EVS Channel Aware Mode Robustness to Frame Erasures

Anssi Rämö[1], Antti Kurittu[2] and Henri Toukomaa[1]
Nokia Technologies[1]. Nokia Networks[2]
e-mail: anssi.ramo@nokia.com, antti.kurittu@nokia.com

April 19, 2016

Abstract

This document contains results taken from a recently made (15.2-23.2.2016) listening test conducted in Nokia Technologies Listening Test Laboratory [1]. The test consisted of two extended range MOS (ACR9) listening tests. The tests contained clean and noisy speech samples as well as samples with musical background. There were four clean speech talkers (two males and females). Noisy speech background samples (ACR9) included four different noisy environments. There were a total of 24 listeners in both listening tests. The test contained clean channel conditions as well as four frame erasure rates (5%, 10%, 20% and 30%) for selected codecs and bitrates. Frame errors were distributed randomly. The results show that EVS CA mode performs better than EVS native modes in high FER rates. For comparison also AMR, AMR-WB and Opus codecs were included to the listening test. EVS and EVS Channel Aware mode were performing significantly better than older 3GPP codecs or Opus at all operation points.

Additionally, objective tests were conducted to test the EVS CA mode with a large number of FER profiles. EVS CA proved to provide better quality in high FER scenarios than the reference codecs in the objective tests. Also, the high FEC offsets vs. low FEC offsets provided better speech quality in bursty FER conditions. The redundancy setting in EVS CA had little effect on performance.

Contents

1. Background

The EVS codec supports four input and output sampling rates (8, 16, 32, and 48 kHz). There are also twelve bitrates ranging from 5.9 kbit/s to 128 kbit/s. The 5.9 kbit/s mode is using VBR (Variable BitRate) with discontinuous transmission (DTX) always enabled [2]. All other bitrates are CBR (Constant BitRate) [3]. Audio and speech coding modes are switched internally in realtime by the EVS codec depending on the input signal characteristics. The EVS codec is designed to be inherently robust to channel errors [4]. Further information about the EVS codec can be found from specification, and ICASSP and GlobalSIP special session papers [5] [6] [7] [8].

There are several different coding modes that help with robustness to frame erasures. There is for example a specific transition mode that codes speech and audio onsets so that individual lost frames do not adversely affect audio quality [9]. Also for tonal music and stable speech vowels there are separate coding modes that handle these kind stable segments very well in case of frame loss [10]. Global gains are also coded independently from previous frames in order to reduce error propagation [11]. LSF spectral coefficients are coded with vector quantizer that takes into account possible lost frames by inserting non-predictive frames at most critical places [12]. Additional information about native EVS PLC can be found in several publications [13] [14] [15] and EVS channel error concealment specification [16].

1.1. CA mode

In addition to the EVS native mode frame error robustness features, EVS [16] contains a special channel aware (CA) mode [13] [17] (Chapter 5). The CA mode is meant to be used in IP based best effort networks such as VoIP, VoWifi and VoLTE, where speech frames could be delayed or lost totally. The error resilience is achieved by using a form of in-band forward error correction (FEC). In case of delayed or lost packet EVS CA uses partially coded frames embedded to the EVS bitstream to conceal the lost frames with less artifacts This requires a novel usage of jitter buffer[18]. Naturally partial redundant frames reduce the full rate bitrate somewhat. Depending on the criticality of the frame, the partial redundancy is dynamically enabled or disabled for a particular frame, while keeping a fixed bit budget of 13.2 kbit/s. The amount of redundant information varies frame by frame and ranges from 0 to 3.6 kbit/s. 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 selfcontained partial copies are used to improve the error robustness in case the original primary frame is lost or discarded due to late arrival. This channel aware functionality is specified only for 13.2 kbit/s wideband (WB) and superwideband (SWB). All other bitrates and signal bandwidths use standard packet loss concealment (PLC) defined in [16]. Also when no redundant information is found for any particular lost frame normal packet loss concealment algorithm is used.

1.2. CA mode details

The difference in time units (in EVS case the frame length is 20 ms) between the transmit time of the primary copy of a frame and the transmit time of the redundant copy of the frame (piggy backed onto a future frame) is called the FEC offset. If the depth of the jitter buffer at any given time is at least equal to the FEC offset, then it is quite likely that the future frame is available in the de-jitter buffer at the current time instance. The FEC offset is a configurable parameter at the encoder which can be dynamically adjusted depending on the network conditions. The redundant copy is only a partial copy that includes just a subset of parameters that are most critical for decoding or arresting error propagation. In general the offset should be longer for a very erroneous channel and shorter for a good channel. Allowed offset values are 2, 3, 5, or 7. The offset value can be changed during the connection and it is sent in the bitstream with 2 bits for every redundant copy containing frame. Since the encoder needs to know the FEC offset in the encoding stage, there has to be some sort of signaling established in the connection, or else a fixed offset is used.

Additionally there is a frame erasure rate (FER) indicator in the encoder having the following values: low for FER rates less than 5% or high for FER higher than 5%. The high setting adjusts the criticality threshold to classify more frames as critical to transmit as compared to the low setting.

The defaults for the encoder are high and FEC offset = 3, if nothing else is specified for the encoder. The RF frame offset information (i.e., offset = 2, 3, 5, or 7) at which the partial copy is transmitted with the primary frame is included in the bit stream. Similarly, the RF frame type with 3 bits is sent with every redundant information containing frame. RF frame type signal are: RF_NO_DATA, RF_TCXFD, RF_TCXTD1, RF_TCXTD2, RF_ALLPRED, RF_NOPRED, RF_GENPRED, and RF_NELP. Due to listening test size limitations tested configurations can be seen in Table 5. Since the specification says that low frame erasure rates use low value and lower FEC offset should be used for low error levels, we set the FEC offset to increase with increasing FER rate and two lowest FER rates used low setting and two highest FER rates used high setting.

2. Description of the Performed Subjective Test

In order to make robustness improvements visible also reference codecs are needed for comparison. In addition to the main test subject EVS channel aware mode the reference codecs for the listening test were EVS native mode as well as standard AMR and AMRWB codecs. Also Opus codec was included to the listening test, since being a relatively new codec and of great interest to various VoIP based systems. All codecs were tested around 12.2-13.2 kbit/s range. In addition EVS-NB was included at 8 kbit/s as well as EVS-FB and Opus (in fullband also) at 24.4 kbit/s. Full list of tested codecs and frame erasure rates can be seen in Table 4.

The codec versions tested were:

- opus-1.1.2.tar.gz [19] [20]

  following command line was used:

  ```
  -e voip 48000 1 xx -cbr input.wav output.pkt
  ```

  Bitrates tested (xx) were 13200, and 24400 bit/s. Constant bitrate mode was used for the listening testing. The Opus decoder was modified so that it read from a file error pattern and inserted lost frames to proper places.

- Latest available fixed point version of EVS codec was used for the testing [21]

  following command line was used:

  ```
  -q -rf [LO/HI] [2/3/5/7] xx [8/16/32/48] input.raw bitstream.pkt
  ```

  Bitrates tested (xx) were 8000 (NB), 13200(WB, SWB), and 24400 (FB) bit/s. LO or HI defines the frame erasure rate indicator. [2/3/5/7] defines the FEC offset. [8/16/32/48] is the sampling rate in kHz.

All other codecs were standardized codecs and latest fixed point version was used for the processing. For error insertion to 3GPP codecs eid-xor utility was used together with some Nokia Networks provided VoIP error insertion tools.

2.1. Extended Range ACR 9-Scale Method

A modified version of the traditional ACR (absolute category rating) [22] testing was used for the listening test. The MOS (mean opinion score) scale was extended to be 9 steps wide with scale shown in Table 1. Absolute category rating is a very good and fast test methodology for quality control purposes, although it is not as discriminative as for example pairtest or DCR (degradation category rating).

Grading value	Estimated Quality
9	Vey good
8	
7	
6	
5	
4	
3	
2	
1	Vey bad

Table 1: 9-step ACR evaluation scale

Only the extremes of the scale are described, since it is very difficult to put 7 distinctive adjectives in multiple languages. Also the distortions may be very different and their descriptions are almost impossible to define exactly with the wide range of codecs used in the testing. The 9-scale ACR scale saturates less easily than the standard 5-scale ACR. In practice this new scale is somewhat between MUSHRA[23] and 5-scale ACR[22]. The voting is not free sliding, but nine different values still helps the listener to discriminate the samples better. Some more information about the usage of 9-scale ACR testing and various listening test results can be found in publications [24] [25] [26] [27].

2.2. Listening

There were a total of 24 listeners in both listening tests. All of them were naive and younger than 20 years old. Each listener listened to all conditions with 8 different samples from eight different voice sample categories. The sample categories are described in Table 2. In each category there were four different sample sequences and they were divided into four sets. Thus all samples were listened to with every condition 8 times over and each condition received a total of 192 votes in combined results. Samples were listened to with Sennheiser HD-650 headphones. Diotic listening was conducted for improved accuracy.

2.3. Test samples

The original voice samples were recorded at 48kHz in a quiet studio environment. The office, car and street noises are the same as used in EVS qualification testing with the exception of increased noise level to -15 dBOv. The music sample originates from a regular CD and it was upsampled to 48kHz with a high quality resampling program[28]. All sample sequences were six to seven seconds long containing a sentence pair and in the noisy test also background noise. The full list of tested sample types can be seen in Table 2.

