Quantifying Quality Degradation of the EVS Super-Wideband Speech Codec

Gabriel Mittag[1], Sebastian Möller[1,2], Vincent Barriac[3], Stéphane Ragot[3]
[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
[3]Orange Labs, Lannion, France
{gabriel.mittag; sebastian.moeller}@tu-berlin.de, {vincent.barriac; stephane.ragot}@orange.com

***Abstract*—Voice transmission networks are commonly planned with the help of computational quality models, which give an estimate of the expected quality that a user will experience. The most popular of these tools is the E-model. When certain parameters are known, such as the applied codec and its bitrate, the model is able to predict the perceived quality of a communication system. Up to now, the E-model is only available for narrowband telephony (300-3400 Hz) and limited also for wideband telephony (100-7000 Hz). With the extension of voice networks to super-wideband telephony (50-14000 Hz), and the introduction of the super-wideband codec EVS to mobile networks and state of the art smartphones, an update of the E-model has become necessary. To this end, we firstly examined the quality improvement of super-wideband over wideband with results from mixed-band listening-only tests, where we found that the quality is improved by 15%. Then, we calculated impairment factors for the EVS codec and analyzed its robustness towards packet loss, by using auditory and instrumental methods.**

***Index Terms*—E-model, EVS, speech quality, SWB**

I. Introduction

Traditionally, in telecommunication systems, speech was band-limited to *narrowband* (NB) with a bandwidth of 300-3400 Hz. This bandwidth limitation results in the typical muffled audio known from telephone conversations. In order to improve the speech quality, *wideband* (WB) transmission was introduced, which offers a bandwidth of 100-7000 Hz and is also marketed as “HD Voice”. Today, in fixed-line and mobile communication systems either NB or increasingly also WB codecs are used. However, the bandwidth of WB is still limited, as the human auditory system is able to perceive frequencies up to 20 kHz. Furthermore, the currently used codecs in mobile telephony perform poorly on non-speech signals (e.g. music). To overcome these problems, the 3GPP Codec for *Enhanced Voice Services* (EVS) [1] was introduced. It is specifically designed for packet-switched mobile voice networks, such as *Voice over LTE* (VoLTE) and general IP telephony. It supports NB, WB, SWB (super-wideband, 50-14000 Hz), and FB (full band, up to 20 kHz) telephony and minimizes the degradation caused by packet loss and delay jitter.[1] Recently, mobile network operators started to support SWB communication through VoLTE and WiFi with the EVS codec, where it is for example promoted as “HD Voice Plus” or “Crystal Clear”. These developments make it necessary to update current network planning models from WB to SWB and to quantify the degradation caused by the EVS codec.

The recommended transmission planning tool by the *International Telecommunication Union* (ITU-T) is the E-Model [2]. It is used to plan future voice networks by taking network parameters that describe specific impairments (e.g. delay, low bitrate codecs, or packet loss) and compute an overall rating of the expected conversational quality as the *transmission rating factor* R. This helps transmission planners to ensure that the users will be satisfied with the perceived quality. The model is based on the impairment factor principle, which assumes that different types of degradations are additive in terms of the perceptual impairment they cause. The rating can be calculated with following basic formula:

$$R = R_0 - I_{\mathrm{s}} - I_{\mathrm{d}} - I_{\mathrm{e,eff}} + A. \tag{1}$$

Essentially, the model assumes a maximum rating R_0, from which impairment factors are subtracted to calculate an overall quality R. R_0 describes the basic signal-to-noise ratio (e.g. caused by circuit noise). If there are no noise sources in the transmission, we can assume $R_0 = R_{\mathrm{max}}$, which is the maximum value that can be achieved on the R-scale. I_{s} represents simultaneous impairments caused by non-optimum loudness or signal-correlated noise, I_{d} stands for impairment caused by delay, and $I_{\mathrm{e,eff}}$ is the effective equipment impairment factor, which represents impairment caused by speech coding and packet loss. A can be used to compensate impairment through other quality advantage factors. In this paper, we only study impairment caused by speech coding, therefore we can write (1) as:

$$R = R_{\mathrm{max}} - I_{\mathrm{e,eff}}, \tag{2}$$

where we simply subtract the impairment caused by speech coding $I_{\mathrm{e,eff}}$ from the highest quality rating possible R_{max}. The R-value is linked to the *mean opinion score* (MOS), which is obtained from auditory experiments, by an S-shaped curve (Fig 1). Through this relationship, new codec impairment factors can be derived by conducting auditory test and transform the MOS obtained from the test participants to the R-scale [3].

