IMPROVED ERROR RESILIENCE FOR VOLTE AND VOIP WITH 3GPP EVS CHANNEL AWARE CODING

Venkatraman Atti, Daniel J. Sinder*, Shaminda Subasingha*, Vivek Rajendran*, Duminda Dewasurendra*, Venkata Chebiyyam*, Imre Varga*, Venkatesh Krishnan*, Benjamin Schubert†, Jeremie Lecomte†, Xingtao Zhang‡, Lei Miao‡*

*Qualcomm Technologies, Inc., †Fraunhofer IIS, ‡Huawei Technologies Co. Ltd.

ABSTRACT

A highly error resilient mode of the newly standardized 3GPP EVS speech codec is described. Compared to the AMR-WB codec and other conversational codecs, the EVS channel aware mode offers significantly improved error resilience in voice communication over packet-switched networks such as Voice-over-IP (VoIP) and Voice-over-LTE (VoLTE). The error resilience is achieved using a form of in-band forward error correction. Source-controlled coding techniques are used to identify candidate speech frames for bitrate reduction, leaving spare bits for transmission of partial copies of prior frames such that a constant bit rate is maintained. The self-contained partial copies are used to improve the error robustness in case the original primary frame is lost or discarded due to late arrival. Subjective evaluation results from ITU-T P.800 Mean Opinion Score (MOS) tests are provided, showing improved quality under channel impairments as well as negligible impact to clean channel performance.

Index Terms— 3GPP EVS codec, channel aware, partial redundancy, packet-switched networks, VoLTE/VoIP/VoWiFi.

1. INTRODUCTION

In packet-switched networks, packets may be subjected to varying scheduling and routing conditions, which results in time-varying end-to-end delay. The delay jitter, is not amenable to most conventional speech decoders and voice post-processing algorithms that typically expect the packets to be received at fixed time intervals. Consequently, a de-jitter buffer (formally referred to as Jitter Buffer Management (JBM) [1][6]) is typically used in the receiving terminal to remove jitter and feed packets in the correct sequential order.

The longer the de-jitter buffer, the better its ability to remove jitter and the greater the likelihood that jitter can be tolerated without discarding packets due to late arrival (or, buffer underflow). However, end-to-end delay is a key determiner of call quality in conversational voice networks. Hence the ability of the JBM to absorb jitter without adding excessive buffering delay is an important requirement. Thus, a trade-off exists between JBM delay and the jitter induced packet loss at the receiver. The JBM designs have evolved to offer increasing levels of performance while maintaining minimal average delay [1]. Aside from delay jitter, the other primary characteristic of packet-switched networks is the presence of multiple consecutive packet losses (error bursts), which are more commonly seen than on circuit switched networks. Such bursts can result from bundling of packets at different network layers, scheduler behavior, edge of the cell, or even a slow-adapting JBM. However, the de-jitter buffer–an essential component for VoIP–can be leveraged for improved underflow prevention and more sophisticated packet loss concealment [1]. One such technique is to use forward error correction by transmitting encoded information redundantly for use when the original information is lost at the receiver ([1] and references therein).

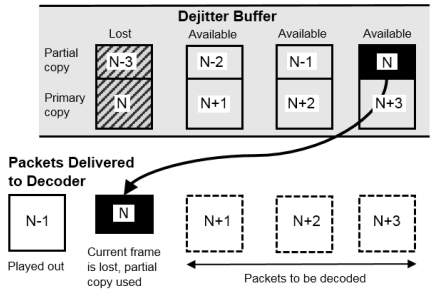

Figure 1. Concept of partial redundancy in channel aware mode.

2. CHANNEL AWARE MODE IN THE EVS CODEC

The EVS Channel Aware mode introduces a novel technique for transmitting redundancy in-band as part of the codec payload in a constant bitrate stream, and is implemented for wideband (WB) and super-wideband (SWB) at 13.2 kbps. This technique is in contrast to prior codecs, for which redundancy is typically added as an afterthought by defining mechanisms to transmit redundancy at the transport layer. For example, the AMR-WB RTP payload format allows for bundling of multiple speech frames to include redundancy into a single RTP payload [2]. Alternatively, RTP packets containing single speech frames can be simply re-transmitted at a later time.

