# PACKET-LOSS CONCEALMENT TECHNOLOGY ADVANCES IN EVS

*Jérémie Lecomte[1], Tommy Vaillancourt[2], Stefan Bruhn[3], Hosang Sung[4], Ke Peng[5], Kei Kikuiri[6], Bin Wang[7], Shaminda Subasingha[8], Julien Faure[9]*

[1]Fraunhofer IIS, [2]VoiceAge Corp., [3]Ericsson AB, [4]Samsung Electronics Co. Ltd., [5]ZTE Corporation, [6]NTT DOCOMO Inc., [7]Huawei Technologies Co. Ltd, [8]Qualcomm Technologies Inc., [9]Orange Labs
Jeremie.lecomte@iis.fraunhofer.de

## ABSTRACT

EVS, the newly standardized 3GPP Codec for Enhanced Voice Services (EVS) was developed for mobile services such as VoLTE, where error resilience is highly essential. The presented paper outlines all aspects of the advances brought during the EVS development on packet loss concealment, by presenting a high level description of all technical features present in the final standardized codec. Coupled with jitter buffer management, the EVS codec provides robustness against late or lost packets. The advantages of the new EVS codec over reference codecs are further discussed based on listening test results.

***Index Terms***— *Concealment, speech coding, audio coding, VoLTE, EVS*

## 1. INTRODUCTION

Designed to meet the needs of packet-switched mobile communication networks, in particular Voice over LTE (VoLTE), the EVS codec [1] has been developed and standardized under the lead of the 3GPP Codec Working Group, 3GPP TSG SA WG4. The EVS codec can also be used in general IP telephony such as Voice over IP (VoIP) and Voice over WiFi (VoWiFi) for speech communications.

Using the RTP/UDP protocol, due to poor radio conditions or congestion in the IP network, an IP packet may be lost or late. The latter case would occur, for example, when a packet is sufficiently late that the receiver declares that the packet is lost. In a low-latency audio over IP transmission, discarding packets that are very late is preferable to having the receiver increase the delay.

This paper describes advanced packet loss concealment (PLC) algorithms designed for the EVS codec. The following sections will introduce the state of the art of modern audio and speech coding and packet loss concealment, followed by a technical description of the advanced concealment techniques in the EVS codec. Finally, performance test results based on the 3GPP Selection Test Plan are presented.

## 2. STATE OF THE ART

The EVS codec was developed based on two major existing codecs, namely G.718 and USAC. The G.718 codec is a narrowband (NB) and wideband (WB) embedded variable bit-rate codec for speech and audio operating in the range from 8 to 32 kbit/s. The Recommendation ITU-T G.718 is designed to be highly robust to frame erasures [2], thereby enhancing the speech quality when used in IP transport applications on fixed, wireless and mobile networks. The EVS codec has been designed ground up to stop error propagation using technologies such as transition coding [3], memory-less line spectral frequency (LSF) and gain quantization.

The ISO/IEC 23003-3:2012 MPEG-D Unified Speech and Audio Coding (USAC) standard [4] is based on enhancements done in the state of the art speech and audio coding technologies, such as AMR-WB+, HE-AAC and MPEG Surround. Although MPEG does not standardize the concealment mechanism, USAC was designed to mitigate packet losses. The Standard Digital Radio Mondial (DRM) describes part of the concealment [5].

## 3. TECHNICAL APPROACH

The EVS codec comprises a suite of advanced packet loss concealment technologies designed to work with signal classification, spectral envelope computation, LP-domain such as algebraic code-excited linear prediction (ACELP) core, modified discrete cosine transform (MDCT) core and bandwidth extension modules. Furthermore, the EVS codec uses a 'guided' PLC scheme for which the encoder provides supplementary data guiding the concealment in case of lost packets and enhancing the convergence and recovery afterwards [6].

Acknowledgment: The authors would like to thank the following people for their valuable contribution to this project: M.Jelinek, S.Ragot, V.Rajendran, D.Dewasurendra, V.Eksler, V.Malenovsky, K.Tsutsumi, R.Spershneider, M.Schnabel, G.Marković, Jonas Svedberg, Erik Norvell

### 3.1. Signal classification

Most of the concealment methods used in the EVS codec are based on signal classification [7]. The frame class is either transmitted and decoded from the bit-stream, or estimated in the decoder. The frame classification, both at the encoder and the decoder, is based on the following parameters: zero-crossing, pitch-synchronous normalized correlation, pitch coherence, spectral tilt, and pitch synchronous relative energy at the end of the frame. The parameters are normalized between 0 and 1 and combined in a figure of merit. The classification is then done by comparing this figure of merit to different thresholds.