Test	Environment	Speaker	Background noise level
1	Clean speech	Male 1	na
1	Clean speech	Female 1	na
1	Clean speech	Male 2	na
1	Clean speech	Female 2	na
2	Speech in street noise	Male 3	-15 dB
2	Speech in cafeteria noise	Female 3	-15 dB
2	Speech over classical music	Male 4	-15 dB
2	Speech in car noise	Female 4	-15 dB

Table 2: Sample types used for ACR9 listening testing

2.4. Test conditions

In addition to the EVS CA mode EVS native mode as well as a selection of older 3GPP coded samples were included in the test for comparison reasons. See Table 4 for full list of conditions. Additional references were bandwidth limited signals using ITU-Tools [29] as well as MNRU noise worsened samples [30] [31]. The reference signal list can be seen in Table 3

Reference	Bandwidth	Notes
Direct FB	24 kHz	Unlimited bandwidth
10 kHz limited	10 kHz	Bandwidth limited signal
WB	8 kHz	Unlimited bandwidth
NB	4 kHz	Bandwidth limited signal
NB MNRU 16 dB	4 kHz	MNRU noise [30]
WB MNRU 18 dB	8 kHz	MNRU noise
FB MNRU 16 dB	24 kHz	P.50 modied MNRU noise [31]
FB MNRU 24 dB	24 kHz	P.50 modied MNRU noise

Table 3: Tested reference conditions

Codec	Bandwidth	Bitrate	FER conditions
AMR	NB	12.2	0, 5, 10, 20, and 30%
AMR-WB	WB	12.65	0 - 30%
AMR-WB	WB	23.85	Only 0%
EVS	NB	8.0	0 - 30%
Opus CBR	MB*	13.2	0 - 30%
EVS	WB	13.2	0 - 30%
EVS CA-mode	WB	13.2	0 - 30%
EVS	WB	24.4	Only 0%
EVS	SWB	13.2	0 - 30%
EVS CA-mode	SWB	13.2	0 - 30%
EVS	FB	24.4	0 - 30%
Opus CBR	FB	24.4	0 - 30%

Table 4: Tested codecs and respective channel frame erasure rates. * MB stands for medium band and is approximately 6 kHz bandwidth limited.

2.5. Channel aware mode conguration

According to the EVS specification [17], the CA mode configuration should be adapted to the channel conditions. Thus we tested the EVS CA-mode so that in lower frame erasure rates, the FER indicator was set to low and FEC offset was set to low values. With increasing frame erasure rate the FEC offset was increased and for FER rates 20 % and 30 % the rate indicator was set to high. Full set of configurations can be seen in Table 5

Frame erasure rate	FER indicator	FEC offset
0 %	low	2
5 %	low	3
10 %	low	5
20 %	high	5
30 %	high	7

Table 5: EVS CA mode tested configurations

3. Listening test results

Since there are so many results available that normal bar graphs and numerical tables are hard to read; a new method of representing listening test results is used. Codec and reference conditions are collected to a X-Y line graph, where bullets point to individual MOS results and line connects the bullets, when relevant scalability can be seen (e.g. Figure 1). The ACR9 MOS scale is shown on the left side of the table. The FER rate is shown in the bottom. All results are represented in linear scale. The colors and markers of individual codecs are also tried to be consistent over different graphs. Additionally, traditional bar graphs with confidence intervals are also presented for overall results in Figure 13.

3.1. Clean speech results

Clean speech results show that the CA mode works as expected. At clean channel or low FER rate it is statistically equivalent to EVS native mode both in WB and SWB. Especially at high FER rates of over 10 %, the channel aware mode improves subjective voice quality significantly. Notably AMR-WB codec performs quite poorly with increasing FER rate, although it shows good performance in clean channel in this listening test. Opus 13.2 kbit/s is significantly worse than AMR-WB 12.65 kbit/s mainly due to reduced bandwidth (6 kHz compared to over 7 kHz of the AMR-WB) in clean channel. At FER rates higher than 10 % Opus is actually better than AMR-WB 12.65 kbit/s. However, even native mode EVS-WB is all the time significantly better than Opus at 13.2 kbit/s. EVS channel aware mode is more than 1.5 MOS points better than Opus at all FER rates.

EVS-NB 8.0 kbit/s is also better than AMR 12.2 kbit/s at every operation point. Although at low FER rates the difference is not significant.

EVS-FB at 24.4 kbit/s and Opus 24.4 kbit/s perform almost identically in clean channel and at all FER rates. The performance of Opus at 24.4 kbit/s is not a surprise, since we have tested it earlier [27]. The similar FER robustness shows that both Opus and EVS-FB 24.4 kbit/s have something to improve since even EVS-WB is working better than either Opus or EVS at 24.4 kbit/s in 20 % and 30 % FER. The reason for the relatively poor performance of fullband codecs in high FER is likely the artifacts caused by the on/off-nature of the high frequency excitation, which degrades overall voice quality. More stable narrower bandwidth probably would benefit the overall voice quality in noisy channel for both codecs.

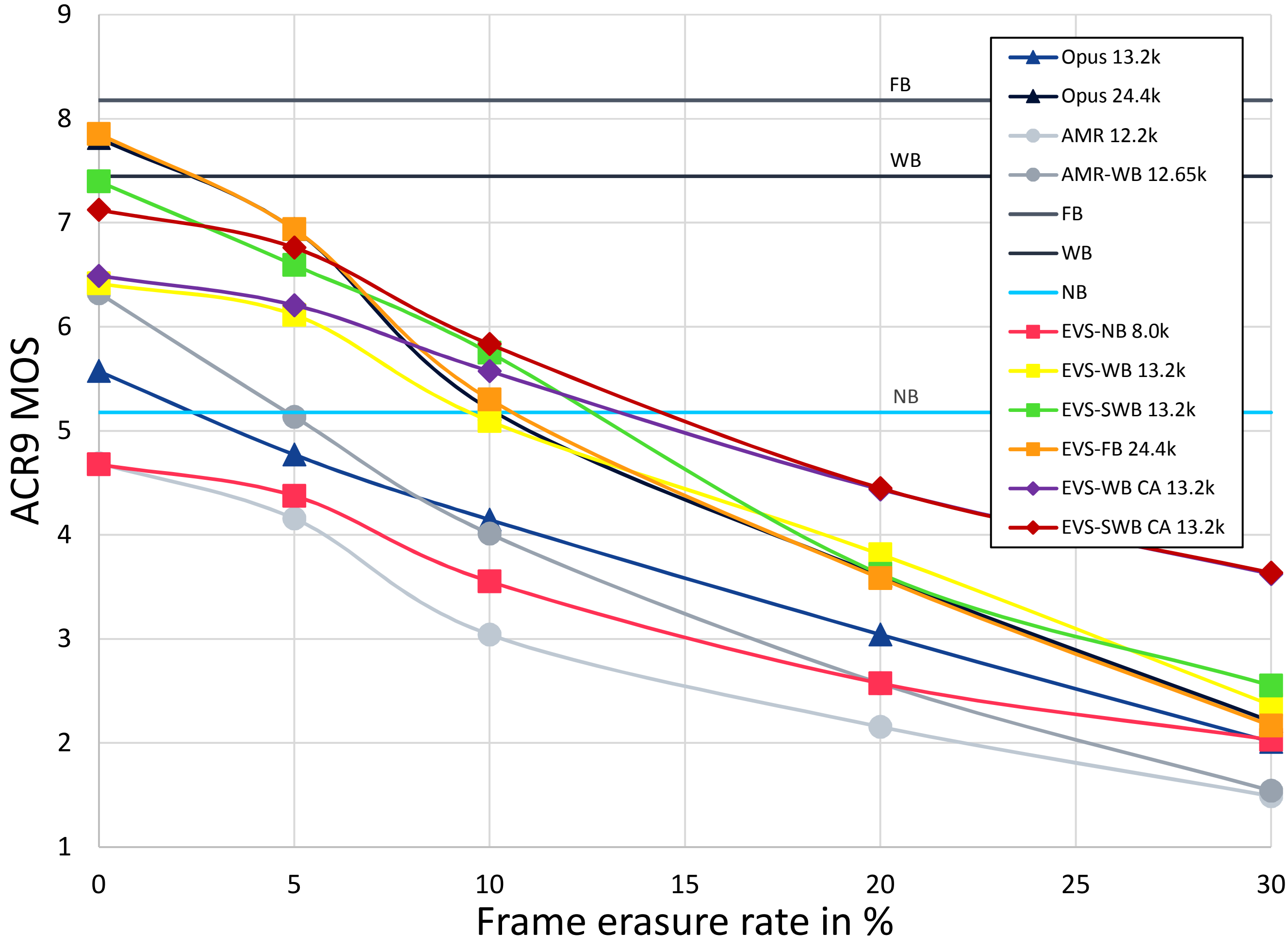

Figure 1: EVS channel aware mode's and references conditions' voice quality with clean speech at increasing frame erasure rate

3.2. Noisy speech results

Noisy speech results show that the CA mode works as expected as can be seen in Figure 2. At clean channel or low FER rates it is statistically equivalent to EVS native mode. Especially at high FER rates of over 10 % channel aware mode improves subjective voice quality. Compared to clean channel results in Chapter 3.1 and in Figure 1 the results are quite similar. The improvement provided by EVS CA mode is slightly less than with clean speech. The reason is that EVS CA mode is heavily optimized for clean speech performance.