Figure 1 depicts the concept of partial redundancy in the EVS channel aware mode. The idea is to encode and transmit the partial redundant copy associated with the N-th frame, along with the primary encoding of the $(N+K)$-th frame. The offset parameter, K, which determines the separation between the primary and partial frames is also transmitted along with the partial copy. In the packet-switched network, if the N-th frame packet is lost, then the de-jitter buffer is inspected for the availability of future packets. If available, then the transmitted offset parameter is used to identify the appropriate future packet for partial copy extraction and synthesis of the lost frame. An offset of 3 is used as an example to show the process in Figure 1. The offset parameter can be a fixed value or can be configured at the encoder based on the network conditions. Including the redundancy in-band in EVS Channel Aware mode allows the

transmission of redundancy to be either channel-controlled (e.g., to combat network congestion) or source-controlled. In the latter case, the encoder can use properties of the input source signal to determine the frames that are most critical for high quality reconstruction and selectively transmit redundancy for those frames only. Furthermore, the encoder can also identify the frames that can be best coded at a reduced bitrate in order to accommodate for the attachment of redundancy while keeping the bit-stream at a constant 13.2 kbps rate. These new techniques significantly improve the performance under degraded channel conditions while maintaining the clean channel quality.

3. CHANNEL AWARE ENCODING

Figure 2 shows a high level description of the channel aware encoder. The input audio that is sampled at either 16 kHz (WB) or 32 kHz (SWB) is segmented into frames of 20 msec. A "pre-processing" stage is used to resample the input frame to 12.8 kHz and perform steps such as voice activity detection (VAD) and signal classification [9]. Based on certain analysis parameters (e.g., normalized correlation, VAD, frame type, and pitch lag), the "Frame criticality configuration" module determines:

1) the compressibility of the current frame, i.e., if the current frame can allow for bitrate reduction, with minimal perceptual impact, to enable the inclusion of a partial copy associated with a previous frame, and
2) the RF frame type classification which controls the number of bits needed to faithfully reconstruct the current frame through the partial copy that is transmitted in a future frame. In Figure 2, the partial copy is transmitted along with a future primary copy at a frame erasure concealment (FEC) offset of 2 frames.

Strongly-voiced and unvoiced frames are suitable for carrying partial copies of a previous frame with negligible perceptual impact to the primary frame quality. If the current frame is allowed to carry the partial copy, it is signaled by setting Rf_{Flag} in the bit stream to 1, or 0 otherwise. If the Rf_{Flag} is set to 1, then the number of bits, $B_{primary}$, available to encode the current primary frame is determined by compensating for the number of bits, B_{RF}, already used up by the accompanying partial copy, i.e., $B_{primary} = 264 - B_{RF}$ at 13.2 kbps constant total bit rate. The number of bits, B_{RF}, can range from 5 to 72 bits depending on frame criticality and RF frame type (Section 3.2).

3.1. Primary Frame Coding

The "primary frame coding" module shown in Figure 2, uses the ACELP coding technology [14] [16] to encode the low band core up to 6.4 kHz while the upper band that is beyond 6.4 kHz and up to the Nyquist frequency is encoded using the Time-domain Bandwidth Extension (TBE) technology [10]. The upper band is parameterized into LSPs, gain parameters to capture both the temporal evolution per sub-frame as well as over an entire frame [10]. The "primary frame coding" module also uses the MDCT-based Transform Coded Excitation (TCX) and Intelligent Gap Filling (IGF) coding technologies [4] [11] to encode the background noise frames and mixed/music content more efficiently. An SNR-based open-loop classifier [15] is used to decide whether to choose the ACELP/TBE technology or the TCX/IGF technology to encode the primary frame.