[1]Although the terminal bandwidth of SWB is usually understood to be limited to 50-14000 Hz, the actual maximum output bandwidth of the EVS codec itself in SWB mode is 14.4 kHz at 9.6 and 13.2 kbit/s and 16 kHz for higher rates.

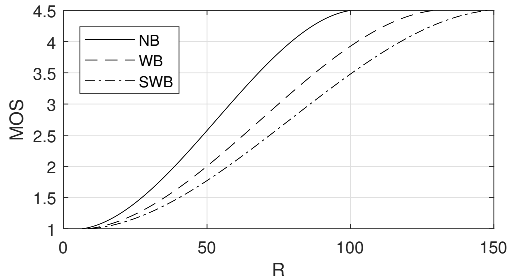

Fig. 1. Transformation rule of the E-model between R-scale and MOS

Because auditory tests are costly and time consuming, another way of deriving new impairment factors has been introduced in [4]. Here, instrumental, signal-based quality models are used to predict MOS scores that can be transformed to the R-scale. The current ITU-T recommendation for perceptual listening quality assessment of SWB speech is POLQA or ITU-T Rec. P.863 [5]. It predicts the perceived quality of a speech signal by comparing the degraded signal with its unimpaired reference. This method is especially useful for determining the robustness of a codec towards packet loss, since many different packet loss rates have to be considered.

Initially, the E-model was only designed for NB telecommunication systems operated with handset telephones and was later extended in [3], [6] to the Wideband E-Model [7]. The maximum value in the WB E-model is $R_{\max,\mathrm{WB}} = 129$. Consequently, for the WB case, the transmission rating scale ranges from $R = 0$ (lowest possible quality) to $R = 129$ (optimum quality). In order to develop a new SWB E-model, the first step is to determine a new SWB maximum R-value $R_{\max,\mathrm{SWB}}$. Then, impairment factors for codecs and their robustness towards packet loss can be defined.

In the following, we will first introduce the methods used for deriving the maximum SWB R-scale value and equipment impairment factors of the EVS codecs. In Section III, the databases and test signals, which were used for the calculations, are presented. In Section, IV the conducted experiments are described and the results are given. Finally, the last section summarizes the main results and presents further steps that have to be undertaken to obtain a full SWB version of the E-model.

II. Method

A. Maximum SWB R-scale value

In auditory speech quality experiments [8], naïve test participants rate speech samples on a five-point *absolute category rating* (ACR) scale. The average rating over all test participants then yield the MOS score. These MOS scores obtained from auditory experiments are also denoted as subjective MOS. In case the "best" speech condition in the experiment is an unimpaired speech file with WB bandwidth, the participants will usually rate this condition with a very high score. However, if the test contains SWB speech conditions as well, the quality of the WB speech conditions will be perceived as inferior.

To estimate the maximum R-scale value in the SWB case, we need to find out how users perceive the quality of a WB speech condition when they are also exposed to SWB speech conditions. To do this, we can use subjective MOS scores from mixed-band experiments and calculate R ratings of the WB conditions, with the S-shaped relation between MOS and R-scale (Fig 1). As a final step to estimate the SWB quality improvement on the R-scale, the WB R-values obtained from mixed-band experiments can be compared to the given R-values from the WB E-model. The method for deriving the SWB maximum R-scale value $R_{\max,\mathrm{SWB}}$ can be described by the following steps:

1) Use conditions from a mixed-band experiment for which the equipment impairment factor $I_{\mathrm{e,WB}}$ are given in Amendment 1 of ITU-T Rec G.113 [9] and calculate their R-value in the WB context R_{WB} with following formula [7]:

$$R_{\mathrm{WB}} = R_{\max,\mathrm{WB}} - I_{\mathrm{e,eff,WB}} = 129 - I_{\mathrm{e,eff,WB}}. \quad (3)$$

