Comparison of Spectrum-based Models for Speech and Audio Quality and Naturalness Estimation

A. Raake[1], M. Wältermann[1], B.C.J. Moore[2], C.-T. Tan[3]

[1] *Quality & Usability Lab, Deutsche Telekom Labs, TU Berlin, Germany, Email: alexander.raake@telekom.de*

[2] *Department of Experimental Psychology, University of Cambridge, UK*

[3] *Department of Otolaryngology, School of Medicine, New York University, USA*

Introduction

Spectral characteristics e.g. of electro-acoustic interfaces in end-devices or network bandpass filters may degrade perceived naturalness and thus yield an impairment of transmitted speech quality. This effect is particularly prominent for wideband (WB) speech (50-7000 Hz) or speech with even wider bandwidth. However, the standardized models typically used to evaluate speech quality in telecommunications deliver accurate predictions only for narrowband (NB) speech (300-3400 Hz), and do not correctly predict the effects of linear distortion. In previous work [8, 3, 10], we suggested an extension of the E-model quality scale [1] for application to WB speech, a framework for the impairment due to WB codecs, and a new impairment factor for quality under linear distortion. For a given spectral shape, this new impairment factor 'I_{bw}' is calculated from a rectangular filter whose effective bandwidth and center frequency match those of the system under test. In this paper, the model is briefly outlined, and it is evaluated on a speech and audio database including ratings of auditory naturalness [6, 7]. The performance of the simple I_{bw}-model is compared to that of a more perception-oriented model of naturalness under spectral distortion [7].

Model 1: I_{bw}

The bandwidth impairment factor I_{bw} has been proposed as an extension to the E-model [1] in [8]. The E-model is a tool recommended by the ITU-T for network planning. It is based on the transformation of a number of quality-relevant technical parameters describing the end-to-end telephone transmission path to a psychological quality scale, the so-called R-scale. The E-model relies on the assumption that different types of degradations are additive in terms of the perceptual impairment they cause. This is reflected by the basic formula of the model:

$$R = R_0 - I_s - I_d - I_{e,eff} \qquad (1)$$

Here, R is the Transmission Rating, the output of the quality model, which ranges from 0 to $R_{0,max}$. For NB speech, the bandwidth the model was developed for initially, $R_{0,max} = 100$. In previous work, we have extended this maximum range to WB, with $R_{0,max} = 129$ [8, 3]. With this extension, NB and WB speech quality can be expressed on a single scale. In Equation (1), R_0 reflects the base-quality that is related to the basic signal-to-noise-ratio; I_s is the simultaneous impairment factor, which expresses the quality impairment due to degradations such as signal-correlated noise; I_d is the delayed impairment factor, which accounts for the effect on quality of pure delay and echo; $I_{e,eff}$ is the effective equipment impairment factor which accounts for the quality impairment due to speech coding and eventual packet loss in VoIP-type systems.

The WB-version of the E-model is currently under development. So far the scale extension in terms of $R_{0,max}$, and the effective equipment impairment factor for wideband $I_{e,eff,WB}$ have been included in the relevant standard [1]. The quality improvement due to super-wideband (SWB, 50-14000 Hz) or fullband (FB, 20-22000 Hz) have not yet been determined, and corresponding E-model scale extensions are not yet available.

For the NB- and WB-case, we have proposed a further extension of the model to include the effect of linear distortion [8], extending Equation (1) with the bandwidth impairment factor I_{bw}; see Equation (2). Note that further considerations on how to cover the contribution of the employed codec to the spectral distortion and thus bandwidth impairment factor I_{bw} can be found in [10].

$$R = R_0 - I_s - I_d - I_{e,eff} - I_{bw} \qquad (2)$$

Listening test

I_{bw} has been developed based on a listening test including 20 different bandpass-filters, with the lower cut-off frequencies chosen from the range $50-600$ Hz and the upper cut-off frequencies from the range $2000-7000$ Hz [8]. 18 combinations of lower and upper cut-off frequencies were selected. Two of the NB-filters were overlaid with an IRS-type filter (Intermediate Reference System, see ITU-T Rec. P.48, 1989), which reflects the linear distortion introduced by an average handset telephone. As source material, 40 shortened speech passages from the Eurom sentences were used [2] (anechoic recording from six speakers, 3 female, 3 male). 20 conditions and six speakers yield a total of 120 samples to be rated by the subjects. The recorded speech was processed with the bandpass filters, and each of the available 40 sentences was used three times in the test. The test items were presented monaurally with a high-quality, circum-aural telephone handset constructed using one earpiece of a STAX professional headphone.

