# Application-Layer Redundancy for the EVS Codec

Najmeddine Majed[1,2], Stéphane Ragot[1], Laetitia Gros[1], Xavier Lagrange[2], and Alberto Blanc[2]

[1]Orange Labs, Lannion, France
Email: {najmeddine.majed, stephane.ragot, laetitia.gros}@orange.com

[2]IMT-Atlantique, Rennes, France
Email: {xavier.lagrange, alberto.blanc}@imt-atlantique.fr

***Abstract*—In this paper, we study the performance of the 3GPP EVS codec when this codec is used in conjunction with 100% application-layer redundancy. The objective of this work is to investigate potential performance gains for Voice over LTE (VoLTE) in bad coverage scenarios. Voice quality for the EVS codec operated in the 9.6–24.4 kbit/s bit range in super-wideband (SWB) is evaluated at different packet loss rates (PLR), using objective and subjective methods (ITU-T P.863 and P.800 ACR). Results show that EVS at 9.6 kbit/s with 100% application-layer redundancy has significantly higher packet loss resilience in degraded channel conditions ($\geq$ 3% PLR), for an overall bit rate (around $2\times$9.6 kbit/s) compatible with VoLTE (assuming a VoLTE bearer configured to a maximum rate of 24.4 kbit/s). We also discuss the relative merit of the partial redundancy mode in the EVS codec at 13.2 kbit/s, known as the channel-aware mode (CAM), and possible RTP/RTCP signaling methods to trigger the use of application-layer redundancy.**

## I. Introduction

Mobile operators have deployed Voice over LTE (VoLTE) to add the support of telephony services in 4G mobile networks. VoLTE is a form of mobile voice over IP (VoIP) using specific network optimizations, with quality of service (QoS) provided by the IP Multimedia Sub-System (IMS) [1]. The enhanced voice services (EVS) codec has been standardized by the 3GPP in 2014 to provide new functionalities and enhancements for VoLTE, such as superwideband (SWB) and fullband voice quality, improved coding efficiency, better music quality, and interoperability to existing wideband voice services [2], [3].

In this paper, we study the performance of the EVS codec in packet loss conditions, resulting for instance from poor radio coverage in VoLTE. This work is motivated by an ongoing feasibility study in 3GPP called enhanced VoLTE performance (eVoLP). The eVoLP work has three main objectives: 1) investigate guidelines or requirements to ensure that VoLTE clients adapt to the most robust codec modes, study performance results for different conditions and adaptation procedures; 2) study how terminals can indicate at setup their ability to send adaptation triggers to robust modes; 3) evaluate the impact of proprietary client implementations of packet loss concealment (PLC) and jitter buffer management (JBM). In this paper we focus on the first two objectives. We quantify the performance of the EVS codec with redundancy; we also briefly review media signaling methods (based on RTP or RTCP) that can be used to trigger the use of application-layer redundancy.

Many approaches have been proposed to address and adapt to packet losses in speech and audio coding, and we classify them as sender/encoder vs. receiver/decoder-based methods. Note that (end-to-end) retransmissions are not considered as an option, because of latency and system constraints in VoLTE.

Encoder-based methods typically consist of adding redundancy [4] or limiting the use of memory (prediction) [5] [6, Sec. 2.1.6] [7]. Two redundant coding approaches are reviewed in [4]: multiple description coding (MDC) and forward error correction (FEC). MDC consists of encoding the signal in complementary descriptions that are sent separately; if some descriptions are lost, a coarser reconstruction is obtained. FEC is based on channel-coding principles. Some FEC variants apply error-correcting codes such as Reed-Solomon codes at the bitstream level [8, Chap. 9]. In this work we refer to 100% application-layer redundancy as an FEC variant where the bitstream of a given speech frame is fully repeated in a subsequent packet [9]. Application-layer redundancy is defined in more details in [10, Sec. 9.2] [1] and in Sec. II-A. The FEC principle can also be applied at the codec parameter level to minimize the bit rate penalty of redundancy [9]. For instance, the ITU-T G.729.1 codec sends frame parameters (signal class, phase, energy) with few bits to guide packet loss concealment [11]. Partial redundancy coding is also used in the LBRR (Low-Bit-Rate Redundancy) of the OPUS codec [6, Sec. 2.1.7] or the Channel-Aware Mode (CAM) of the EVS codec [12] – see Sec. II-B for more details on CAM. It can be noted that some sender-based adaptation strategies to packet losses may rely on rate/congestion control mechanisms, for instance by reducing codec bit rate or even packet rate to deal with insufficient throughput.

