TOWARDS A NEW E-MODEL IMPAIRMENT FACTOR FOR LINEAR DISTORTION OF NARROWBAND AND WIDEBAND SPEECH TRANSMISSION

Marcel Wältermann and Alexander Raake

Quality and Usability Lab, Deutsche Telekom Laboratories, Berlin University of Technology, Germany
{marcel.waeltermann, alexander.raake}@telekom.de

ABSTRACT

The E-Model, a tool for network planning recommended by the ITU-T, suffers from the lack of predicting linear distortions as they may occur from channel filtering, codecs, and user interfaces. In order to face this deficiency, a new impairment factor is introduced in this paper which is estimated on the basis of two simple parameters from the linear portion of a system. Therewith, conventional Equipment Impairment Factors of narrowband and wideband speech codecs are decomposed into a linear and a residual part, allowing to quantify both magnitudes separately on the so-called R-scale. By extending the concept of distortion classes in the E-Model, the proposed scheme provides a plausible picture of the linear effect of transmission systems. The advantage of wideband causing a 36% quality gain is well reflected. Further, the decomposition leads to a reduction of error when impairment factors are added on the R-scale. Examples for instrumentally estimating the residual impairment are given.

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

1. INTRODUCTION

In telephone connections, be it traditional PSTN or packet-based VoIP, linear distortions of the speech signal may occur due to channel filtering, codec-dependent restrictions in audio-bandwidth, or due to imperfect electro-acoustical interfaces [1].

As it was frequently shown in the related literature, the bandwidth and spectral shape of the "transfer function" of a transmission path is of great importance for speech quality perception (e.g., [1][2]). In particular, it can lead to an increase of quality when the bandwidth is extended from narrowband (300-3400 Hz) to wideband (50-7000 Hz). In turn, linear distortion of both wideband and narrowband speech can cause perceptual impairment. This type of impairment is not explicitly taken into account by current speech quality models [3][4].

In this paper, we propose an extension of the so-called E-Model, a parameter-based tool for speech quality estimation recommended by the ITU-T for network planning [3]. In a recent work, the model framework has been extended to wideband, and a set of wideband impairment factors has been defined for wideband and narrowband speech codecs [5]. Whereas these impairment factors capture both the effect of the codec non-linearity and the audio bandwidth, we base our approach on a decomposition of coding artifacts into a linear and residual component. This strategy allows to employ the E-Model for the quantification of linear distortions, stemming, e.g., from channel filters or user interfaces. Thus, the scope of the model is significantly extended.

To this aim, we here propose to introduce an additional impairment factor for linear distortions. The residual component results from the difference of conventional Equipment Impairment Factors and the Bandwidth Impairment Factor described in this paper. It is promising that the strict separation of linear and residual impairment strengthens the core property of the E-Model, namely the additivity of distortion classes on a one-dimensional scale.

In Sec. 2, the framework of the E-Model is briefly described. The state-of-the-art extension to wideband speech and its restrictions are discussed in Sec. 3. A new impairment factor comprising end-to-end linear distortions of arbitrary shape is introduced in Sec. 4, followed by an approach for decomposing conventional Equipment Impairment Factors into a linear and residual part (Sec. 5) in order to incorporate the bandwidth impairment in the E-Model framework. Sec. 6 gives examples on how the residual impairment may be estimated instrumentally. In Sec. 7, conclusions and an outlook on future work are provided.

2. THE E-MODEL FRAMEWORK

The E-Model [3] is a network planning tool for the prediction of conversational and listening speech quality. By means of an algorithmic combination of the parametric description of the underlying network like the packet loss rate, transmission delay, speech codec, overall loudness, room and circuit noise, et cetera, subjective speech quality judgments are estimated. The distortion types are categorized into so-called *Impairment Factors*, describing the amount of impairment due to the basic signal-to-noise ratio (R_0), the signal-simultaneous distortions (I_s), and the delayed impairments

such as transmission delay or echo (I_d). Distortion originating from codecs are subsumed under the *Equipment Impairment Factor* I_e which is extended to the *Effective Equipment Impairment Factor* $I_{e,eff}$, taking packet loss effects into account. These factors are assumed to be additive on a psychological scale, the so-called *Transmission Rating Scale*, or R-Scale, reflecting an overall quality estimate:

$$R = R_0 - I_s - I_d - I_{e,eff} + A, \text{ with } R \in [0; R_{0,max}]. \quad (1)$$

The factor A denotes the quality-advantage related with a given technology as perceived by the user.