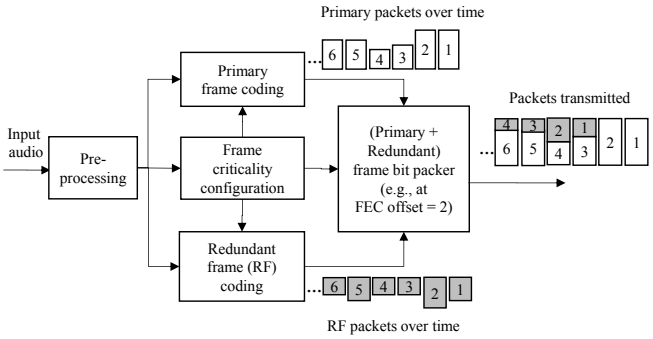

Figure 2. Channel aware encoder framework

Dietz *et al.*, [9] give an overview of various advancements to the EVS primary modes that further improve the coding efficiency of the ACELP technology beyond the 3GPP AMR-WB [14]. The EVS Channel Aware mode leverages these ACELP and TCX core advancements for primary frame encoding. Additionally, as the partial copy uses varying number of bits across frames, the primary frame encoding also needs to correspondingly accommodate for an adaptive bit allocation.

3.2. Redundant Frame Coding

The "redundant frame (RF) coding" module performs compact re-encoding of only those parameters that are critical to protect. The set of critical parameters are identified based on the frame's signal characteristics and are re-encoded at a much lower bitrate (e.g., less than 3.6 kbps). The "bit packer" module arranges the primary frame bit-stream and the partial copy along with certain RF parameters such as RF frame type and FEC offset (Table 1) at fixed locations in the bit-stream.

A frame is considered as critical to protect when loss of that frame would cause significant impact to the speech quality at the receiver. The threshold, to determine whether a particular frame is critical or not, is a configurable parameter at the encoder which can be dynamically adjusted depending on the network conditions. For example, under high FER conditions it may be desirable to adjust the threshold to classify more frames as critical. The criticality also depends on the previous frames losses. For example, a frame may get classified from being non-critical to critical if the previous frames were also lost.

3.2.1. ACELP Partial Frame Encoding

For ACELP frames, the partial copy encoding uses one of the four RF frame types, RF_NOPRED, RF_ALLPRED, RF_GENPRED, and RF_NELP depending on the frame's signal characteristics. Parameters computed from the primary frame coding such as frame type, pitch lag, and factor τ are used to determine the RF frame type and criticality where,

$$\tau = 0.25\left(\frac{E_{ACB} - E_{FCB}}{E_{ACB} + E_{FCB}} + 1\right)$$

where E_{ACB} denotes the adaptive codebook (ACB) energy and E_{FCB} denotes the fixed codebook (FCB) energy. A low value of τ (e.g., 0.15 and below) indicates that most of the information in the current frame is carried by the FCB contribution. In such cases, the RF_NOPRED partial copy encoding uses one or more FCB parameters (e.g., FCB pulses and gain) only. On the other

TABLE I
BIT ALLOCATION FOR CHANNEL AWARE CODING AT 13.2 KBPS

	Core coder	ACELP		TCX/IGF
	Bandwidth	WB	SWB	
Signalling information ($bwidth, coder\ type, Rf_{Flag}$)		5		
Primary frame	Core	181-248	169-236	232-254
	TBE	6	18	
Partial frame	Core	0-62	0-62	0-22
	TBE	0-5	0-5	
	FEC offset	2		
	RF frame type	3		

hand, a high value of τ (e.g., 0.35 and above) indicates that most of the information in the current frame is carried by the ACB contribution. In such cases, the RF_ALLPRED partial copy encoding uses one or more ACB parameters (e.g., pitch lag and gain) only. If τ is in the range of [0.15, 0.35], then a mixed coding mode RF_GENPRED uses both ACB and FCB parameters for partial copy encoding. For the UNVOICED frames, the low bitrate noise-excited linear prediction (NELP) [9] is used to encode the RF_NELP partial copy. The upper band partial copy coding relies on coarse encoding of gain parameters and extrapolation of LSF parameters from the previous frame [4].

3.2.2. TCX Partial Frame Encoding

In order to get a useful TCX partial copy, many bits would have to be spent for coding the MDCT spectral data, which reduces the available number of bits for the primary frame significantly and thus degrades the clean channel quality. For this reason, the number of bits for TCX primary frames is kept as large as possible, while the partial copy carries a set of control parameters, enabling a highly guided TCX concealment.