Decoder-based methods are mainly based on packet loss concealment (PLC) techniques to fill and recover from missing frames [13]. Examples are given for instance by the pitch repetition method of ITU-T G.711 App. I [14], signal extrapolation based on linear-predictive coding (LPC), classification, adaptive muting in ITU-T G.722 App. IV [15], or more advanced methods in EVS [16]. It can be noted that in voice over IP (VoIP) PLC may be integrated with JBM [17], and concealment and recovery may be implemented by JBM expand and merge operations at the reconstructed signal level.

In this paper, we rely on the PLC and JBM algorithms defined for the EVS codec and focus on the performance of EVS with and without redundancy in different packet loss conditions. The paper is organized as follows. In Sec. II we review existing mechanisms to use redundancy with EVS.

ISBN 978-90-827970-1-5 

In Sec. III, we describe the modifications made to the EVS source code required to support application-layer redundancy in simulations. In Sec. IV, we describe the experimental setup. In Sec. V, we present the objective and subjective test results. Before concluding in Sec. VII, we discuss possible mechanisms to send adaptation requests to activate application-layer redundancy in Sec. VI.

## II. Review of existing redundancy mechanisms for the EVS codec

We review here two approaches to use redundancy for the EVS codec: the channel-aware mode (CAM) and application-layer redundancy.

### A. Application-layer redundancy

Application-layer redundancy can be used with any codec [10, sec. 9.2]. In normal operation, when redundancy is not used, it is assumed that an RTP packet transports a single codec frame. When application-layer redundancy is used, a packet will transport the bitstream of the current frame ($N$) as well as the bitstream of one or several past (redundant) frames ($N-k$), where $k<0$. We focus in this paper on 100% redundancy on single frames, where only one redundant frame is added; however in principle it is possible to use more redundancy (e.g., 200% with two redundant frames per packet) and several frames per packet (e.g., 2 frames per packet). The RTP payload format of codecs such as 3GPP AMR, AMR-WB [18] or EVS [2, Annex A] includes a 'max-red' media type parameter, which restricts the maximal time interval (offset) for redundancy. It is noted in [10, sec. 9.2] that this type of redundancy may not be an appropriate solution in scenarios with packet losses due to limited throughput or congestion.

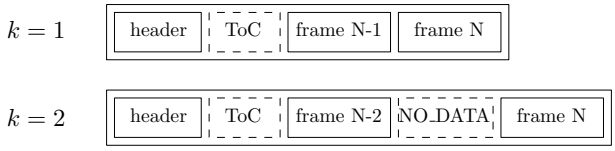
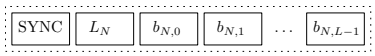



Fig. 1. Example structure of an RTP packet with 100% application-layer redundancy for offset $k=1$ or 2: RTP header followed by the RTP payload – including an optional table of content (ToC) and dummy (NO_DATA) frame.

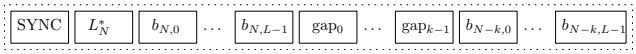
As shown in Fig. 1, in the 100% redundancy case, an offset $k=1$ implies that an RTP packet includes the RTP header followed by an optional table of content (ToC) and frames $N$ and $N-1$; when $k=2$ a dummy frame $N-1$ (referred to as NO_DATA) has to be inserted between frames $N$ and $N-2$, typically this insertion is done implicitly in the ToC part [18]. The bit rate overhead depends on the redundancy level (e.g. 100%) and the ToC length.

In principle end-to-end delay may be increased if redundancy is used; in practice for small offset values (e.g., $k=1$ or 2) and assuming a sufficient jitter buffer depth, the redundant frame may often be available if it was received as a future packet before decoding and playing out the current frame.

Performance results for 100% application-layer redundancy with 3GPP AMR were reported in [9], [19].