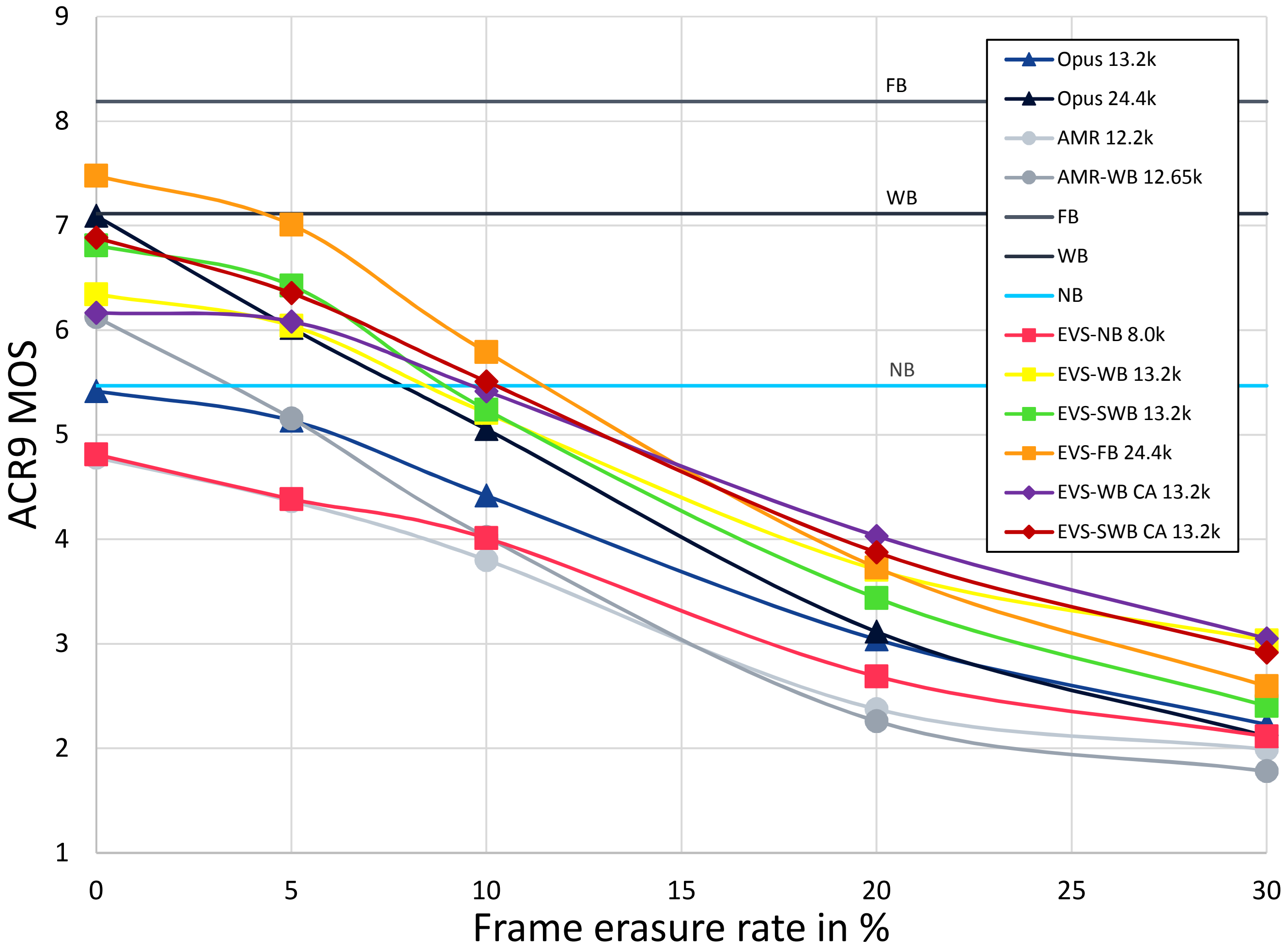

Figure 2: EVS channel aware mode's and references conditions' voice quality with background noise added speech at increasing frame erasure rate

3.3. Clean and noisy speech combined results

When we combine both clean and noisy speech results to the same graph we can see how the codecs compare against each other at varying FER rates. Figure 3 shows nicely that EVS channel aware mode improves robustness to frame erasures at higher frame erasure rates. Mostly similar conclusion can be drawn like in Chapter 3.1.

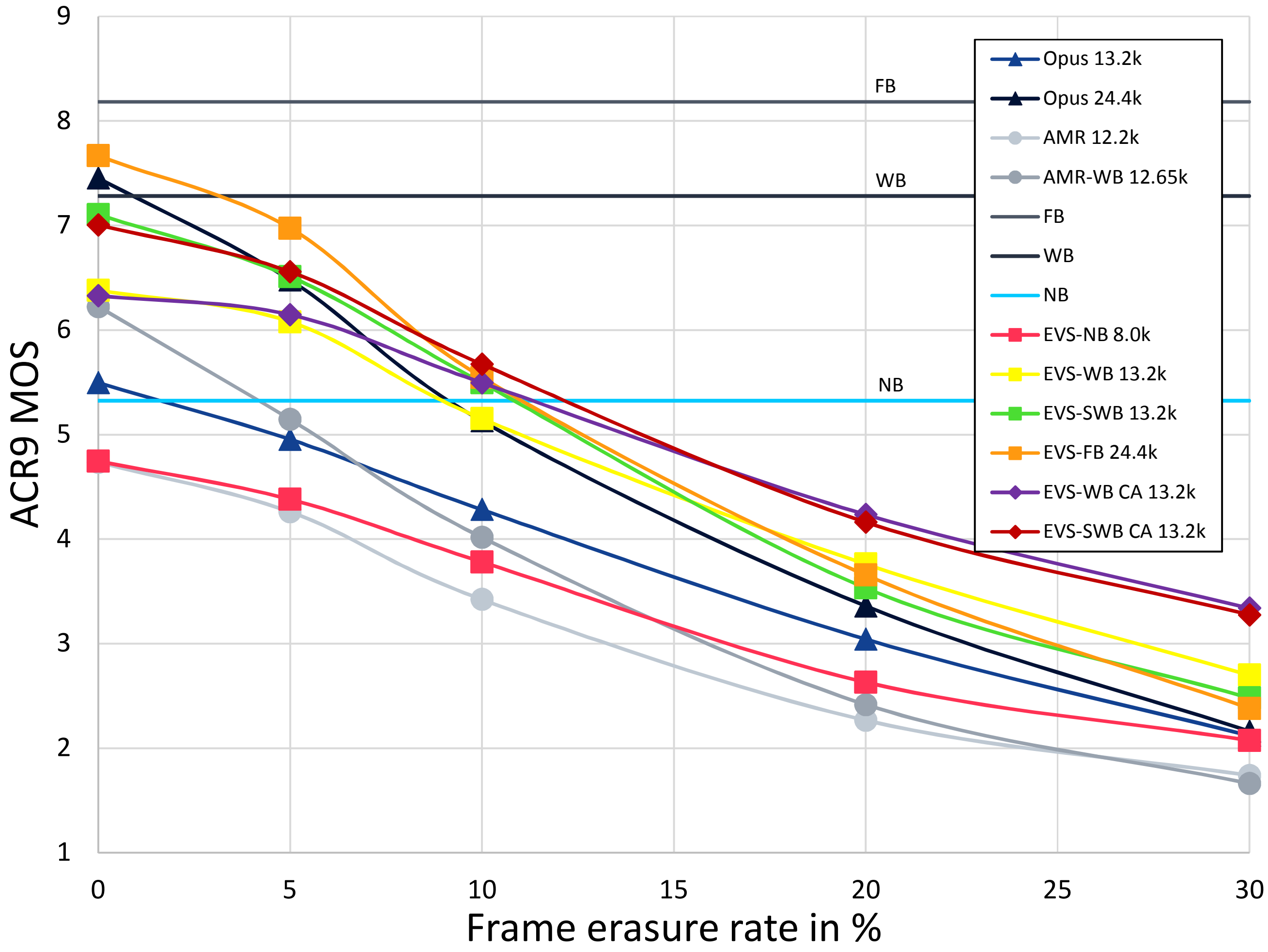

Figure 3: EVS channel aware mode's and references conditions' overall voice quality at increasing frame erasure rate

3.4. Differential comparison of EVS channel aware mode

When we subtract the reference codec results from the EVS CA mode results we get the following difference Figure 4 for NB/WB conditions and Figure 5 for SWB/FB conditions respectively. As can be seen in both figures, the EVS CA mode is significantly better than reference codecs: in NB/WB Figure 4 at FER rates starting at 10 % and in SWB/FB figure starting at around 15 %. The confidence interval in this listening test was around 0.2 for combined results as can be seen in Figure 13, where confidence interval is shown for all conditions. Thus any difference larger than 0.2 can be considered significant.

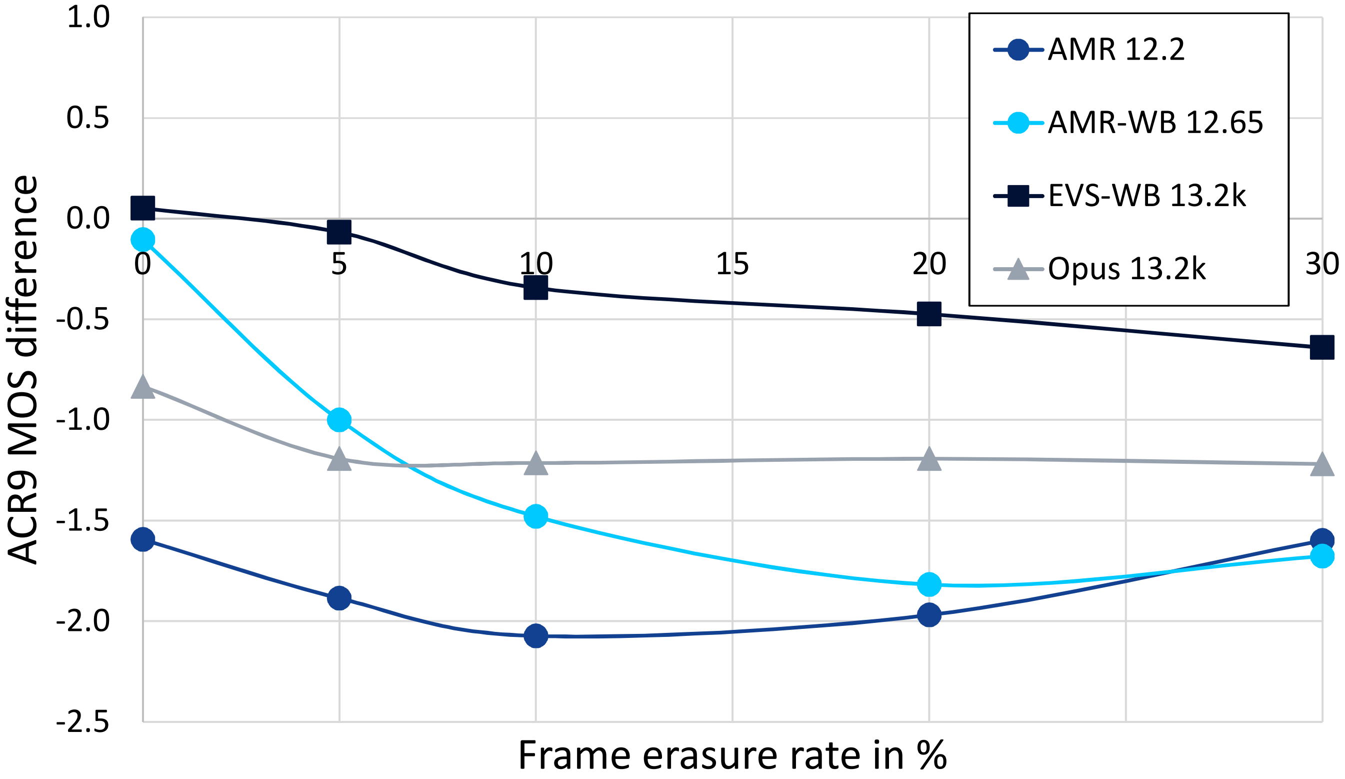

Figure 4: Difference of narrowband and wideband reference codecs to EVS-WB CA mode at 13.2 kbit/s