The TCX partial copy encoding uses one of the three RF frame types, RF_TCXFD, RF_TCXTD1, and RF_TCXTD2. While the RF_TCXFD carries control parameters for enhancing the frequency-domain concealment, the RF_TCXTD1 and RF_TCXTD2 are used in time-domain concealment [13]. The TCX RF frame type selection is based on the current and previous frame's signal characteristics, including pitch stability, LTP gain and the temporal trend of the signal. Certain critical parameters such as the signal classification, the LSPs, the TCX gain and pitch lag are encoded in the TCX partial copy.

In background noise or in inactive speech frames, a non-guided frame erasure concealment is sufficient to minimize the perceptual artifacts due to lost frames. An RF_NO_DATA is signaled indicating the absence of a partial copy in the bit-stream during the background noise. In addition, the first TCX frame after a switch from ACELP frame, also uses an RF_NODATA due to lack of extrapolation data in such a codec switching scenario.

4. CHANNEL AWARE DECODING

Figure 3 presents a high level description of the channel aware decoder. At the receiver, if the current frame is not lost, the JBM provides the packet for "primary frame decoding" and disregards any RF information present in the packet. In case the current frame is lost, and a future frame is available in the de-jitter buffer, then the JBM provides the packet for "partial frame decoding". If a future frame is not available in the de-jitter buffer, then a non-guided erasure concealment [13] is performed.

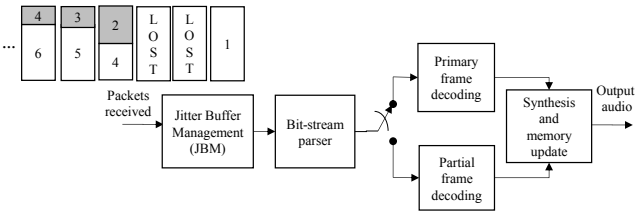

Figure 3. Channel Aware decoder framework

4.1. Interface with JBM

As described earlier, if the N-th frame is not available (lost or delayed) at the play-out time, JBM is checked for the availability of a future $(N + K)$-th frame that contains the partial redundancy of the current frame where $K \in \{2, 3, 5, 7\}$. The partial copy of a frame typically arrives after the primary frame. JBM delay adaptation mechanisms are used to increase the likelihood of availability of partial copies in the future frames, especially for higher FEC offsets of 5 and 7. The EVS JBM conforms to the delay-jitter requirements specified by the 3GPP TS 26.114 [3] for all the EVS modes including the channel aware mode.

In addition to the above described functionality, the EVS JBM [6] computes the channel error rate and an optimum FEC offset, K, that maximizes the availability of the partial redundant copy based on the channel statistics. The computed optimum FEC offset and the channel error rate can be transmitted back to the encoder through a receiver feedback mechanism (e.g., through a codec mode request (CMR) [2]) to adapt the FEC offset and the rate at which the partial redundancy is transmitted to improve the end user experience.

4.2. ACELP and TCX Partial Frame Decoding

The "bit-stream parser" module in Figure 3 extracts the RF frame type information and passes the partial copy information to the "partial frame decoding" module. Depending on the RF frame type, if the current frame corresponds to an ACELP partial copy, then the RF parameters (e.g., LSPs, ACB and/or FCB gains, and upper band gain) are decoded for ACELP synthesis. ACELP partial copy synthesis follows similar steps to that of the primary frame decoding except that the missing parameters (e.g., certain gains and pitch lags are transmitted in alternate subframes) are extrapolated.

Furthermore, if the previous frame used a partial copy for synthesis, then a post-processing is performed in the current frame for a smoother evolution of LSPs and temporal gains. The post-processing is controlled based on the frame type (e.g., VOICED or UNVOICED) and spectral tilt estimated in the previous frame. If the current frame corresponds to a TCX partial copy, then the RF parameters are used to perform a highly-guided concealment.