In case the conditions contain packet loss, the impairment factor $I_{\mathrm{e,eff,WB}}$ can be calculated as follows [7]:

$$I_{\mathrm{e,WB,eff}} = I_{\mathrm{e,WB}} + (95 - I_{\mathrm{e,WB}}) \cdot \frac{P_{\mathrm{pl}}}{P_{\mathrm{pl}} + B_{\mathrm{pl}}}. \quad (4)$$

2) Then, the subjective MOS values of all conditions need to be transformed to the R-scale. In the MOS transformation rule of the E-model, the MOS range is limited to [1;4.5]. Because of this, the actual range of MOS values appearing in the respective subjective test, first has to be linearly transformed to a maximum of 4.5 [10]:

$$MOS_{\mathrm{norm},i} = \frac{MOS_i - 1}{MOS_{\max} - 1} \cdot 3.5 + 1, \quad (5)$$

where $MOS_{\max}$ is the maximum MOS obtained in one experiment (usually the reference SWB condition).

3) The MOS to R transformation according to the S-shaped curve of Fig. 1 can now be applied (formula in [2]). With this calculation the MOS values are transformed to the non-extended NB R-scale (range[0;100]):

4) Next, the linear transform to the WB R-scale given in the WB E-Model [7] is applied:

$$R_{\mathrm{WB/SWB}} = 1.29 \cdot Rx. \quad (6)$$

5) As a result, we have pairs of R-values for each condition: In the WB context R_{WB} (derived from the equipment impairment factors) and in the mixed-band context $R_{\mathrm{WB/SWB}}$ (derived from the subjective MOS). As a consequence of the mixed-band test design, a clean WB signal will have a lower $R_{\mathrm{WB/SWB}}$ value, as its quality is inferior compared to a SWB signal. In order to make the SWB E-model compatible with the WB E-Model, the R-scale now has to be decompressed, such that a clean WB speech signal again receives a value of 129, and the SWB clean reference signal consequently a higher value, which is our maximum SWB R-scale value $R_{\max,\mathrm{SWB}}$ to be determined.

6) To decompress the mixed-band R-values $R_{\mathrm{WB/SWB}}$, we use a simple linear regression to fit the compressed values $R_{\mathrm{WB/SWB}}$ onto the uncompressed values R_{WB}. This approach has been used before in the WB extension of the E-model [3]. The linear regression was calculated without intercept term because both R-scales have a lower quality boundary of zero:

$$R_{\mathrm{WB}} = \beta \cdot R_{\mathrm{WB/SWB}}. \tag{7}$$

7) The maximum SWB R-value $R_{\mathrm{max,SWB}}$ can then be calculated by applying the linear regression model to the $R_{\mathrm{WB/SWB}}$ value of the clean SWB reference signal:

$$R_{\mathrm{max,SWB}} = \beta \cdot R_{\mathrm{WB/SWB}}(\mathrm{SWB\,clean}). \tag{8}$$

B. Equipment impairment factors derivation

After $R_{\mathrm{max,SWB}}$ is determined, equipment impairment factors $I_{\mathrm{e,SWB}}$ can be calculated on the new SWB R-scale. They can be derived by either auditory methods, using listening-only test, according to ITU-T Rec. P.833 for NB [11] and P.833.1 for the WB E-model [10], or by instrumental methods, according to ITU-T Rec. P.834 for NB [12] and P.834.1 for the WB E-model [13]. Since POLQA has already been validated in the ITU-T for the new EVS codec, we used both methods for deriving impairment factors of EVS.[2]

1) Auditory derivation: In order to determine impairment factors for EVS we need to calculate their R-scale value from subjective MOS values. The calculation was conducted with following steps:

1) The MOS values from subjective experiments need to be normalized according to (5).
2) Transform MOS values to the NB R-scale and then extend them to the SWB R-scale as follows:

$$R_{\mathrm{SWB}} = R_{\mathrm{max,SWB}} \cdot Rx. \tag{9}$$

3) From these R-values, raw impairment factors $I_{\mathrm{e,SWB,sub}}$ can be calculated by subtracting the $R_{\mathrm{SWB}}(\mathrm{condition})$ value of the condition under consideration from the clean SWB reference R-value $R_{\mathrm{SWB}}(\mathrm{clean\,SWB}) = R_{\mathrm{max,SWB}}$:

$$I_{\mathrm{e,SWB,sub}} = R_{\mathrm{max,SWB}} - R_{\mathrm{SWB}} \tag{10}$$

4) Individual test setups of auditory experiments often lead to a bias in the ratings. In order to remove experiment specific offsets from the raw impairment factors, a normalization is applied. To this end, reference conditions for which the impairment factors are known are used to perform a linear mapping. So far, for diotic listening experiments, only impairment factors of WB codecs are available [9]. Additionally, the impairment factors of clean NB, WB and SWB signals are known. After mapping the observed impairment factors $I_{\mathrm{e,SWB,sub}}$ to the expected impairment factors from [9] $I_{\mathrm{e,SWB,exp}}$ we obtain the normalized impairment factors $I_{\mathrm{e,SWB}}$.