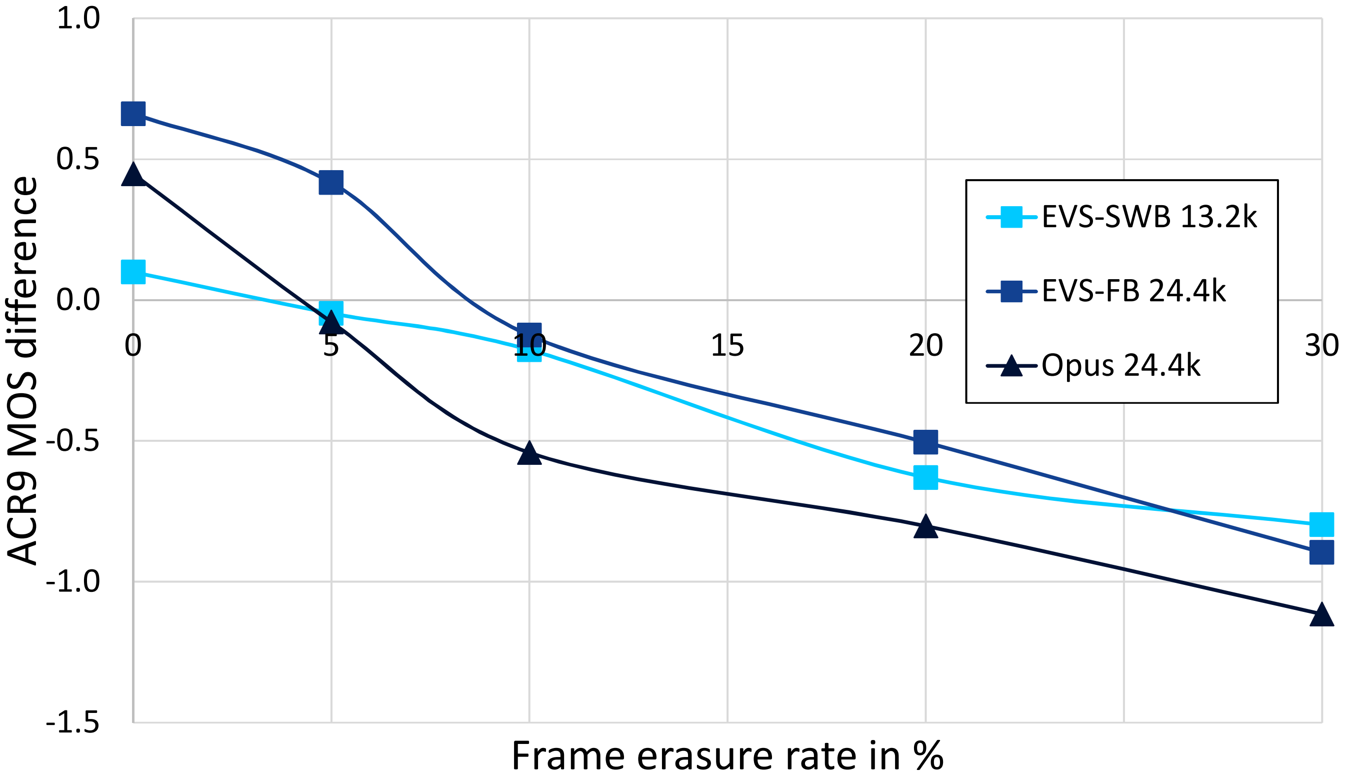

Figure 5: Difference of superwideband and fullband reference codecs to EVS-SWB CA mode at 13.2 kbit/s

If we just look at EVS native mode against the same EVS CA mode we get following Figure 6. From the Figure it can be seen that in clean channel conditions there is a negligible quality degradation for the CA mode compared to the EVS native mode both in WB and SWB mode. With increasing frame erasure rate the relative performance of the EVS CA mode improves. Already at 10 % frame erasure rate the CA mode is significantly better than native mode. In 30 % frame erasure rate CA-mode is on average 0.7 ACR9 MOS points better than EVS native mode.

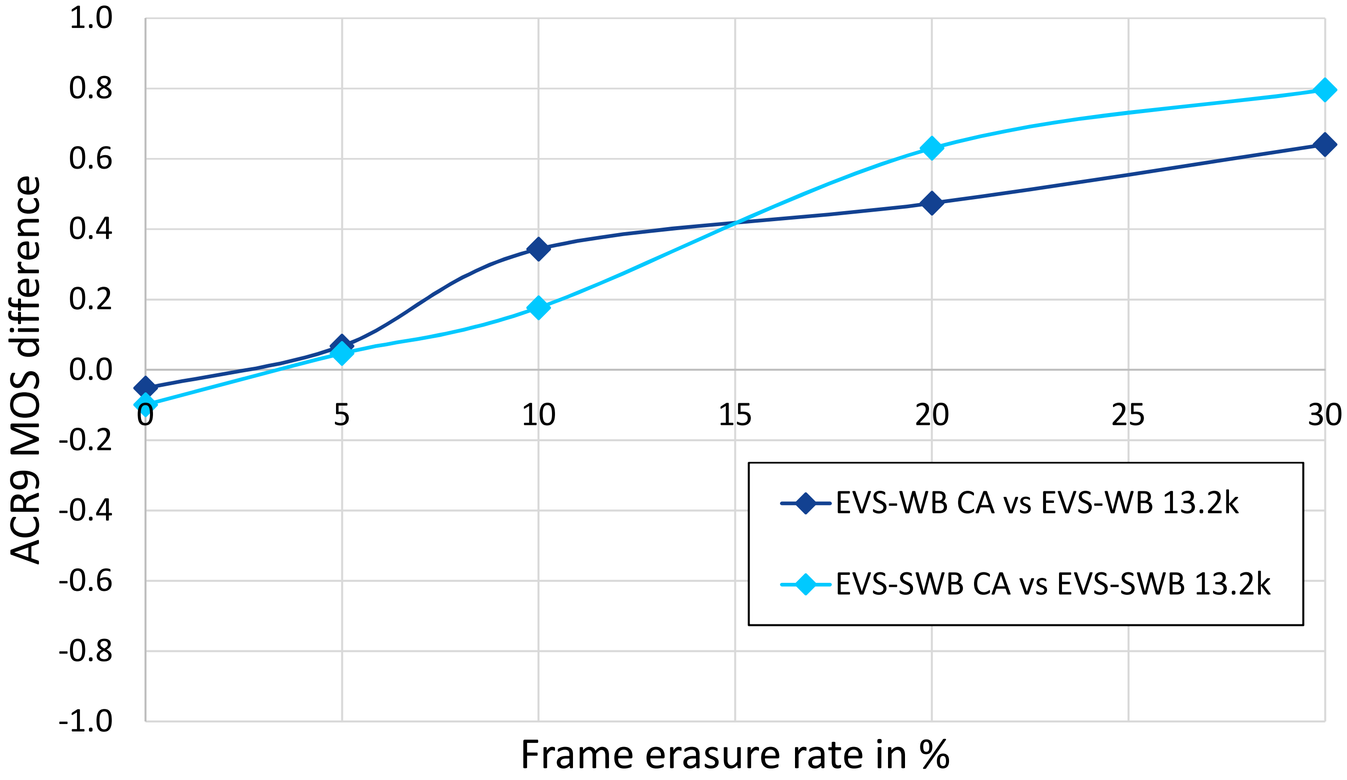

Figure 6: EVS at 13.2 kbit/s the difference between CA mode enabled and EVS native mode with increasing frame erasure rate. Both WB and SWB modes shown.

4. Objective test results

In addition to the subjective listening test also objective measurements were conducted with P.863 a.k.a POLQA [32]. Recommendation ITU-T P.863 describes an objective method for predicting overall listening speech quality from narrowband (NB) (300 to 3400 Hz) to super-wideband (SWB) (50 to 14 000 Hz) telecommunication scenarios as perceived by the user in an absolute category rating (ACR) listening-only test. The super-wideband mode of POLQA was used, providing predicted scores on a MOS 5-scale ACR super-wideband scale for narrowband, wideband and super-wideband scenarios. The recommendation does not support fullband speech. The maximum objective MOSLQOswb score is 4.75. POLQA enables fast processing of large amount of files, making it possible to assess the performance of codecs in more conditions than is possible in a subjective test.

