OPEN ACCESS Freely available online

T. Biberger,1 J.-H. Fleßner, and S. D. Ewert,
"Audio Quality Prediction of Non-Waveform Preserving Distortions in Realistic Complex Scenes,"
J. Audio Eng. Soc., vol. 73, no. 11, pp. 707–721 (2025 Nov.).
https://doi.org/10.17743/jaes.2022.0236.

Audio Quality Prediction of Non–Waveform Preserving Distortions in Realistic Complex Scenes

THOMAS BIBERGER,[1,*] JAN-HENDRIK FLEßNER,[1,2] AND STEPHAN D. EWERT[1]
(thomas.biberger@uni-oldenburg.de) (hendrik.flessner@uni-oldenburg.de) (stephan.ewert@uni-oldenburg.de)

[1]*Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, Oldenburg, Germany*
[2]*Hörtech gGmbH and Cluster of Excellence Hearing4All, Oldenburg, Germany*

Intrusive audio quality models typically compare "internal representations" of a reference and a test signal. These models are often optimized for the prediction of small signal degradations, where the test and reference signals are still highly correlated (waveform preserving distortions). However, differences between uncorrelated signals like two Gaussian-noise tokens or, for example, more complex, realistic signals in spatial audio reproduction schemes, that show only partial correlation (non–waveform preserving distortions) are not necessarily easy to distinguish by listeners. Despite this, current audio quality models typically predict large perceptual differences between such signals. Here, the decision back-end of a reference-based audio quality model was modified to account for this overestimation of signal quality differences. The suggested modifications were intended to effectively mimic short-term memory limitations by analyzing similarities in the differences between the internal representations of reference and test signals across time frames, auditory channels, and modulation channels. The modified model was evaluated with data based on different audio reproduction and room simulation methods and was compared to other state-of-the-art audio quality models. Results support the need for modifications of state-of-the-art audio quality models to accurately predict the perceptual effects of non–waveform preserving distortions.

0 INTRODUCTION

Subjective assessment of audio quality is often conducted using pairwise comparisons of a reference signal and one or more test signals. Two example methods are the International Telecommunication Union recommendations [1] and [2] [MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)], recommended for the "assessment of small impairments in audio systems" and "assessment of intermediate quality level of audio systems," respectively. Test signals, including a hidden reference and anchor, are compared to a reference, and listeners can switch freely between them as often as they like. It is possible to repeat the signals in loops with a minimum duration of 500 ms.

Similar to such a perceptual procedure, intrusive audio quality models also compare a test and reference signal. State-of-the-art audio quality models (e.g., [3–8]) often exploit auditory models or processing stages to compute internal representations of the audio signals, meant to reflect the relevant perceptual cues. The internal representations of the reference signal and test signal are compared using a distance or correlation measure, typically considering all details of the internal representation (e.g., on a sample-by-sample basis). This assumes accurate short-term memory of the listeners. However, the signals are typically presented to the listeners successively, so that the listeners need to store the signal information in memory to perform the comparison. It is unlikely that listeners are able to store and access each "sample" of the internal representation in an optimal way, as often assumed in the models.

Accordingly, perceptually motivated audio processing technologies take advantage of the inability of listeners to perceive every detail of an audio signal. The processed signals are often non–waveform preserving. For example, the spectral band replication technology [9] used for low bitrate coding reconstructs the high-frequency components of an audio signal mainly by replicating low-frequency components. Listening tests showed very high audio quality of spectral band replication–encoded signals. However, a detailed comparison of the original and encoded signals would reveal substantial differences. Another example technology is the spatial decomposition method [10], which is used to

*To whom correspondence should be addressed, email: thomas.biberger@uni-oldenburg.de.

encode spatial room impulse responses based on acquisition via microphone arrays. The method computes a pressure value and spatial location for each sample of the room impulse response. The spatial location is determined by the average spatial location of the arriving sounds at a given time, leading to clear deviations from the original room impulse response. However, perceptual evaluation of the method shows that it is very difficult for listeners to distinguish between signals convolved with the original room impulse response and with the encoded room impulse response.

Goossens et al. [11] investigated the ability to discriminate Gaussian-noise tokens from one another and random tone-burst complexes from one another. The experiments showed a decreased ability to discriminate Gaussian-noise tokens with increasing duration above 40 ms. The authors concluded that the higher amount of information in a signal with longer duration combined with limitations inherent to the auditory short-term memory are responsible for the decrease in discrimination performance. The access to a small segment of the signals seems to be impossible—a signal is rather perceived and stored in the form of auditory objects.

Wallaert et al. [12] accounted for amplitude modulation detection thresholds for low amplitude modulation rates and stimulus durations above 1–2 s by modeling limitations of auditory short-term memory. It appears plausible that for complex natural signals such as those involving multiple sound sources, reverberation, or extended durations, similar processing limitations come into play when differences are perceptually rated. This hypothesis is supported by the results of Brinkmann et al. [13]. In their study, the authenticity of individual dynamic binaural synthesis was subjectively rated highest (fewest differences detected) for reverberant conditions and rated lowest for anechoic conditions.

Based on this evidence, intrusive audio quality models that compare (internal representations of) reference and test signals in a detailed way with "perfect memory" are likely to overestimate perceived differences between specific, complex scenario signals that are non–waveform preserving. Although the capabilities of short-term auditory memory seem to be limited, based on the work presented above, several studies (e.g., [14–16]) showed high short-term auditory memory performance: These studies indicate that humans can learn random temporal sequences if the sequences are repeated. However, this task differs from typical audio quality assessments, where two signals are compared and differences are rated.

Recently, Delgado and Herre [17] proposed a method for predicting the effects of non–waveform preserving distortions on audio quality perception by weighting the model output values of the Perceptual Evaluation of Audio Quality (PEAQ) method [3, 18] with the outputs of a cognitive salience model that included aspects of the nature of the signals (e.g., probability of speech). Delgado and Herre [17] trained and evaluated their method with signals subjected to audio coding distortions.

In the current study, it was investigated whether state-of-the-art audio quality models that employ the internal representation comparison approach are able to predict the quality of signals with complex, non–waveform preserving properties [e.g., when comparing reverberant signals based on measured and simulated binaural room impulse responses (BRIRs)]. Modifications to the monaural generalized power spectrum model for quality (GPSM^q) [8] backend were introduced to restrict the precision of the comparison of the extracted auditory features based on spectro-temporal signal properties. Such restrictions are purely based on the analysis of features provided by GPSM^q. This is different from the approach of Delgado and Herre [17], where not all metrics of the cognitive salience model are based on model features. Instead, in their approach, the probability of speech is directly estimated from the reference signal by using a speech-music classifier.

The hypothesis underlying the approach suggested in the current study is that a consistent difference between the auditory features of the reference and test signals (i.e., a difference that is similar over time or across frequencies) increases discrimination ability. A simple example of a difference with such spectro-temporal similarities would be a time-invariant boost of high frequencies, whereas replacing a measured room impulse response with a similar simulated one might introduce time-varying and inconsistent "noise-like" differences across frequencies.

Further modifications to GPSM^q were introduced: The "dual monaural" approach for GPSM^q [19] to reduce the sensitivity to binaural distortions was replaced with a processing stage that combines the binaural input signals earlier, so that GPSM^q mainly captures monaural distortions. This aspect is particularly important when combining monaural and binaural models for predicting overall quality. The suggested approach was evaluated and compared to four state-of-the-art quality models: PEAQ [3, 18], the perception model of quality assessment (PEMO-Q) [5], the audio quality model of Moore et al. [4], and GPSM^q [8], using spatial audio quality data. Given that the extension of Delgado and Herre [17] was not publicly available, only PEAQ without the extension was used in this study. To enable a fair comparison between the four monaural comparison models, none of them included a processing scheme to reduce their sensitivity to binaural distortions.

First, all models were calibrated with the data containing monaural quality degradations from Fleßner et al. [19]. The models were then applied to the spatial audio quality data of Grosse and van de Par [20] and Wendt [21]. Grosse and van de Par [20] subjectively compared their approach of a room recording and playback simulation with several other rendering methods, such as a conventional stereo setup or a multichannel setup. The signals contained different amounts of reverberation, and the subjective ratings included cases with only small perceived differences, although the waveform of the reference had not been preserved. Wendt [21] simulated three rooms with different methods, and the overall difference from the measured room was subjectively assessed. The waveforms of the measured and synthesized room signals differed considerably. It was further investigated how the monaural quality models predict the perceptual difference of two uncorrelated white noises and the difference between signals with interaural

time differences (ITDs) but without monaurally perceptible differences.

