EXTENSION OF THE E-MODEL TOWARDS SUPER-WIDEBAND SPEECH TRANSMISSION

Marcel Wäldtermann, Izabela Tucker, Alexander Raake, and Sebastian Möller

Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany

ABSTRACT

In this paper, the quality gain of super-wideband (SWB) speech, transmitted in the much wider frequency range of 50-14000 Hz compared to the standard 300-3400 Hz narrowband, is quantified employing the E-model framework, a parametric tool for speech quality prediction. Based on two listening experiments, a linear extrapolation of the E-model transmission rating scale was found that leads to a maximum quality advantage of 39% relative to wideband (50-7000 Hz) transmission, and 79% relative to narrowband. Furthermore, narrowband, wideband, and super-wideband conditions can be quantified on this universal quality scale.

Equipment Impairment Factors were derived and discussed for several SWB codecs. It will further be shown that a model quantifying the quality impact of linear distortions, reflected by the Bandwidth Impairment Factor, can successfully applied to SWB conditions. The correlation between the overall impairment and the model predictions amounts to \( r = 0.977 \) for linearly distorted speech samples.

Index Terms— Modeling, Speech codecs, Linear systems, Nonlinear systems, Nonlinear distortion

1. INTRODUCTION

In traditional circuit-switched telecommunication networks such as PSTN and ISDN, speech is typically transmitted in a relatively narrow frequency range of 300-3400 Hz (narrowband, NB). It is well-known [1] and intuitively clear that an increase in transmission bandwidth leads to better quality perceived by the user. Although the G.722 coding scheme which is capable of transmitting wideband (WB) speech (50-7000 Hz) has been standardized by the ITU-T already in the late eighties, only the transition to packet-based transmission made beyond-NB speech transmission economically feasible at larger scale.

However, distortions introduced by the codec as well as packet loss, noise, or the acoustic design of the handset can severely reduce the quality advantage of WB transmission. In order to quantify the overall quality of a transmission path end-to-end, instrumental estimation models have been developed that predict the result of the judgment process of a listener taking part in a subjective experiment. Prominent models are PESQ [2] (signal-based) and the E-model [3] (parametric), both of which were recently extended to also capture relevant effects in WB speech transmission. In [1], a particularly useful extension of the WB E-model, the Bandwidth Impairment Factor, has been introduced which is capable of predicting the quality impact due to bandwidth limitations.

Relatively new is the usage of codecs that work beyond the WB frequency limits and allow super-wideband (SWB, 50-14000 Hz) speech transmission. The development of instrumental quality models for SWB transmission is in a similar early stage as the technology itself. Neither the quality advantage of SWB transmission nor SWB degradations are quantified by current speech quality models. However, there are ongoing efforts in ITU-T Study Group 12 to define a SWB-capable signal-based model as a successor for PESQ. Correspondingly, it is the purpose of the present study to extend the parameter-based E-model towards SWB in such a way that its validity for WB and NB conditions is maintained.

The remainder of this paper is organized as follows: Sect. 2 reviews the current version of the E-model. In Sect. 3, the procedure for extending the E-model quality scale towards SWB will be presented. Two subjective experiments will be described that were conducted for the scale extension. Using data from these experiments, degradation quantifiers for a number of SWB codecs will be derived. In Sect. 4, the Bandwidth Impairment Factor will be applied and tested for the SWB conditions. Conclusions will be given in Sect. 5.

2. THE E-MODEL

The E-model [3] is a transmission planning tool which has proven useful for predicting conversational and listening speech quality. The output of the model is a scalar transmission rating value \( R \) which combines transmission parameters relevant for a transmission path and is linked to the overall speech quality, expressed by the Mean Opinion Score (MOS), \( MOS \in \{1; 5\} \), through an S-shaped curve [3]. The E-model is based on summing impairment factors which subsume different types of distortion from end to end of the transmission path. The resulting transmission rating \( R \) is obtained by the following formula:

\[
R = R_0 - I_s - I_d - I_{\text{eff}} + A, \quad \text{with} \quad R \in [0; R_{0,\text{max}} + A].
\]

(1)

Here, the advantage factor \( A \) denotes the quality-advantage related to a given technology, with \( A = 0 \) without loss of generality for this entire paper. \( R_0 \) describes the basic signal-
to-noise ratio, including circuit and ambient noise. As long as a transparent transmission can be assumed, \( R_0 = R_{0,\text{max}} \), where \( R_{0,\text{max}} = R_{0,\text{max},N\text{B}} = 93.2 \) for NB conditions including a G.711 codec [3] (theoretically, \( R_{0,\text{max},N\text{B}} = 100 \) for a direct NB channel), and \( R_{0,\text{max}} = R_{0,\text{max},W\text{B}} = 129 \) for WB conditions [4]. The factor \( I_e \) combines signal-simultaneous distortions, \( I_e \) represents degradations occurring delayed to the speech signal, while codec distortions are subsumed by the Effective Equipment Impairment Factor \( I_{e,\text{eff}} \), including impairment due to packet loss. If no packet loss is present, this factor is replaced by \( I_e \). Values \( I_e \) for different codecs are tabulated in ITU-T Rec. G.113.