4.1. Comparison between objective POLQA and subjective listening test results

The same speech samples that were used in the listening test were processed with the objective processing device POLQA. The fullband conditions were omitted as POLQA does not support fullband speech. It can be seen in Table 6 that POLQA ranks the 0 % FER conditions almost in the same order as the listeners in the performed listening test. There are some differences however. For example POLQA ranks AMR-WB 23.85 kbit/s as better than EVS-SWB CA 13.2 kbit/s, whereas in the subjective test it's vice versa Also EVS-NB 8.0 kbit/s is significantly worse than AMR-NB 12.2 kbit/s in the objective measurement, whereas in the listening test they were scored equally. Moreover, it seems that POLQA is not able to predict differences in audio bandwidth as in the subjective test. POLQA scores the WB and 10 kHz limited reference conditions equally, while there is a clear difference in subjective scores. POLQA also scores most of the super-wideband conditions similarly, while there is a difference of over 0.7 MOS in the subjective test. The Opus 13.2 kbit/s medium band codec is scored equally to EVS-WB CA 13.2 kbit/s, while there is a difference of 0.8 MOS in the subjective score.

The subjective test was conducted on an extended MOS scale and it included fullband conditions. The POLQA model is trained with data from 5-scale ACR test where superwideband conditions are the highest bandwidth. Thus the basis of the subjective and objective scores are different here. Based on the results, it seems that POLQA scores saturate in the high end of the MOS scale with good quality codecs. This could in part be a characteristic of the 5-scale ACR test for a subjective test containing a large variety of bandwidths (narrowband to super-wideband). It could also be feature of POLQA. For an apples-to-apples comparison, the subjective test should also be on a 5-scale ACR. As is claimed in Chapter 2.1 the ACR9 gives much more accurate results than ACR5 especially with wide variety of bandwidths and that is really seen here. Naturally the objective scores also depend heavily on the used speech samples so the result may have been different with a different sample set.

Condition	MOS	MOS-LQOswb
10kHz limited	7.90	4.75
WB	7.28	4.74
EVS-WB 24.4k	7.17	4.70
EVS-SWB 13.2k	7.10	4.51
EVS-SWB CA 13.2k	7.01	4.28
AMR-WB 23.85k	6.40	4.43
EVS-WB 13.2k	6.38	4.31
EVS-WB CA 13.2k	6.33	4.16
AMR-WB 12.65k	6.22	4.26
Opus 13.2k	5.49	4.18
NB 4kHz	5.32	3.84
EVS-NB 8.0k	4.74	3.27
AMR 12.2k	4.73	3.61

Table 6: Subjective MOS vs. objective MOS-LQOswb at 0 % FER

Figure 7 shows the objective MOS scores for the same conditions as in the subjective test in Figure 1. The EVS CA mode's better quality compared to reference codecs at higher FER (>5 %) ratios is apparent. One slight difference is that POLQA actually evaluates the quality degradation between clean channel and 5 % FER larger than it actually is in subjective testing.

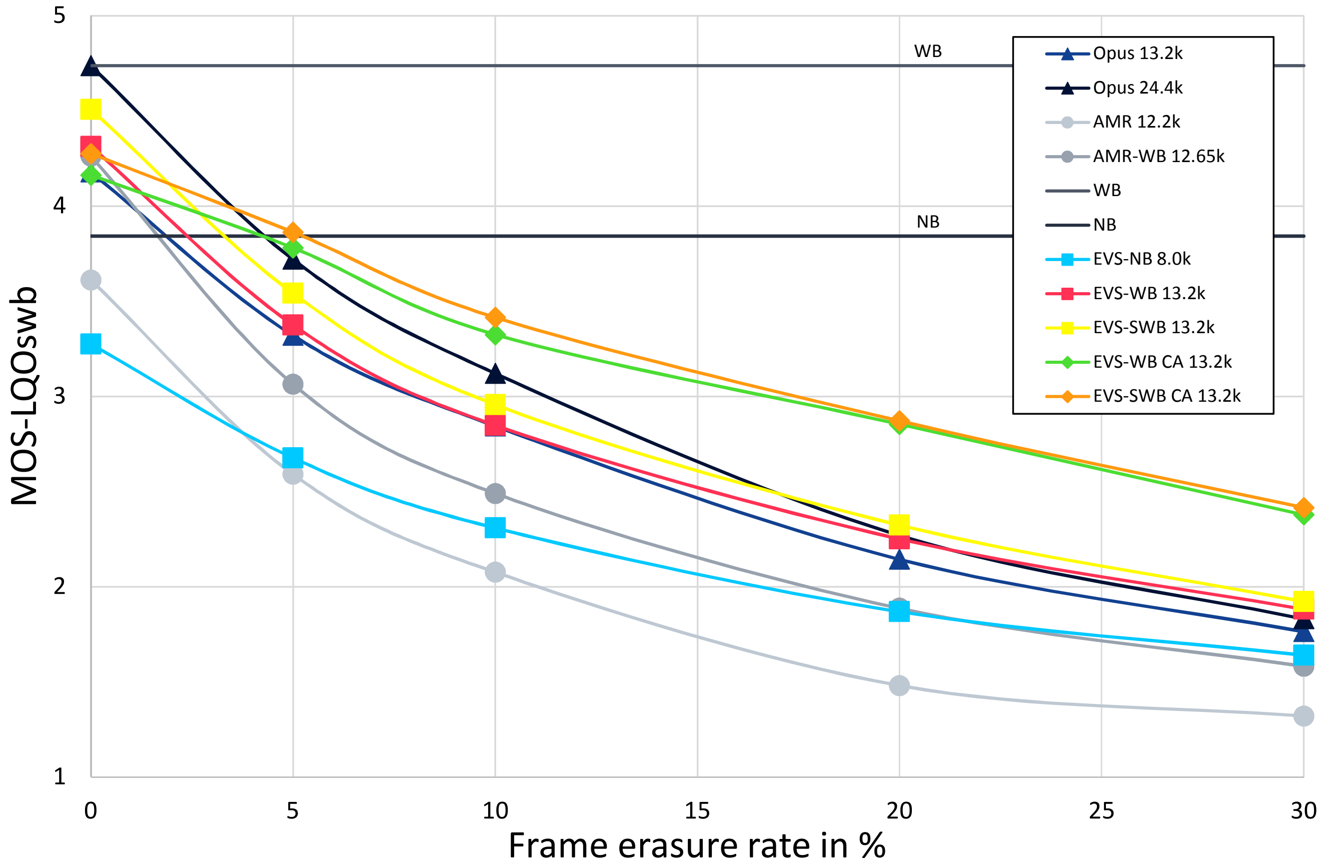

Figure 7: Voice quality with clean speech at increasing frame erasure rate

Finally both subjective and objective results are shown in the same figure 8. The solid lines are showing subjective results with scale shown on the left side ranging from 1 - 9. Objective results are shown with dotted lines with scale shown on the right side ranging from 1 - 5.

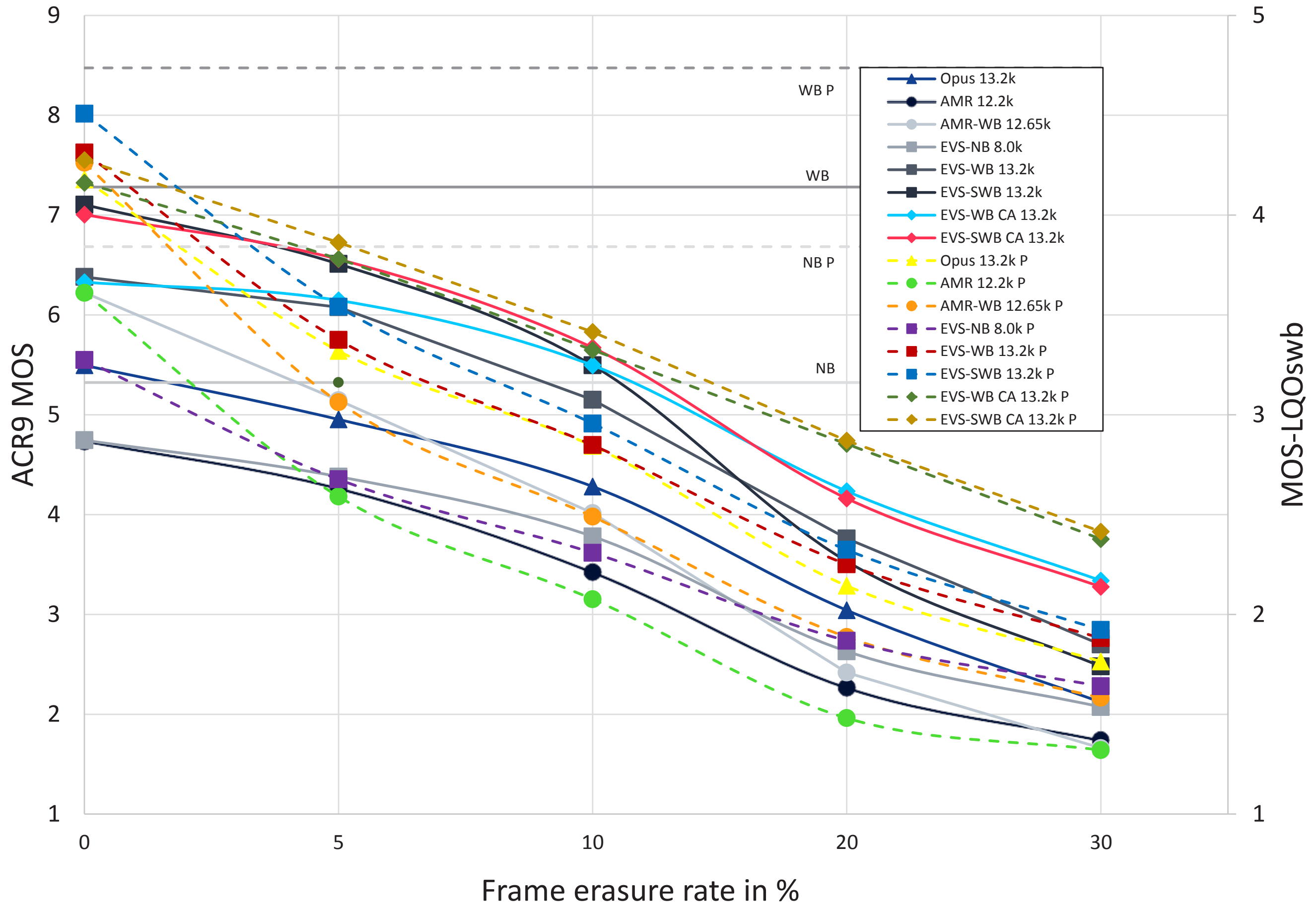

Figure 8: Subjective and objective voice quality with increasing frame erasure rate. Solid line represent subjective results with scale on the left side. Objective results are shown with dotted line with scale on the right side.

