INTELLIGIBILITY EVALUATION OF SPEECH CODING STANDARDS IN SEVERE BACKGROUND NOISE AND PACKET LOSS CONDITIONS

Emma Jokinen[1,2]*, Jérémie Lecomte*[3]*, Nadja Schinkel-Bielefeld*[3]*, Tom Bäckström*[2,3]

[1]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
[2] International Audio Laboratories Erlangen, Friedrich-Alexander Universität (FAU), Germany
[3] Fraunhofer Institute of Integrated Circuits, IIS, Erlangen, Germany
emma.jokinen@aalto.fi

ABSTRACT

Speech intelligibility is an important aspect of speech transmission but often when speech coding standards are compared only the quality is evaluated using perceptual tests. In this study, the performance of three wideband speech coding standards, adaptive multi-rate wideband (AMR-WB), G.718, and enhanced voice services (EVS), is evaluated in a subjective intelligibility test. The test covers different packet loss conditions as well as a near-end background noise condition. Additionally, an objective quality evaluation in different packet loss conditions is conducted. All of the test conditions extend beyond the specification range to evaluate the attainable performance of the codecs in extreme conditions. The results of the subjective tests show that both EVS and G.718 are better in terms of intelligibility than AMR-WB. EVS attains the same performance as G.718 with lower algorithmic delay.

Index Terms— Speech intelligibility, packet loss concealment, adaptive multi-rate wideband, enhanced voice services, G.718

1. INTRODUCTION

In modern speech transmission systems, the intelligibility of the communication can be jeopardized by many factors. Phone calls are increasingly moving to IP-based networks which creates new challenges, such as varying delay and lost packets, for speech coding and processing. This can also happen to such a degree that the intelligibility of the speech is severely compromised. Another common problem is environmental noise in one or both ends of the communication channel for which many pre- and post-processing stages can be used in the mobile devices. For instance, in the speaker's end, the effects of the far-end noise can be diminished by utilizing noise suppression as a pre-processing step. Additionally, in the receiving device, the intelligibility of the speech can be increased over the near-end noise in the listener's surroundings with the utilization of post-processing techniques. However, it depends highly on the phone manufacturer whether additional enhancement techniques are used. Therefore, the performance of the speech codec alone in the presence of degradations is very important.

Speech coding standards are usually rigorously evaluated in terms of subjective speech and audio quality before they even become standards. For instance, the qualification tests for the new ETSI 3GPP Enhanced Voice Services (EVS) [1] included test conditions with far-end background noise as well as different frame error rates and jitter profiles [2]. However, the test conditions used in the qualification tests do not cause notable decrease in intelligibility, because the focus is on typical operating conditions where degradations are relatively small. Furthermore, intelligibility is often seen only as one aspect of quality [3, 4] along with other attributes such as naturalness, brightness and pleasantness. However, quality and intelligibility can also be considered as separate concepts [5]. Especially in the context of speech communication in severely degraded conditions, the quality of the signal is a secondary concern whereas the intelligibility of the speech signal is a high priority.

In this study, three different speech coding standards, the adaptive multi-rate wideband (AMR-WB) [6], the G.718 [7], and the recently standardized EVS, are evaluated in terms of intelligibility in near-end background noise and packet loss conditions. In development and testing of the EVS standard, the requirement was that it is better in quality than the current state-of-the-art codecs. In contrast, to assess the performance of the coding standards in terms of intelligibility, the test conditions used in this study fall outside of the normal test specification range [2]. This also gives information on the maximum amount of degradation that the coding standards can tolerate before the communication completely breaks down. The performance of the coding standards is evaluated using both objective measures and subjective intelligibility tests. The utilized objective measure is designed mainly for quality evalu-

The International Audio Laboratories Erlangen (AudioLabs) is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

ation and here it is used to close the gap between the standard test conditions and the ones used in the subjective evaluation conducted in this study.

The evaluated standards have different ages: the AMR-WB has been standardized already in 2001 whereas the G.718 is relatively new (2008) and the EVS was accepted as standard in September 2014. One of the interest points of the study was to evaluate how the age of the standard effects the performance, i.e., have the newer standards improved also in terms of intelligibility although it is not commonly considered as an evaluation metric. Additionally, evaluation of AMR-WB is especially interesting now because it has only recently been deployed as the usage of wideband telephony has increased.