The classifier used for EVS is based on G.718 with the following two adaptations. First, if the actual frame is in MDCT mode, only one long-term prediction (LTP) lag information is available for each frame so that the pitch coherence cannot be computed. Second, when the core coding mode is different than generic then the decoder classification is skipped and the signal is classified according to the coding mode (voiced, unvoiced or inactive).

### 3.2. Spectral envelope representation

The spectral envelope is estimated by means of Linear Prediction (LP) filters and quantized by means of the LSFs. The LSF parameters are coded in active frames based on two analysis windows (mid-LSFs and end-LSFs). At the decoder, the LSF parameters of a lost frame are extrapolated using the last frame's LSF parameters. The general idea is to fade the last good frame's LSF parameters towards an adaptive mean of the LSF vector. Compared to state of the art, a second LSF vector is derived based on the previous frame LSF but faded to the LSF representation of the comfort noise estimate done in the decoder side [1]. In the context of multiple frame losses, this second set of LSFs is used to slowly fade from the last active frame characteristics to comfort noise only. It prevents complete silence observed in standard muting mechanisms in case of burst losses.

In the decoder, the quantized end-frame LSFs of the current and the previous frame are combined using an unconstrained weighing vector to interpolate the mid-frame LSFs [1]. Hence, the extrapolated end-LSFs of a concealed frame could affect the mid-LSF interpolation (of the successive good frame) and could potentially create an LSF clustering that results in an unstable LSF synthesis filter. This can lead to severe artifacts in the output audio. To avoid such degradation, the decoder tests the mid-LSF stability by checking if the computed mid-LSFs are ordered correctly in increasing order with a minimum gap. If this is not the case, it uses a fixed weighing factor (biased toward the good frame) for mid-LSF interpolation, and discards the received weighing vector.

The LSFs in each subframe are computed in the decoder by combining the end-LSFs of the current and the previous frame and the mid-LSFs of the current frame, using fixed interpolation factors known to both the encoder and the decoder. However, the extrapolated end-LSFs of a concealed frame may deviate significantly from the end-LSFs at the encoder, and may adversely affect the audio quality of the successive good frame. Hence, an output gain ratio between the concealed frame and the successive good frame is estimated by computing the energies of the impulse responses of the corresponding synthesis filters. This ratio is used to adapt the subframe interpolation factors, along with other criteria. For example, a very low output gain ratio indicating an abrupt decrease in energy of the good frame compared to the concealed frame triggers the use of interpolation factors which give more weight to the end-LSFs of the good frame.

The concealment and interpolation of the LP parameters will lead to a change of the overall gain of the signal, which is unwanted in case of targeting a certain background noise level. Therefore the energy of the LP synthesis filter is measured and stored for the received frames, and it is used to compensate for the LP filter energy differences of the subsequently concealed frames. The LP filter might become unstable, creating resonant peaks in the spectrum, in case of voiced recovery, or a lost onset, because it is differentially coded and strongly reliant on concealed predictor memories. In such cases, the unstable LP filter is detected at the encoder and a flag is transmitted to the decoder. Consequently the decoder modifies the LP filter to suppress the spectral peak.

### 3.3. Concealment of LP-based coding

In case of frame erasures, the ACELP concealment strategy can be summarized as a convergence of the signal energy and the spectral envelope to the estimated parameters of the background noise. Similar principle holds also for the other LP-based modes of the EVS. The gain of the long term predictor is adjusted such that it converges to zero. The speed of the convergence is dependent on the parameters of the last correctly received frame, the number of consecutive lost frames, and on the stability of the LP synthesis filter. In general, the convergence is slow if the last good frame belongs to a stable segment and rapid if the frame belongs to a transition segment. The following section describes the novel features of the ACELP PLC scheme [1].

*3.3.1. Concealment of lost frame(s)*

The accurate estimation of the end-of-frame pitch of a concealed frame is essential to keep the adaptive codebook synchronized with the encoder to achieve fast recovery from the frame losses. The pitch extrapolator assumes that the encoder uses a smooth pitch contour. To overcome the glottal pulse position drifting inside a concealed frame during a voiced segment and to improve the decoder convergence, glottal pulse positions are adjusted in the concealed frame similar to method described in [8] using estimated end-of-frame pitch information. This estimate is based on weighted straight line fitting of past pitch values.