4.2. POLQA in FER

An additional objective test was performed for similar conditions as in the subjective test. The test was performed to get more variation to the FER profiles. The subjective test is limited in size due to practical restrictions. No such restrictions apply to objective speech quality testing. The test was conducted with one reference speech sample, chosen from ITU-T P.501 Amendment 3 [33]. The recommendation contains speech files for use with perceptual based objective speech quality prediction. FER profiles were generated with Itu-tools genn-patt. The random FER model with gamma parameter 0.4 was used to get a bit more burstiness to the FER profiles. The FER profiles spanned from 0 % to 50 % with one percent intervals. A hundred profiles were generated per FER percentage, adding up to 5100 total FER profiles. The distribution of FER burst lengths can be seen in the Figure 9.

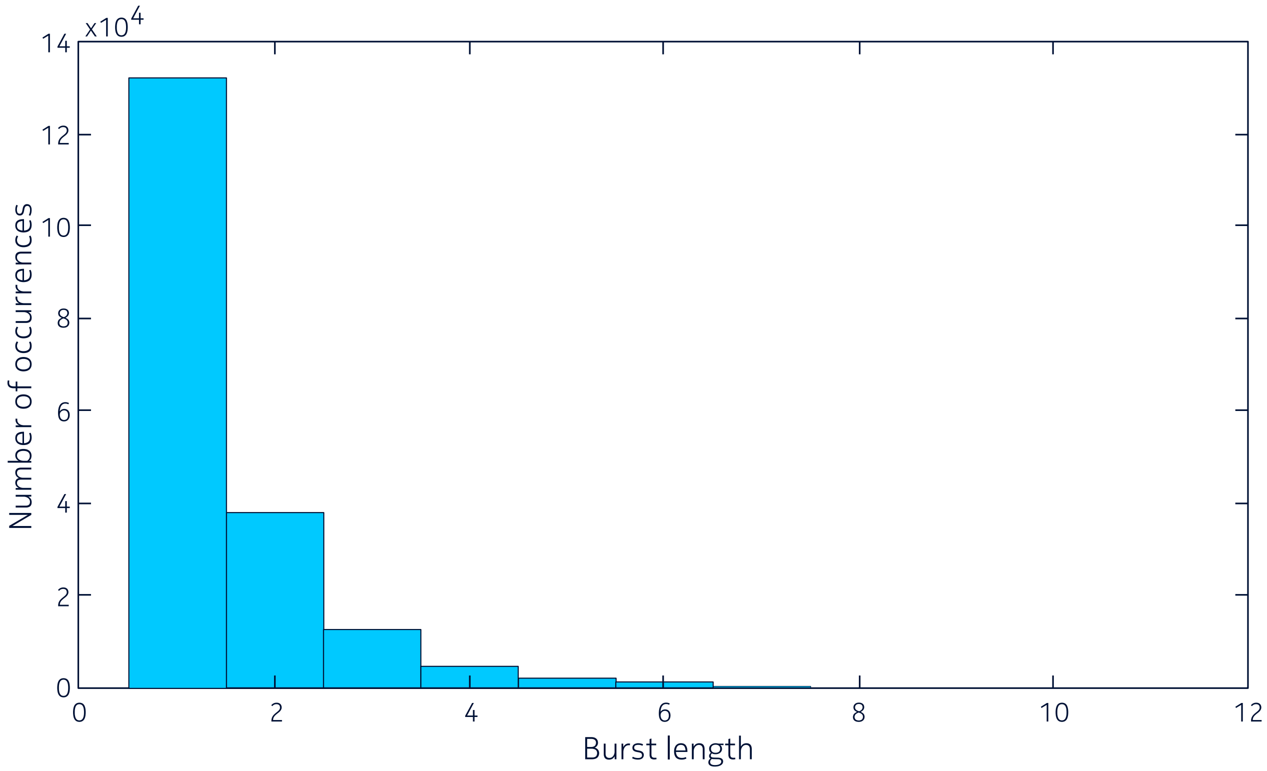

Figure 9: Frame erasure burst length distribution in FER profiles

For EVS CA mode, low FER indicator was used if FER of the used profile was <5 %. The high frame erasure rate indicator was used otherwise. Opus was used in variable bitrate mode in this test. The higher Opus bitrate was coded in super-wideband mode and the lower in wideband mode. A 6th order polynomial was then fitted to the data to obtain a MOS-LQOswb vs FER curve for each condition.

As can be seen the results are quite similar to the ones obtained in the subjective listening test. This is a quite good achievement since EVS, AMR-WB and Opus all use relatively quite different technologies to achieve their quality and POLQA has sometimes shown to be quite unreliable with different technologies utilizing codecs [34]. Of course not all details are perfectly in line, as different FER profiles and a different speech sample was used. But for example EVS channel aware mode's high quality at higher FER ratios shows up nicely in both results (compare e.g. Figure 1 and 7).

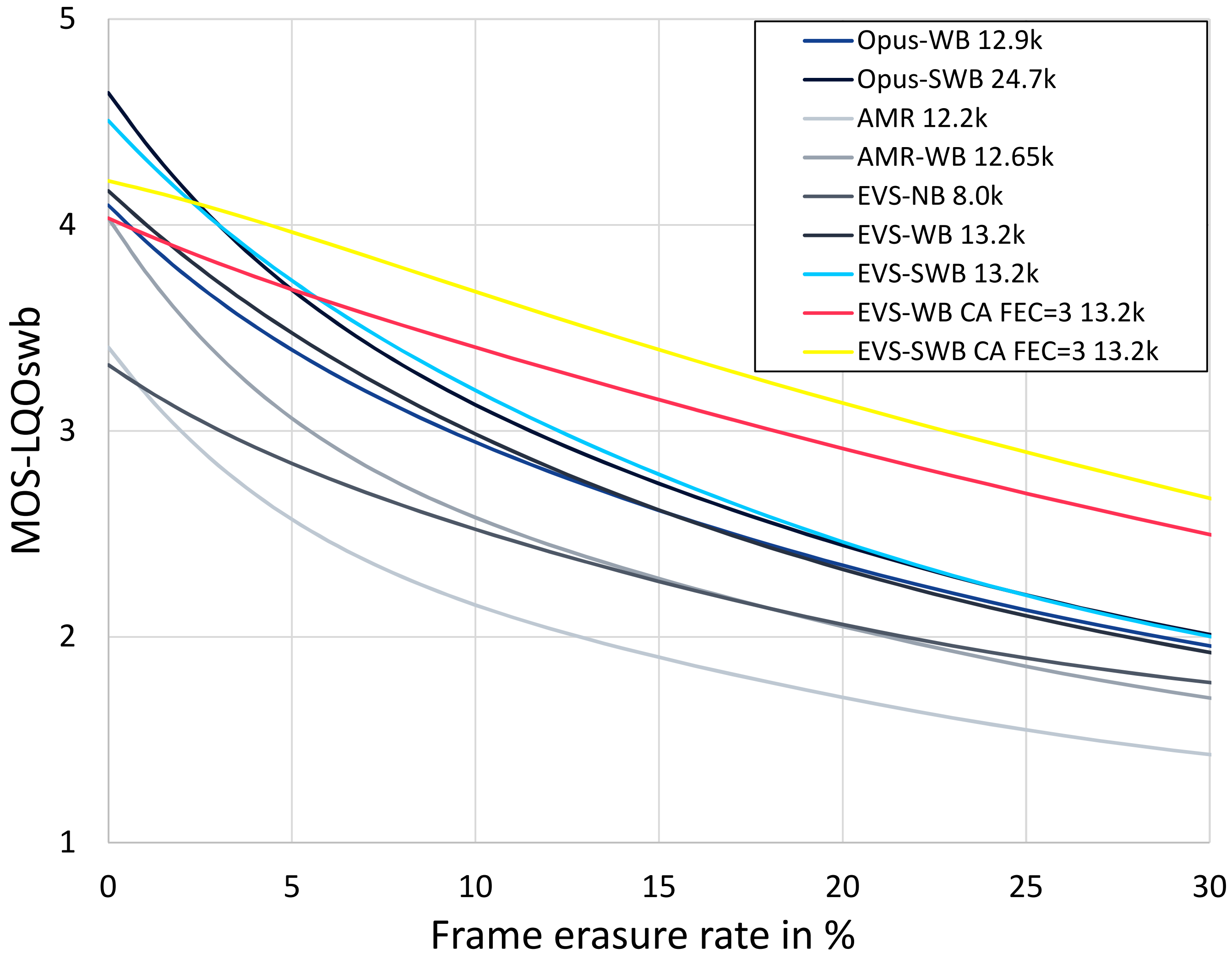

Figure 10: Objective voice quality with clean speech at increasing frame erasure rate

4.3. POLQA in bursty FER

FER profiles were again generated with Itu tools genn-patt. This time the bursty FER model was used to get heavily bursty FER profiles. The FER profiles spanned from 0 % to 30 % with one percent intervals. A hundred profiles were generated per FER percentage, adding up to 3100 total FER profiles. The same speech samples was used as in 4.2. Following figures illustrate how the performance changes with different EVS CA mode parameters in bursty FER scenarios. The assumption was that with increasing FEC offset the performance should be better in high FER scenarios. Also it was expected that the high FER indicator setting would be better than the low setting in high FER. The results show that with higher FER increasing the FEC offset helps the voice quality. The downside of increasing FEC offset is increasing delay in decoding. The FER indicator setting low/high had little effect on the results. It is possible that with some other samples the FER indicator setting would have an effect, as the redundancy depends on the used sample as well.