[2]See also [14], [15], where it was shown that POLQA / P.863 SWB (rev 2.4) predicts speech signals coded with EVS in SWB mode reliably, even without mapping.

2) Instrumental derivation: For the instrumental derivation of $I_{\mathrm{e,SWB}}$ we applied the same steps as for the auditory calculations; however, replacing the subjective MOS scores by POLQA estimated MOS scores. Because the maximum POLQA score in SWB mode is $MOS = 4.75$ the POLQA results were firstly normalized with $MOS_{\mathrm{max}} = 4.75$, using (5). A normalization of the impairment factors was not conducted, as the predicted MOS scores are not exposed to test specific setups. Influence by speakers and sentences should be ruled out by using a variety of 32 different speech files (see Section III).

C. EVS packet loss robustness derivation

In modern voice networks, one of the main impairments of transmitted speech are lost packets, which lead to interruptions in the speech signal and/or unnatural voices, caused by packet loss concealment algorithms. To take this degradation into account the NB and WB E-model calculate an effective equipment impairment factor $I_{\mathrm{e,eff}}$ based on the impairment factor of a codec and a corresponding packet loss robustness factor B_{pl} according to (4).

Only instrumental methods were used for the calculation of the codec-specific robustness factor B_{pl}, since only one of the available databases contains packet loss conditions for EVS. We applied several packet loss ratios for each bitrate of the EVS codec and then used POLQA to estimate MOS values. As in the previous section (II-B), we firstly transformed the MOS values to the R-scale, resulting in $I_{\mathrm{e,SWB,eff}}$ values. B_{pl} can then be determined by fitting the $I_{\mathrm{e,SWB,eff}}$ values of each EVS bitrate mode with a curve in a least-squares sense according to (4), using B_{pl} as the independent parameter [3].

III. Databases

A. Auditory derivation

a) Orange Databases: These two databases were kindly provided by Orange and contain speech samples in French under clean and packet loss conditions. They were rated by 24 test participants and the speech files were presented diotically with the frequency response limited to SWB. In total, there are 20 conditions (10% WB, 90% SWB) for O1, and 54 conditions (39% WB, 61% SWB) for O2.

b) Rhode & Schwarz Database: The database was kindly provided by Rhode & Schwarz [15] and contains speech samples in German under clean and packet loss conditions. The speech files were presented diotically and rated by 24 test participants. In total, there are 52 conditions (23% NB, 37% WB, 19% SWB, 21% FB) with each 4 different sentences. The direct FB condition received approximately the same rating as the direct SWB condition.

c) Qualcomm Database: The database contains speech samples in American English and the results are published in the P.863 Implementer's Guide for EVS [14]. There are 60 conditions with varying bit rates and other internal test conditions. The internal conditions are left out in the implementer guide, so that the results of 49 conditions are available (17%

TABLE I
AVAILABLE IMPAIRMENT FACTORS OF ALL DATABASES

Codec	**Ie,eff,WB**	**O1**	**O2**	**RS**	**QC**
Direct SWB		x	x	x	x
Direct WB	0.0				x
Direct NB	35.8				x
AMRWB 06.60	56.0			x	
AMRWB 08.85	41.0				x
AMRWB 12.65	20.0	x	x	x	x
AMRWB 15.85	17.0				x
AMRWB 23.85	10.0	x	x	x	x
AMRWB 23.85 / PL 3.3%	44.2		x		
AMRWB 23.85 / PL 5.0%	52.9		x		
AMRWB 23.85 / PL 6.2%	57.5		x		

TABLE II
ESTIMATED MAXIMUM SWB R-SCALE VALUES

Database	O2	RS	QC	Average
Rmax,SWB	154.5	151.0	145.1	150.2

NB, 38% WB, 45% SWB). There were 24 different sentences for each condition that were rated by 32 test participants.

