Method of Packet Errors Cancellation Suitable for any Speech and Sound Compression Scheme

Balázs Kövesi and Dominique Massaloux
France Telecom R&D, 2, Av Pierre Marzin, 22300 LANNION – France
{balazs.kovesi,dominique.massaloux}@rd.francetelecom.com

ABSTRACT

The present paper presents a novel method that has been successfully employed to cope with high packet loss rate of encoded speech and sound signals transmitted over IP, using several audio codecs.
This technique relies on an A-b-S paradigm such as found in CELP coders, where the spectrum characteristics of the signal prior to the packet loss are used to generate the synthesis parameters. The estimation of the parameters being performed on the output signal, the algorithm is independent from the type of codec used to produce this output signal. However, better results can be obtained by adapting the nature of the generated excitation signal to that of the codec. For predictive coders it is also important to feed back correctly the decoder memories using the signal generated during the packet loss.
The algorithm pays a special attention to the control of the synthesized signal energy, using an adaptive gain control. Several techniques have also been introduced to avoid the "robot voice" produced by a too static (and oversimplified) model.
This method had been adapted to a MDCT coder [1] running in a visioconferencing over IP application, in conjunction with optimized reception buffers handling. It has been observed that the conversation still remains possible in severe transmission conditions (i.e. people can forget the effects of the errors and focus on the discussion).

1. INTRODUCTION

In applications like Voice Over Internet Protocol (VoIP) or audio streaming of radio programs, packet losses may generate very annoying artifacts, especially if the system does not handle carefully the problem.
Efficient VoIP systems generally cope with packet loss by combining clever buffer management and error concealment techniques employed at the decoder side. Such error concealment techniques extrapolate the samples of the missing signal relying on the signals and data available from the neighboring valid frames.
Most of these techniques were developed in the case of frames based coders (such as CELP and transform codecs), and rely on the algorithm used in the codec, constituting de facto an extension of the decoder.
For the CELP coding algorithm [2], missing samples are obtained using the CELP synthesis model: an excitation signal filtered through an LPC (Linear Prediction Coding) filter and, in the case of a voiced signal, using a Long Term Prediction (LTP) filter. The generation of the LPC and LTP parameters is deduced from past valid frames. The excitation signal is generally generated in a different way for voiced and unvoiced signals.
For the transform coders, the techniques of rebuilding of the missing frames are also based on the hosting coding structure: the algorithms, such as [3], aim at regenerating the lost transform coefficients from the last received coefficients.
Finally, sample-based coders generally use waveform substitution techniques [4][5]. This type of methods reconstructs the signal by selecting portions of the past-decoded signal before the lost period. Smoothing techniques are applied to avoid the discontinuity artifacts produced by the concatenation of the various signals.

The present paper proposes an error concealment technique able to run in conjunction with any coding technology. This technique can be used for example in the case of sample-based coders (like G.711) where the decoder structure is not directly usable for error concealment method.
The presented technique is based on the general idea of the method previously described in [6], which proposed to apply to a transform coder a frame erasure concealment paradigm equivalent to that used in CELP coders. The concept was proven to be valid yet suffered from drawbacks: synthetic voice, unpleasant resonances…
It was identified that these artifacts were due to an insufficient control of LTP synthesis filters and of the energy. The present paper describes a new technique that allows the concealment of erased frames with smaller perceptual distortion even for higher error rates. This is obtained on one hand by an efficient generation of the excitation signal, and on the other hand by a sophisticated adaptive gain control that especially improves the performances in case of long erased zones.
Another feature is the existence of a smoothing procedure avoiding audible discontinuities between the signal rebuilt during the erased period and the next decoded signal.
Finally this technique ensures the update of the decoder memory. This update is particularly important in the case of recursive or predictive coding structures: it limits the de-synchronization between the encoder and the decoder.
The remainder of this paper is organized as follows: Section 2 describes the principle of the proposed method. Section 3 describes the way to implement it in the context of a particular coding structure. Section 4 generalizes the paradigm to the case of other types of coding structures. Finally Section 5 concludes this paper.

2. PROPOSED METHOD

The figure 1 presents the connections between the frame error concealment (FEC) module and the decoder.