Quality ratings were collected using the 5-point absolute category rating scale (ACR-scale, the so-called MOS-scale according to ITU-T Rec. P.800, 1996). The test

results were transformed onto the NB-R-scale using the MOS-to-R-transformation provided in [1]. To obtain values on the WB-extended R_{WB}-scale, the linear extrapolation proposed in [8, 3] was used, with $R_{WB} = 1.29 \cdot R_{NB}$. Then, for a given bandpass i the bandwidth impairment factor I_{bw} is obtained as

$$I_{bw,i} = R_{0,max} - R_{WB,i} \tag{3}$$

Parameters from amplitude spectrum

In our parametric approach, the shape of the amplitude spectrum is described by an equivalent rectangular bandpass filter. To derive this bandpass, the amplitude spectrum is given in dB and on a hearing-appropriate frequency scale. In our approach, we have used the Bark-scale for this purpose, which provides a linear mapping between Bark units and the location of excitation on the Basilar membrane, choosing numbers reflecting the concept of critical bands according to [11]. An approximation of the experimentally found relations is given in Equation (4).

$$z/\text{Bark} = 13\arctan(0.76 f/\text{kHz}) + 3.5\arctan\left[(f/7.5\text{kHz})^2\right] \tag{4}$$

Note that due to the conceptual similarity of the two scales, the alternative usage of the equivalent rectangular bandwidth scale ERB_N (N for normal hearing) according to [4] leads to similar prediction results. An approximation of the experimental data found in [4] for ERB_N numbers is given by Equation (5):

$$ERB_N\text{-number} = 21.4 \cdot \log_{10}(4.37 \cdot f/Hz + 1). \tag{5}$$

In the $I_{bw} - model$, the bandwidth z_{bw} is derived using

$$z_{bw} = \frac{\text{area}(curve)}{\max(curve)}. \tag{6}$$

The center of gravity of the area is defined as the center-frequency z_G in Bark. The upper and lower cut-off frequencies are obtained using

$$z_l = z_G - z_{bw}/2 \tag{7}$$

$$z_u = z_G + z_{bw}/2 \tag{8}$$

The center frequency f_c is obtained by transforming the Bark rates z_l and z_u to Hertz using Equation (9) [8]

$$f_j/Hz = 1285.93\left(\frac{e^{(z_j/Bark)^{2.64}}}{1836.93} - 1\right) + 93.3\frac{z_j}{Bark}, \tag{9}$$

(with $j \equiv l$ or $j \equiv u$), which yields a good approximation to the Hertz-to-Bark transformation given in [11]. The center frequency f_c is calculated as it is typically done in the context of technical applications:

$$f_c = \sqrt{f_l \cdot f_u}. \tag{10}$$

Model

The two simple parameters z_{bw} and f_c serve as the input parameters to our I_{bw}-model. Based on the subjective test described above, the following relation between I_{bw} and the input parameters has been determined:

$$\begin{aligned} I_{bw} = \quad & a_1 \cdot \left|\tfrac{f_c}{\text{Hz}} - a_6 \cdot \left(\tfrac{z_{bw}}{\text{Bark}} + a_5\right)\right| \\ - \quad & a_2 \cdot \left(\tfrac{f_c}{\text{Hz}} - a_6 \cdot \left(\tfrac{z_{bw}}{\text{Bark}} + a_5\right)\right) \\ - \quad & a_3 \cdot \tfrac{z_{bw}}{\text{Bark}} + a_4 \end{aligned} \tag{11}$$

The curve-fitting parameters a_1 to a_6 were obtained as:

a_1	a_2	a_3	a_4	a_5	a_6
$3.5 \cdot 10^{-2}$	$6.7 \cdot 10^{-3}$	7.4	129.2	101.8	9.9

With these settings, the model yields a linear correlation of $\rho = 0.992$ and a root mean squared error of $RMSE = 3.39$ (on the model-scale $(0, 129)$).