The \( R \)-scale is universal in the sense that the transmission rating \( R \) for a given condition is unique, independent of the bandwidth context. For instance, a G.711-coded condition obtains \( R = 93.2 \) in both the NB and WB context.

In the next section, this framework will be extended towards SWB transmission. Therefore, the basic quality \( R_{0,\text{max},SW\text{B}} \) will be determined by means of two subjective experiments, and Equipment Impairment Factors \( I_{e,SW\text{B}} \) for a number of SWB codecs will be derived.

3. SUPER-WIDEBAND E-MODEL

3.1. General overview

The extension of the current WB-version of the E-model towards SWB can be outlined as follows (analogous to the approach presented in [4]):

1. \( MOS \) data of two subjective experiments are available: Exp. 1 was conducted in a WB context, whereas Exp. 2 included SWB conditions together with a subset of the WB conditions of Exp. 1. The experiments will be described in Sect. 3.2.

2. The \( MOS \) data of both experiments are transformed to the \( R \) domain. The \( R \) values of the conditions common to both experiments are compared. The degree of compression of the NB/WB conditions in Exp. 2 allows to derive \( R_{0,\text{max},SW\text{B}} \) (see Sect. 3.3).

3. SWB Equipment Impairment Factors for conditions included in Exp. 2 can be derived according to \( I_{e,SW\text{B}} = R_{0,\text{max},SW\text{B}} - R(\text{condition}) \) (see Sect. 3.4).

3.2. Experiments

Exp. 1 comprises 125 NB and WB conditions such as clean WB PCM, WB codecs like the ITU-T Rec. G.722.2, the G.722, the G.729.1, and NB-codecs such as the G.711, the G.729A, and the G.726. Apart from that, several codec tandems were investigated, as well as different packet loss rates and additional background noise at send side. This experiment serves multiple research purposes and is employed as a reference experiment in the present study. A detailed description can be found in [5].

In Exp. 2, besides a direct PCM condition (50-14000 Hz), four SWB codecs at three different bitrates each were tested: The AMR-WB+, the open source speech codec Speex (version 1.2b3), the G.722.1 Annex C, and one anonymous codec prototype. The test conditions also included several bandpass-filtered SWB PCM conditions. In addition, NB and WB codecs from Exp. 1 were included, as well as conditions with packet loss or background noise. Altogether, Exp. 2 comprises 31 conditions.

Using tools from ITU-T Rec. G.191, the NB and WB speech samples were pre-/post-processed with sending and receiving filters according to ITU-T Rec. P.830 and P.341, the SWB samples were bandpass filtered (50-14000 Hz), and all samples were normalized to -26 dBV.

Both tests were conducted according to [6]. Source speech material from four different speakers (2 f, 2 m) was used, each of the speakers uttering different sentence pairs per condition. The source material was recorded at \( f_s = 48 \) kHz under anechoic conditions. Two different groups of paid naïve listeners (a number of 100 and 27 in Exp. 1 and 2, respectively) took part in the experiments, approximately balanced in age and gender. The participants gave their ratings on a slider-based version of the 5-point Absolute Category Rating scale ("MOS scale") [6]. The conditions were presented in randomized order and the experiments were divided into different sessions in order to avoid listener fatigue. In Exp. 2 (SWB), the subjects were screened for normal hearing ability using pure-tone audiometry. As a consequence, 6 listeners had to be excluded from the further analysis. For the tests, headphones were used (Sennheiser HD 25 in Exp. 1, STAX Lambda Pro in Exp. 2). The diotic presentation level was set to 73 dB SPL.

