ENHANCED TIME DOMAIN PACKET LOSS CONCEALMENT IN SWITCHED SPEECH/AUDIO CODEC

Jérémie Lecomte[1], Adrian Tomasek[1], Goran Marković[1], Michael Schnabel[1], Kimitaka Tsutsumi[2], Kei Kikuiri[2]

[1]Fraunhofer IIS, Erlangen, Germany, [2]NTT DOCOMO, INC., Yokosuka, Japan
jeremie.lecomte@iis.fraunhofer.de

ABSTRACT

This paper describes new time domain techniques for concealing packet loss in the new 3GPP Enhanced Voice Services codec. Enhancements to the existing ACELP concealment methods include guided, improved pitch prediction, increased flexibility and accuracy of pulse resynchronization. Furthermore, the new method of separate linear predictive (LP) filter synthesis aims for sound quality improvement in case of multiple packet loss, especially for noisy signals. Another enhancement consists of a guided LP concealment approach to limit the risk of creating artifacts during recovery. These enhancements are also used in the presented advanced TCX concealment method. Subjective listening tests show that quality is significantly increased with these methods.

Index Terms— *EVS, Packet Loss Concealment, guided concealment, ACELP, TCX*

1. INTRODUCTION

The Enhanced Voice Services codec (EVS) [1] is the next generation 3GPP real-time communications codec. It is based on an architecture that allows seamless switching between a frequency domain and an LP-domain core [2]. The EVS codec is designed for packet-switched networks such as LTE. Even the LTE network is known to be prone to errors; therefore, an important design criteria is error robustness [3]. This paper focuses on concealment technologies applied in the time domain (TD).

Section 2 gives an overview on the state of the art methods. Section 3 describes the improvements done on the ACELP concealment and presents a guided concealment approach that calculates the future pitch on the encoder side as well as a novel scheme based on separate synthesis of the periodic and the noisy excitation. In state of the art methods, most of the MDCT core related concealment algorithms are applied in the MDCT domain. One of the main factors limiting the quality of frequency domain based technologies is phase mismatch on the frame borders that is clearly audible for monophonic signals. To overcome this problem, a new technique developed to enhance the concealment of speech like signals in transform coding is described in section 4. The improved recovery during the first valid frame after a packet loss is presented in section 5. Subjective evaluation results in section 6 demonstrate the improved performance of the proposed methods.

2. STATE OF THE ART

There are two time domain concealment approaches known from the literature: waveform and parameter based. Waveform based approaches like Time Scale Modification [4] are out of scope for this paper and will not be described further. The most commonly used parameter based time domain concealment approaches are described in ITU-T G.718 [5] and AMR-WB+ [6].

In G.718 the ACELP concealment method is based on the previous frame class, which is either transmitted and decoded from the bitstream, or estimated in the decoder. Each valid frame is classified as unvoiced, voiced, onset or transition. No periodic excitation is generated for the lost frame after a valid unvoiced frame, otherwise the periodic excitation is constructed by repeatedly copying the last low-pass filtered pitch period of the previous frame. The CELP adaptive codebook used in the next frame is updated only with this periodic excitation. The length of the segment that is copied is $T_r = \lfloor T_c + 0.5 \rfloor$, where T_c is the last adaptive codebook lag with fractional precision. Since the pitch may change during the lost speech frame, the position of glottal pulses may be wrong near the end of the constructed excitation. This would produce problems in the correctly received ACELP frame after the concealed frame. To overcome this problem a resynchronization method adjusts the positions of the glottal pulses to the estimated glottal pulse positions, that are estimated in the decoder based on the result of a pitch extrapolation method [5]. A uniformly distributed random noise, filtered with a linear phase high pass FIR filter, is used as the noisy excitation. The gain is progressively reduced to an averaged gain, obtained over the last 20 correctly received unvoiced frames.

AMR-WB+ [6] uses a time domain concealment method when the previous frame is transform coded. There the adaptive codebook and the pitch lag are derived from the

synthesis signal for every correctly received TCX frame and are reused in case of packet loss. The concealment is performed in the excitation domain and operates at 12.8 kHz. The LP filter available from the bit-stream is reused for LP filtering the extrapolated adaptive codebook.

3. ACELP CONCEALMENT

In EVS, the concealment of packet loss after an ACELP frame is similar to the case described in [5] and [7], where neither the last pulse position is known nor is the future frame available.