The R-values may be converted to *Mean Opinion Scores* (MOS), reflecting the judgments subjects typically give in speech quality experiments ($MOS \in [1; 4.5]$). $R_{0,max}$ corresponds to the highest quality.

3. FROM NARROWBAND TO WIDEBAND

In the context of traditional narrowband (NB) speech (300-3400 Hz), $R_{0,max}$ is fixed to $R_{0,NB,max} = 100$, corresponding to a MOS-value of 4.5. For wideband speech, $R_{0,max}$ has recently been extended to $R_{0,WB,max} = 129$, empirically proven by extensive auditory tests [5] and adopted by the ITU-T in Rec. G.107, App. II. Thus, a clean WB channel is approximately 36% "better" than a standard G.711 narrowband channel, represented by $R_{G.711} = 93.2$. Additionally, I_e-values for WB-capable speech codecs have been derived in order to quantify them on the R-scale (adopted by the ITU-T in Rec. G.113, App. IV). It has been shown that by adding 35.8 to $I_{e,NB}$, i.e.

$$I_{e,WB} = (129 - 93.2) + I_{e,NB}\,, \quad (2)$$

the quality of NB codecs can still be covered with the extended scale. Therewith, NB and WB speech codecs can directly be compared with each other.

The additivity property, however, is no longer valid if, e.g., two NB codecs are connected in series in a wideband context. Since the bandwidth disadvantage of the NB codecs is inherently expressed by a single factor $I_{e,WB}$ which is assumed to be additive, the NB disadvantage sums up to $2 \cdot 36$, leading to unrealistic high overall impairment factors. This problem could be avoided by expressing the NB-effect by a separate impairment factor which only takes the bandwidth impairment into account, which is introduced in the next section.

4. BANDWIDTH IMPAIRMENT FACTOR

In the current version of the E-Model, the impairment of bandwidth distortion is implicitly taken into account by $I_{e,WB}$. This, however, prohibits the explicit quantification of pure bandwidth impairment originating from codecs, user interfaces, or channel filters and goes along with additivity problems described in the preceding section. In order to face this drawback, a dedicated impairment factor describing the perceptual effect of linear frequency distortions has been introduced by the second author, which gives intuitive insight on the influence of audio bandwidth [1]. This *Bandwidth Impairment Factor* I_{bw} was derived from auditory tests mainly including bandwidth restricted stimuli in a WB context. By mapping the resulting MOS-values onto the R-Scale, the following formula was found by curve fitting:

$$I_{bw} = 0.035 \cdot |s| - 0.0067 \cdot s - 7.4 \cdot \frac{z_{bw}}{\text{Bark}} + 129.2, \quad \text{with } s = \frac{f_c}{\text{Hz}} - 9.9 \cdot \left(\frac{z_{bw}}{\text{Bark}} + 101.8\right). \quad (3)$$

In Eq. (3), z_{bw} denotes the transmission bandwidth in Bark [6], and f_c denotes the center frequency in Hertz, for which a bandwidth dependent optimum exists. A correlation of $r = 0.992$ between model estimates and auditive test results could be achieved. For practical reasons, z_{bw} is approximated by the *Equivalent Rectangular Bandwidth (ERB)* [1], estimated from the amplitude spectrum. More complex filter shapes are represented sufficiently well by ERB [1], such that $z_{bw} \approx ERB$. The center frequency f_c can then be obtained by calculating the geometric mean of the cut-off frequencies of ERB, previously transferred to the Hz-scale.