All databases comply with ITU-T Rec. P.800 [8] and are rated on 5-point ACR MOS scales. The conditions with known impairment factors for diotically conducted test are shown for each database in Table I. These conditions can be used as reference points.

B. Instrumental derivation

The 32 SWB test signals in 8 different languages from Annex C of ITU-T Rec. P.501 were used for the instrumental methods. These speech files are prepared for use with ITU-T P.800 conformant applications and speech quality prediction models [16] (e.g. POLQA).

IV. EXPERIMENTS AND RESULTS

A. Maximum SWB R-scale value

The method from Section II-A was applied to databases O2, RS and QC. Because database O1 mostly contains SWB conditions and only two reference points, it was left out for the calculation of $R_{\max,\mathrm{SWB}}$. The results in Fig. 2 show that there are relatively few data points with low R-values, especially for database QC. However, by setting the regression line through the origin, we can partly compensate for the missing low R-values. The resulting $R_{\max,\mathrm{SWB}}$ are presented in Table II, the average value over all three databases is 150.2. This means a SWB quality improvement of 15% compared to WB and 59% compared to NB on the R-scale. These results were brought up at the last ITU-T SG12 meeting, where it was decided together with results contributed by NTT, to define the new maximum SWB R-scale value as $R_{\max,\mathrm{SWB}} = 148$. Consequently, we use this value to calculate the equipment impairment factors of the EVS codec.

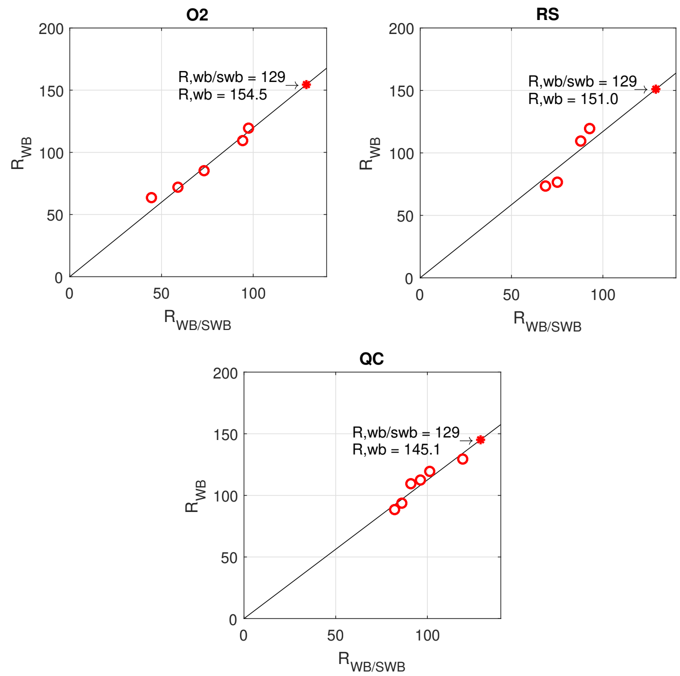

Fig. 2. Comparison of R-values derived in a WB and in a mixed WB/SWB context for 3 different databases.

TABLE III
EQUIPMENT IMPAIRMENT FACTORS $I_{\mathrm{e,SWB}}$ ESTIMATED BY AUDITORY AND INSTRUMENTAL METHODS FOR EVS IN SWB MODE

Codec	**O1**	**O2**	**RS**	**QC**	**Average**	**POLQA**
EVS 9.6	25.9	21.0	36.6	22.5	26.5	34.3
EVS 13.2	18.8	14.9	26.5	17.7	19.5	24.8
EVS 16.4	15.9	10.0	14.7	8.3	12.2	16.3
EVS 24.4	5.6	0.0		8.3	4.6	8.7
EVS 32.0		11.6		11.8	11.7	17.2
EVS 48.0				13.3	13.3	2.2
EVS 64.0						7.1
EVS 96.0						0.0
EVS 128.0						0.0

B. Equipment impairment factors

In total, EVS supports 10 different SWB bitrate modes.[3] All six EVS conditions that were included in the four databases were used to calculate equipment impairment factors for the auditory derivation. But not all bitrates were available in every database (see Table III).