Model 2: Naturalness

Another, more perception-oriented approach to modeling the impact of linear distortion on speech and music quality was proposed in [7], in which perceived naturalness was predicted.

Listening tests

Two extensive naturalness tests were used for model development, one for music and one for speech [6]. In the music test, a Jazz excerpt was used as source material, and in the speech test a concatenation of two sentences, one uttered by a female and one by a male speaker. A total of 168 linear distortion conditions were used in each of the two tests, employing different response-modifying filters (see [6] for details), with:

- Spectral ripples, sinusoidal in dB on an ERB_N scale, of different rates and depths, either extended over the range 3–32 ERB_N ($\equiv$ 87 to 6981 Hz), or limited to a restricted frequency range, with a flat response outside that range.
- Spectral tilts, linear in dB/ERB_N, with tilts of $\pm$0.1, 0.2, 0.5, and 1 dB/ERB_N. Tilts either extended over the frequency range 87 to 6981 Hz, or were limited to a restricted frequency range, with a flat response outside that range.
- Spectral ripples combined with spectral tilts, either extending over the frequency range 87 to 6981 Hz, or limited to a subrange.
- Bandpass filtering with all combinations of different lower and upper cut-off frequencies, and flat in the passband: Lower — 2, 4, 6, and 8 ERB_N ($\equiv$ 55, 123, 208, and 313 Hz); upper — 26, 28, 30, 32, 36, and 40 ERB_N ($\equiv$ 3547, 4455, 5583, 6981, 10869, and 16854 Hz).

For each test type, 169 processed audio files (the filtered samples and the clean reference) were presented diotically via diffuse-field equalized Sennheiser HD580 headphones. After each presented sample, the subjects gave ratings of perceived naturalness using a 10-point scale with the

labels 10: "very natural – uncolored" and 1: "very unnatural – highly colored". Ten subjects participated in the test, and both the music and the speech test were conducted twice to assess the test/re-test reliability.

Model

The test results were used for developing a naturalness prediction model, using the following modeling steps:

1. Determine excitation patterns using speech-shaped or pink noise, both for the original signal and the signal processed with the bandpass filter under consideration; excitation patterns are determined as in [5], with an additional sharpening parameter s of the auditory filters as a first free model parameter.
2. Excitation patterns expressed on an ERB_N-number scale (see Equation 5), with $ERB_{N,max} = 40$ ($\equiv$ 16800 Hz), and ERB_N number i sampled at 0.5-ERB_N intervals. Result: Excitation levels $EO(i)$ for the original, and excitation levels $ED(i)$ for the degraded signal.
3. Thresholding with threshold g: If $EO(i) < g$, set $EO(i) = g$; if $ED(i) < g$, set $ED(i) = g$ (f is a second free model parameter).
4. Calculate first-order differences for each i:
$$EO(i) - ED(i) \tag{12}$$
5. Calculate second-order differences for each i:
$$\{EO(i+1) - ED(i+1)\} - \{EO(i) - ED(i)\} \tag{13}$$
6. Weight first- and second-order differences according to their position on the ERB_N-number scale:
$$\begin{array}{rcll} W(i) & = & 1, & i < 17.5 \\ W(i) & = & 1 - w_s(i - 17.5)/46, & i \geq 17.5 \end{array} \tag{14}$$
Here, w_s is a third free model parameter.
7. Calculate standard deviations (SD) for first- and second-order differences.
8. Calculate weighted sum of standard deviations for first- and second-order differences:
$$\begin{array}{rcl} D & = & w \cdot SD\left(W(i)(EO(i) - ED(i))\right) + \\ & & (1-w) \cdot SD\left(W(i)\left(EO(i+1)\right.\right. \\ & & \left.\left. -ED(i+1) - EO(i) + ED(i)\right)\right) \end{array} \tag{15}$$
Here, w is the fourth and last free model parameter, and D is the final model output.