2. EVALUATED CODING STANDARDS

The three selected standards, AMR-WB, G.718, and EVS are shortly described in the following. All of them are based on the algebraic code excited linear prediction (ACELP) speech coding paradigm but EVS and G.718 additionally have a modified discrete cosine transform (MDCT) based coding mode which can be used for generic audio. All of the codecs utilize a 20-ms frame length and are used in this study with 16-kHz sampling frequency. EVS with 13.2 kbps was selected as a reference mode because it corresponds to the most commonly used rate. For the other codecs, the bit-rates were selected around the reference mode to provide a fair comparison.

2.1. AMR-WB

The AMR-WB [6] is a multi-rate coder with bit-rates ranging from 6.60 kbps to 23.85 kbps. For this study, two bit-rates from the mid-range were selected for the tests: 12.65 kbps and 14.25 kbps. These will be referred to as AMR_12 and AMR_14. The AMR-WB is designed for solely speech coding and has only two modes, speech and silence, where the decision is based on a voice activity detection (VAD) [8]. The algorithmic delay of the coder is 25.9375 ms.

The decoder contains also packet loss concealment (PLC) functions for both speech and silence frames that have been lost or received erroneously. The erroneous speech frames are concealed with either extrapolation or repetition of the previous, correctly received speech frames [9]. For instance, the gains of the long-term predictor (LTP) and the fixed codebook are computed as a median over the last 5 frames and the parameters describing the vocal tract, the immittance spectral frequencies (ISFs), are estimated as the past ISFs that have been shifted towards their partly adaptive mean. As more consecutive frames are lost, the parameters estimated for the substituted frames are gradually muted such that after 6 or more consecutive frames have been lost the output is almost completely muted. For instance, the attenuation factors for the LTP gain reduce from 0.98 for the first bad frame to 0.23 for the fourth consecutive bad frame.

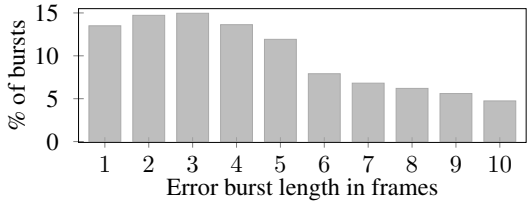

Fig. 1. The distribution of the error burst lengths in frames with average packet loss rate of 15%.

2.2. G.718

The G.718 [7] is a coding standard designed to handle both speech and generic audio. It has a layered coding structure where the two inner layers are based on the ACELP technique and the higher layers utilize MDCT to encode the residual of the lower layers. The core layer has several encoding modes for different types of frames, such as voiced, unvoiced and transient. According to the five-tier layered structure, the bit-rates are 8, 12, 16, 24 and 32 kbps where for this study only the rates 12 and 16 kbps were selected. These will be referred to as G718_12 and G718_16, respectively. The algorithmic delay of G.718 for wideband signals is 42.875 ms which is considerably higher than the delay of AMR-WB and EVS.

In case of frame erasures, the general solution is to let the speech parameters gradually approach the background noise parameters. In concealment, the parameters are estimated from past and future frames using the classification of the erased frame. This classification is a part of the side information which is transmitted in layer 3. The exact information varies according to the frame classification but can contain for instance parameters describing the glottal pulse position or the spectral envelope in the previous frame which can be used in the reconstruction. Additionally, in the default version a 10-ms delay, which corresponds to half of the frame, allows a smooth transformation from the first half of the reconstructed frame to the correctly received following frame.

2.3. EVS

The EVS is a multi-rate codec optimized for both speech and generic audio [1] with constant bit-rates from 7.2 kbps to 128 kbps and 5.9 kbps with source controlled varying bit-rate operation. For this study, 13.2 bkps bit-rate, referred to as EVS_13, was selected. The codec has three encoding strategies, ACELP, MDCT, and comfort noise generation (CNG), which are employed based on analysis of the input frame [10]. The overall algorithmic delay of the codec is 32 ms.

The functionality of the packet loss concealment in EVS depends on the mode but generally the coder parameters slowly approach the background noise parameters [11]. Parameters that have been extrapolated from correctly received frames are attenuated according to the amount of consecutive bad frames and the frame classification. For instance, the