Each of the 32 SWB test signals from P.501 was coded with the 9 different EVS SWB bitrate modes for the instrumental derivation. Then POLQA was computed for each file, therefore we receive 32 POLQA MOS scores per EVS bitrate. Since the degradation of the speech file caused by the codec varies depending on the speaker, sentence, and the language, we will obtain different MOS predictions for each file. However, the 95% confidence interval of the MOS predictions of each EVS codec condition was below 0.01. The results of both

[3]For the bitrate 13.2 kbit/s a *channel aware* mode is available that we don't consider in this paper, since random/bursty profiles generated by G.191 are not directly applicable to measure its performance in terms of speech quality.

methods are shown in Table III for all 9 considered EVS-SWB bitrate modes. First, the $I_{\mathrm{e,SWB}}$ results of each database are presented, then the average over all databases, and in the last column the results from the instrumental method with POLQA. There is some deviation in the results between the different databases. However, the average values are quite close to the ones predicted by POLQA. The EVS bitrate 24.4 kbit/s seems to perform better than 32 kbit/s, in that it produces higher quality. This counter-intuitive result is in line with the results of the EVS performance evaluation in the characterization phase [17]. The bitrate 48 kbit/s performs better than 32 kbit/s in the instrumental experiments, but worse in the auditory results (48 kbit/s receives $I_{\mathrm{e,SWB}} = 13.3$ and 32 kbit/s, $I_{\mathrm{e,SWB}} = 11.7$). That being said, there is only one database available for 48 kbit/s and the difference between both bitrates in MOS is only 0.028.

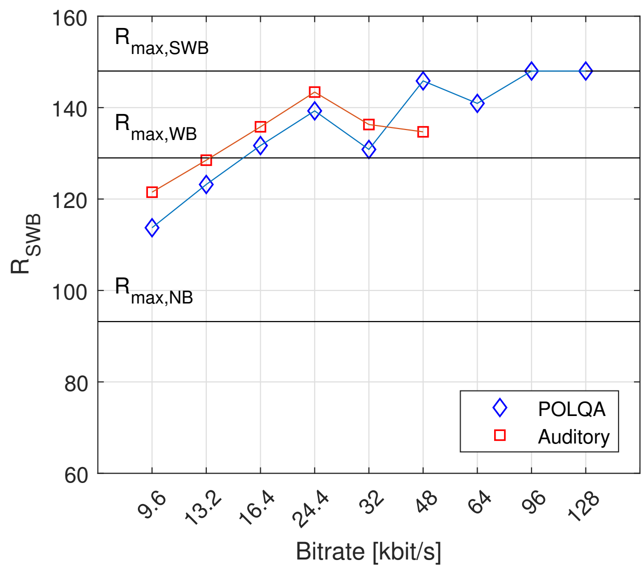

Fig. 3. Quality rating of EVS codec in SWB mode over different bitrates.

Fig. 3 presents the quality rating of EVS for different bitrates, derived from POLQA predictions and from the averaged results of auditory experiments. It can be seen that the trend of the curves of both methods are very similar, with POLQA giving slightly lower quality estimates. Here, it should be noted that the POLQA ratings were predicted with different signals and the average of the auditory results were calculated over different numbers of databases. From the results we can conclude that the quality rating of EVS in SWB mode with bitrates of 16.4 kbit/s or higher are superior to a clean reference WB speech signal. The lowest bitrate 9.6 kbit/s still outperforms a clean NB speech signal and the higher bitrates, from 48 kbit/s on, undergo almost no quality degradation.

C. EVS packet loss robustness derivation

To derive the robustness parameter of the EVS codec we applied random and bursty packet loss with 7 different error rates on the 32 test signals from ITU-T Rec. P.501 (the 7 error rates are: 1%, 3%, 5%, 10%, 15%, 20%, and 30%). The packet loss was applied with the ITU-T STL Toolbox [18], which uses a Bellcore model [19] implementation for generating bursty error patterns. The random error patterns were generated with the same tool, by applying the Gilbert model using $\gamma = 0$. The two different pattern modes were then applied to each of the 9 EVS bitrate modes and 7 error rates. This results in 126 conditions for each speech file. Next, we applied POLQA to receive MOS predictions for each of the different conditions. The quality impact of packet loss depends greatly on the location of the lost packet; for example, if the lost packet occurs during a silent segment of the speech signal, it may not even be noted by the user. Because of that, we repeated the calculations for each condition 20 times with newly and randomly generated error patterns. Thus, in total, our experiments yield 80,640 different MOS scores. Again, the 95% confidence interval of the MOS predictions of each condition was below 0.01.