Aiming at the ultimate goal of a combined monaural and binaural audio quality, a binaural extension for the suggested monaural GPSM^q is also proposed. The binaural extension is based on the binaural auditory model for audio quality (BAM-Q) [7], adequately adapted to incorporate the modifications introduced here to the monaural model component. The same concepts applied in the modified GPSM^q were used in the suggested binaural extension. The combined approach was evaluated and compared to [19] and to a straightforward combination of the modified GPSM^q proposed here with the original binaural stage of [7], using the same spatial audio quality databases as used for the monaural quality models.

1 MODEL DESCRIPTION

First, a modified version of the monaural audio quality model GPSM^q [8] is described. The description focuses on the suggested back-end modifications. MATLAB implementations of GPSM^q and the suggested model are publicly available.[1] Finally, comparison models from the literature are briefly described.

1.1 GPSM^q

The monaural GPSM^q is based on the GPSM [22–23], which combines the power spectrum model [24–25] and envelope power spectrum model [26]. The combined model was shown to account for the results of various monaural psychoacoustic and speech intelligibility experiments. GPSM^q typically uses monaural reference and test signals as input. In this study, the left and right ear signals for binaural inputs were independently processed by GPSM^q (i.e., model features were separately calculated for the left and right ear signals; see Fig. 1).

First, a fourth-order gammatone filterbank [27–28] filters the input signal for each ear. The filterbank consists of 17 third-octave spaced filters with center frequencies from 315 Hz to 12.5 kHz and a width of one equivalent rectangular bandwidth (ERB_N) [29]. The third-octave filter spacing was chosen because of a slightly lower computational cost than the ERB-spacing, without presenting a significant performance disadvantage, as tested in [8]. The choice of 315 Hz as the center frequency of the lowest auditory channel is discussed in SEC. 5.4. The filter outputs are used to derive the Hilbert envelope E. Given that the left and right channels are processed separately, interaural level differences (ILDs) in the reference and test signals potentially interfere with monaural quality predictions.

Thus, in the next stage, compensation of long-term ILDs in each auditory channel is performed to reduce the ILD sensitivity of the monaural GPSM^q, following the concept in [19], so that the GPSM^q mainly captures monaural distortions. For this, the levels of test and reference signals in each left and right auditory channel are compared. The signal with the lower level is then adjusted to match the larger level (either test or reference) in each left and right auditory channel, resulting in the same long-term ILD for the adjusted test and reference signals. The ILD-compensated Hilbert envelopes E_{comp} are low-pass filtered at 150 Hz. The power SNRs (DC path; right path in Fig. 1) and envelope power SNRs (left path in Fig. 1) are derived from these outputs E_{lp}.

The first step of the DC path is the calculation of the local DC power $P_{\text{DC},j}(p)$ in 12.7-ms frames (in contrast to 375-ms frames in [8]), where $\bar{E}_{\text{lp},j}(p)$ refers to the mean value calculated from the samples within the 12.7- ms frame. The

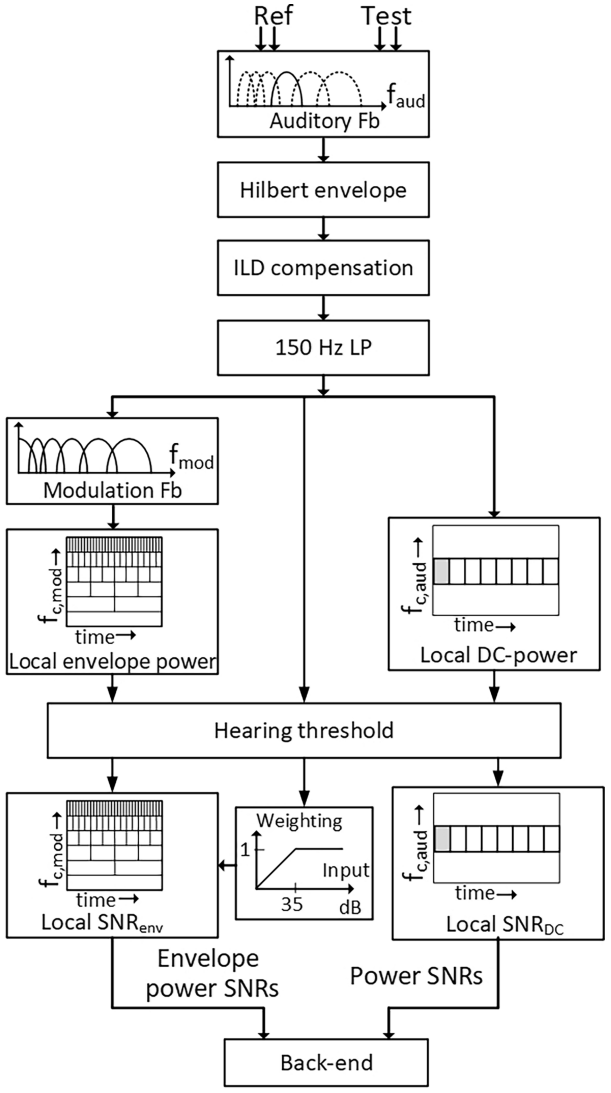


Fig. 1. Block diagram of the front-end of the monaural quality model. The right path shows the power processing, whereas the envelope power processing is shown in the left path. In contrast to [8], the input of the model is binaural and all depicted stages exist for the left and right ears. The model includes an ILD compensation processing stage to reduce the ILD sensitivity. Finally, the back-end combines the power and envelope power SNRs of the left and right ears.

[1] MATLAB implementations of GPSM^q and the suggested model can be found at www.faame4u.com. The suggested model can be downloaded from Zenodo: https://doi.org/10.5281/zenodo.17201035.

frame length of 12.7 ms was empirically determined from results of the synthetic stimuli experiment (see SEC. 2.2.1). Rectangular windows j (with 33.3 % overlap) are used in each channel p:

$$P_{\mathrm{DC},j}(p) = \frac{\left[\overline{E_{\mathrm{lp,j}}}(p)\right]^2}{2}. \tag{1}$$

The hearing threshold [30] is applied by setting local DC power values below threshold to the threshold value. The power SNRs are calculated from the local DC power for the increment and decrement (reflecting cases where the test signal has more and less power than the reference signal, respectively). For the left channel:

$$\mathrm{SNR}_{\mathrm{DC,left,inc},j}(p) = \frac{P_{\mathrm{DC,left,test},j}(p) - P_{\mathrm{DC,left,ref},j}(p)}{P_{\mathrm{DC,left,ref},j}(p)}, \tag{2}$$

$$\mathrm{SNR}_{\mathrm{DC,left,dec},j}(p) = \frac{P_{\mathrm{DC,left,ref},j}(p) - P_{\mathrm{DC,left,test},j}(p)}{P_{\mathrm{DC,left,test},j}(p)}. \tag{3}$$

The same processing steps are applied to the right channel. The SNR values are limited by an empirically determined range with a lower and upper bound of 0 and 20 (∼13 dB), which provided the best choice in terms of audio quality predictions, as described in [8]. Thus, increments and decrements span a dynamic range of 26 dB, which is similar to typical values utilized in speech intelligibility models (e.g., ESII) [31].