Furthermore, using the encoder look-ahead signal for the LP estimation, the LTP lag of the next frame is predicted and transmitted to improve the pitch estimation [9].

*3.3.2. Recovery after erasure*

When a frame erasure occurs, the adaptive excitation of the lost frame is extrapolated from the previous frame using the past pitch information. As the extrapolated pitch information is often incorrect, the encoder transmits the glottal pulse position of the last subframe of the previous frame as supplementary information. In particular, the glottal pulse position is used to correct and rapidly resynchronize the memory of the adaptive codebook prior to decoding the first good frame after an erasure. This correction can also bring transition artifacts due to a rapid resynchronization. Pitch synchronous waveform interpolation between the last prototype pitch period of the first good frame after erasure (synthesized using the corrected adaptive codebook memory) and the last prototype pitch period of the previous lost frame eliminates such potential boundary artifacts and at the same time keeps the gains at correct value due to faster resynchronization.

*3.3.3. Concealment for bandwidth extension*

The codec includes both time domain bandwidth extension (TBE) and frequency domain bandwidth extension (BWE) schemes on top of the LP-based cores. For BWE, simple frame repetition and attenuation are employed. For TBE, the upper band signal is reconstructed using three key parameters; temporal sub-frame gains, global frame gain, and high band LSF.

The TBE PLC utilizes the inter-frame dependency and the correlations between the lower band and the upper band to achieve smoothly reconstructed spectrum for the upper band. The high band LSF is copied from the previous frame. In super wideband TBE mode, the gain shapes are calculated based on the two previous ones and their gradients or just generated by attenuating the gain values from the previous frame. Starting from initial values, the gain shapes and the global frame gain are then further adjusted depending on the coder type, the frame class, the number of the consecutive lost frames, the energy and the tilt of the lower band. In wideband TBE mode, the gain shapes are set to a constant value, the global frame gain is calculated by attenuating the gain of the previous frame.

**3.4. MDCT domain**

Optimal performance under frame losses is obtained through selecting the most suitable PLC method for a given operating mode and signal dependent parameters such as coder type, classification, and the length of the error burst.

*3.4.1. TCX MDCT*

If the last good frame was coded with MDCT based TCX, four different optimized PLC techniques are used. They are selected based on criterions such as the number of consecutively lost frames, the last LTP gain, the number of detected tonal components and the waveform adjustment flag.

First a time domain PLC technique that is considerably different from the other three frequency domain concealment techniques is presented. This novel time domain TCX concealment operating in the excitation domain is used to improve the concealment of lost active speech frames and single instrumental music segments. LP analysis and then inverse filtering is done on the pre-emphasized synthesized time domain signal of the last frame to obtain the local LP parameters and the corresponding residual signal. Those are then used to conceal following single and multiple frame losses. The subsequent processing resembles the concealment in ACELP [9].

For non-periodic noise like signals, a low complexity technique, called sign scrambling, has been found to be effective. It is based on repeating the last frame and multiplying the spectral coefficients with a randomly generated sign to conceal the lost frame.

For tonal signals, a third method is used which is based on predicting the phase of the spectral coefficients of the detected tonal components. This method shows a consistent improvement for stationary tonal signals. A tonal component consists of a peak that existed in the last 2 received frames, and within its 6 surrounding bins. The pitch information available in the bitstream is used to improve the detection of the tonal components. The phase of the spectrum coefficients belonging to the tonal components is determined from the power spectrum of the second to last received frame. For the spectrum coefficients not belonging to tonal components, sign scrambling is used.

Finally, the last method uses the output of the sign scrambling as an initially compensated signal. Based on that a waveform adjustment is performed to obtain the concealed signal of the current lost frame by periodically extending the last pitch period of the signal of the last frame. This method shows advantage for speech or speech-like signals at high bit-rates.

*3.4.2. HQ MDCT*

In case the last good frame prior to a frame loss was coded with HQ MDCT, one of the following specifically optimized PLC methods is chosen.