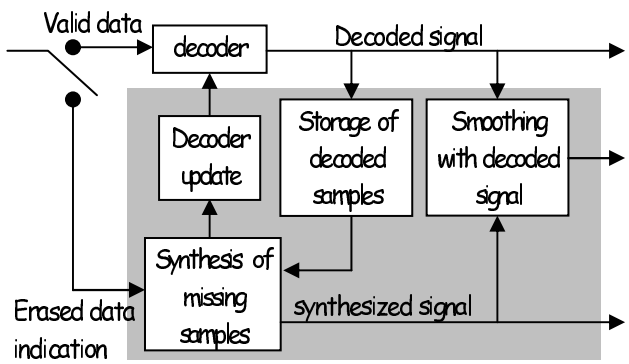

Figure 1: Main functions of the method

The frame error concealment module receives as input the indication of erased data and the decoded signal during valid periods. Its outputs are the reconstructed signal during erased periods, the smoothed signal after an erasure and data transmitted to the decoder for its memory update.
The FEC module consists of the following main parts:

1. Storage of the decoded samples when the transmitted data are valid,
2. During an erased block, synthesis of the samples corresponding to the lost data,
3. Memory update of the decoder. This update can be carried out either just after the generation of the erased samples, or when the transmission is restored.
4. When the transmission is restored, smoothing between the synthesis samples produced during the erased period and the decoded samples.

3. IMPLEMENTATION IN A MDCT CODEC

The algorithm will be presented via its implementation in a wide band transform codec based on the MDCT (Modified Discrete Cosine Transformation) transform. The length of the MDCT analysis and synthesis windows is 40ms with 50% (20ms) overlap, hence the frame length is 20ms (320 samples at 16kHz sampling frequency).
Figure 2 illustrates the overlap-add synthesis method where the nth 20ms output frame is obtained by addition of the second half of the $(n-1)^{th}$ decoded windowed signal and first half of the current one.

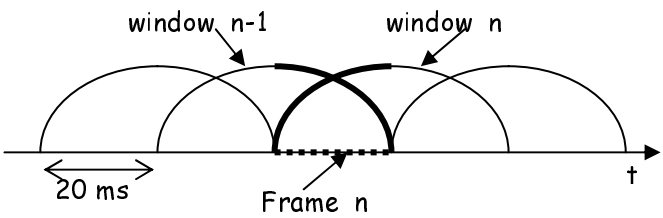

Figure 2: Overlap-add of an inverse MDCT transform

The second half of the windowed signal obtained after the IMDCT transform is kept in the decoder memory for the next frame.

3.1. Illustration of the frame erasures effect

Figure 3 illustrates the fact that if the nth (bitstream) frame is erased, both the nth and the $(n+1)^{th}$ output frames are disturbed. Generally speaking, the loose of x bitstream frames affects $x+1$ output frames.

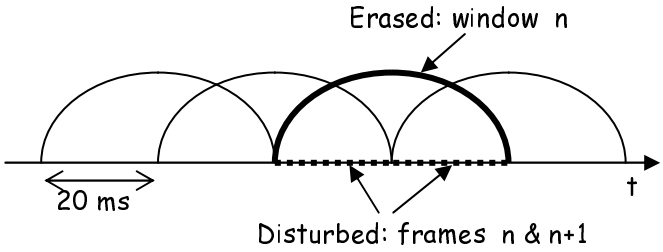

Figure 3: Disturbed frames in case of a single frame erasure

The frame error concealment technique will now be explained according to the sequence of operations presented in section 2.

3.2. Storage part

After the decoding of a valid frame, the frame error concealment module updates its memory with the last decoded samples. In the given example the decoded signal buffer contains 640 samples (40ms) necessary to estimate properly low frequency pitch lags (cf. paragraph 3.3.1.2). The energy of the valid frames is also calculated and the energy values for the last 5s are also kept in memory.

3.3. Synthesis part

The synthesis module is based on the CELP synthesis model. The synthesis method is different for the first erased frame after a valid period and for the next erased frames. This is due to the fact that analyses need to be performed for the first frame that will also be used for the next frames.

3.3.1. First erased frame after a valid period

The operations performed in case of a first erased frame are illustrated by figure 4.

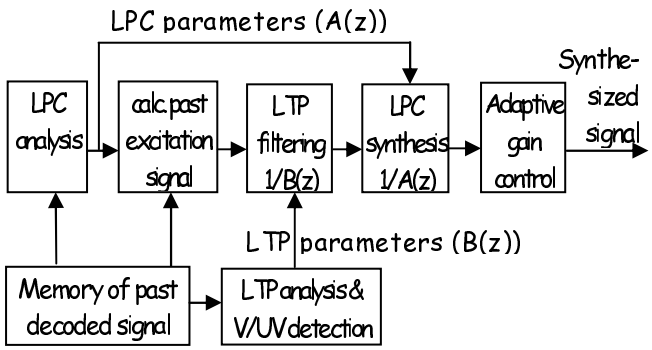

Figure 4: Synthesis of the missing samples

3.3.1.1. LPC analysis:

The spectral envelope is modeled by an LPC filter *A(z)* [7]. A common way to estimate the spectral envelope of an erased frame is to take that of the signal preceding the erasure. The analysis is carried out on the last stored 20ms decoded signal by a classical method, using an asymmetrical Hamming window. Since the obtained coefficients need not to be transmitted, the LPC analysis order can be higher than

that usually used in CELP coders (32 coefficients for the presented example). This allows better performances on music.