In another step, the subjective test data S are transformed according to:

$$T = 2 \cdot \arcsin\left(\sqrt{(S - \min(S)) \,/\, (\max(S) - \min(S))}\right) \tag{16}$$

A linear relation was found between the transformed listening test results T and the model predictions D. By curve-fitting of the test results, the following settings were derived in [7] for the four free model parameters: $s = 1.5$, $g = 32$ dB, $w_s = 0.5$, and $w = 0.4$.

Model performance comparison

In spite of the fact that the two presented models both were developed for assessing audio signals under linear distortion, the models differ in three key aspects:

- Predicted quality measures: The I_{bw}-model predicts a quality impairment on the E-model WB-extended transmission rating scale, while the naturalness model predicts perceived naturalness.
- Target application: The I_{bw}-model was developed for wide- and narrowband speech (monaural presentation), the naturalness model for up to fullband speech and music (diotic presentation).
- The degree to which auditory perception is modelled: The I_{bw}-model only employs a hearing-related frequency scale, while the naturalness model employs aspects of auditory perception more explicitly.

In addition to the speech- and music-naturalness tests conducted by Moore and Tan [6] to develop their naturalness model, two further tests are described in [7], which were conducted for model validation. These tests are used in the following for comparing the performance of the two models.

Listening tests

The three databases available from [6, 7] both in terms of naturalness ratings and test samples are:

1. Model development test from [6], as described above: 168 conditions.
2. Model validation test (a) [7]: 57 conditions, including 43 realistic transfer functions from actual telephones (provided for the studies in [6, 7] by Nokia Corporation), and 14 conditions as already used in the model development test 1.
3. Model validation test (b) [7]: 63 conditions, including all conditions from validation test (a) (test 2.), and six additional conditions reflecting loudspeaker playback captured on axis.

As for test 1., all tests were conducted once for speech and once for music, using headphones with diotic presentation, and collecting naturalness ratings from the subjects on a 10-point naturalness scale.

Performance and Conclusions

We have applied the I_{bw}-model to the source and processed speech and music files corresponding to tests 1.-3. To extract the bandwidth z_{bw} and the center of gravity z_G (for deriving the center frequency f_c, see Equations (6)-(10)), we have derived an estimate of the gain-function characterizing the filter, based on the power density spectrum Φ_{xx} of each input signal $x(k)$, and the cross-power-density spectrum Φ_{xy} of $x(k)$ and the output signal $y(k)$ (see [9]).

In Figure 1, the predictions obtained with the I_{bw}-model, Equation (11), are compared with the subjective naturalness results. The key observations are:

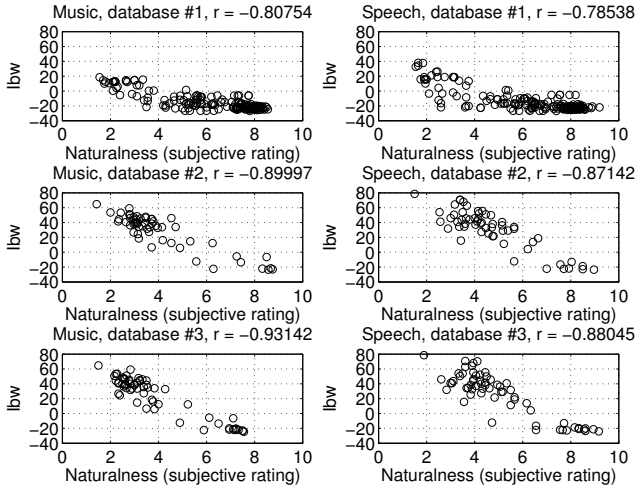

Figure 1: Comparison of I_{bw}-model predictions with the subjective test results of tests 1. to 3.; the figures in each line correspond to one test (1.-3.); the figures in the left column show the results for music, the figures in the right column the results for speech. The correlations between predictions and test results are given in the title of each plot.