The envelope path (left path in Fig. 1) starts with a modulation filterbank comprising band-pass filters with center frequencies from 2 to 128 Hz (in contrast to 64 Hz in [8]) and a Q-value of 1. A modulation channel is disregarded if its center frequency is higher than one-fourth of the center frequency of the corresponding auditory filter [32]. The modulation filter outputs e are used to calculate the local envelope power (normalized by the local DC power) in each auditory channel p, modulation channel n, and temporal frame i:

$$P_{\mathrm{env},i}(p,n) = \frac{1}{\left[\overline{E_{\mathrm{lp,i}}}(p,n)\right]^2/2}\overline{[e_i(p,n,t) - \overline{e_i}(p,n)]^2}. \tag{4}$$

The length of each temporal frame i is determined by the inverse of the corresponding modulation filter center frequency (e.g., a modulation filter with center frequency of 8 Hz has a frame length of 125 ms). The hearing threshold is applied by setting the local envelope power to the envelope power limit (–30 dB) if the local DC power is below the hearing threshold. The envelope power SNRs are calculated from the local envelope power for the increment and decrement and for the left and right channels in the same way as for the power SNRs, Eqs. (2) and (3). The dynamic range of the SNR values is limited by applying the same interval as for the power SNRs, ranging from 0 to 20 as suggested in [8]. A linear weighting (from 0 to 1) of local envelope power SNRs is applied for auditory channels with local DC power between 0 and 35 dB. This lowers the contribution of envelope power SNRs for levels below 35 dB (e.g., in an auditory channel with local DC power of 17.5 dB, the local envelope power SNRs are multiplied by a weighting factor of 0.5).

The back-end used here is significantly different to that described in [8], as it is aimed at effectively mimicking short-term memory limitations by analyzing spectro-temporal regularities in the differences between the internal representations of reference and test signals. The suggested back-end modifications are not directly based on perceptual or physiological processes but represent a functional simulation of the effects of limited auditory short-term memory.

First, to reduce the two monaural paths to one path, the (power and envelope power) SNRs are averaged across the left and right channels in each time frame (in contrast to the "dual monaural" approach in [19] where the quality measures for the left and right channels were averaged). Consequently, the similarity analysis does not need to be calculated separately for the left and right channels. Such averaging of the SNRs for the left and right ears assumes that the auditory images of the left and right ears are combined into one auditory image. For the DC path, the differences between the increment and decrement power SNRs are calculated to provide a joint representation of increments and decrements for each time frame j over which a running average is calculated using 250-ms temporal windows tw (the window length corresponds to $L = 32$ SNR frames). Hereby, the 250-ms windows lead to a temporal smoothing of distortion statistics. The window duration has been empirically adjusted to account for the decreased ability of listeners to discriminate longer Gaussian-noise tokens (e.g., [11]) using the synthetic stimuli (see SEC. 2.2.1):

$$\Delta\mathrm{SNR}_{\mathrm{DC},tw}(p) = \frac{1}{L}\sum_{l=0}^{L-1}\left(\mathrm{SNR}_{\mathrm{DC,inc},j+l}(p) - \mathrm{SNR}_{\mathrm{DC,dec},j+l}(p)\right). \tag{5}$$

Then a correlation matrix based on the averaged SNR values is calculated to estimate the consistency of the power SNRs:

$$C_{\mathrm{DC}} = \begin{bmatrix} & \rho_{12} & \rho_{13} & \rho_{14} & \cdots & \rho_{1W} \\ \rho_{21} & & \rho_{23} & \rho_{24} & \cdots & \rho_{2W} \\ \rho_{31} & \rho_{32} & & \rho_{34} & \cdots & \rho_{3W} \\ \rho_{41} & \rho_{42} & \rho_{43} & & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & & \rho_{(W-1)W} \\ \rho_{W1} & \rho_{W2} & \rho_{W3} & \cdots & \rho_{W(W-1)} & \end{bmatrix} \tag{6}$$

with the entries being Pearson correlation coefficients:

$$\rho_{xy} = \frac{1}{M-1}\sum_{p=1}^{M}\left(\frac{X_p - \mu_x}{\sigma_X}\right)\left(\frac{Y_p - \mu_Y}{\sigma_Y}\right). \tag{7}$$

The correlation coefficients ρ_{xy} are calculated between $\Delta\mathrm{SNR}_{\mathrm{DC},tw}(p)$ in different time windows tw [$X_p = \Delta\mathrm{SNR}_{\mathrm{DC},tw=x}(p)$, $Y_p = \Delta\mathrm{SNR}_{\mathrm{DC},tw=y}(p)$]. The correlation matrix C_{DC} consists of correlations across auditory channels for different combinations of time windows ($x = 1...\mathrm{W}$, $y = 1...\mathrm{W}$). It provides an estimate of the sim-

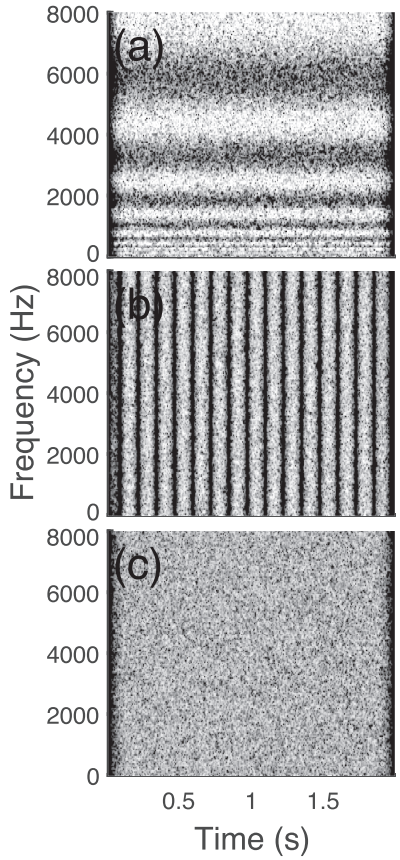


Fig. 2. (a)–(c) are intended to visualize the dependence of the weightings w_{DC} and w_{env} on spectro-temporal regularities in the SNR patterns. The x-axis and y-axis represent time and frequency, whereas white and black colors represent higher and lower sound levels, respectively. Spectral modulations (i.e., variations in level of a sound across the frequency spectrum), as shown in (a), result in high w_{DC} but rather low w_{env} values, while SNR patterns with temporal modulations (i.e., variations in level of a sound over time), as shown in (b), result in high w_{env} but rather low w_{DC} values. Irregular SNR patterns, as occur for two independent Gaussian-noise tokens, are shown in (c) and lead to low w_{DC} and w_{env} values. The higher weighting for SNR patterns with spectro-temporal regularities effectively increases the sensitivity to signal distortions.

ilarity of the SNR patterns [i.e., regularities in the spectral envelope differences as shown in Fig. 2(a)] in all time windows. Negative correlations ($\rho_{xy} < 0$) are set to zero to avoid cancelation of negative and positive values in the averaging process, which would distort the result. A weighting factor w_{DC} is computed by averaging the correlation matrix C_{DC} across the two dimensions. Hence, the weighting decreases if the power SNR pattern (across auditory channels) changes over time (i.e., if there are temporal irregularities in the spectral envelope). The final DC path computation step is

$$\Delta\mathrm{SNR}_{\mathrm{DC}} = \left(\frac{1}{M}\sum_{p=1}^{M}\sqrt{\sum_{tw=1}^{W}\Delta\mathrm{SNR}_{\mathrm{DC},tw}^{2}(p)}\right)\cdot w_{DC}^{\alpha}. \quad (8)$$

The exponent α is a free parameter.

Likewise, the difference between the increment and decrement is calculated for the envelope SNRs:

$$\Delta\mathrm{SNR}_{\mathrm{env},i}(p,n) = \mathrm{SNR}_{\mathrm{env,inc},i}(p,n) - \mathrm{SNR}_{\mathrm{env,dec},i}(p,n). \quad (9)$$

For the calculation of the modulation-based consistency weighting measure, $\Delta\mathrm{SNR}_{\mathrm{env},i}(p,n)$ is summed over frequency channels and modulation channels, respectively:

$$\Delta\mathrm{SNR}_{\mathrm{env,FreqSum},i}(n) = \sum_{p=1}^{M}\Delta\mathrm{SNR}_{\mathrm{env},i}(p,n) \quad (10)$$

$$\Delta\mathrm{SNR}_{\mathrm{env,ModSum},i}(p) = \sum_{n=1}^{N}\Delta\mathrm{SNR}_{\mathrm{env},i}(p,n). \quad (11)$$

Similar to the computation of the correlation matrix C_{DC}, a correlation matrix is built for $\Delta\mathrm{SNR}_{\mathrm{env,FreqSum},i}(n)$ and for $\Delta\mathrm{SNR}_{\mathrm{env,ModSum},i}(p)$. In contrast to the power SNR case, the correlation coefficients ρ_{xy} are calculated between time windows in either different auditory or different modulation channels. Thus, the correlation matrices $C_{\mathrm{env,Freq}}$ and $C_{\mathrm{env,Mod}}$ estimate the similarity of the temporal evolution across auditory and modulation channels, respectively. Accordingly, for comodulated SNR patterns, as shown in Fig. 2(b), elements of the correlation matrix have high values. A consistency weighting $\mathrm{w}_{\mathrm{env}}$ is then calculated by averaging the two correlation matrices $C_{\mathrm{env,Freq}}$ and $C_{\mathrm{env,Mod}}$ separately and averaging the two resulting average consistency measures. Hence, the weighting w_{env} decreases if the temporal evolution of the envelope SNRs changes across auditory or modulation frequency. A visualization of the relationship between w_{DC} and w_{env} and their sensitivity to different specto-temporal regularities in the SNR pattern is provided in Fig. 2.