5. SUBJECTIVE QUALITY TESTS

Extensive testing of the EVS channel aware mode has been conducted via subjective ITU-T P.800 Mean Opinion Score (MOS) tests conducted at an independent test laboratory with 32 naïve listeners. The tests were conducted for both WB and SWB, using absolute category rating (ACR) and degradation category rating (DCR) test methodologies [17], respectively. Since the

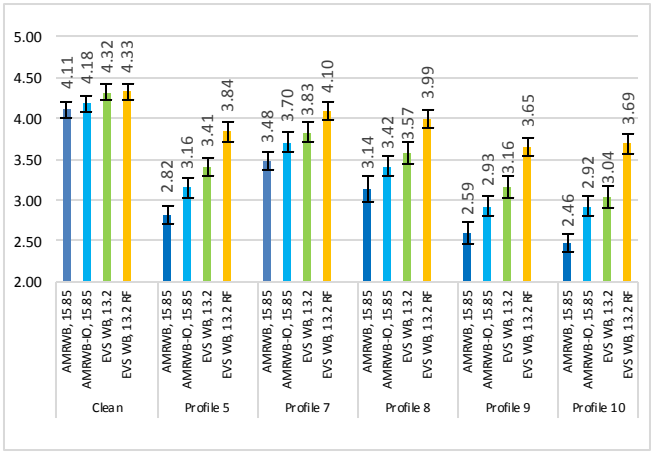

Figure. 4. Wideband clean speech ITU-T P.800 ACR test results

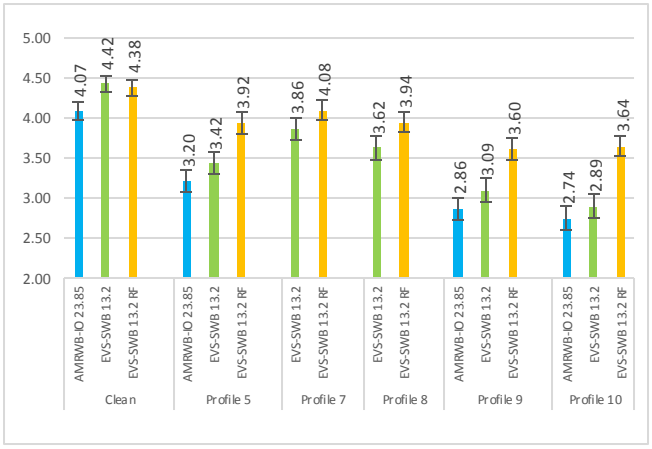

Figure 5. Super-wideband clean speech ITU-T P.800 DCR test results

channel aware mode is specifically designed to improve performance for VoLTE networks, evaluating the performance under these conditions is critical for establishing its potential benefits. Therefore, testing was conducted using codec outputs from simulations in which VoLTE-like patterns of packet delays and losses were applied to received RTP packets before insertion into the de-jitter buffer. Four of these patterns – or, delay-loss profiles – were derived from real-world call logs of RTP packet arrival times collected in different networks in South Korea and the United States.

The resulting profiles mimic closely VoLTE network characteristics under different channel error conditions. In deriving the profiles, characteristics such as jitter, temporal evolution of jitter, and burstiness of errors were considered. These four profiles are identified as profiles 7, 8, 9, and 10 in the results below, and correspond to frame erasure rates (FER) at the decoder of approximately 3%, 6%, 8%, and 10%, respectively. These same four profiles have also been selected by 3GPP for use by that body for its own characterization testing of the EVS channel aware mode (using FEC offset $K = 3$) under channel impairments.

In addition to the VoLTE profiles, all codecs considered here were tested for error-free conditions and also for an HSPA profile included in the 3GPP MTSI specification [3] that yields about 6% frame erasure rate at the decoder. In all of the experiments, the EVS conditions used the reference EVS de-jitter buffer [6]. The AMR-WB conditions used a fixed delay buffer to convert delay-loss profiles to packet-loss profiles, such that packets experiencing a delay greater than a fixed threshold are discarded as described in EVS performance requirements specification [7].

