a frame be lost during transmission), the encoder simulates the concealment of the frame. This correction is delayed by one frame so that it is not lost with the lost frame (which is symbolized by the operator $z^{-1}$), then quantized and sent to the decoder as side information.

In terms of complexity this is not a very demanding operation. The concealment operation is generally less complex than the normal decoding operation, which in turn is also much less complex than the encoding operation. In addition, only the excitation needs to be computed (synthesis filters of the decoder are not needed to update the past excitation samples).

4. Extracting the Correction Information at the Encoder

We compared a number of actual “good” and “bad” past excitation signals computed on a speech database, and observed that the concealment procedure introduces three distinct types of errors that induce significant error propagation:

- **Type A: lost onset.** The concealment procedure produces a “bad” past excitation signal that does not contain the necessary elements (generally, a pulse) for the decoder to build a fully voiced signal during subsequent frames. This type of error has very long repercussions on the decoded signal which are thoroughly discussed in [3].

- **Type B: lost alignment.** The shape of the glottal pulse does not change much during the lost frame, but the concealment procedure is unable to reproduce the correct pitch contour. The resulting “bad” past excitation signal is consequently not aligned with the good one.

- **Type C: waveform mismatch.** The concealment procedure fails to keep track of changes in the shape of the glottal pulse.

To demonstrate our concept, we have chosen to concentrate on errors of types A and B only. Those types of errors are readily corrected with the addition of a small amount of side information. Solutions for mitigating the impact of errors of type C, as well as improvements for handling errors of types A and B, are discussed in section 7.

Although the technique is applicable to any CELP coder, the following description is based on our implementation in the AMR-WB codec [11].

First, the input audio frame is classified as either voiced or unvoiced/onset using a voicing classification technique inspired by that of the VMR-WB coder [5]. This class information is encoded using one bit and serves to indicate to the decoder how the correction information is encoded.

When the frame is classified as unvoiced or onset, the last segment of the “good” past excitation is searched for the sample which has the largest absolute value (Figure 2). The position of that sample is encoded with eight bits, while its sign and amplitude are jointly encoded with four bits. The total bit budget for side information is in this case 13 bits per 20 ms frame. Note that in this scenario the correction is not computed with respect to the good past excitation signal and is very similar to what is done for onsets in VMR-WB.
For stationary voiced segments, the gain ($g$) and shift ($\Delta$) that minimize the total quadratic error between the “good” and “bad” past excitation signals is determined by correlating the two signals and transmitted by the encoder as shown in Figure 3. To be able to shift the bad excitation in both directions (to adjust for delays or advances), the concealment operation was slightly modified to generate more excitation samples than would normally be needed to simply update the past excitation (specifically, the excitation signal is extended by half the maximum possible pitch value). Alternatively, the delay search could be done by assuming that the bad excitation signal is periodic. When the shift $\Delta$ is between -32 and +32, it is encoded on six bits and the gain $g$ is linearly quantized on four bits between 0.4 and 3.4. Otherwise no correction is transmitted. The bit budget for the correction is in this case 11 bits per frame.

The “good” and “bad” excitation signals were purposely represented with different lengths on Figure 2 and 3 because the good (decoded) and bad (concealed) pitch values normally differ. We use the concealed value to compute the correction as it is available at both the encoder and the decoder. However using the first pitch value of the frame that follows the lost one is also a legitimate alternative.

For the correction to be valid, the concealment performed at the encoder must exactly match the one done at the decoder. In the AMR-WB decoder, a small random value (between -1 and +1) is added to the concealed pitch value. Therefore, to ensure that the above condition requiring the concealed values to match remains true, we reset the random number generator at every frame using the frame number (both at the encoder and the decoder).

5. Performing the Resynchronization at the Decoder

The resynchronization operation takes place at the very beginning of the decoding of the frame that immediately follows the lost one. For onset frames, a single pulse is added to the past excitation signal at the location and with the amplitude specified by the side information. We noticed that slightly de-emphasizing that pulse (to lower its high-frequency content) improved the resulting subjective quality. For voiced frames, the extended “bad” past excitation is simply shifted and scaled in amplitude before building the adaptive codebook.

6. Evaluation Results

6.1. Objective results

Figure 5 and 6 give example signals for the cases of a lost onset and a lost pitch pulse alignment, respectively. Audio frames are delimited by vertical bars, the lost one being designated by solid lines instead of dotted lines. On both figures, the output of the decoder when no frame is lost (a) is given as a reference. During the lost frame, the build-up in the error signal as shown in (d) and (e) is obviously the same for the standard (b) and the modified (c) decoders because the resynchronization operation takes place only right before decoding the next frame. In Figure 6 the vertical scale is the same for all signals, while in Figure 5 it is twice as much for error signals (d) and (e).

Figure 5 shows that inserting one pulse in the past excitation signal is very efficient at restoring lost onsets. Figure 6 shows that resynchronizing the past excitation signal after a stable voiced frame has been lost also significantly reduces error propagation.
6.2. Subjective results

To assess the performance of the resynchronization technique, we performed an AB-like comparison test on a database of 32 sentence pairs (2 male and 2 female talkers, English language). We simulated one lost frame every 10 frames (10% frame erasure rate). Six experienced listeners participated in the test.

On average, a substantial 48% and 12% of the votes expressed a small or a strong preference towards the standard decoder while none expressed a strong preference towards the modified decoder, respectively. In contrast, only 17% of the votes expressed a small preference towards the standard decoder while none expressed a strong preference for it. No preference towards either of these decoders was expressed by 23% of the votes.

Two weak points of the modified codec were identified when listening to the test samples. For voiced segments, the resynchronization is sometimes too abrupt. For onsets, there are cases where the reconstructed signal sounds unnatural. Those limitations are discussed in the following section and some solutions are proposed.

7. Conclusions and Perspectives

We have presented a novel resynchronization technique which uses very low bit rate side information to correct the past excitation signal used to build the adaptive codebook of CELP coders. The novelty is that the correction (to be applied before decoding the frame that immediately follows the lost one) is computed in a closed loop manner, based on the actual error introduced by the concealment. Subjective test results showed that this approach is a promising area for future research on frame loss recovery.

An obvious limit of this technique is that the correction is valid only for single (isolated) frame losses. We believe however that this condition is most likely to be satisfied at typical frame erasure rates of 3 to 6%. If necessary, frame interleaving can be used to decorrelate bursty frame losses.

For errors of type A (lost onset), seeding the voice segment with a low-pass pulse has a very positive effect but does not always sound very natural. Better results would be obtained with a small ACELP codebook, a codebook of glottal pulses (vector quantizer), or an approach like the one proposed in [12] to code pitch pulses. For errors of type B (lost alignment), the abrupt changes in the pitch contour after the resynchronization is sometimes problematic. Smoothing the pitch contour as described in [8] or [10] should alleviate this problem. Various solutions can be considered to deal with errors of type C (waveform mismatch). A Wiener filtering approach could be used to restore the glottal pulse shape when it is not distorted too much by the concealment. Alternatively, when the error is too large to be corrected, it might be better to reset the past excitation signal (or at least the main part of the glottal pulse) using a small codebook as proposed for errors of type A. This approach would also have the advantage of providing a fallback solution in case of classification errors.

8. References