The final envelope path computation step is

$$\Delta\mathrm{SNR}_{\mathrm{env}} = \left(\frac{1}{I}\sum_{i=1}^{I}\sum_{n=1}^{N}\left|\Delta\mathrm{SNR}_{\mathrm{env,FreqSum},i}(n)\right|\right)\cdot w_{\mathrm{env}}^{\beta}. \quad (12)$$

The values of the free parameters $\alpha = 2$ in Eq. (8) and $\beta = 3$ in Eq. (12) were empirically determined with the monaural calibration database and a subset of the room-in-room database (see SEC. 2.2.2). The objective perceptual measure OPM_{Δ} combines the power path and envelope power path:

$$\mathrm{OPM}_{\Delta} = 10\log_{10}(\Delta\mathrm{SNR}_{\mathrm{DC}} + \Delta\mathrm{SNR}_{\mathrm{env}}). \quad (13)$$

The OPM_{Δ} is transformed by linear regression to a MUSHRA [2] scale using the monaural calibration database described in SEC. 3:

$$\mathrm{OPM}_{\Delta}^{*} = -3.23\cdot\mathrm{OPM}_{\Delta} + 69.84. \quad (14)$$

1.2 Comparison Models

Four comparison models, partly publicly available, were applied. The first comparison model, PEAQ [3, 18] is recommended by the International Telecommunication Union

for most audio signal processing with a focus on low bitrate coded audio signals. PEAQ (basic version) is a combination of several (Fast Fourier Transform–based) models and computes several intermediate "submeasures" based on masking thresholds and comparison of internal representations, respectively. PEAQ submeasures include partial noise loudness, amount of linear distortion, relative frequency of audible distortion, changes in temporal envelope, noise detection probability, noise-to-mask ratio, and harmonic structure in the error signal. The submeasures are combined by an artificial neural network to produce a quality estimate.

The second comparison model is the audio quality model of Moore et al. [4]. It combines a model for linear distortion [33] and a model for nonlinear distortion [34]. The model for linear distortion computes long-term excitation patterns of the reference and test signals and compares them by computation of first-order and second-order differences. The differences are weighted depending on the ERB number [28]. The overall weighted excitation-pattern difference D is calculated by combining the first-order and second-order differences using a weighted sum of the standard deviation of the first-order and second-order differences. A transformation to linearize the relationship between D and subjective ratings is also included. The model for nonlinear distortion processes reference and test signals by an outer/middle ear filter and a gammatone filterbank. Short-term (30-ms frames) cross-correlations are computed between the outputs of the gammatone filterbank for the reference and test signals for lags from –10 to 10 ms. The maximum of the short-term cross-correlations is determined, and a power weighting is applied. The values are summed across filters and then averaged across frames to obtain R_{nonlin}, which is also transformed to linearize its relationship to subjective data.

The third comparison model, PEMO-Q [5], is based on the PEMO [35–37], which successfully accounts for psychoacoustic masking in various conditions. The processing stages in PEMO include a gammatone filterbank, hair-cell transduction, adaptation, and a modulation filterbank. The internal representations computed by PEMO are used in the PEMO-Q back-end to calculate the perceptual similarity measure (PSM). PSM is calculated by the power-weighted linear cross-correlation between the internal representations of the reference and of the test signal. Besides PSM, PEMO-Q computes two other quality measures, which were not used in this study as they showed lower correlations with the monaural calibration database described in SEC. 3.

Finally, GPSM^{q} [8] without the modifications described above was used as a comparison model. It should be mentioned that GPSM^{q} with dual-monaural approach [19] to reduce the influence of ILD differences in the reference and test signals on predicted monaural distortions was also tested for the databases considered in this study. Given that its prediction performance was very similar to that of GPSM^{q} [8], it was decided to only report results of the original model [8], referred to as GPSM^{q} (2018) in the following.

For all of the four comparison models, audio quality was calculated individually for the left and right ear channels of the binaural signals. The final quality measure was obtained by averaging the quality predictions across ears.

To increase the comparability of the different models, the outputs of the models were fitted to the subjective data of the monaural calibration database (see SEC. 3). For the calibration of PEAQ, PEMO-Q, and GPSM^{q} (2018) the best results were obtained with linear regression. The model of Moore et al. [4] was calibrated with the procedure described in [4] and [8]. The "calibrated" models are denoted PEAQ*, Moore et al.*, PEMO-Q*, and GPSM^{q}* (2018).

2 DATABASES

2.1 Calibration Database

In Fleßner et al. [19], a database with monaural, binaural, and combined manipulations was generated to investigate the relation between monaural and binaural aspects of audio quality. The monaural subset of this database was used in this study. The monaural manipulations were separated into linear distortions, nonlinear distortions, and their combination and were inspired by Moore et al. [4]. The linear distortions were generated by spectral ripples through sinusoidal modulation of the spectral envelope, and the nonlinear distortions were generated by raising the absolute value of the instantaneous amplitude of the broad-band time signal to a power (for details, see [19]). Additionally, anchor signals were generated by low-pass–filtering the reference signal (cut-off frequencies 3.5 and 7 kHz). The source signals included female speech, guitar, a pop song, and pink noise.

The subjective quality ratings were performed by 16 normal-hearing listeners with a method similar to MUSHRA [2]. The listeners could listen as long as they liked to the samples (duration: 10 s), switch between them seamlessly, and select loops (minimum loop duration: 1 s). This database is referred to here as the "monaural calibration database".

2.2 Evaluation Databases

2.2.1 Synthetic Signals

Signals with monaurally nearly indistinguishable differences (when listening to their full duration) between reference and test were used to assess the model predictions. One signal was uncorrelated white noise for the reference and test, each with a duration of 3 s. Goossens et al. [11] showed decreasing discriminability of Gaussian-noise tokens with increasing noise duration. Best discriminability was reported for durations between 16.1 and 102.4 ms. The longest duration tested was 409.6 ms, and it was very difficult for the listeners to discriminate these noise tokens. It was therefore assumed that noise signals with a duration of 3 s, as used here, would be very difficult to discriminate, thus resulting in high subjective audio quality scores. Another test signal was binaurally, but not monaurally, distinguishable [7] from the reference signal, referred to as "alternating ITD" in the following. It was generated by

introducing alternating ITDs (±400 μs in every second third-octave frequency band, with a smooth transition in the intermediate bands) to the source signals. The 3-s–long source signals included speech, pop music, pink noise, and guitar music.

2.2.2 Spatial Audio

Grosse and van de Par [20] suggested a method to reproduce recordings made in one room in a different playback room without changing the impression of the original sound source and room acoustics. The method employs two front loudspeakers and two dipole loudspeakers for the diffuse sound field in the playback room and was subjectively compared with a conventional stereo method, a surround method (4.0), and rendering of the source signals directly into the playback room without any room-influence of the recording room. The source signals were 12 anechoic recordings, including female speech, violin, guitar, clarinet, piano, saxophone, snare drum, trumpet, and a Beethoven piece. The sample duration was between 5 and 10 s.

Two recording rooms and two playback rooms were used for evaluation. The recording rooms had reverberation times of $T60{,}rec1 = 699$ ms and $T60{,}rec2 = 3040$ ms, and the playback rooms had reverberation times of $T60{,}play1 = 371$ ms and $T60{,}play2 = 697$ ms. The signals were assessed by 12 normal-hearing listeners with a method similar to MUSHRA [2]. The listening test was conducted with headphones, using BRIRs from the recording and playback rooms to obtain binaural signals. The recording room signals served as the reference, and the various reproduction methods generated the test signals. Additionally, an anchor signal was generated by low-pass filtering (3.5 kHz) of the reference signal. The listeners could listen as long as they liked to the samples. The samples started from the beginning when switching between them, and the listeners could not select loops. Therefore, only the first 3 s of the samples were used in the current study, since the listeners most likely used the initial seconds of the signals to form their assessment because of brief auditory storage [38]. This database is referred to as the "room-in-room database."