Generating a repetitive harmonic signal tends to sound artificial. Thus, in case of a long burst of errors the periodic excitation fades towards silence and the synthesized noisy signal fades towards a comfort noise level. As EVS is a switched codec with a speech and a transform coder it is not possible to trace the innovative codebook gains continuously and to use the average as target noise level during packet loss concealment (PLC). The comfort noise level is derived from the comfort noise generator (CNG) system that is featured in the EVS codec [8]. During the clean channel decoding, the CNG system is continuously estimating the FFT spectrum and the RMS level of the background noise. The later is used as the long-term target RMS level of the noise part during PLC. Informal experiments have shown that this gives a more pleasant sound than muting in case of burst of errors. The speed of the convergence to the comfort noise is controlled by an attenuation factor. The latter depends on the number of consecutively lost packets and on the parameters of the last received frame. Those parameters being the Euclidian distance between the last two line spectral frequencies (LSFs) pairs, the coder type and the signal class of the last good frame. In contrast to the prior art, in the EVS codec also the shape of the high pass FIR filter used on the noisy excitation is changing towards white noise during a consecutive loss of packets

3.1. Pitch extrapolation

A novel pitch extrapolation based on straight line fitting [9] is utilized in the EVS Codec. As pointed out for example in [10] and [11], representing a pitch contour with linear interpolation of the pitch coded at the frame borders does not affect the quality. The main benefit of the proposed algorithm is, that it uses a weighted error function for the linear fitting. Stable and more recent pitch lags contribute more to the extrapolated pitch. Coefficients of the linear function are determined by minimizing the error function defined by the equation:

$$err(a,b) = 0.125 \sum_{i=-1}^{-5} g_p^i (11+i)\left((a+bi) - \left(d^i\right)\right)^2 \quad (1)$$

where g_p^i and d^i are the past adaptive codebook gains and lags for each previous sub-frame. Note that $(11+i)$ is acting as a factor that puts more weight on the more recent pitch lags and g_p^i puts more weight on pitch lags associated with higher gains.

The minimization is done by solving the linear equations obtained by setting:

$$\frac{\partial err}{\partial a}(a,b) = \frac{\partial err}{\partial b}(a,b) = 0 \quad (2)$$

The predicted pitch lag at the end of the concealed frame is then calculated using:

$$T_{ext} = a + b(M-1) \quad (3)$$

where M is the number of sub-frames in a frame.

3.2. Pulse resynchronization

As in [5][7][12], the pulse resynchronization is done by adding or removing samples in the minimal energy regions between glottal pulses.

In contrast to [5][7][12], the proposed pulse resynchronization algorithm; in line with the linear pitch extrapolation; assumes that the number of samples to be removed or added in each pitch cycle is linearly changing. The pitch change per sub-frame is given by:

$$\delta = \frac{T_{ext} - T_c}{M} \quad (4)$$

Based on the expectation to add $(p[i] - T_r)\frac{L}{MT_r}$ samples in the i-th sub-frame, where $p[i] = T_c + (i+1)\delta$ and L is the frame length, the total number of samples to be removed or added in the concealed frame is:

$$d = \delta \frac{L}{T_r}\frac{M+1}{2} - L\left(1 - \frac{T_c}{T_r}\right) \quad (5)$$

The index of the last glottal pulse that will be present after the resynchronization is:

$$k = \left\lceil \frac{L - d - T[0]}{T_r} - 1 \right\rceil \quad (6)$$

where $T[0]$ is the location of the first glottal pulse in the constructed periodic excitation, found by searching for the absolute maximum.

In contrast to the iterative calculations in [5][12], assuming linearity allows direct calculations. Furthermore it allows modifications before the first and after the last pulse (single pulse case included), which are incorrectly handled and introduce abrupt pitch changes in [5][12]. The number of samples to be added or removed is calculated as:

$$\Delta_0^p = (|T_r - T_{ext}| - (k+1)a)\frac{T[0]}{T_r} \quad (7)$$

$$\Delta_i = |T_r - T_{ext}| - (k+1-i)a, 1 \le i \le k \quad (8)$$

$$\Delta_{k+1}^p = |d| - \Delta_0^p - \sum_{i=1}^{k} \Delta_i \quad (9)$$

where Δ_0^p is the number of samples before the first pulse, Δ_i between two pulses and Δ_{k+1}^p after the last pulse.
a is calculated as:

$$a = \frac{|T_r - T_{ext}|(L-d) - |d|T_r}{(k+1)\left(T[0] + \frac{k}{2}T_r\right)} \quad (10)$$

3.3. Guided pitch extrapolation

On top of prior art, where the last valid pulse position might be transmitted in the bitstream [7], in the EVS codec at 24.4 kbps the pitch lag of the future frame is calculated within the look-ahead buffer at the encoder side and transmitted to the decoder to assist the pitch extrapolation in the case of packet loss. In order to reduce the average bitrate of the side information the pitch lag is coded differentially to the previous sub-frame pitch lag and transmitted only for onset and voiced frames. Since the look-ahead necessary for LP filter analysis can be exploited for the pitch estimation, no additional delay is required.