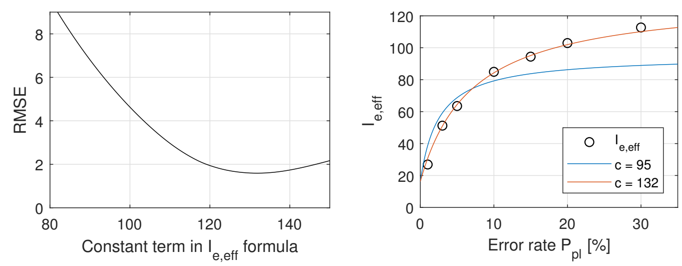

Fig. 4. LEFT: Average RMSE between measured $I_{\mathrm{e,SWB,eff}}$ and their estimates according to (11) over different constants. RIGHT: $I_{\mathrm{e,SWB,eff}}$ derived with POLQA (circles) and estimates by the E-model using (11) (red and blue line) for EVS (16.4 kbit/s).

TABLE IV
ROBUSTNESS FACTOR B_{Pl}

Codec	**Bursty**	**Random**
EVS 9.6	9.2	14.9
EVS 13.2	8.6	14.0
EVS 16.4	7.0	11.1
EVS 24.4	6.4	9.9
EVS 32.0	7.0	11.5
EVS 48.0	6.0	10.3
EVS 64.0	5.8	9.9
EVS 96.0	5.2	8.3
EVS 128	5.2	8.2

After applying the method from Section II-C, we obtain a packet loss robustness factor B_{pl} for each EVS bitrate mode. As can be seen on the right-hand side of Fig. 4, Eq. (4) does not model the packet loss impairment very well. Especially for large error rates the model estimates (blue line) do not match the measured values (black circles). Therefore, the constant term in the formula does not seem to be suitable for the SWB E-model. Because of that, we propose to use 132 as a constant (red line) in the formula instead. In Fig. 4 (left side), which shows the average RMSE over different constant values, it can be seen that 132 yields the best fit. In the calculation of the RMSE we used random and bursty conditions but neglected

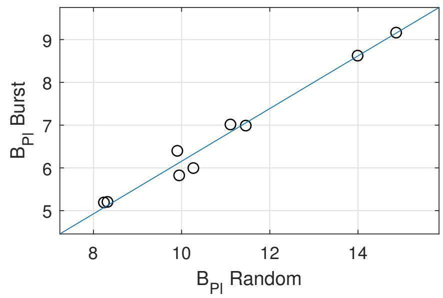

Fig. 5. Relation between robustness of the EVS codec for random and bursty packet loss of the 9 EVS SWB modes.

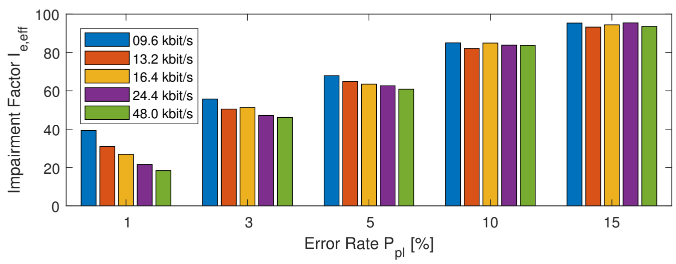

Fig. 6. Quality degradation of EVS for bursty packet loss over different error rate percentages.

the error rates 20% and 30%, as these rates result in very low MOS values and are hence not important in practice. Also, we only used the bitrates 9.6, 13.2, 16.4, and 24.4 kbit/s, as higher bit rates are not commonly implemented. Thus, the effective impairment factor can be calculated as follows:

$$I_{\mathrm{e,SWB,eff}} = I_{\mathrm{e,SWB}} + (132 - I_{\mathrm{e,SWB}}) \cdot \frac{P_{\mathrm{pl}}}{P_{\mathrm{pl}} + B_{\mathrm{pl}}}. \quad (11)$$

The resulting robustness factors are presented in Table IV. As expected, the robustness towards bursty packet loss is lower than for random loss. This is due to the fact that only a single lost packet, with a duration of 20ms, is not perceived as such a strong interruption as multiple consecutive packets. Furthermore, packet loss concealment algorithms use information from previous packets (e.g. pitch) to synthesize a signal that supposedly sounds similar to the missed packets. Therefore, the more consecutive packets are lost, the higher the chance for the algorithm to synthesize an unnatural sounding signal.