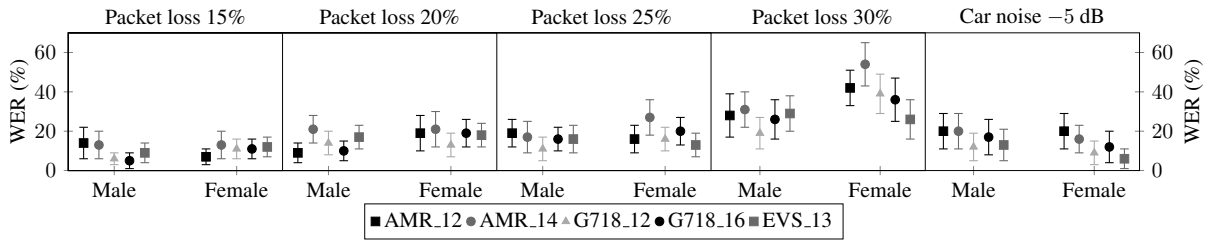

Fig. 2. The mean word-error rates (WERs) and their 95% confidence intervals for the codecs in each of the test conditions. The codecs under evaluation were AMR-WB with 12.65 kbps (AMR_12) and with 14.25 kbps (AMR_14) bit-rates, G.718 with 12 kbps (G718_12) and 16 kbps (G718_16) bit-rates, and EVS with 13.2 kbps bit-rate (EVS_13).

attenuation factor for a voiced onset after three or more lost frames is 0.4. Some additional information, e.g., parameters helping to estimate the pitch lag or the excitation signal, can be transmitted to the decoder depending on the bit-rate used. At 13.2 kbps, only the frame class is sent as side information.

3. SUBJECTIVE EVALUATION

The subjective intelligibility evaluation of the codecs consisted of a word-error rate (WER) test in different packet loss and background noise conditions. The background noise refers here to a near-end noise condition which means that the degrading environmental noise is on the listener's side of the communication channel. Thus, the encoding and decoding are not affected by the noise as they would be in the far-end noise condition where the additive noise is introduced on the sending side of the communication channel. The near-end noise scenario was selected over the far-end noise scenario because far-end noise conditions are commonly tested as a part of the quality evaluations of the coding standards.

The test conditions were selected such that in addition to the degradation in quality, the intelligibility would also be negatively affected. In the case of packet loss, this means quite high loss rates in comparison to the ones used for instance in the evaluation of the EVS codec. However, in [12] and [13], bursty packet loss with maximum loss rates from 35% to 50% were used for the evaluation of voice over IP (VoIP) services. Based on the loss rates used in these studies and on informal listening, bursty packet loss with loss rates 15%, 20%, 25% and 30% was selected for the subjective evaluation. Although the packet loss concealment algorithms of the coding standards converge either to an estimate of the background noise or silence after many consecutive lost frames (i.e., a long burst), the bursty loss was found more suitable for intelligibility testing. Based on the distribution of the burst lengths, majority of the bursts are short, less than 4 frames long, and therefore, the differences in the performance of the packet loss algorithms should have an impact. An example of the burst length distribution with 15% loss rate is shown in Fig. 1. For the noisy condition, near-end car noise with -5 dB signal-to-noise ratio (SNR) was chosen.

The speech data consisted of high-quality recordings of the Berlin and Marburg sentence lists [14, 15]. Both of the sentence lists contain short, meaningful sentences of 3-7 words in German. The recordings of the Marburg sentences were obtained from the NTT-AT super wideband stereo speech database [16] whereas the Berlin sentences had been recorded previously in an anechoic room with several native speakers. For the subjective evaluation, the recordings from one male and one female speaker were selected from both databases resulting in altogether four speakers for the evaluation set.

The original recordings are sampled with 48-kHz rate and contain the full speech spectrum. The preprocessing stages were done according to the guidelines provided for the EVS qualification phase [17]. In short, the original samples were filtered at 48-kHz rate with the HP50 filter, which is a high-pass filter simulating mobile device input characteristics [18]. Then the samples were downsampled to 16 kHz and level adjusted to -26 dBov with SV56 [19]. After this, the samples were encoded with one of the codecs under evaluation and the error insertion device (EID) was used to corrupt the bit stream with the desired error pattern in the packet loss conditions. Several error patterns were generated for each packet loss condition and a random pattern was selected each time. In the noisy condition, no errors were introduced at this stage. Finally, the obtained bit stream was decoded with the correct codec and the background noise was added to the signal in the noisy condition. The speech samples were then upsampled to 48 kHz which was the presentation rate in the test.