(a) G.192 format for frame $N$: sync word, frame length ($L_N$) and frame bits ($b_{N,0},\ldots,b_{N,L_N-1}$)




(b) extended G.192 format for frame $N$ with redundancy offset $k$: sync word, frame length ($L_N^*=L_N+L_{N-k}+k$), frame bits, gap ($k$ bits), redundant bits

Fig. 2. Bitstream format.

### B. Channel-aware mode (CAM) of EVS

The EVS-CAM [12], [2, section 5] is a partial redundancy mode supported at a single bit rate (13.2 kbit/s) in wideband (WB) and in super-wideband (SWB). The redundant data is defined in the form of partially coded frame $N-k$ embedded within the bit stream of the current frame $N$. The offset is restricted to $k=2,3,5$ or 7. The higher offsets ($k=5$ or 7) may be used to deal with long bursts of losses, however they typically imply significantly higher receiver delay.

A key feature of EVS-CAM is that it keeps a fixed bit rate (13.2 kbit/s) at the application level, so the activation of CAM is transparent to the network. Typically, when CAM is activated, the partial copy of the past frame $N-k$ is coded using about 3 kbit/s, therefore the remaining bit rate budget to code the current frame $N$ is reduced to about 10 kbit/s. The performance of CAM has been reported in [20] and [21]. In clean channel conditions (i.e., no packet loss) the intrinsic quality of CAM is close to the 9.6 kbit/s mode of EVS, however CAM is significantly better than the regular EVS modes at 9.6 or 13.2 kbit/s for packet loss rate (PLR) typically $\geq 3\%$ - see also results reported in Sec. V.

The EVS RTP payload format [2, Annex A] defines a media type parameter 'evs-ch-aw' to control the use of EVS-CAM. The signaling methods to trigger EVS-CAM relies on RTP CMR (Codec Mode Request) or RTCP-APP [10].

## III. Modifications to the EVS encoder/decoder

We describe here how the source code of the EVS codec was modified to simulate application-layer redundancy. The EVS bitstream is compliant with the serial bitstream format in ITU-T G.192 [22, App. I.2]. This format is convenient to simulate transmission of frames over a synchronized noisy channel between the encoder and decoder. As shown in Fig. 2 (a), each encoded frame is represented by a block of 16-bit words starting with a synchronization word (of value 0x6B21) and a frame length indication ($L_N$), which are followed by $L_N$ softbits (0x007F for 0 and 0x0081 for 1). Packet losses are simulated by applying loss profiles (with values 0x6B21/0x6B20 for good/bad frames) by an EID (Error Insertion Tool) tool performing an XOR operation on individual synchronization words – in other words, in case of packet loss, the synchronization word is changed from 0x6B21 to 0x6B20.

In this work, we only modified the G.192 bitstream formatting in the EVS encoder part; the actual EVS encoding

ISBN 978-90-827970-1-5 

algorithm was not changed. We added a buffer of encoded frames (bitstreams) outside the main EVS encoding loop to produce an extended G.192 bitstream including a gap ($k$ zero words where $k$ is the redundancy offset), as shown in Fig. 2 (b). It is important to note that, in case of discontinous transmission (DTX), redundant frames were not allowed in inactive periods, i.e., if the current frame was either a silence description (SID) frame or not transmitted (NO_DATA frame).

At the decoder side, a receiving buffer was added as a pre-processing step to bitstream decoding, with an extra decoder delay to allow detecting if the (lost) current frame is available as a redundant frame in a future packet at a given offset.

## IV. Experimental setup

We tested four EVS bit rates (9.6, 13.2, 16.4 and 24.4 kbit/s) with or without application-layer redundancy, together with EVS-CAM at 13.2 kbit/s. For all tests, we used the latest fixed-point version of EVS (v14.1.0) [20] with DTX activated.

### A. Generation of processed audio samples

To generate the audio samples required for subjective and objective tests, we reused EVS qualification scripts [23] to automate the encoding and decoding tasks. The original clean speech samples in the French language (16-bit linear PCM) sampled at 48 kHz were high-pass filtered, downsampled to 32 kHz, normalized to -26 dBov and encoded with EVS. When application-layer redundancy was used, we refer to the $2 \times X$ bit-rate where $X$ is the regular EVS mode and we used the modified EVS codec described in Sec. III. We used a redundancy offset $k = 2$.