Wendt [21] subjectively evaluated the perceptually plausible room acoustics simulator [39, 40] using three choices of feedback-delay-network optimizations for the simulation of late reverberation. Signals convolved with measured BRIRs were compared to signals convolved with the synthesized BRIRs and presented over headphones. The measured BRIRs were obtained in three rooms with reverberation times of about 0.3, 2.5, and 4.7 s. The source signals were speech and a bass guitar. Eight normal-hearing listeners participated in the experiment. The task was to rate perceived differences between measured and synthesized rooms. Several attributes taken from the Spatial Audio Quality Inventory [41] were assessed, with the overall difference rating being of interest for this study. As for the data of [20], the first 3 s of the signals were used for the model evaluations. This database is referred to as "room simulation database."

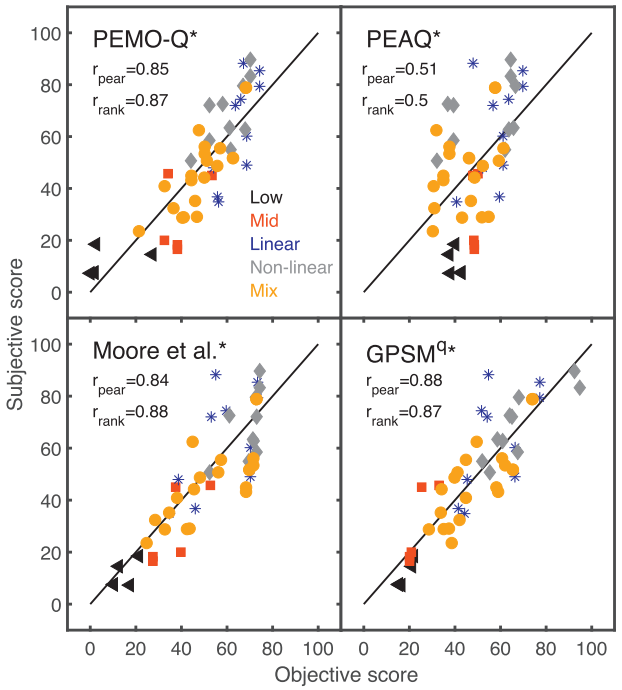


Fig. 3. Predictions of three (comparison) monaural quality models and the here-suggested model (lower right panel) for the monaural calibration database. Signals "Low" and "Mid" were low-pass–filtered anchor signals, "Linear" and "Non-linear" refer to the respective signal manipulations. The "Mix" signals had combinations of "Linear" and "Non-linear." The results ($r_{pear} = 0.86$) for the comparison model GPSM^q* (2018) are not shown here but were included in [19].

Both, the room-in-room database and room simulation database mainly included processed signals without preserved waveforms. A subset of the room-in-room database was used to determine the exponents of the reliability weightings of GPSM^q*, as described in SEC. 1. The subset included the signals of the recording room with the reverberation time of $T60{,}rec1 = 699$ ms in combination with the playback room with the reverberation time of $T60{,}play1 = 371$ ms.

3 QUALITY PREDICTIONS

First, the audio quality predictions are shown for the monaural calibration database, which can be thought of as the baseline. Second, the predictions for the synthetic test data and the room-in-room and room simulation datasets are compared.

3.1 Calibration Data

The audio quality predictions for the monaural calibration database are shown in Fig. 3. The signals "Low" and "Mid" were low-pass–filtered anchor signals. Signals "Linear" and "Non-linear" contained linear and nonlinear differences, respectively, and "Mix" contained combinations of "Linear" and "Non-linear." Prediction performance is characterized by the Pearson linear correlation coefficient r_{pear} and the Spearman rank correlation coefficient r_{rank} be-

Table 1. Comparison of the monaural model outputs (on a rating scale where 100 denotes "no difference" and 0 denotes "big difference") for signals with monaurally, perceptually nearly indistinguishable differences between reference and test signal. The outputs are averaged across four alternating ITD signals.

	Models				
Signals	PEMO-Q*	PEAQ*	Moore et al.*	GPSM^q* (2018)	GPSM^q*
White noise	0	32	34	33	100
Alternating ITD	51	66	43	74	100

tween subjective and objective quality scores. The highest correlation was found for the here-suggested GPSM^q*, with $r_{pear} = 0.88$ compared to $r_{pear} = 0.86$ for GPSM^q* (2018), as shown in [19]. A Hotelling-Williams Test [42] did not show a significant difference between the prediction performances of GPSM^q* and GPSM^q* (2018). PEMO-Q* and Moore et al.* also achieved good prediction performance indicated by $r_{pear} = 0.85$ and $r_{pear} = 0.84$, respectively. PEAQ* showed moderate prediction performance ($r_{pear} = 0.51$).

3.2 Synthetic Test Data

Table 1 shows the objective scores of the monaural models for the uncorrelated white noise and the alternating ITD signals. The scores for the ITD signals are averaged across the four signals. For both signals, the here-suggested, modified GPSM^q* predicted quality scores of 100, implying no perceptual (monaural) differences between reference and test signals. The Moore et al.* model, PEMO-Q*, PEAQ*, and GPSM^q* (2018) predicted lower quality scores. An informal listening test showed very high subjective rating scores, close to 100.

3.3 Spatial Audio

The results of the audio quality predictions for the room-in-room database are shown in Fig. 4. The PEMO-Q* and PEAQ* predictions were negatively correlated with the subjective scores. This is because, in contrast to the listeners' poor ratings for the anchor signals, both models predicted higher ratings for the anchor signals than for the other signals. The prediction performance of Moore et al.* was rather poor, as indicated by $r_{pear} = 0.33$, whereas GPSM^q* (2018) showed a somewhat better but still weak performance ($r_{pear} = 0.71$). As shown in Fig. 4, the quality of signals that received high subjective ratings was consistently underestimated by Moore et al.* and GPSM^q* (2018). Overall, the GPSM^q* predictions agreed well with subjective quality scores ($r_{pear} = 0.88$). This suggests a dominance of monaural quality differences like timbre or coloration. The results for the room simulation database (Fig. 5) resemble the results for the room-in-room database. The comparison models overestimated the signal differences for signals with subjective ratings near the top of the subjective rating scale. The predictions of the here-suggested GPSM^q* showed the highest correlation with subjective data ($r_{pear} = 0.96$).

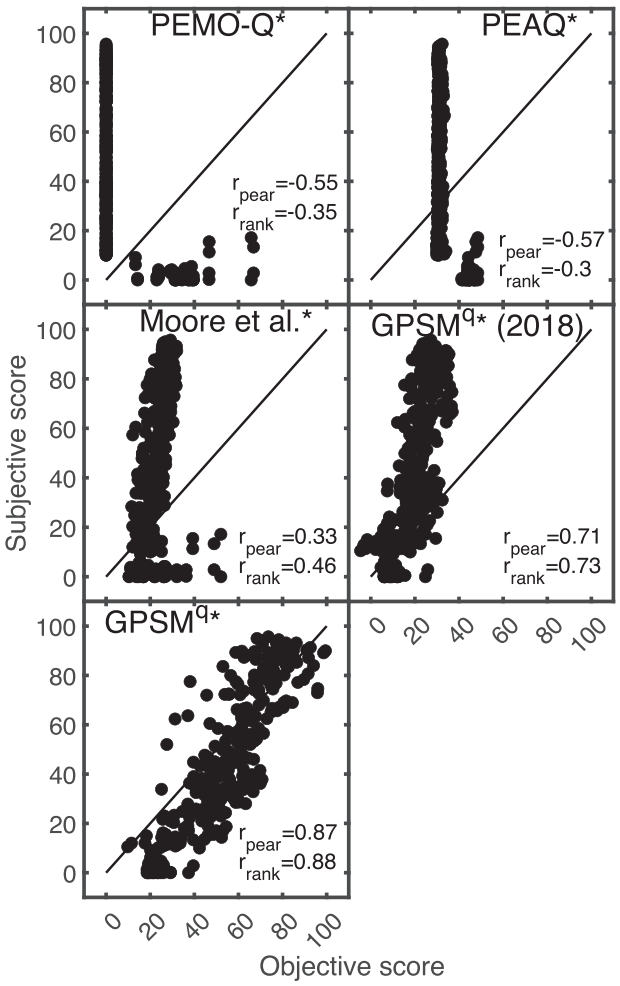


Fig. 4. Audio quality predictions of the comparison models PEMO-Q*, PEAQ*, Moore et al.*, GPSM^q* (2018), and the here-suggested GPSM^q* for the room-in-room database.