3.3. Scale extension

As a prerequisite to the following analysis, the \( MOS \) values [1;5] were linearly transformed in order to obtain a minimum value of 4.5 as required for the transformation to the \( R \)-scale (i denotes the condition per experiment) [7]:

\[
MOS_{\text{norm},i} = \frac{MOS_i - 1}{MOS_{\text{max}} - 1} \cdot 3.5 + 1. 
\]  

The transformation \( MOS_{\text{norm}} \rightarrow R \), with \( R \in [0;100] \) and \( MOS \in [1;4.5] \), according to the S-shaped curve (3rd order polynomial) given in [3][7] is valid for NB data. However, \( MOS \) ratings are dependent on the best quality condition presented to the listeners, i.e., they are context-dependent. This results in a MOS-scale-compression of the NB conditions in Exp. 1 which was carried out in a WB context. Analogously, the NB and WB conditions in Exp. 2 are compressed due to the presence of SWB conditions. Consequently, the values \( MOS_{\text{norm}} \) were provisionally transformed to values \( R_{WB,100} \in [0;100] \) for Exp. 1 and \( R_{SWB,100} \in [0;100] \) for Exp. 2, where the subscripts indicate the compression of
the conditions of the WB Exp. 1 and SWB Exp. 2, both in the range [0;100].

For Exp. 1, this compression is compensated by employing the expansion [4]
\[ R = R_{WB,129} = 1.29 \cdot R_{WB,100} , \]  
(3)
leading to valid R-values for the NB and WB conditions contained in Exp. 1 and extending the R-scale to the WB range [0;129].

A modified version of Eq. (3) holds for the Exp. 2 data:
\[ R_{SWB,129} = 1.29 \cdot R_{SWB,100} , \]  
(4)
where \( R_{SWB,129} \in [0;129] \) denotes compressed SWB R-values in the WB range of the R-scale. This remaining compression can be seen from Fig. 1, where the (compressed) R-values of conditions common to both experiments are plotted on the x-axis (\( R_{SWB,129} \), Exp. 2) and y-axis (\( R = R_{WB,129} \), Exp. 1).

In order to achieve a complete decompression for the SWB conditions and thus obtain valid R-values for NB, WB, and SWB conditions in the SWB context, the R-values of Exp. 2, \( R_{SWB,129} \), are mapped onto those of Exp. 1 using a linear relation:
\[ R = a \cdot R_{SWB,129} + b . \]  
(5)

The coefficients \( a \) and \( b \) are determined by curve fitting in a least-squares sense. A root mean squared error of \( RMSE = 3.02 \) can be achieved with \( a = 1.29 \) and \( b = 7.68 \). An only slightly worse fitting model can be obtained with the constraint \( b = 0 \), forcing the curve to go through the origin (\( a = 1.39, RMSE = 3.47 \)). The latter relation will be used in the remainder of this paper since it is simple and more intuitive.

Eq. (5) represents a rule transforming R values obtained in a SWB context from the WB-range of the R-scale to an extended R-scale for SWB, retaining the R-values for NB and WB conditions, analogous to Eq. (3). In this way, the new SWB-range of the R-scale can be determined. It amounts to \( R_{0,max,SWB} = 39 \cdot R_{0,max,WB} = 179 \). This corresponds to a quality gain with respect to WB of 39%, and of 79% relative to NB. In spite of the little amount of speech energy between 7 kHz and 14 kHz the observed advantage of 39% appears to be plausible when listening to the speech files oneself. It is assumed to partly stem from the improved spectral balance as compared to the pure WB case [1][5].

### 3.4. SWB Equipment Impairment Factors

The R-values of codecs and linearly distorted conditions contained in Exp. 2, as obtained by Eq. (5), were transformed to Equipment Impairment Factors according to Eq. (1) (with \( I_e = I_d = A = 0 \)): \( I_e,SWB = 179 - R(\text{condition}) \). Note that in general, NB and WB codecs can be expressed on the extended R-scale via the expression
\[ I_e,SWB = \begin{cases} R_{0,max,SWB} - R_{0,max,WB} + I_e,WB & \text{(WB)} \\ R_{0,max,SWB} - R_{0,max,NB} + I_e,NB & \text{(NB)} \end{cases} \]  
(6)
accounting for the impairment due to bandwidth reduction from SWB to NB or NB, respectively. Eq. (6) results from the constraint of consistent R-values for given conditions across NB, WB, and SWB contexts (cf. Sect. 2) and allows for a direct comparison of NB, WB, and SWB impairment factors on a universal scale.

---

**Fig. 1.** R-scale extension based on the listening tests 1 and 2. See text for details. Abbreviations: BGN .. Background noise, PL .. Packet loss. Bitrates are given in brackets.

**Fig. 2.** Impairment Factors for conditions of Exp. 2.

In Fig. 2, the values \( I_e,SWB \) of a subset of the Exp. 2 conditions are displayed in ascending order (black bars). The horizontal lines at \( I_e,SWB = 179 - R_{0,max,WB} = 50 \) and \( I_e,SWB = 179 - R_{0,max,NB} \approx 86 \) indicate the transitions to the R-scale-ranges for WB and NB conditions, respectively.