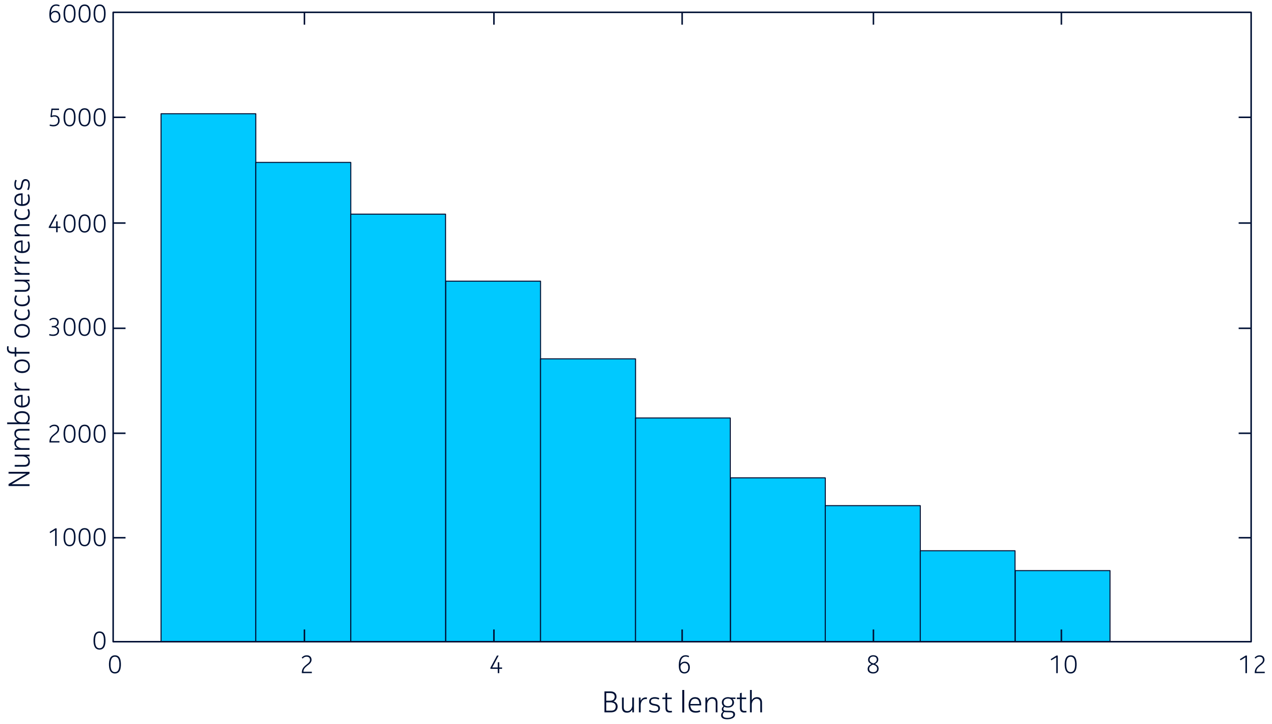

Figure 11: Frame erasure burst length distribution in bursty FER profiles

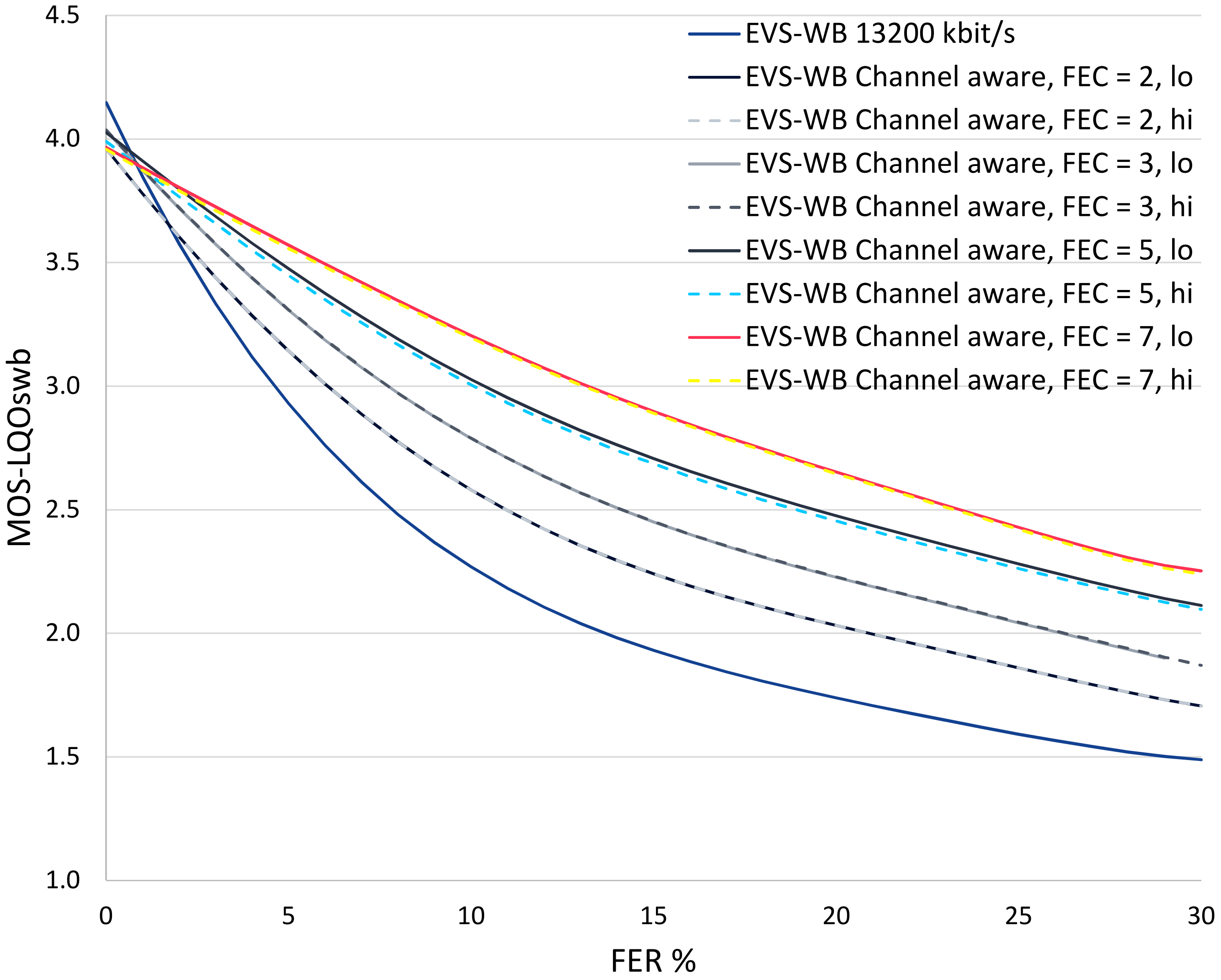

Figure 12: Effect of EVS CA-mode FEC offset and FER indicator setting on objective voice quality

5. Conclusion

In conclusion EVS-CA (Channel Aware) mode provides nice boost in quality with high FER ratios. Currently EVS-CA mode only supports bitrate of 13.2 kbit/s with either WB or SWB bandwidth (NB is not supported). It also requires that packet transmission network together with jitter buffer in the decoder side is used. This means that channel aware benefits cannot always be realized, but still 13.2 kbit/s SWB will likely be the most used mode anyways in VoLTE. Comparison to older generation 3GPP codecs or Opus at the same bitrate shows that there is significant improvement in both subjective and objective voice quality at all FER rates.

Finally Figure 13 shows all conditions in bar format with confidence intervals, allowing more detailed comparison of various pairs of codecs and FER conditions. Especially interesting is that reasonable communications quality (MOS>3 similar to e.g. AMR 12.2 kbit/s in 10% FER) is achieved with EVS-CA mode with 30% FER allowing communications even when almost one third of the frames are lost or delayed in the network.

The objective tests showed that with bursty FER, increasing the FEC offset helps the voice quality. The tests also revealed that the FER indicator setting low / high has very little effect on voice quality. The objective and subjective results correlate quite nicely, although POLQA had some trouble consistently predicting the quality of speech between different bandwidths.