For conditions with packet losses, we used the ITU-T 'gen-patt' tool to generate loss profiles. We used two different channel models. The first one consists of random packet loss profiles (i.i.d Bernoulli variables), the associated results are given in Sec. V-A. The second model is a Gilbert model with memory [24], which consists of a 2-state Markov chain, one state without errors (state *Good*) and the other one with errors (state *Bad*); we used a transition probability from *Good* to *Bad* equal to PLR and a transition probability from *Bad* to *Good* equal to $\frac{1}{2} - PLR$. Each channel profile contained one entry per packet indicating whether the packet is received or not. These loss profiles were applied to the bitstream by the ITU-T 'eid-xor' tool.

For EVS-CAM conditions, we used the network simulator developed in 3GPP [25] to simulate VoIP transmission, using delay/loss profiles. We used a CAM offset of 2 with default settings for the FEC indicator (HI). The EVS decoder operated in VoIP mode in conjunction with the EVS JBM algorithm. The loss-only profiles generated by 'gen-patt' were converted to delay/loss profiles with a fixed delay and identical packet loss distribution.

### B. Subjective test (P.800 ACR)

We used the P.800 ACR methodology [26] to allow potential comparisons with P.863 predictions. In ACR tests, groups of listeners evaluate series of processed audio files using a five-category scale. The experimenter allocated the following categories to scores: Excellent=5, Good=4, Fair=3, Poor=2, Bad=1. We recruited 48 naive listeners for the subjective test.

TABLE I
Subjective test plan settings.

| | |
|---|---|
| Voice codec | EVS [28] with DTX on |
| Rating scale | Absolute category rating (ACR) [26] |
| Listening | 73 dB SPL, naive listeners, 6 panels of 8 listeners |
| Loss profiles | static with no jitter, only random losses are applied<br>0%, 3%, 6%, 9%, and 12% |
| Coding mode | 9.6, 13.2, 16.4, and 24.4 kbit/s<br>2×9.6, 2×13.2, 2×16.4, 2×24.4 kbit/s with offset=2<br>13.2 kbit/s CAM with offset=2 |
| Calibration | P.50 MNRU: 10, 16, 22, 28, 34, and 40 dB |
| Talkers | 4 (two males and two females) |
| Samples | 6 clean speech samples (8-s double sentences ) per talker |

Table IV-B describes the settings used for the subjective test. This resulted in $1152$ processed sequences with $24$ blocks for 6 panels (of 8 listeners), 4 blocks per panel. Each block contained $48$ conditions and 4 talkers equally. The number of votes per condition is $4 \times 48 = 192$. The overall listening/scoring duration is around 42 minutes for each subject ($192 \times 13$ s). The list of conditions and randomizations can be found in [27]. The statistical analysis was based on independent-group t tests.

### C. Objective tests (P.863)

The objective quality evaluation used ITU-T Rec. P.863 [29] using the commercial implementation known as POLQA v2.4 in SWB mode (with no level adjustment). MOS-LQO$_s$ scores were computed by providing the reference audio sequences (SWB) and the degraded ones. For the random channel, the audio files used in the objective test were the same as in the subjective test.

## V. Subjective and objective test results

### A. Results for random channel (no channel memory)

Fig. 3 shows a bar chart with the average subjective scores, including $95\%$ confidence intervals (in the order of $\pm 0.1$ MOS). Fig. 4 (a) and 4 (b) present subjective and objective test results, as a function of packet loss rate (PLR). Note that Fig. 4 (a) is just an alternative representation of the same scores as in Fig. 3, and confidence intervals are not shown in Fig. 4 (a) to improve readibility. Subjective and objective results show similar trends, however P.863 predictions emphasized the intrinsic quality difference between EVS bit rates; one can observe that subjective scores are more compressed, and this may be explained by that fact that subjects may have focused their assessment on artefacts related to packet losses more than on the impact of codec rate. Assuming VoLTE bearers configured for EVS up to 24.4 kbit/s, application–layer redundancy at 2×9.6 kbit/s gives the best performance for PLR $\geq 3\%$ for all compatible EVS operation modes. The MOS score stays close to 4 even at a high packet loss rate (12%); this could be predicted, given that application-layer redundancy in the random channel case reduces packet loss