3.4. Separate LP filter Synthesis

This method aims to keep speech/music quality high, even when background noise is present. This technique improves the subjective quality mainly for burst packet loss.

Separate sets of LP filter coefficients are used for the periodic and the noisy excitation. Each excitation is filtered by its corresponding LP filter and afterwards added up to obtain the synthesized output, as shown in Figure 1. In contrast, other known techniques [5] add up both excitations and feed the sum to a single LP filter.

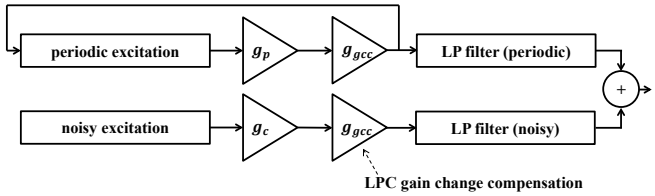

Figure 1 – *TD PLC using separate LP filter synthesis.*

The energy during the interpolation is precisely controlled by compensating for any gain that is introduced by the change of the LP filters. Using a separate set of LP filter coefficients for each excitation has the advantage that the voiced signal part is played out almost unchanged (e. g. desired for vowels), while the noise part is being converged to the background noise estimate [8].

4. TIME DOMAIN TCX CONCEALMENT

A frame will often be coded with TCX, even if the signal contains speech. This happens because TCX is usually more suited for speech with background noise or for music. However, in many cases frequency domain concealment has poor performance for speech signals. For example a long transform length makes it hard to conceal quickly varying harmonic structures while keeping the pitch contour smooth within one transform window. The relatively low performance of concealment for speech coded with TCX was improved by introducing concepts from ACELP.

In contrast to prior art [6], TD TCX PLC in EVS operates at the output sampling rate (up to 48 kHz) and derives the 16th order LP filter parameters from the past synthesized signal. The past excitation is obtained by filtering the past pre-emphasized time domain signal through the LP analysis filter. The first order pre-emphasis filter coefficient depends on the sampling rate and is in the range from 0.68 to 0.9. In case of consecutively lost packets, the LP filter parameters and the excitation are not recalculated, but the last computed ones are reused.

Furthermore, unlike [6], TD TCX PLC uses the same procedure as the EVS ACELP concealment for constructing the periodic excitation, including low-pass filtering, improved pitch extrapolation and pulse resynchronization. TD TCX PLC also includes the noise addition with the adaptive high pass filtering.

Pitch information for a TCX frame, consisting of the pitch lag T_c and the pitch gain, is computed on the encoder side and transmitted in the bit-stream. TD TCX PLC uses the pitch information from the previously received TCX frame. At low bitrates, the pitch information is also used for the long term prediction (LTP) post-filter [2], whereas at high bitrates it is used solely for the concealment.

For all frames classified other than unvoiced, the gain of the periodic excitation G_p is computed using a normalized autocorrelation with delay T_r directly on the past pre-emphasized synthesized signal *syn* rather than on the excitation signal, as done in ACELP:

$$G_p = \frac{\sum_{i=0}^{L/2-1}(syn(i-L/2)\cdot syn(i-L/2-T_r))}{\sum_{i=0}^{L/2-1}(syn(i-L/2-T_r))^2} \qquad (11)$$

This avoids the drawback of imprecise modeling of the formants with the low order LP filter at high sampling rates. Similar to ACELP concealment, G_p will determine the amount of tonality that will be created. For unvoiced frames, no periodic excitation is generated.

As in state of the art ACELP concealment, a random noise generator is used to create the noisy excitation, which is then high pass filtered to prevent addition of rumbling noise in the lower frequency region. Like in the ACELP concealment, the noisy excitation is slowly being converged towards white noise for consecutive packet loss. After that, the noisy excitation is pre-emphasized for voiced and onset frames to avoid adding disturbing noise in between the harmonic frequency structure.