The ACR scores for the WB case are shown in Figure 4. For each profile, starting with the error-free ("Clean") profile, the chart compares (from left to right) AMR-WB, EVS AMR-WB IO mode, EVS baseline WB, and EVS WB channel aware ("RF"). The AMR-WB and EVS AMR-WB IO conditions used a higher bit rate of 15.85 kbps, whereas both EVS conditions used the same 13.2 kbps rate. From these results, it is clear that despite maintaining quality consistent with the EVS non-channel-aware mode under error-free conditions, the channel aware mode provides a clear advantage under all frame erasure conditions. Notably, the channel aware mode quality degrades much more gracefully even out to the 10% FER of profile 10. Compared to the AMR-WB and AMR-WB-IO conditions, the dramatic quality benefit at these FER rates restores intelligibility under periods of high loss as might be encountered during a handoff, poor radio conditions, edge of the cell scenarios or even on best-effort networks [1].

The performance advantage of the channel aware mode is similarly compelling in the super-wideband mode, the results for which are shown in Figure 5. As with WB, the channel aware mode does not degrade performance under error-free conditions, but has a statistically significant performance benefit under each of the lossy profiles, with the degree of improvement increasing as error rate increases.

6. CONCLUSIONS

The Channel Aware coding mode of the new 3GPP EVS codec offers users and network operators a highly error resilient coding mode for VoLTE at a capacity operating point similar to the most widely deployed bit rates of existing deployed services based on AMR and AMR-WB. The mode gives the codec the ability to sustain high quality WB and SWB conversational voice quality even in the presence of high frame erasure rates that may occur during network congestion, poor radio frequency coverage, handoffs, or in best-effort channels. Even with its graceful quality degradation under high loss, the impact to quality is negligible under low loss or even no-loss conditions. This error robustness offered by the Channel Aware mode further allows for relaxing certain system level aspects such as frequency of re-transmissions and reducing scheduler delays. This in turn has potential benefits for increased network capacity, reduced signaling overhead and power savings in mobile handsets. The mode can therefore be used in most networks without capacity impact to insure high quality communications.

7. REFERENCES

[1] D. J. Sinder, I. Varga, V. Krishnan, V. Rajendran and S. Villette, "Recent Speech Coding Technologies and Standards," in *Speech and Audio Processing for Coding, Enhancement and Recognition*, T. Ogunfunmi, R. Togneri, M. Narasimha, Eds., Springer, 2014.

[2] J. Sjoberg, M. Westerlund, A. Lakaniemi and Q. Xie, "RTP Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs," April 2007. [Online]. Available: http://tools.ietf.org/html/rfc4867.

[3] 3GPP TS 26.114, Multimedia Telephony Service for IMS, V12.7.0, September 2014.

[4] 3GPP TS 26.445: "EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)", 2014.

[5] 3GPP TS 26.447: "EVS Codec Error Concealment of Lost Packets"

[6] 3GPP TS 26.448: "EVS Codec Jitter Buffer Management".

[7] 3GPP Tdoc S4-130522, "EVS Permanent Document (EVS-3): EVS performance requirements", Version 1.4.

[8] S. Bruhn, *et al.*, "Standardization of the new EVS Codec", submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[9] M. Dietz, *et al.*, "Overview of the EVS codec architecture," submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[10] V. Atti, *et al.*, "Super-wideband bandwidth extension for speech in the 3GPP EVS codec," submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[11] G. Fuchs, *et al.*, "Low delay LPC and MDCT-based Audio Coding in EVS," submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[12] S. Disch *et al.*, "Temporal tile shaping for spectral gap filling within TCX in EVS Codec," submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[13] J. Lecomte *et al*., "Packet Loss Concealment Technology Advances in EVS", submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[14] B. Bessette, *et al*, "The adaptive multi-rate wideband speech codec (AMR-WB)," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 620-636, November 2002.

[15] E. Ravelli, *et al*., "Open loop switching decision based on evaluation of coding distortions for audio codecs, submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[16] M. Jelínek, T. Vaillancourt, and Jon Gibbs, "G.718: A New Embedded Speech and Audio Coding Standard with High Resilience to Error-Prone Transmission Channels," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 117-123, October 2009.

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