3.3.1.2. LTP analysis and voiced / unvoiced detection

A precise estimation of the voicing degree and of the pitch value in case of voiced sound is crucial for the good performance of the error concealment method. Errors like halving the pitch period in a voiced frame or using a harmonic excitation for an unvoiced frame can result in very annoying artifacts. That is why special care was taken for the estimation of these parameters.

Only integer pitch values are examined in the range [26, 320] (600Hz – 50Hz) by computing for each value a kind of normalized correlation. To better model the evolution of the decoded signal just before the erased period, this correlation for a pitch *P* is computed on the last *L=max(2P, 80)* stored samples (at least 80 samples are involved for the small pitch values) in the following manner:

$$\mathrm{Corr}(P) = \frac{2\sum_{i=640-L+P}^{639} \tilde{s}_i \tilde{s}_{i-P}}{\sum_{i=640-L}^{639} \tilde{s}_i^{\,2} + \sum_{i=640-L+P}^{639-P} \tilde{s}_i^{\,2}}$$

where $\tilde{s}_0 \ldots \tilde{s}_{639}$ is the past decoded signal. At the first stage the value *P'* where *Corr(P)* is maximum is considered as the pitch value. This decision can be modified by further verifications: for example to avoid the detection of a multiple or a fraction of the true pitch.

Then the voicing decision is taken essentially in function of the selected *Corr(P')* value but other parameters like the energy of the frame with respect to the stored energy of the last 5 seconds, or the energy evolution of the pitch periods are also taken into account.

3.3.1.3. Computation of the past excitation signal

Knowing the LPC coefficients, the past excitation signal is computed by filtering the past decoded signal through the inverse LPC filter *A(z)*.

3.3.1.4. Generation of the current excitation signal

The excitation samples are generated by LTP filtering the past excitation signal. The procedure is different for voiced and unvoiced frames.

<u>Voiced frames</u>

To avoid synthesizing too regular and too harmonic sounds, the voiced excitation signal is the sum of two components. One, called exc_L, highly harmonic but limited to the lower frequency bands, is obtained by the following combination of LTP filtering and low pass FIR filtering:

$$\mathrm{exc_L}(i) = 0.15\mathrm{exc}(i-P'-1) + 0.7\mathrm{exc}(i-P') + 0.15\mathrm{exc}(i-P'+1)$$

The second component is obtained by a LTP filtering where the pitch value is randomly modified for each sample. Furthermore, this filter is combined by a fifth order FIR high pass filter.

<u>Unvoiced frames</u>

The unvoiced excitation is also obtained by a LTP filtering where the same low pass FIR coefficients are used as for exc_L, but to randomize the resulting signal the pitch value is increased regularly and the sign of the signal is randomly changed.

Another procedure was also designed to avoid large and sudden energy variation in the unvoiced excitation signal.

The unvoiced excitation is often generated as a random signal. It was found that using an LTP like filtering as described above instead of a white noise helps to preserve the nature of the signal in the erased periods. For example, using a Gaussian noise as unvoiced excitation in a CELP codec introduces a different ("smoother") sound in the output signal.

3.3.1.5. LPC synthesis:

The synthesized samples are obtained by LPC filtering the generated excitation signal with the filter *1/A(z)*.

3.3.2. Next erased frames

For the next erased frames, no new analysis is needed.

The missing samples are synthesized by generating the excitation signal as described in paragraph 3.3.1.4 and by filtering this excitation according to the paragraph 3.3.1.5.

3.3.3. Adaptive gain control:

In case of long erased periods (> 20ms), the perceived quality of the frame error concealment largely depends on the energy management of the synthesized output signal. Generally, the output energy is slowly decreased towards a minimum energy level. In the present example this level is the lowest energy level memorized during the last 5s. When a VAD decision is available this can be also the background noise level.

The energy adaptation has to be different for stationary signal where the energy is kept high for a longer interval than for non-stationary parts when it is better to decrease the energy rapidly. The adaptation depends also on the pitch value as it was found that the energy level of a low pitch voice could be maintained longer without annoying artifact.

In this spirit, two different adaptation laws are applied depending on the stationary / non-stationary decision. This decision is available from the LTP analysis module (paragraph 3.3.1.2).

The adaptive gain control constitutes the last part of the general synthesis module.

3.4. Recoverable information

Due to the particular structure of the MDCT synthesis algorithm, information is available from both sides of the erasure that can be exploited.

As illustrated in figure 5 in the case of a two-frames erasure, the first and the last disturbed output frames (the $(n-1)^{th}$ and the $(n+1)^{th}$ output frames in the figure) are not completely lost: half of the needed information is received in the neighboring valid bitstream frames (in the $(n-2)^{th}$ and the $(n+1)^{th}$ bitstream frames in figure 5). After synthesis of the missing samples as presented above, the following operations are done:

On the first two synthesized output frames an MDCT transform is performed followed by the inverse transform. This allows using the available memory of the transform.