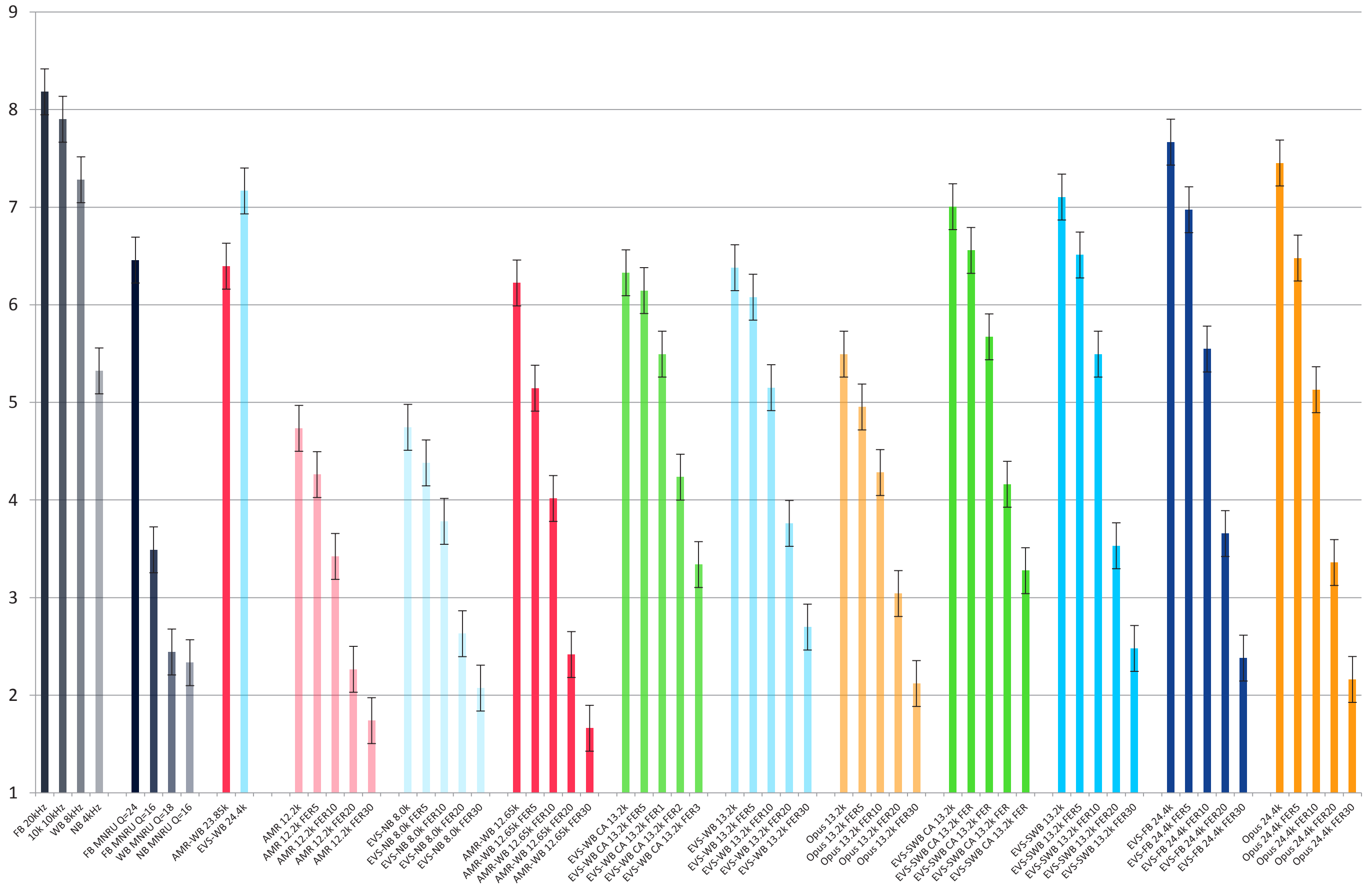

Figure 13: Combined complete results in bars format with confidence intervals.

6. References

1. Mikko Kylliäinen, Heikki Helimäki, Nick Zacharov, and John Cozens, “Compact high performance listening spaces,” in Proc. Euronoise, Naples, Italy, May 2003.
2. Kari Järvinen, Anssi Rämö, Jari Hagqvist, Olli Kirla, Adriana Vasilache, and Lasse Laaksonen, “Nokia networks white paper: The 3GPP Enhanced Voice Services (EVS) codec,” June 2015.
3. 3GPP TS 26.450, “Codec for enhanced voice services (evs); discontinuous transmission (dtx),” 2015.
4. Jeremie Lecomte, Tommy Vaillancourt, Stefan Bruhn, Hosang Sung, Ke Peng, KeiKikuiri, Bin Wang, Shaminda Subasingha, and Julien Faure, “Packet loss concealment technology advances in EVS,” in Proc. ICASSP, Brisbane, Australia, April 2015.
5. 3GPP TS 26.441, “Codec for Enhanced Voice Services (EVS); General overview,” 2015.
6. Stefan Bruhn, Harald Ploboth, Markus Schnell, Bernhard Grill, Jon Gibbs, Lei Miao, Kari Järvinen, Lasse Laaksonen, Noboru Harada, Nobuhiko Naka, Stéphane Ragot, Stéphane Proust, Takako Sanda, Imre

Varga, Craig Greer, Milan Jelinek, Minjie Xie, and Paolo Usai, “Standardization of the new 3GPP EVS codec,” in Proc. ICASSP, Brisbane, Australia, April 2015.

7. Martin Dietz, Markus Multrus, Vaclav Eksler, Vladimir Malenovsky, Erik Norvell, Harald Ploboth, Lei Miao, Zhe Wang, Lasse Laaksonen, Adriana Vasilache, Yutaka Kamamoto, Kei Kikuiri, Stéphane Ragot, Hiroyuki Ehara, Vivek Rajendran, Venkatraman Atti, Hosang Sung, Eunmi Oh, Hao Yuan, and Changbao Zhu, “Overview of the EVS codec architecture,” in Proc. ICASSP, Brisbane, Australia, April 2015.

8. Stefan Bruhn, Tomas Frankkila, Frédéric Gabin, Karl Hellwig, and Maria Hultström, “System aspects of the 3GPP evolution towards enhanced voice services,” in Global-SIP 2015, Orlando, Florida, USA, December 2015.

9. Václav Eksler, Milan JelÃnek, and Redwan Salami, “Efficient handling of mode switching and speech transitions in the EVS codec,” in Proc. ICASSP, Brisbane, Australia, April 2015.

10. Ralph Sperschneider, Janine Sukowski, and Goran Markovic, “Delay-less frequency domain packet-loss concealment for tonal audio signals,” in Global-SIP 2015, Orlando, Florida, USA, December 2015.

11. Vladimir Malenovsky and Milan Jelinek, “Memory-less gain quantization in the evs codec,” in Global-SIP 2015, Orlando, Florida, USA, December 2015.

12. Adriana Vasilache, Anssi Rämö, Hosang Sung, Sangwon Kang, Jonghyeon Kim, and Eunmi Oh, “Flexible spectrum coding in the 3GPP EVS codec,” in Proc. of ICASSP-2015, Brisbane, Australia, April 2015.

13. Venkatraman Atti, Daniel J. Sinder, Shaminda Subasingha, Vivek Rajendran, Duminda Dewasurendra, Venkata Chebiyyam, Imre Varga, Venkatesh Krishnan, Benjamin Schubert, Jérémie Lecomte, Xingtao Zhang, and Lei Miao, “Improved error resilience for VoLTE and VoIP with 3GPP EVS channel aware coding,” in Proc.ICASSP, Brisbane, Australia, April 2015.

14. Jérémie Lecomte, Adrian Tomasek, Goran Markovic, Michael Schnabel, Kimitaka Tsutsumi, and Kei Kikuiri, “Enhanced time domain packet loss concealment in switched speech/audio codec,” in Proc. ICASSP, Brisbane, Australia, April 2015.

15. Anssi Rämö, Adriana Vasilache, and Henri Toukomaa, “Robust speech coding with EVS,” in Proc. of Global-SIP 2015, Orlando, Florida, USA, December 2015.

16. 3GPP TS 26.447, “Codec for Enhanced Voice Services (EVS); Error concealment of lost packets,” 2015.

17. 3GPP TS 26.445, “Codec for Enhanced Voice Services (EVS); Detailed algorithmic description,” 2015.

18. 3GPP TS 26.448, “Codec for enhanced voice services (evs); jitter buffer management,” 2015.

19. Jean-Marc Valin and Koen Vos, “Definition of the opus audio codec,” http://tools.ietf.org/html/draft-ietf-codec-opus-10, October 2011.

20. Jean-Marc Valin, Koen Vos, and Timothy Terriberry, “Opus homepage,” http://www.opus-codec.org/.

21. 3GPP TS 26.442, “Codec for Enhanced Voice Services (EVS); ANSI C code (fixed point),” 2015.

22. ITU-T Rec. P.800, Methods for subjective determination of transmission quality, ITU, August 1996, online: http://www.itu.int/rec/T-REC-P.800-199608-I/en.

23. ITU-R Rec. BS.1534-1, “Method for the subjective assessment of intermediate quality level of coding systems,” June 2014.

24. Anssi Rämö, “Voice quality evaluation of various codecs,” in Proc. of ICASSP-2010, Dallas, TX, USA, March 2010, pp. 4662-4665.

25. Anssi Rämö and Henri Toukomaa, “Voice quality evaluation of recent open source codecs,” in Proc. of

Interspeech-2010, Tokyo, Japan, September 2010, pp. 2390-2393.

26. Anssi Rämö and Henri Toukomaa, “Voice quality characterization of IETF Opus codec,” in Proc. of Interspeech-2011, Florence, Italy, August 2011, pp. 2541-2544.
27. Anssi Rämö and Henri Toukomaa, “Subjective quality evaluation of the 3GPP EVS codec,” in Proc. of ICASSP-2015, Brisbane, Australia, April 2015, pp. 5157-5161.
28. McGill University, “AFsp programs and routines,” http://www-mmsp.ece.mcgill.ca/documents/Software/Packages/AFsp/AFsp.html.
29. ITU-T G.191, “Software tools for speech and audio coding standardization,” November 2009.
30. ITU-T P.810, Telephone transmission quality: methods for objective and subjective assessment of quality. Modulated noise reference unit (MNRU), ITU, February 1996, online: https://www.itu.int/rec/T-REC-P.810-199602-I/en.
31. ITU-T P.50, Telephone transmission quality, telephone installations, local line networks: Objective measuring apparatus, ITU, September 1999, online: http://www.itu.int/rec/T-REC-P.50-199909-I.
32. ITU-T Rec. P.863, “Perceptual objective listening quality assessment,” September 2014.
33. ITU-T Rec. P.501 Amendment 3, “Speech files with male/female sentences prepared for use with perceptual based objective speech quality prediction,” June 2012.
34. Hannu Pulakka, Ville Myllylä, Anssi Rämö, and Paavo Alku, “Speech quality evaluation of artificial bandwidth extension: Comparing subjective judgements and instrumental predictions,” in Proc. of Interspeech-2015, Dresden, Germany, September 2015.

Nokia Oyj
Karaportti 3
FI-02610 Espoo
Finland
Tel. +358 (0) 10 44 88 000

Product code SR1701005522EN