ISBN 978-90-827970-1-5 

rate from $p$ to $p^2$; the MOS score at 12% PLR with $2\times9.6$ kbit/s is theoretically the same as the MOS score at $1.44\%$ PLR for 9.6kbit/s. Results also confirm that EVS-CAM at 13.2 kbit/s is significantly better than EVS from 9.6 to 16.4 kbit/s for PLR $\geq 3\%$, however its performance is significantly worse than $2 \times 9.6$ kbit/s for PLR $\geq 3\%$. Note that application-layer redundancy used an extra decoder delay of 40 ms (due to the offset $k = 2$). As discussed in Section II-B there may be no impact on receiver delay when using EVS-CAM with offset 2 (assuming a minimal jitter buffer depth of 2 extra frames).

### B. Results for Gilbert channel model

Objective test results for the Gilbert model are provided in Fig. 5. Compared to Fig. 4 (b), we can see that in a bursty channel the performance decreases faster at a given PLR and the use of redundancy is less efficient. Indeed, due to a longer burst of losses, the redundant frames can only be exploited to compensate for losses at the end of a burst.

## VI. Signaling methods to trigger application-layer redundancy in VoLTE

The media handling of voice over IMS is specified in 3GPP TS 26.114 [10], which defines two methods to signal adaptation requests for speech: RTP Code Mode Request (CMR) and application-specific RTCP (RTCP-APP). RTP CMR consists of sending adaptation requests in-band within the codec payload. It was defined for AMR and AMR-WB in [18] and for EVS in [2, Annex A]. Several CMR codes were left reserved or unused and there are no CMR codes specified to request application-layer redundancy. RTCP-APP is a form of out-of-band feedback [30] used for speech adaptation in [10, clause 10.2.1] to more general signal adaptation requests than

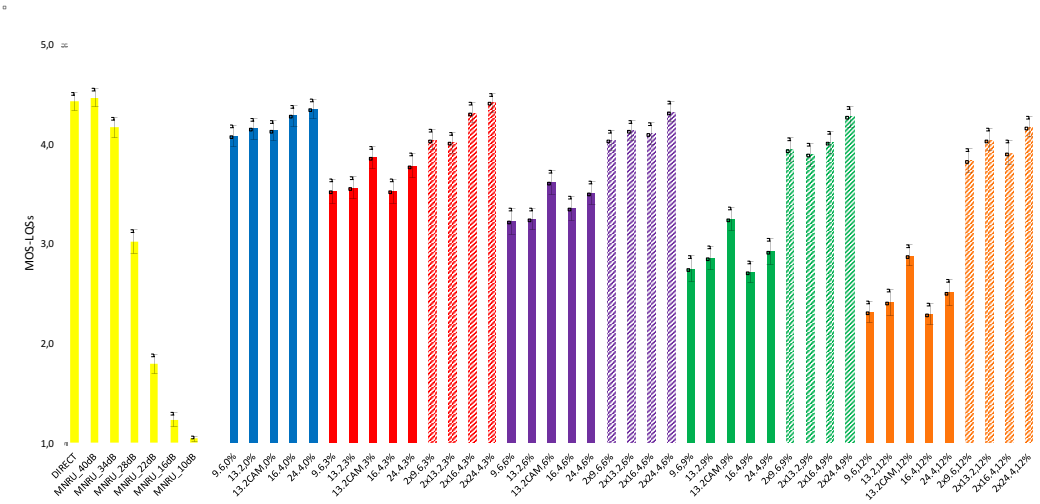
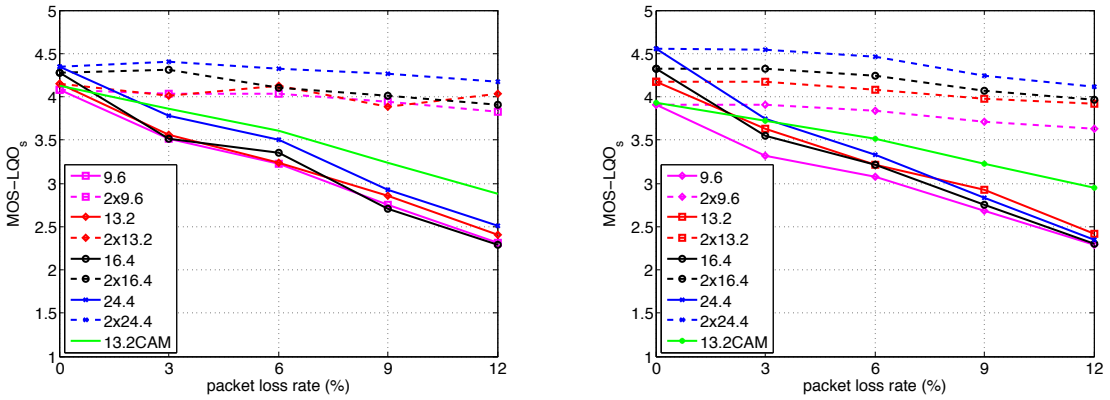