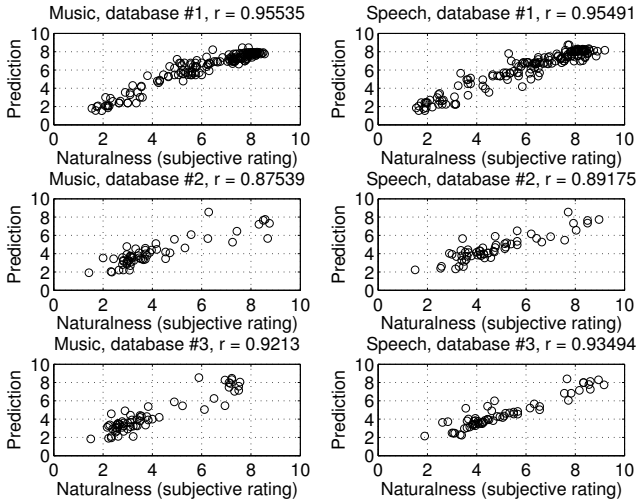

Figure 2: Comparison of naturalness model predictions with the subjective test results of tests 1. to 3. [6, 7].

- The model predictions are quite highly correlated with the subjective ratings.
- Performance of the I_{bw}-model is worst for test 1.: The model was developed for bandpass filters; it predicts naturalness less well for artificial ripple.
- The untrained model performs better for tests 2. and 3.: The real-life response-modifying filters used here are more similar to the ones used for model development.
- The model shows an even higher performance for music than for speech, indicating a validity of the I_{bw}-approach for general audio.
- I_{bw} often has negative values, since it ranges from 0 to 129 in case of WB speech (50-7000 Hz). Instead, the test database contains files up to fullband speech and music: Negative values of I_{bw} indicate a quality improvement over WB, and the values at highest naturalness of $I_{bw} \approx -30$ imply that corresponding quality ratings on the E-model scale will lie around $R_{0,max} \approx 160$. This indicates an advantage of about 30% of FB over WB (expressed on the NB-scale of the initial E-model with $R_{0,max} = 100$), in addition to the 30% advantage of WB over NB.

Figure 2 shows that the more perception-oriented naturalness model performs better than the I_{bw} model, especially for its training data set (test 1.). The correlations for tests 2. and 3. are still higher than for the I_bw model, but more similar in terms of range. In summary, it can be said that the I_{bw} model is suitable for a simple modeling of linear distortions in real-life speech- and audio-transmission scenarios.

In future work, we plan to retrain the I_{bw} model, and to extend it to include further technical parameters such as the slope of the gain function in the passband. This and other parameters are already considered in our previous work [9], but are used only implicitly in the simple model employed for this paper. Moreover, we are continuing our work on the WB E-model extension, and the E-model scale-extension for super-wideband and fullband.

References

[1] ITU–T Rec. G.107. The E-Model, a Computational Model for Use in Transmission Planning. International Telecommunication Union, Geneva, CH, 2008.

[2] D. Gibbon, R. Moore, R. Winski. Handbook on Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, Germany, 1997.

[3] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, M. Wältermann. Impairment Factor Framework for Wideband Speech Codecs. IEEE Trans. Audio Speech and Lang. **14**, 2006.

[4] B.C.J. Moore, B.R. Glasberg. Derivation of Auditory Filter Shapes from Notched-noise Data. Hear. Res. **47**, 1990.

[5] B.C.J. Moore, B.R. Glasberg, T. Baer. A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. J. Audio Eng. Soc. **45**, 1997.

[6] B.C.J. Moore, C.-T. Tan. Perceived Naturalness of Spectrally Distorted Speech and Music. J. Acoust. Soc. Am. **114**, 2003.

[7] B.C.J. Moore, C.-T. Tan. Development and Validation of a Method for Predicting the Perceived Naturalness of Sounds Subjected to Spectral Distortion. J. Audio Eng. Soc. **52**, 2004.

[8] A. Raake. Speech Quality of VoIP – Assessment and Prediction. John Wiley & Sons Ltd, Chichester, UK, 2006.

[9] K. Scholz, M. Wältermann, L. Huo, A. Raake, S. Möller, U. Heute. Estimation of the Quality Dimension "Directness/Frequency Content" for the Instrumental Assessment of Speech Quality. In: Proc. INTERSPEECH 2006, Pittsburgh, USA, 2006.

[10] M. Wältermann, A. Raake. Towards a new E-Model Impairment Factor for Linear Distortion of Narrowband and Wideband Speech Transmission. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, USA, 2008.

[11] E. Zwicker and H. Fastl. Psychoacoustics: Facts and Models. Springer, Berlin, DE, 1999.