The first possible method consists in a concealment based on sinusoidal phase evolution. It is based on sinusoidal analysis and synthesis paradigm operated in DFT domain [10][11][12]. It is expected that an audio signal is composed of a limited number of individual sinusoidal components. In the analysis step the sinusoidal components of a previously synthesized audio frame are identified. In the synthesis step these sinusoidal components are phased evolved to the time instant of the lost frame. Subsequently the frame is transformed into the time domain and further into the windowed time-aliased domain of the HQ MDCT where it is used instead of a regularly decoded and inversely

transformed MDCT frame. Unlike earlier methods of this paradigm, interpolative sinusoidal frequency refinement is done to increase the frequency resolution over that of the DFT. Instead of zeroing or magnitude adjusting DFT coefficients not belonging to spectral peaks, the original DFT magnitudes are retained while adaptive phase randomization is used. This, together with an effective adaptation control, completely alleviates any need for energy compensation of DFT coefficients or rescaling of the frame after IDFT, and still no tonal artifact are observed. Also unlike prior art the method performs very well even in case of burst errors. In case of multiple frame loss, phase randomization is used with an increasing degree. Also, the reconstructed sinusoids are increasingly attenuated and replaced by a suitably shaped additive noise signal.

The second possible concealment method is based on sinusoidal synthesis with adaptive noise filling. The pitch cycle from the past synthesis is extracted, interpolated to a length corresponding to a power of 2 and analyzed by FFT. Sinusoidal components are then selected based on local amplitude peaks and sinusoids are generated for the selected components and added to each other at the output sampling frequency. A residual signal is also computed by subtracting the past decoded synthesis and the sinusoidal synthesis; this residual is repeated with an adaptive overlap-and-add method. In case of multiple frame loss the method selector switches to the first sinusoidal concealment method.

Finally, it is possible to do the concealment by repeating the MDCT coefficients of the last frame while optimizing the signs. This concealment is used for HQ MDCT in NB and it consists of frequency and time domain approaches. In the frequency domain approach, the synthesized spectral coefficients of the last good frame are repeated for the current frame with signal modification such as a gain scaling and a combination of sign prediction and sign randomization. When burst of errors occur, an adaptive fade-out by regression method is used. In this method, a grouped average norm value of a lost frame is predicted using K-grouped average norm values of the previous good frame through linear regression analysis.

In the time-domain approach, repetition and smoothing techniques are used for almost stationary signals, and a phase matching technique is used on really stationary signals. When the first method is selected, a smoothing window is applied between the repeated signal of the previous frame and the signal of the current concealed frame. In the later method, a matching frame is selected for concealment from a buffer of past two decoded good frames.

## 4. TEST RESULTS

Extensive testing of the EVS PLC performance has been done via subjective ITU-T P.800 [13] Mean Opinion Score (MOS) tests in three independent testing laboratories.

Figures 1 to 3 show the performance of the EVS under various transmission conditions (3% and 6% FER) compared to the reference codecs defined in [14], AMR-WB/G.718 IO (referred as Ref) for wideband (WB) clean speech, AMR-WB and G.722.1 for WB mixed and music and G.722.1C and G.719 for super wideband (SWB) clean speech. For every test point EVS quality outperforms the reference. The robustness of EVS demonstrated in WB and SWB was also confirmed in NB, and is in the same range of improvement as the one reported from the Global Analysis Laboratory (GAL) for the Selection Phase [15].