5. INTEGRATION OF BANDWIDTH IMPAIRMENT INTO THE E-MODEL FRAMEWORK

With the new impairment factor introduced in Sec. 4, arbitrary transfer functions of a transmission chain might be quantified on the WB R-scale. Since the difference between NB and WB transmission is included in the Equipment Impairment Factors in the current version of the E-Model (cf. Sec. 3), it is proposed to decompose the $I_{e,WB}$ values into a linear portion, reflected by the bandwidth impairment I_{bw}, and a residual portion I_{res}, such that

$$I_{e,WB} = I_{bw} + I_{res}\,. \quad (4)$$

From a system theoretical perspective, Eq. (4) corresponds to the notion of modeling a system by a linear time-invariant (LTI) component, characterized by its impulse response $h(k)$, and an additional (non-white and system-/signal dependent) noise $n(k)$, modeling the non-linear distortion (cf. [7]), such that the output signal $y(k)$ arises from $y(k) = s(k) + n(k)$. Here, $s(k)$ can be obtained via convolving the input signal $x(k)$ with $h(k)$. If $n(k)$ and $x(k)$ are uncorrelated and $n(k)$ or $x(k)$ unbiased, the transfer function $H(e^{j\Omega})$ of the LTI system can be described with

$$H(e^{j\Omega}) = \frac{\Phi_{xy}(e^{j\Omega})}{\Phi_{xx}(e^{j\Omega})}, \text{ with } \Omega = 2\pi\frac{f}{f_S}\,, \quad (5)$$

where Φ_{xy} and Φ_{xx} are the power density spectra of $x(k)$ and $y(k)$, respectively, f is the frequency, and f_S is the sampling frequency. In order to obtain I_{bw}, ERB and f_c are

estimated from the amplitude spectrum $20\log_{10}|H(e^{j\Omega})|$ (cf. [8]). Although most codecs are highly non-linear systems, rendering the prerequisites of the above mentioned assumption invalid, it will now be shown that the amplitude spectrum provides plausible estimates of I_{bw}. For the moment, we assume that $I_{e,WB}$ is known. Thus, the residual portion I_{res} directly arises from Eq. (4). If $I_{e,WB}$ is unknown, e.g., for new codecs, $I_{e,WB}$ may determined through subjective tests as it was done, e.g., in [5], or through instrumental estimates discussed in the next section.

In the following, Equipment Impairment Factors of a multitude of NB and WB codecs, following a variety of coding principles are decomposed into their linear and residual components:

- NB: G.711 A-law, G.723.1 (5.3, 6.3), G.726 (16, 24, 32, 40), G.729A, GSM-EFR, and GSM-FR
- WB: G.722 (48, 56, 64), G.722.1 (24, 32), G.722.2 (6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85), and plain PCM

If more than one codec-mode is considered, the bitrate in kbit/s is indicated in parentheses.

The NB and WB samples were pre-processed with sending and receiving filters according to ITU-T Rec. P.830 and P.341, respectively. Differences in delay and RMS levels between the files were compensated. The spectral amplitudes of each condition were determined according to the approach discussed in this section. The extracted parameters were averaged across four different speakers, each of them uttering 2 different phrases between 5 and 10 s.

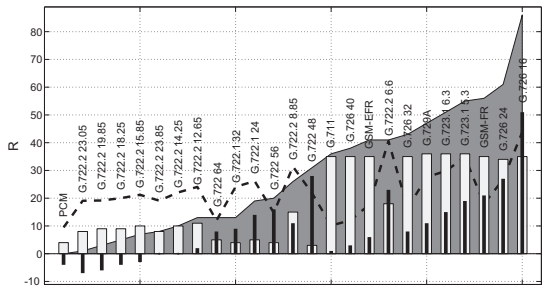

Fig. 1. $I_{e,WB}$ (dark gray), I_{bw} (light gray), I_{res} (black), and d_{SYM} (re-scaled, dashed).