Fig. 3. Subjective test results (random losses).




(a) Subjective results ($MOS - LQS_s$)

(b) Objective results ($MOS - LQO_s$)

Fig. 4. EVS codec performance (random losses), as a function of packet loss rate (PLR) and operation mode (bit-rate, redundancy).

ISBN 978-90-827970-1-5 © EURASIP 2018

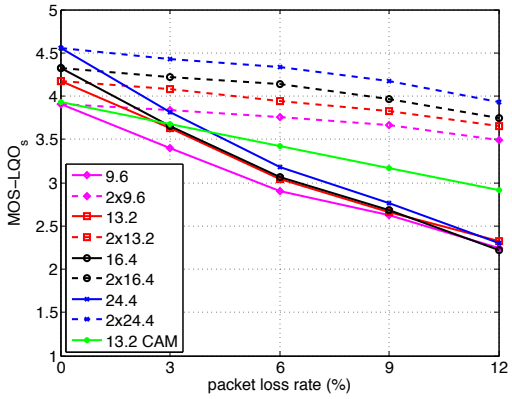



Fig. 5. Objective results ($MOS-LQO_s$) with Gilbert channel model, as a function of packet loss rate (PLR) and operation mode (bit-rate, redundancy).

CMR, in particular it can be used to request the activation of application-layer redundancy. On the other hand, VoLTE deployments are based on a profile (subset) of TS 26.114 defined in GSMA IR.92 [31] where only the RTP Audio Video Profile (AVP) [32] profile is allowed. The use of the RTP Audio Video Profile with Feedback (AVPF) [33] profile is forbidden for speech. Therefore, there are two possible options that may be used in VoLTE to request the use of application-layer redundancy. The first option consists of using the RTP CMR codes that are currently left 'reserved' or 'not used', if this was negotiated at call setup with an appropriate SDP parameter. The second option consists of using RTCP-APP requests with AVP. At the time of writing the decision between these two options is still open in 3GPP.

## VII. Conclusion

In this paper, we presented performance results for the 3GPP EVS codec with or without application-layer redundancy. We evaluated speech quality with objective and subjective tests as function of PLR and operation mode. We demonstrated that application-layer redundancy ($2\times9.6$ kbit/s) has better quality than the 13.2 kbit/s CAM mode in adverse conditions (PLR$\geq$ 3%). Future work will investigate performance delay/loss profiles that reflect measured VoLTE conditions. Moreover, the actual relationship between payload bit rate, radio path loss and PLR will be further studied to be able to map test results as a function of radio path loss.

## Acknowledgment

The authors would like to thank Thierry Moal and Caroll Rattazzi (Orange) for their help in conducting the subjective test, and Fabrice Plante (Intel) for helpful discussions in 3GPP.