4 BINAURAL EXTENSION

For the assessment of monaural and binaural aspects of audio quality, an overall quality measure as proposed in [19], combining GPSM^q and BAM-Q outputs, is highly desirable. Toward such a combined model, the combination of the revised monaural and binaural parts potentially requires recalibration, and the binaural part requires changes determined by the modified monaural part.

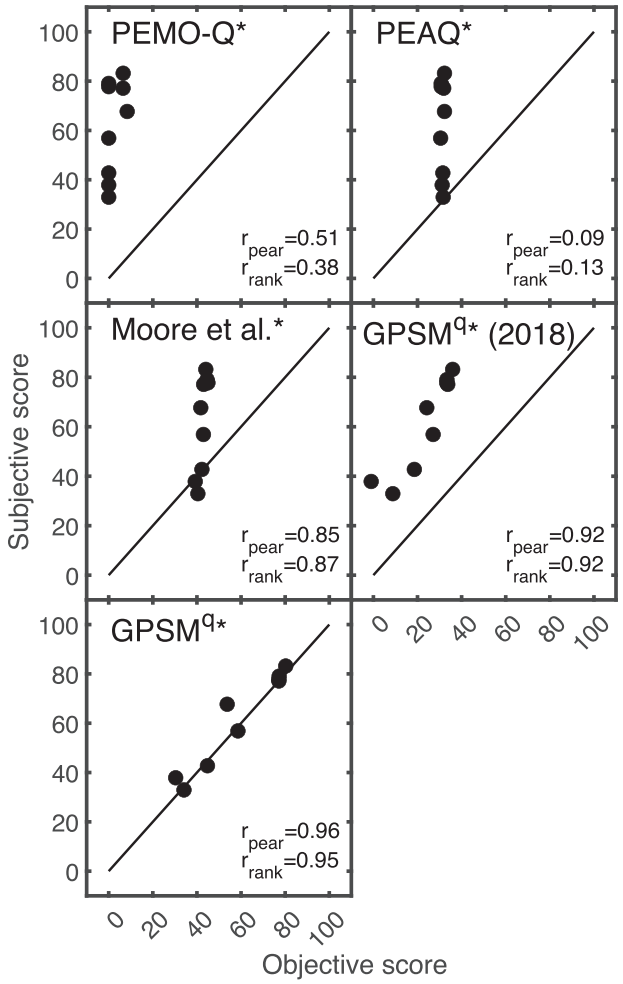


Fig. 5. Audio quality predictions of the comparison models PEMO-Q*, PEAQ*, Moore et al.*, GPSM^q* (2018), and the here-suggested GPSM^q* for the room simulation database.

4.1 Model Description

The BAM-Q is based on the binaural psychoacoustic model front-end of Dietz et al. [43], from which the submeasures ILD, ITD, and interaural vector strength were calculated. In the suggested modified BAM-Q, the concept of the consistency weighting stage, intended to effectively mimic short-term memory limitations by analyzing similarities in the differences between the internal representations of reference and test signals, as introduced in SEC. 1.1 for the GPSM^q*, was also applied to each of the three binaural submeasures. The outcomes for the submeasures were then combined by a nonlinear regression method providing the final output measure binQ*. A detailed description of the modified BAM-Q (BAM-Q*) and its calibration procedure are provided in the supplementary material.

Following [19], GPSM^q* and BAM-Q* outputs were combined using a minimum-based function,

$$Q_{\text{overall}} = \min\left(0.55 \cdot \text{OPM}_{\Delta},\ 0.53 \cdot \text{binQ}\right), \tag{15}$$

where OPM_{Δ} and binQ are the transformed outputs of the suggested, modified GPSM^q* and BAM-Q*, respectively.

4.2 Quality Predictions

The prediction performance of the proposed overall audio quality model GPSM^q* & BAM-Q* was compared to that of the original GPSM^q & BAM-Q (2019) [19] and the GPSM^q* & BAM-Q for the room-in-room [20] and room simulation [21] databases. The last mentioned version combines the modified GPSM^q* in a straightforward way with the original BAM-Q, so that differences with respect to the suggested GPSM^q* & BAM-Q* can be attributed to the current modifications to the binaural BAM-Q. As shown in Fig. 6, the modified GPSM^q* & BAM-Q* resulted in more accurate predictions than GPSM^q & BAM-Q (2019) and GPSM^q* & BAM-Q for both databases.

5 DISCUSSION

5.1 Calibration Data

The results for the monaural calibration database showed the highest correlations for the models following a rather straightforward approach (Moore et al.*, PEMO-Q*, and GPSM^q*). The PEAQ* measure is probably too strongly optimized for the prediction of low bit-rate–coded audio signals and does not account well for other sorts of quality differences. The correlation between Moore et al.* predictions and subjective scores was $r_{\text{pear}} = 0.84$, which is similar to the correlations (r_{pear} of 0.85 to 0.95) reported in [4]. In Biberger et al. [8], the model of Moore et al. [4] gave r_{pear} values ranging between 0.56 and 0.79 for the four tested databases (audio coding, audio source separation, noise reduction, and loudspeaker reproduction). In the current study, a higher correlation was expected, because the signal manipulations in the calibration database were similar to the signal manipulations in [4]. PEMO-Q* showed high correlations ($r_{\text{pear}} = 0.85$) for the monaural calibration database, although Huber and Kollmeier [5] noted that the model is not suited for the prediction of linearly distorted signals, while the calibration database contained highly linearly distorted signals.

Fleßner et al. [19] reported good prediction performance ($r_{\text{pear}} = 0.86$) of the original GPSM^q* (2018) for the monaural calibration database. Thus, the r_{pear} of 0.88 obtained using the here-suggested GPSM^q* for the monaural calibration database indicates that the modifications suggested in this study slightly improved the prediction performance for this database.

5.2 Spatial Audio Data

Subjective ratings of the signals in the room-in-room and room simulation databases were based on binaural signal presentations. The listeners did not specifically assess monaural or binaural audio quality, but overall audio quality. Therefore, it might seem inappropriate to compare the results of the listening test to the quality predictions of the monaural models. However, the results of Fleßner et al. [19] and Eurich et al. [44] indicate that overall quality is dominated by whichever quality aspect is lower, either monaural or binaural. The current results show a strong impact of monaural audio quality, since the suggested GPSM^q* pre-