Fig. 1 depicts the $I_{e,WB}$ values, taken from ITU-T Rec. G.113 for the considered codecs, in increasing order (dark gray). The light gray boxes represent the estimated bandwidth impairment I_{bw}, and the resulting residual impairment I_{res} is represented by black boxes.

Observing the linear and residual components of the codecs, a good picture is given on which kind (and which ratio) of degradations leads to a quality decrease. A grouping between NB and WB conditions can be recognized: NB conditions show a bandwidth impairment of $I_{bw} \approx 35$, which is nearly constant for all NB codecs. It is striking that for the best NB codec, G.711, $I_{bw} \approx I_{e,WB} = 36$, resulting in almost no residual impairment ($I_{res} = 1$). Hence, the perceptual difference between a clean NB and a clean WB ($I_{bw} = 0$) transmission is completely covered by the model of Eq. (3) and perfectly confirming Eq. (2). Thus, additional (non-linear) distortion caused by specific NB codecs can be subsumed under the residual impairment I_{res}, reflecting perceptual effects that can be different in nature (depending on, e.g., the coding principle and the bitrate).

The bandwidth impairment for WB is generally lower, however, more differences in linear distortions are salient in this case. For instance, the impairment due to G.722.2 codecs is higher than for the other codecs. The negative residual components reflect a benefit of the non-linearity of the codec in partly ruling out the increased bandwidth impairment. The pure PCM coding obtains an impairment of $I_{bw} = 4$. In [1], however, a bandwidth of 200-7000 Hz was considered as best quality ($I_{bw} = 0$) which is reflected by Eq. (3).

A number of 24 codec tandems were considered in a second step:

- WB×NB (and vice versa, avv.): {G.711, G.726 (32), G.729A} × G.722.2 {12.65, 23.05} avv., GSM-EFR × G.722.2 (23.05) avv., G.722 (64) × {G.711, G.726 (32)}
- WB×WB: G.722.2 {12.65, 23.05} × G.722.2 {12.65, 23.05}, G.722 (64) × G.722.2 (23.05) avv.
- NB×NB: G.726 (32) × G.726 (32), G.729A × G.729A

Although the impairments on the R-scale are claimed to be additive, a proof failed in many cases if an attempt was undertaken to simply sum the $I_{e,WB}$ of corresponding codecs connected in series (e.g., [5]). Following the notion that the $I_{e,WB}$ values are composed of a linear and residual portion, one source of error may stem from adding the linear parts implicitly (cf. Sec. 3). Indeed, an analysis of the resulting factors $I_{bw,1}$ and $I_{bw,2}$ of codecs 1 and 2 connected in series reveals that the bandwidth impairment is approximately determined by the codec of smaller bandwidth, i.e.:

$$I_{bw,\text{Tandem}} \approx \max\{I_{bw,1}, I_{bw,2}\} \,. \tag{6}$$

This relation is intuitively clear, since the frequency distortion of a series of codecs does certainly not correspond to the sum of bandwidth impairments of the single codecs. Thus, the sum $I_{e,WB,1} + I_{e,WB,2}$ provokes a minimum error of $\min\{I_{bw,1}, I_{bw,2}\}$ (assuming that no interactions between the respective components I_{bw} and I_{res} exist). However, a subtraction of the error term from $I_{e,WB,1} + I_{e,WB,2}$ only leads to a reduction of error in most cases, indicating that I_{res} might be composed of further dimensions for which additivity does not necessarily need to be valid.

6. INSTRUMENTAL ESTIMATION OF THE RESIDUAL COMPONENT

Assuming that I_{bw} is validly estimated by Eq. (3) and that there are no interactions between I_{bw} and I_{res}, the residual component I_{res} may be estimated via instrumental methods. In a recent study [9], it has been shown that I_e-values can be predicted with signal-based models achieving a high correlation (e.g., with WB-PESQ, a correlation of $r = 0.85$ is achievable). Then, provided that I_{bw} is given, I_{res} can be determined via Eq. (4).