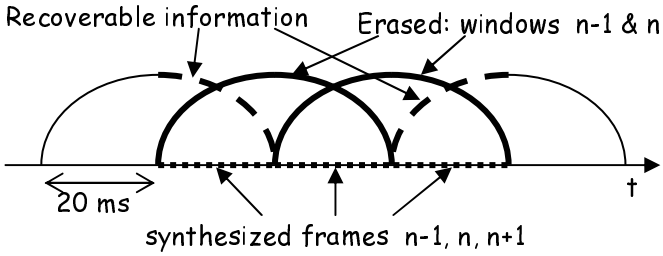

Figure 5: Recoverable information

3.5. Decoder memory update part

In the same way as in the previous paragraph, an MDCT / IMDCT transform pair done on the last two synthesized output frames permits to update the memory for the next IMDCT transform.

3.6. Smoothing part

In the example illustrated by the figure 5, the synthesis module synthesizes the frames *n-1*, *n* and *n+1*. Then the $(n+2)^{th}$ frame is correctly decoded from the valid $(n+1)^{th}$ and $(n+2)^{th}$ bitstream frames. In the absence of any smoothing operation the continuity at the border of these two parts is not ensured. If the decoder memory is updated as described in paragraph 3.5, by doing an IMDCT transform with the first received frame, the overlap-add also guarantees the smooth transition between the FEC synthesized domain and the error free domain.

4. GENERALISATION TO OTHER CODEC TYPES

Though the proposed frame error concealment method was presented through the example of an MDCT transform codec, it can be adapted to any coding scheme. It was successfully implemented in sample-based codecs (G.711, G.721, G722), in a CELP codec (G.723.1) and in a scaleable codec composed of a CELP and a transform codec.
Notice first that the decoded samples storage and the synthesis module do not depend on the codec type. On the opposite, the memory update and smoothing techniques may be general or adapted to the codec to be more efficient.

4.1. Memory update

The memory update can be generalized by doing a coding and decoding operation pair on the synthesized frames, as it is partly done in the given example. The drawback of this solution is its complexity for example in the case of a CELP codec. Simpler backtracking of the memories can also be envisaged.

4.2. Smoothing

A codec independent smoothing solution is that along with the first received and decoded output packet the synthesized output signal is generated in parallel for a limited period (for example 5ms) and a crossfading is applied on these two signals.
Some coding schemes allow the implementation of a more efficient smoothing. This was the case of the given example (paragraph 3.6)
In the case of CELP coding the mentioned memory update of the LPC synthesis filter ensures the output signal continuity.

5. CONCLUSION

We have presented an efficient frame error concealment method that can be implemented in any coding scheme. Though this method is optimal for speech, it also gives good performances on music. Special attention was paid to avoid too synthetic sounds for erased voiced frame and to keep the nature of the unvoiced frames unchanged. The energy management plays an important role in the performance of the proposed method as well as the careful update of the decoder memory and the smoothing after an erased period.
Several informal subjective tests have shown the good behavior of this method compared to other existing and standardized methods [8]. Notice also that the technology is adapted to real time applications: it has been successfully implemented into group communication tools.

6. REFERENCES

[1] Y. Mahieux, J.P. Petit. "High quality aaudio transform coding At 64 kbit/s". Trans IEEE. on Com., Vol.42-11, nov.1994, PP.3010-3019.
[2] B.S. Atal and M.R. Schroeder. "Predictive coding of speech signal and subjective error criteria". Trans IEEE. on Acoustics, Speech and Processing Signal, 27:247-254, June 1979.
[3] Y. Mahieux, "Dissimulation errors of transmission", patent 92 06720 deposited on June 3, 1992.
[4] AT&T (D.A. Kapilow, R.V. Cox) "A high quality low-complexity algorithm for frame erasure concealment (FEC) with G.711", Delayed Contribution D.249 (WP 3/16), ITU, may 1999.
[5] D.J. Goodman et al, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications". Trans IEEE. one Acoustics, Speech and Processing Signal, Flight. ASSP-34, December 1986, PP. 1440-1448.
[6] P.Combescure et al, "A 16,24,32 kbit/s Wideband Speech codec Based one ATCELP". Proc. of ICASSP conference, 1998.
[7] L.R. Rabiner, R.W. Schafer. "Digital processing of speech signals". Bell Laboratories Inc, 1978.
[8] F. Bouteille, P. Scalart, B. Kovesi. "Packet Loss Concealment Using Audio Morphing". ETSI STQ Workshop, February 2003, Sophia-Antipolis.