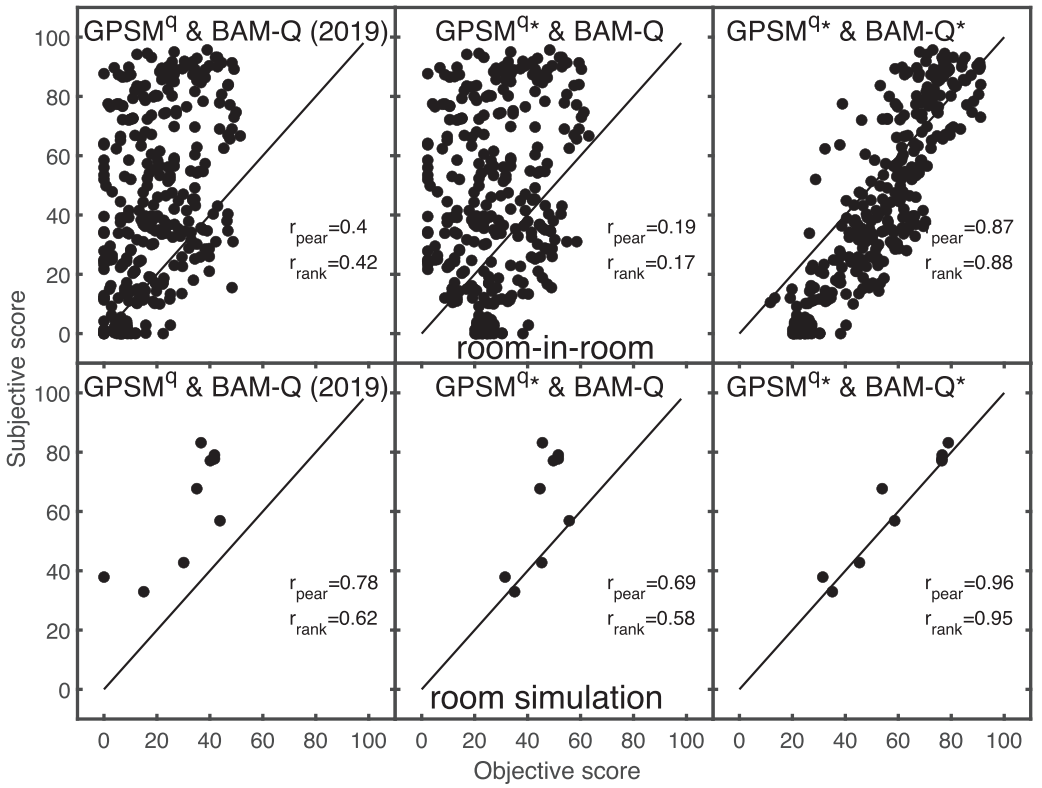


Fig. 6. Audio quality predictions of the overall audio quality models GPSM^{q} & BAM-Q (2019) [19], GPSM^{q}* & BAM-Q, and the proposed GPSM^{q}* & BAM-Q* for the room simulation (lower panels) and room-in-room (upper panels) databases.

dicted the subjective scores well. A similar observation was made in [45] for binaural signals with distortions that typically occur in smart headphones or hearables. Informal listening tests using example signals from the current study indicated that regardless of whether the signals were presented monaurally (only left or right channel active) or binaurally, the same overall quality was perceived by the listeners.

Prediction results showed that the state-of-the-art comparison models overestimated the differences between reference and test signals when the waveform was not preserved. For signals processed by the room-in-room reproduction methods, PEMO-Q* and PEAQ* predicted lower quality scores than for the anchor signals, which did not agree with the listeners' quality ratings. Thus, an alternative to the linear fitting would not improve the predictions. As shown in Fig. 4, PEMO-Q* did not predict the linear anchor signal distortions well.

The Moore et al.* model also showed poor prediction performance ($r_{\text{pear}} = 0.33$), as the perceptual differences between recorded and reproduced signals were overestimated. The submodel for linear distortions of Moore et al.* considers long-term differences of the excitation patterns (i.e., it does not use a detailed comparison, e.g., sample-by-sample, method). Hence, the submodel for nonlinear distortions is mainly responsible for the overestimation of signal differences, as it performs a sample-wise comparison of reference and test signals in short time frames.

Although the original, unmodified GPSM^{q}* (2018) also overestimated the perceptual differences between reference and test signals, it showed higher prediction performance ($r_{\text{pear}} = 0.71$) than the comparison models Moore et al.*, PEMO-Q*, and PEAQ*, which is statistically significant according to the Hotelling-Williams Test [42]. GPSM^{q}* (2018) uses less signal information than PEMO-Q*, as envelope power and power SNRs are based on a frame-wise calculation instead of a correlation-based (sample-by-sample) comparison. Nevertheless, the reduced amount of information available in GPSM^{q}* (2018) seems not sufficient to accurately predict audio quality for signals in the spatial audio databases.

As shown in Fig. 4, the suggested GPSM^{q}* sometimes also overestimated and underestimated the subjective quality ratings for the room-in-room database, but the deviations were less pronounced than for the comparison models, and accordingly, GPSM^{q}* showed better prediction performance ($r_{\text{pear}} = 0.87$, $r_{\text{rank}} = 0.88$). According to a Hotelling-Williams Test [42], the prediction performance of GPSM^{q}* was significantly different than that of GPSM^{q}* (2018) and the other comparison models.

The sensitivity of GPSM^{q}* was reduced for the spatial audio databases by the introduction of the model modifications without a decrease in prediction performance for the monaural calibration database. The consistency weightings introduced in GPSM^{q}* are different for power and the envelope power SNRs, with the weighting in the power SNR path representing the consistency of the power SNR pattern over time. Thus, the weighting would decrease if the power SNR pattern (across auditory channels) based on differences between reference and test signals changed over

time. For example, low-pass filtering of the reference signal (as, e.g., in the anchor signals) results in the same changes of power SNR pattern over time, so the weighting is high. In contrast, the application of the room-in-room reproduction method of [20] leads to varying power SNR patterns over time (i.e., irregular spectral envelope differences over time), and thus, the weighting is low.

The weighting in the envelope power SNR path includes two consistency measures, where the first measure examines the consistency of the temporal evolution of the envelope power SNRs across auditory channels, and the second measure examines the consistency of the temporal evolution of the envelope power SNRs across modulation channels. Differences in the pattern of the temporal envelope power SNRs across frequency and/or modulation channels would reduce the weighting of the envelope power SNR path. For example, the introduction of nonlinear distortions such as hard clipping or brief interruptions in signal transmission often increases envelope power SNRs across the whole range of frequency/modulation channels at the same time, resulting in similar evolutions of envelope power SNRs and a high consistency weighting. In contrast, the comparison of two Gaussian-noise tokens leads to envelope power SNR changes across (modulation) frequencies at different points in time, resulting in different evolutions of envelope power SNRs and a low consistency weighting.

The importance of across-auditory-channel processing has also been shown in other contexts: For speech intelligibility prediction, Chabot-Leclerc et al. [46] introduced the sEPSM^{X}, which includes a processing stage to weight the envelope power SNRs with the across-channel variance of the envelope power. In addition to marginal moments (mean, variance, skew, and kurtosis, respectively), McDermott and Simoncelli [47] employed correlations of temporal evolutions of modulation envelopes across auditory frequencies and across modulation frequencies for sound texture synthesis. An across-channel process relevant to the modeling of comodulation masking release in psychoacoustic experiments was implemented by Dau et al. [48].

The predictions for the synthetic stimuli confirmed the results for the spatial audio databases. PEMO-Q*, PEAQ*, GPSM^{q}* (2018), and the Moore et al.* approach predicted large differences between statistically independent noise tokens, while no differences (reflecting the results of informal listening tests) were observed for the suggested GPSM^{q}*. Differences in the synthetic stimuli without and with alternating ITDs are only perceivable when the signals are binaurally presented to the listeners but not when they are presented monaurally. Accordingly, monaural models should not account for such purely binaural differences, while most of the here-considered monaural models were sensitive to these ITD differences, as shown in Table 1.

5.3 Comparison With Sound Texture Perception

In the context of non–waveform preserving distortions, findings on sound texture perception may be relevant. Sound textures can have strong differences in their waveforms but sound very similar to listeners, as long as they show similar time-averaged signal statistics in the auditory periphery (e.g., [47]).

McWalter and McDermott [49] systematically investigated the perceptual integration of sound texture statistics by comparing textures with subtle shifts in stimulus statistics over time relative to a given reference texture (e.g., ocean waves, shaking coins, or rain in woods). They found that listeners appear to average texture statistics over a multisecond window, depending on the variability of the textures. Shorter integration windows of about 1.4-s duration were estimated for textures showing less variability (e.g., heavy rain, waterfall, applause), while longer windows of about 3.2-s duration were estimated for textures with higher variability (e.g., ocean waves, chewing carrots, galloping horses).

In this respect, the Gaussian noise used in this study for model evaluation (see SEC. 2.2.1) represents a signal with very low variability, and the relatively short integration time of 250 ms used in this study seems to be sufficient. However, it should be stressed that the consistency weightings in this study represent spectro-temporal similarities of signal differences (distortions) between reference and test signals, while in [49], the statistical similarities between sound textures were measured, which is conceptually different. On the other hand, the auditory texture model used in [49] to measure texture statistics had similar properties to the model used in this study to calculate consistency weightings. Applying the longer integration times found in [49] directly to the power, Eq. (1), and envelope-power, Eq. (4), cues used in the GPSM^{q}* would correspond to a more direct implementation of their findings, but would probably limit this approach to the prediction of perception of sound texture differences, as it does not provide any information about spectro-temporal regularities in the distortions.