## References

[1] S. Chakraborty, T. Frankkila, J. Peisa, and P. Synnergren, *IMS Multimedia Telephony over Cellular Systems*. John Wiley & Sons, 2007.
[2] 3GPP TS 26.445, "Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description."
[3] M. Dietz et al., "Overview of the EVS codec architecture," in *Proc. ICASSP*, 2015.
[4] J. Skoglund and al., "Voice over IP: Speech Transmission over Packet Networks," in *Handbook of Speech Processing (eds. J. Benesty and M.H. Sondhi and Y.A. Huang)*. Springer, 2008, pp. 307–330.
[5] S. Andersen and al., "Internet Low Bit Rate Codec (iLBC)," IETF RFC 3951, Dec. 2004.
[6] J. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," IETF RFC 6716, Sep. 2012.
[7] V. Eksler and M. Jelinek, "Transmission mode coding for source controlled CELP codecs," in *Proc. ICASSP*, 2008.
[8] C. Perkins, *RTP: Audio and Video for the Internet*. Addison-Wesley, 2003.
[9] I. Johansson, T. Frankkila, and P. Synnergren, "Bandwidth efficient AMR operation for VoIP," in *Proc. Speech Coding Workshop*, 2002.
[10] 3GPP TS 26.114, "IP Multimedia Subsystem (IMS); Multimedia Telephony; Media handling and interaction."
[11] S. Ragot and al., "ITU-T G.729.1: An 8-32 Kbit/S Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice Over IP," in *Proc. ICASSP*, 2007.
[12] V. Atti and al., "Improved error resilience for VoLTE and VoIP with 3GPP EVS channel aware coding," in *Proc. ICASSP*, 2015.
[13] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, Sept.–Oct. 1998.
[14] R. Cox, D. Malah, and D. Kapilow, "Improving upon toll quality speech for VOIP," in *Proc. 38th Asilomar*, 2004.
[15] B. Kövesi and S. Ragot, "A low complexity packet loss concealment algorithm for ITU-T G.722," in *Proc. ICASSP*, 2007.
[16] J. Lecomte and al., "Packet-loss concealment technology advances in EVS," in *Proc. ICASSP*, 2015.
[17] US Patent 7733893, "Method and receiver for determining a jitter buffer level."
[18] J. Sjoberg and al., "RTP Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs," IETF RFC 4867, Apr. 2007.
[19] 3GPP Tdoc S4-060448, "Results from Subjective Listening Test with Redundancy," Aug 2006, Source: Ericsson.
[20] 3GPP TR 26.952, "Codec for Enhanced Voice Services (EVS); Performance Characterization."
[21] A. Ramo, A. Kurittu, and H. Toukomaa, "EVS Channel Aware Mode Robustness to Frame Erasures," in *Proc. Interspeech*, 2016.
[22] ITU-T Rec. G.192, "A common digital parallel interface for speech standardisation activities," March 1996.
[23] 3GPP Tdoc S4-130155, "EVS Permanent Document EVS-7a: Processing functions for qualification phase, v1.3," Jan.-Feb. 2013, Source: Editor (Fraunhofer IIS).
[24] E.N. Gilbert, "Capacity of a Burst-Noise Channel," *Bell System Technical Journal*, no. 39, pp. 1253–1265, 1960.
[25] 3GPP Tdoc AHEVS-197, "Updated network simulator for JBM," Sept. 2012, Source: Fraunhofer IIS.
[26] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," July 2006.
[27] 3GPP Tdoc S4-180150, "Subjective test results for EVS with application-layer redundancy," February 2018, Source: Orange.
[28] 3GPP TS 26.442, "Codec for Enhanced Voice Services (EVS); ANSI C code (fixed-point)."
[29] ITU-T Rec. P.863, "Perceptual objective listening quality assessment," March 2016.
[30] H. Schulzrinne and S. Casner and R. Frederick and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," IETF RFC 3550, Jul. 2003.
[31] GSMA IR.92, "IMS Profile for Voice and SMS," version 11.0, 15 June 2017.
[32] H. Schulzrinne and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control," IETF RFC 3551, Jul. 2003.
[33] J. Ott and al., "Extended RTP Profile for Real-time Transport Control Protocol (RTCP)-Based Feedback (RTP/AVPF)," IETF RFC 4585, Jul. 2006.

ISBN 978-90-827970-1-5 © EURASIP 2018