The gain of the noise is chosen to be equivalent to the energy of the LTP residual in the last half frame of the past excitation signal, exc, using the delay T_r and the gain G_p:

$$G_c = \sqrt{\frac{\sum_{i=0}^{L/2-1}\left(exc(i-L/2)-G_p\cdot exc(i-L/2-T_r)\right)^2}{L/2}} \qquad (12)$$

For consecutive frame loss, the gain is progressively faded to a value that causes the RMS level to match with the CNG level. The CNG level derivation is the same as for ACELP.

Finally, the synthesized signal is obtained by filtering the total excitation through the derived LP synthesis filter followed by the first order de-emphasis filter.

5. RECOVERY

Since the excitation and the synthesis memories are updated during the concealment, the transition to the first good ACELP frame after packet loss is seamless.

For transition to the first valid TCX frame, the overlap-add buffer is constructed using the same procedure as for a concealed frame during a consecutive packet loss, followed by the artificial construction of the time domain aliasing [13].

In the case of the first frame after packet loss featuring significantly different content than before the loss, e. g. for onset frames, the LP filter spectra sometimes feature an extremely sharp peak due to wrong concealed LSF in the lost frame and its application to the LSF extrapolation at the subsequent recovery frame. Then the peak causes a sudden power increase in the decoded speech and severe quality degradation. To mitigate the power fluctuation, the spectrum is modified to eliminate the peak by forcing wider LSF gaps compared to the clean channel LSF decoding. In case of sharp peaks being present, the encoder transmits a flag indicating the necessity of this spectral power diffusing.

6. PERFORMANCE EVALUATION

To show the performance of the concealment tools proposed in this paper a MUSHRA [14] test with 9 expert listeners was conducted in an acoustically controlled environment using STAX headphones.

The EVS codec was evaluated under clean and impaired channel conditions (6% FER), for wide band at 9.6 kbps and 24.4 kbps against the corresponding reference codecs identified for the 3GPP selection test [15]. The reference is AMR-WB/G.718 IO (RefCodec) at 12.65 kbps and 23.85 kbps for noisy speech under impaired channel conditions. A restricted EVS decoder (EVS VC) was added to the test, where the guided PLC, TD TCX PLC and fading to background noise were disabled. Furthermore, in EVS VC the pitch prediction and the pulse resynchronization from G.718 were used instead of the one proposed above. The following test items known from USAC development [16] were used: es03 (English female, clean speech), te1_mg54_speech (German male, clean speech), Alice_short (English female between/over classical music), lion (English male between effects), SpeechOverMusic_1_short (English female over noise) and phi4_short (English male over music).

Figure 2 and Figure 3 show the average absolute scores with 95% confidence intervals for each codec at the two tested bitrates. For better visualization, the 3.5 kHz anchor (rated on average with 25) and the hidden reference (always recognized correctly) are not displayed.

The results show that the EVS codec is significantly better than the reference codec, namely AMR-WB/G.718 IO, for clean channel as well as for the noisy channel. Moreover the tests show that the overall quality of the impaired EVS codec improves with the proposed PLC techniques. Based on T-test measures, in both listening tests the difference between the restricted and the standardized EVS codec is statistically significant. Furthermore, the proposed PLC techniques allow the EVS codec with 6% packet loss to compete with the clean channel AMR-WB/G.718 IO at bitrates around 24 kbps.