Fig. 5 presents the robustness B_{pl} to bursty packet loss over the robustness to random packet loss. It can be seen that the relation is quite linear and the EVS codec is approximately 0.6 times as robust to bursty packet loss as to random loss, when using the standard STL Toolbox implementation. Fig. 6 shows how bursty packet loss affects the perceived degradation of EVS coded speech. The quality rating for error rates higher than 5% is almost equal for different bit rates. This means that a high bitrate mode cannot handle strong packet loss better than a mode with a lower bitrate. Even for an error rate of 3% the quality rating is similar across the different bitrate modes.

V. Conclusion

In this paper, we analyzed mixed-band speech databases from different sources to determine the quality improvement of SWB over WB. We found that the quality of a clean SWB speech signal is perceived as 15% higher than the one of a clean WB signal. As a consequence of these and other results, the ITU-T defined the new SWB maximum R-scale value as $R_{\mathrm{max,SWB}} = 148$. Based on this value, we calculated impairment factors for the EVS codec by applying auditory and instrumental methods and following basically the same methods that were used for the WB E-model extension [3], [4]. We obtained consistent results throughout the different methodology that will be brought forward to the ITU-T. The work in this paper presents a good basis for a full SWB E-model, where further degradations need to be analyzed. These comprise effects of circuit noise, ambient noise, sound level, echo, and delay. Also, impairment factors of other SWB codecs, such as Opus, need to be added.

References

[1] 3GPP TS 26.441, "Codec for Enhanced Voice Services (EVS); General overview."

[2] ITU-T Rec. G.107, "The E-model: A computational model for use in transmission planning," 2015.

[3] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment factor framework for wide-band speech codecs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1969–1976, Nov 2006.

[4] S. Möller, N. Côté, V. Gautier-Turbin, N. Kitawaki, and A. Takahashi, "Instrumental estimation of E-model parameters for wideband speech codecs," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, p. 782731, Oct 2010.

[5] ITU-T Rec. P863, "Perceptual objective listening quality assessment," 2014.

[6] A. Raake, S. Möller, M. Wältermann, N. Côté, and J. P. Ramirez, "Parameter-based prediction of speech quality in listening context," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, June 2010, pp. 182–187.

[7] ITU-T Rec. G.107.1, "Wideband E-model," 2015.

[8] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 1996.

[9] ITU-T Rec. G.113 Amendment 1, "Revised appendix IV – Provisional planning values for the wideband equipment impairment factor and the wideband packet loss robustness factor," 2009.

[10] ITU-T Rec. P.833.1, "Methodology for the derivation of equipment impairment factors from subjective listening-only tests for wideband speech codecs," 2009.

[11] ITU-T Rec. P.833, "Methodology for derivation of equipment impairment factors from subjective listening-only tests," 2001.

[12] ITU-T Rec. P.834, "Methodology for the derivation of equipment impairment factors from instrumental models," 2015.

[13] ITU-T Rec. P.834.1, "Extension of the methodology for the derivation of equipment impairment factors from instrumental models for wideband speech codecs," 2015.

[14] ITU-T P.Imp863, "Implementer's guide on assessment of EVS coded speech with recommendation ITU-T P.863," 2016.

[15] ITU-T Contribution SG12-C022 (Source: R&S), "A subjective ACR LOT testing fullband speech coding and prediction by P.863," 2017.

[16] ITU-T Rec. P.501, "Test signals for use in telephonometry," 2017.

[17] 3GPP TR 26.952, "Codec for Enhanced Voice Services (EVS); Performance characterization."

[18] ITU-T Rec. G.191, "ITU-T software tool library 2009 user's manual," 2009.

[19] V. K. Varma, "Testing speech coders for usage in wireless communications systems," in *Proceedings., IEEE Workshop on Speech Coding for Telecommunications,*, 1993, pp. 93–94.