Alternatively, I_{res} can be estimated directly by determining the noise component $n(k)$ (cf. Sec. 5). Solving this problem, however, is inherently addressed in all attempts of signal-based quality prediction models (e.g., [10]), since it breaks down to capturing the codec distortion itself, albeit without linear distortion. Recent models mostly rely on computing a spectral distance between input and output signals after preprocessing and transforming the spectra into a perceptual domain. Hence, as a side information, the spectral distance according to the processing in PESQ, the so-called *symmetric disturbance* d_{SYM} [4][10], is calculated and correlated with $I_{res} = I_{e,WB} - I_{bw}$. Here, perceptual representations of the spectra of $s(k)$ and $y(k)$ are compared to each other, resulting in a difference reflecting the noise component. With this approach, pure bandwidth distortions lead to equal values and are therefore unaffected by these measures, showing that linear distortions are orthogonally assessed. However, the lack of a sophisticated pre-/post-processing for the given task results in worse estimates than usual signal-based models give for overall quality estimates.

The dashed curve in Fig. 1 represents the spectral distortion d_{SYM} which roughly follows the residual impairment I_{res}, especially for the NB codecs. The correlation between these values amount to $r = 0.66$ ($RMSE = 10.24$), showing that the residual distortion can principally be captured. Proposals for improvement, however, will be given elsewhere in the future.

7. CONCLUSIONS AND OUTLOOK

In this study, a new E-Model impairment factor I_{bw}, capturing linear distortions of a transmission chain, has been introduced. Conventional Equipment Impairment Factors were decomposed into linear and residual components, where the linear part is reflected by I_{bw} and estimated on the basis of the amplitude spectrum. Examples were given for estimating the residual part instrumentally.

As it was shown, splitting off the linear part from the E-Model Equipment Impairment Factors leads to a decrease of errors that usually occur when making use of the additivity property. In future investigations, the error may further be reduced when partitioning the residual impairment into further dimensions it may be composed of, and from which simple additivity must not be assumed.

With the presented approach, it has been shown that the influence of frequency distortions is reflected by I_{bw} in a plausible way. Thus, the proposed extension of the E-Model is foreseen to constitute a basis for predicting end-to-end speech quality including the frequency distortion of codecs, channel filters, and user interfaces.

8. ACKNOWLEDGMENT

This study was supported by the Deutsche Forschungsgemeinschaft (DFG), grant MO1038/5-2.

9. REFERENCES

[1] Alexander Raake, *Speech Quality of VoIP – Assessment and Prediction*, Wiley, UK-Chichester, 2006.

[2] B.C.J Moore and C.T. Tan, "Perceived naturalness of spectrally distorted speech and music," *J. Acoust. Soc. Am.*, vol. 114(1), pp. 408–419, 2003.

[3] ITU–T Rec. G.107, *The E-Model, a computational model for use in transmission planning*, CH–Geneva, 2005.

[4] ITU–T Rec. P.862, *Perceptual evaluation of speech quality (PESQ)*, CH–Geneva, 2005.

[5] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment factor framework for wideband speech codecs," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 14(6), pp. 1969–1976, 2006.

[6] Eberhard Zwicker and Hugo Fastl, *Psychoacoustics: Facts and Models*, Springer, D–Berlin, 1999.

[7] H.W. Schüssler, "An objective method for measuring the performance of weakly non-linear and noise systems," *Frequenz*, vol. 41(6), pp. 147–154, 1987.

[8] K. Scholz, M. Wältermann, L. Huo, A. Raake, U. Heute, and S. Möller, "Estimation of the quality dimension "directness/frequency content" for the instrumental assessment of speech quality," in *Proc. ICSLP 2006*, USA–Pittsburgh PA, 2006, pp. 1523–1526.

[9] S. Möller, N. Côté, V. Gautier-Turbin, N. Kitawaki, and A. Takahashi, "Instrumental estimation of equipment impairment factors for wideband speech codecs," *subm. to IEEE Trans. Audio, Speech and Language Proc.*

[10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP '01*, 2001, pp. 749–752.