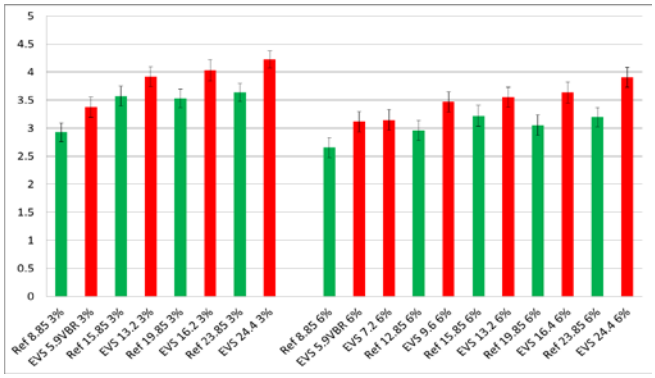


Figure 1 - MOS results for WB clean speech

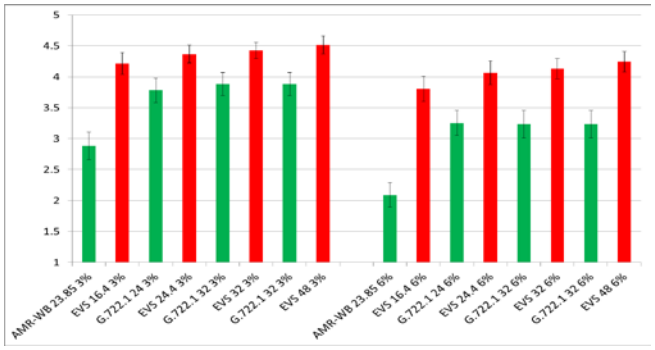


Figure 2 - MOS results for WB mixed and music

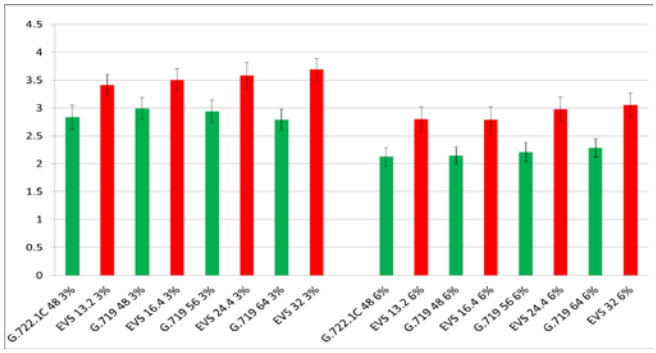


Figure 3 - MOS results for SWB clean speech

## 5. CONCLUSION

This paper presents the concealment methods available in the Codec for Enhanced Voice Services standardized by 3GPP. It was demonstrated that the performances of these methods meet and often significantly outperform the state of the art codecs in every aspect.

## 6. REFERENCES


[1] 3GPP Spec., Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description, TS 26.445, v.12.0.0, Sep. 2014.

[2] Jelinek, M., and al, "G.718: A new embedded speech and audio coding standard with high resilience to error-prone transmission channels", IEEE communication magazine, Oct. 2009, vol. 47, no. 10, pp. 117-123.

[3] V. Eksler and M. Jelínek, "Glottal-Shape Codebook to Improve Robustness of CELP Codecs", IEEE Trans. on Audio, Speech and Language Processing, vol. 18, no. 6, pp. 1208-1217, Aug. 2010.

[4] M. Neuendorf and al.,"The ISO/MPEG Unified Speech and Audio Coding Standard — Consistent High Quality for All Content Types and at All Bit Rates" Journal of the AES, 61(12): 956—977, Dec. 2013.

[5] ETSI ES 201 980, "Digital Radio Mondial (DRM), system specification", V4.1.1., Jan. 2014.

[6] 3GPP Spec., Codec for Enhanced Voice Services (EVS); Error concealment of lost packets, TS 26.447, v.12.0.0, Sep. 2014.

[7] M. Jelínek and R. Salami, "Wideband Speech Coding Advances in VMR-WB standard," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1167-1179, May 2007.

[8] T. Vaillancourt, et al, "Efficient Frame Erasure Concealment in Predictive Speech Codecs Using Glottal Pulse Resynchronisation," in Proc. IEEE ICASSP, Honolulu, HI, USA, Apr. 2007, vol. 4, pp. 1113-1116.

[9] J. Lecomte, et al, "Enhanced time domain packet loss concealment in switched speech/audio codec", submitted to IEEE ICASSP, Brisbane, Australia, Apr. 2015.

[10] R.J. McAulay and T.F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation", IEEE Trans. on Acoust., Speech and Signal Processing, ASSP-34, (4), pp. 744-754, 1986.

[11] Vipul N. Parikh, Juin-Hwey Chen, and Gerard Aguilar, "Frame Erasure Concealment Using Sinusoidal Analysis-Synthesis and Its Application to MDCT-Based Codecs", in Proc. IEEE ICASSP, 2000.

[12] Huan Hou, Weibei Dou, "Real-time audio error concealment method based on sinusoidal model", International Conference on Audio, Language and Image Processing, 2008.

[13] ITU-T P.800, Methods for Subjective Determination of Transmission Quality. International Telecommunication Union (ITU), Series P., Aug. 1996.

[14] 3GPP, Tdoc S4-130522, EVS Permanent Document (EVS-3): EVS performance requirements, Version 1.4, Apr. 2013.

[15] 3GPP, Tdoc S4-141065, GAL report for EVS Selection Phase, Aug. 2014.