The subjective evaluation was completed by 10 listeners between the ages 22 and 31 who were all native German speakers with normal-hearing. The test was done in a listening room conforming to the ITU R BS.1116-1 [20] with Sennheiser HD 650 headphones. In the test, each test sample was played once through the headphones and the listeners were then asked to write what they had heard. The test was divided into five sections each of which was preceded by a short training where the listener could become familiar with the type of degradation. Each section contained 40 test samples, i.e., altogether 200 samples were used for each listener. To avoid learning effects, all of the sentences in these 200

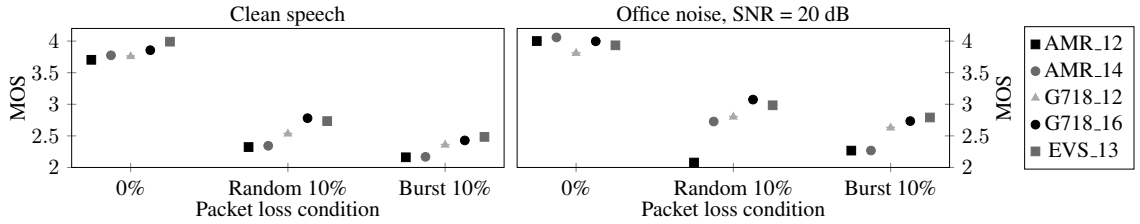

Fig. 3. The mean opinion scores (MOS) from the objective evaluation with POLQA for different packet loss conditions for clean speech and 20 dB far-end office noise. The packet loss conditions are 0% (no packet loss), 10% with random distribution (mostly single frames lost) and 10% with burst (groups of frames lost). The codecs under evaluation were AMR-WB with 12.65 kbps (AMR_12) and with 14.25 kbps (AMR_14) bit-rates, G.718 with 12 kbps (G718_12) and 16 kbps (G718_16) bit-rates, and EVS with 13.2 kbps bit-rate (EVS_13).

test samples were different. During the first training section, the listeners were asked to adjust the playback volume to a comfortable listening level after which the volume setting was kept constant for the remainder of the test. Overall, one test session lasted approximately one hour and the listeners were encouraged to take short breaks between the test sections.

3.1. Results

Before the data was analyzed, all obvious spelling errors were corrected. Additionally, for some words, e.g., given names, multiple spellings were accepted. The WER score was obtained by computing the percentage of erroneous words in each sentence.

The results of the WER test were analyzed with a four-way analysis of variance (ANOVA) test with the method (AMR_12, AMR_14, G718_12, G718_16, EVS_13), speaker gender (male, female) and condition (packet loss 15%, 20%, 25%, 30% and noisy -5 dB) modelled as fixed factors and the listener modelled as a random factor. According to the ANOVA results, the WER score was significantly affected by the method [$F(4, 156) = 7.67, p < 0.001$], the condition [$F(4, 156) = 53.36, p < 0.001$] and the speaker gender [$F(1, 39) = 10.51, p < 0.01$] as well as the interaction between the condition and speaker gender [$F(4, 156) = 6.16, p < 0.001$].

Post-hoc tests were conducted using the Tukey method with 95% confidence level. In the following, only the relevant statistically significant results will be reported. EVS_13, G718_12, and G718_16 were overall significantly better than both AMR_12 and AMR_14. From the test conditions, the 15% packet loss was the easiest and the 30% packet loss the most difficult, as expected. However, the 20% and 25% packet loss as well as the noise condition were equally challenging in terms of intelligibility. These observations are visualized in Fig. 2 where the means and the 95% confidence intervals of the word-error rates of the codecs in each of the test conditions are shown.

4. OBJECTIVE EVALUATION

In addition to the subjective tests, the codecs were evaluated objectively using POLQA [21] in different packet loss conditions. Because the POLQA is a quality measure, the packet loss rates used for the subjective evaluation are unnecessarily high. On the other hand, using lower packet loss rates for the subjective intelligibility test would have caused ceiling effects. Therefore, the objective evaluation was used to bridge the gap between the specification tests and the tests conducted in this study and the test conditions were selected accordingly. Both random and bursty packet loss with 10% loss rate were used in clean and noisy condition. In this case, the noisy condition was far-end office noise with 20 dB SNR. The results of the POLQA are shown in Fig. 3. In the bursty packet loss, EVS received higher scores than G.718 and AMR-WB was last. For the random loss case, G718_16 was slightly better than EVS_13 followed by G718_12, AMR_14 and AMR_12.