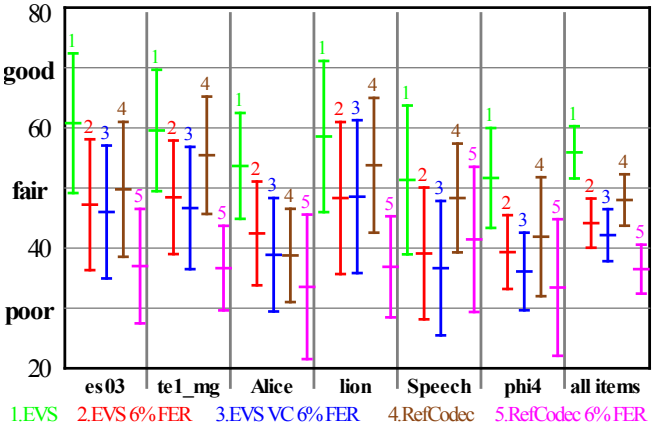

Figure 2 - *Result of the 9.6/12.65 kbps listening test.*

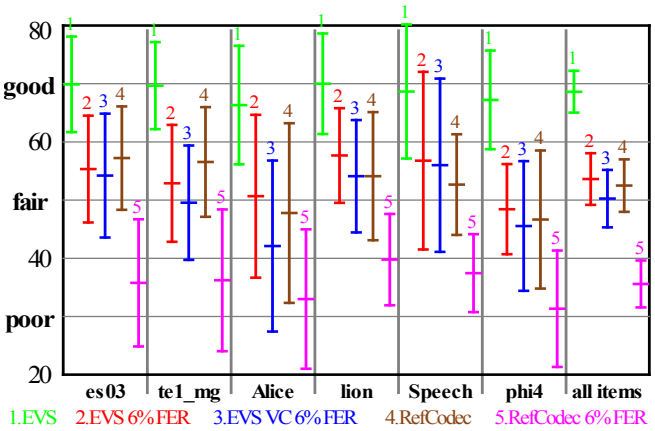

Figure 3 - *Result of the 23.85/24.4 kbps listening test.*

7. CONCLUSION

In this paper various advanced approaches to error concealment in the time domain were discussed. In the ACELP part of the EVS concealment, the main improvements have been achieved by altering the pitch prediction and the pulse resynchronization, including the encoder assisted pitch extrapolation. Furthermore a new technique for generating the synthesis signal using the periodic excitation and the noise like excitation was described. The time domain TCX concealment method is introduced to compensate the relatively low performance of frequency domain concealment for speech signals. The guided LP filter concealment reduces the risk of creating artifacts during recovery. All these changes lead to an increase of quality under erroneous channel conditions, as shown by the listening tests.

8. REFERENCES

[1] 3GPP, TS 26.441, “Codec for Enhanced Voice Services (EVS); General Overview (Release 12),” 2014.

[2] 3GPP, TS 26.445, “Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description (Release 12),” 2014.

[3] 3GPP, TS 26.447, “Codec for Enhanced Voice Services (EVS); Error Concealment of Lost Packets (Release 12),” 2014.

[4] S. Roucos, A. Wilgus, “High quality Time-Scale Modification of Speech”, ICASSP, pp. 236-239, 1985.

[5] ITU-T Recommendation G.718, "Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s," ITU-T, Geneva, 2008.

[6] 3GPP, TS 26.290, “Audio codec processing functions; Extended Adaptive Multi-Rate – Wideband (AMR-WB+) codec; Transcoding functions (Release 12),” 2014.

[7] T. Vaillancourt, M. Jelinek, R. Salami and R. Lefebvre, "Efficient Frame Erasure Concealment in Predictive Speech Codecs using Glottal Pulse Resynchronisation,” in Proc. IEEE Int. Conference on Acoustic, Speech and Signal Processing (ICASSP) vol. 4, pp. 1113-1116, April 2007.

[8] 3GPP, TS 26.449, “Codec for Enhanced Voice Services (EVS); Comfort Noise Generation (CNG) aspects (Release 12),” 2014.

[9] C. L. Lawson, R. J. Hanson, “Solving Least Squares Problems. Series in Automatic Computation", Prentice-Hall, Englewood Cliffs, USA, 1974.

[10] W. B. Kleijn, R. P. Ramachandran and P. Kroon, “Interpolation of the pitch-predictor parameters in analysis-by-synthesis speech coders,” in Proc. IEEE Int. Conference on Acoustic, Speech and Signal Processing (ICASSP) vol. 2, pp. 42-54, January 1994.

[11] M. Leong, P. Kabal, “Smooth Speech Reconstruction Using Waveform Interpolation,” in Proc. IEEE Workshop on Speech Coding for Telecommunications, pp. 39-40, October 1993.

[12] ITU-T Recommendation G.729.1, "G.729 based Embedded Variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," ITU-T, Geneva, 2006.

[13] J. Lecomte, P. Gournay, R. Geiger, B. Bessette, M. Neuendorf, “Efficient cross-fade windows for transitions between LPC-based and non-LPC based audio coding, ” in 126th Audio Eng. Soc. Convention, number 7712, Munich, May 2009.

[14] International Telecommunication Union, “Method for the subjective assessment of intermediate sound quality (MUSHRA)," 2001, ITU-R, Recommendation BS. 1534-1, Geneva, Switzerland.

[15] 3GPP Tdoc S4-130522, EVS Permanent document (EVS-3): EVS performance requirements, Version 1.4, April 2013.

[16] USAC Verification Test Report ISO/IEC JTC1/SC29/WG11 MPEG2011/N12232, July 2011, Torino, Italy.