As it can be seen, the direct SWB condition corresponds to an impairment of zero. Compared to a transparent SWB transmission, the best codec contained in Exp. 2, the G.722.1 Ann. C at 48 kbps, introduces a remarkable degradation. Al-
most equivalent to this codec in quality, but more economic in bandwidth consumption, is the AMR-WB+ (36 kbps).

In general, the quality impairment increases with decreasing bitrate and audio bandwidth, as expected. At low bitrates, SWB codecs may degrade the quality even more severely than WB or NB codecs, completely destroying the actual quality advantage of SWB. This is partly due to non-linear codec distortions, partly due to the internal bandwidth reduction at low bitrates, depending on the coding scheme.

4. BANDWIDTH IMPAIRMENT FACTOR

It has been shown in [8] that the Impairment Factor $I_e$ can be split into two parts, one describing the degradations caused by linear distortion, reflected by the Bandwidth Impairment Factor $I_{bw}$, and one describing residual degradations stemming from codec non-linearities (residual impairment $I_{res}$); thus $I_e = I_{bw} + I_{res}$. For purely linear distortions, $I_{res} = 0$, and thus $I_e = I_{bw}$.

In this section, the applicability of the model predicting $I_{bw}$, as introduced in [1], is shown for the SWB case. The model is based on a non-linear combination of the equivalent rectangular bandwidth (ERB) in Bark and the center frequency in Hz of a transmission channel, which in turn can be estimated by the input and output signals of a transmission path following the method described in [9]. It has been shown in [8] that it also provides plausible estimates for audio bandwidth restrictions introduced by speech codecs.

The $I_{bw}$ model was originally derived for WB conditions (in the following referred to as $I_{bw,WB}$). Thus, an impairment value around zero is the output for a direct WB condition. Hence, when applying the prediction model to the SWB conditions of Exp. 2, negative values $I_{bw,WB}$ are obtained.

In order to adapt this model to SWB transmission, a first order polynomial mapping function was determined on the basis of the linearly filtered SWB conditions of Exp. 2, i.e., where $I_{bw,WB} < 0$ (see Fig. 2, gray bars). For these conditions, it further holds $I_{e,SWB} = I_{bw,SWB}$, since $I_{res} = 0$. The correlation of $r = 0.977$ and $RMSE = 5.639$ show that the $I_{bw}$ model is principally applicable to SWB conditions. For $I_{bw,WB} > 0$, NB and WB Bandwidth Impairment Factors can be expressed on the SWB $R$-scale by evaluating $I_{bw,SWB} = R_{0,max,SWB} - R_{0,max,SWB} + I_{bw,WB} = 50 + I_{bw,WB}$, in analogy to Eq. (6).

Fig. 2 depicts the transformed $I_{bw,SWB}$ predictions for the conditions of Exp. 2 (gray bars). They were obtained by estimating $I_{bw,SWB}$ for four different speakers out of the sample set of Exp. 2 and calculating the mean. As it can be seen, the “overall” impairment of the linearly distorted samples $I_{e,SWB}$ is well matched by $I_{bw,SWB}$. Furthermore, $I_{bw,SWB}$ provides valuable diagnostic insight into the composition of the overall codec distortion (see Fig. 2). For direct SWB, WB, and NB conditions, evaluating $R = R_{0,max,SWB} - I_{bw,SWB}$ corresponds approximately to $R = R_{0,max,SWB} = 179$ ($I_{bw,SWB} = 0$), $R = R_{0,max,SWB} = 129$ ($I_{bw,SWB} = 50$), and $R = R_{0,max,NB} = 93$ ($I_{bw,SWB} = 86$), respectively, as expected.

5. CONCLUSION AND OUTLOOK

In the present paper, the current WB version of the E-model has been extended to SWB transmission on the basis of two subjective experiments. Therefore, the E-model transmission rating scale has been extended to $R_{0,max,SWB} = 179$. On this universal quality scale, NB, WB, and SWB distortions can be quantified and directly compared with each other, keeping the E-model downwards-compatible. The extension corresponds to a quality gain of 39% as compared to WB for transparent channels, and 79% as compared to a clean NB channel.

Equipment Impairment Factors for several SWB codecs have been derived and put into relation to WB and NB codecs. It has been shown that even the best tested codecs may degrade the quality of SWB remarkably. Furthermore, an existing quality measure for linear WB distortions has been successfully applied to SWB conditions.

The scale extension and the impairment factors derived here will constitute an input to ITU-T Study Group 12. In order to obtain fully robust results, additional experiments by other laboratories are required to include factors of other languages and other cultures.

6. REFERENCES