5. CONCLUSION

Three speech coding standards of different ages, AMR-WB, EVS and G.718, were evaluated in terms of intelligibility in difficult packet loss and near-end noise conditions. Additionally, an objective measure was used to quantify their performance. The results of the subjective test show that the newer standards, EVS and G.718, were better in terms of intelligibility than AMR-WB. The objective evaluation also supports this conclusion. The result is not surprising because although AMR-WB is being currently deployed, it is over 10 years old. The difference between EVS and G.718 is less clear since there are no significant differences in the subjective test. The error patterns cause large variance in the results which reduces the statistical power of the test. In the objective evaluation, EVS had higher scores than G.718 in case of bursty loss but in the random packet loss condition, G.718 with 16 kbps was only slightly better than EVS at 13.2 kbps. Additionally, G.718 has much more delay which is used to improve error concealment. Especially in the case of single lost frames, the delay allows half of the frame to be recovered completely.

6. REFERENCES

[1] 3rd Generation Partnership Project, Valbonne, France, *Specification TS 26.441: EVS codec general overview*, 2014, version 2.0.0.

[2] 3rd Generation Partnership Project - Meeting S4-72, Valencia, Spain, *EVS Permanent Document EVS-8a:Test plans for qualification phase including host lab specification*, 2013, version 1.3, available: http://www.3gpp.org/DynaReport/TDocExMtg–S4-72–30006.htm, accessed: 26.09.2014.

[3] S. Möller and A. Raake, "Telephone speech quality prediction: towards network planning and monitoring models for modern network scenarios," *Speech Commun.*, vol. 38, no. 1–2, pp. 47–75, 2002.

[4] M. Wältermann, A. Raake, and S. Möller, "Quality dimensions of narrowband and wideband speech transmission," *Acta Acustica united with Acustica*, vol. 96, pp. 1090–1103, 2010.

[5] P.C. Loizou, "Speech quality assessment," in *Multimedia Analysis, Processing and Communications*, pp. 623–654. Springer Verlag, 2011.

[6] 3rd Generation Partnership Project, Valbonne, France, *Specification TS 26.171: Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description*, 2012, version 11.0.0.

[7] Int. Telecommun. Union, Geneva, Switzerland, *Recommendation ITU-T G.718: Frame error robust narrowband and sideband embedded variable bit-rate coding of speech and audio from 8–32 kbit/s*, June 2008.

[8] 3rd Generation Partnership Project, Valbonne, France, *Specification TS 26.190: Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions*, 2012, version 11.0.0.

[9] 3rd Generation Partnership Project, Valbonne, France, *Specification TS 26.191: Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Error concealment of erroneous or lost frames*, 2012, version 11.0.0.

[10] 3rd Generation Partnership Project, Valbonne, France, *Specification TS 26.445:EVS codec detailed algorithmic description*, 2014, version 1.0.0.

[11] 3rd Generation Partnership Project, Valbonne, France, *Specification TS 26.447:EVS codec error concealment of lost packets*, 2014, version 1.0.0.

[12] A.M. Gomez, J.L. Carmona, A.M. Peinado, V. Sánchez, and J.A. Gonzalez, "Intelligibility evaluation of Ramsey-derived interleavers for Internet voice streaming with the iLBC codec," in *Proc. Interspeech*, 2008, pp. 707–710.

[13] A. Janicki and B. Księżak, "Packet loss concealment algorithm for VoIP transmission in unreliable networks," in *Proc. 2008 Conf. New Trends in Multimedia and Network Information Systems*, 2008, pp. 23–33.

[14] W. Niemeyer and G. Beckmann, "A speech-audiometric test (in German)," *Arch. Ohren-, Nasen-, Kopfheilk*, vol. 180, pp. 742–749, 1962.

[15] J. Sotscheck, "Sentences for speech-quality measurements and their phonological adaptation to German language (in German)," in *Proc. DAGA*, 1984, pp. 873–876.

[16] NTT-AT, "Super wideband stereo speech database," available: http://www.ntt-at.com/product/widebandspeech, accessed 29.09.2014.

[17] 3rd Generation Partnership Project - Meeting S4-72, Valencia, Spain, *EVS Permanent Document EVS-7a: Processing functions for qualification phase*, 2013, version 1.3, available: http://www.3gpp.org/DynaReport/TDocExMtg–S4-72–30006.htm, accessed: 26.09.2014.

[18] Int. Telecommun. Union, Geneva, Switzerland, *Recommendation ITU-T G.191: Software tools for speech and audio coding standardization*, September 2005.

[19] Int. Telecommun. Union, Geneva, Switzerland, *Recommendation ITU-T P.56: Objective measurement of active speech level*, March 1993.

[20] Int. Telecommun. Union, Geneva, Switzerland, *Recommendation ITU-R BS.1116-2: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, June 2014.

[21] Int. Telecommun. Union, Geneva, Switzerland, *Recommendation ITU-T P.863: Perceptual objective listening quality assessment*, January 2011.