In the study of McWalter and Dau [50], listeners showed poor abilities in discriminating between measured and synthesized sound textures like insects, campfire, helicopter, or frogs and crickets. The synthesized textures were generated by modifying Gaussian-noise to match the statistics of the recorded real-world textures, while the waveforms of recorded and synthesized textures were different. An informal test showed that the proposed modified GPSM^{q}*, despite its relatively short integration time of 250 ms, predicted smaller differences between recorded and synthesized textures (supplementary material provided in [50]) than the original GPSM^{q} (2018) [8] and, thus, better matched listeners' discrimination abilities for sound textures. With regard to the above-mentioned stimulus-dependent integration times for sound textures, it would be interesting to examine whether combinations of consistency weightings based on shorter and longer integration times can improve audio quality prediction performance for the types of distortions considered here. However, this is out of the scope of this work.

5.4 Limitations and Future Extensions

The suggested modifications to GPSM^{q}* are applied to distance metrics (e.g., SNR) to reduce the impact of dif-

ferences in auditory features with low consistency over time or auditory and modulation channels. This approach is not expected to directly reflect auditory processing in humans, given that in the current modeling approach, a difference of the internal representations of test and reference is available with a high amount of information, from which, in a subsequent processing step, the consistency weightings are calculated. This appears unlikely to happen in the auditory system, and an approach more strongly related to physiology would already reduce the information available in the internal representation, which would more directly reflect the effects of limited auditory short-term memory.

Wallaert et al. [12] used additive noise applied to the internal representation after modulation filtering to simulate the effect of limited auditory short-term memory. Such an approach would, however, not be directly applicable to state-of-the-art audio quality back-ends, since the comparison of internal representations of two Gaussian-noise tokens already leads to underestimation of audio quality (see SEC. 3.2). Thus, additive memory noise would lead to a further underestimation of audio quality. If an additive memory noise is used in the front-end, the audio quality back-end needs to be redesigned. For example, the back-end could compare specific characteristics and statistical measures of the internal representations, which is probably closer to physiology [47].

In summary, approaches that include the simulation of auditory short-term memory in the model front-end also seem to require an adjustment of the back-end. The suggested approach in GPSM^{q}* can be interpreted as a functional simulation of the effects of limited auditory short-term memory. The correlation matrices of the modifications give an estimate of the similarity of the distance measures across the different dimensions (time, auditory, and modulation channels), depending on the distance measure. If the similarity across a dimension is high, the distance measure receives a high weight. This could reflect the reduced information that has to be stored in auditory memory, since the information is repeated. If the similarity across a dimension is low, the auditory memory requirement is high, because different information has to be stored. Since auditory short-term memory is limited, not all information can be stored, and accordingly, the discriminability of the signals is reduced. Hence, the distance measure receives a low weight. The joint consideration and analysis of several auditory/modulation channels could also be regarded as a first step toward auditory object formation, where similar difference patterns across channels are likely to originate from one auditory object [51].

The suggested GPSM^{q}* back-end assumes that the auditory images of the left and right ears are combined into one auditory image by averaging the SNRs for the left and right ears. This assumption might be incorrect for some scenes, for example, if a listener is presented with different sound sources in the left and right ears via headphones. Thus, if multiple auditory objects are distinctly spatially separated, a quality measure based on individual, separated auditory objects might be required.

In the suggested GPSM^{q}*, the center frequency of the lowest auditory channel is 315 Hz, which is higher than the lowest auditory channel typically used in other audio quality models (e.g., [3–6, 52]). However, due to redundant information in neighboring auditory channels, a center frequency of 315 Hz often captured sufficient low-frequency differences, as shown in Fig. 5 in [44]. Nevertheless, depending on the sounds considered, it may be necessary in future experiments to extend the frequency range (with a lowest auditory channel center frequency of, e.g., 63 Hz) in the modified GPSM^{q}*. Accordingly, it might be also necessary to adapt integration times (e.g., use longer integration times for the low-frequency channels for the calculation of consistency weightings).

In this study, overestimation of signal differences occurred for signal assessments obtained without the ability to select and listen to signal loops of short duration (<500 ms). When it is possible to listen to small segments of signals, it is easier to distinguish the signals (e.g., [11]). This is probably the reason why the suggested GPSM^{q}* showed poorer prediction performance than the original GPSM^{q} (2018) for the database used in [8]. Therefore, it is probably necessary to adjust the proposed modifications according to the signal duration (e.g., adjust the weights of the modifications).

5.5 Binaural Extension

For both evaluation databases, the overall audio quality model GPSM^{q}* & BAM-Q* proposed here achieved better prediction performance than the original GPSM^{q} & BAM-Q (2019) [19] and GPSM^{q}* & BAM-Q. These differences were significant according to a Hotelling-Williams Test [42]. However, a comparison of Fig. 6 with Figs. 4 and 5 shows that for the tested databases, GPSM^{q}* & BAM-Q* never performed better than the GPSM^{q}* alone. Additionally, predictions based on the combination of GPSM^{q}* and BAM-Q without modifications were less accurate than those resulting from GPSM^{q}* & BAM-Q*. This observation reflects the fact that in the current room simulation and room-in-room databases, monaural aspects are dominant, but it also indicates that the modified binaural BAM-Q* is less sensitive to monaural audio quality aspects than the original BAM-Q. Accordingly, BAM-Q* made only a minor contribution to the overall audio quality predictions.

Although the binaural extension was not required for successfully predicting the types of distortions considered here, the suggested combined model provides a working revision of the overall audio quality model of [19]. Because of the negligible impact of binaural distortions in the current databases, the suggested binaural extension needs to be evaluated in more detail with purely binaurally distorted sounds.

6 CONCLUSION

Limitations of audio quality models that exploit the comparison of internal representations in the auditory pathway were investigated for two spatial audio databases and syn-

thetic stimuli, where reference and test signals typically showed strong differences in their waveforms. A modified monaural audio quality model, GPSM^q*, was suggested to cope with such non-waveform preserving distortions. The following conclusions can be drawn:

- State-of-the-art audio quality models fail to predict the audio quality of complex signals like spatially processed audio signals or Gaussian-noise tokens.
- Incorporation of a consistency weighting stage into the back-end of a monaural audio quality model improved prediction performance considerably for signals in spatial audio reproduction schemes and Gaussian-noise tokens. In the here suggested GPSM^q*, this additional stage is intended to effectively mimic short-term memory limitations by analyzing similarities in the differences between the internal representations of reference and test signals across time frames, auditory channels, and modulation channels.
- Further psychoacoustic experiments are required to investigate the relationship between the complexity of signals and their discriminability and to develop model modifications or fine-tune the suggested model modifications. In this respect, sound textures are an interesting class of signals.
- The tested monaural audio quality models were able to distinguish binaural signal differences that are monaurally not detectable and should therefore not be used in conditions in which binaural distortions occur.
- A binaural extension was suggested, applying a consistency weighting stage to the submeasures in the binaural audio quality model BAM-Q [7]. Combining the modified monaural and binaural audio quality models resulted in more accurate predictions than a combination without modifications.

7 ACKNOWLEDGMENT

The authors thank Brian C. J. Moore and an anonymous reviewer of an earlier version of this paper for their helpful comments. The authors would also like to thank the members of Medizinische Physik and Birger Kollmeier for continued support. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330 A2.

8 REFERENCES

[1] ITU-R, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," *Recommendation ITU-R BS.1116-3* (2015 Feb.).

[2] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," *Recommendation ITU-R BS.1534* (2014 Jun.).

[3] ITU-R, "Method for Objective Measurement of Perceived Audio Quality," *Recommendation ITU-R BS*. 1387 (1998 Dec.).

[4] B. C. J. Moore, C.-T. Tan, N. Zacharov, and V.-V. Mattila, "Measuring and Predicting the Perceived Quality of Music and Speech Subjected to Combined Linear and Nonlinear Distortion," *J. Audio Eng. Soc.*, vol. 52, no. 12, pp. 1228–1244 (2004 Dec.)

[5] R. Huber and B. Kollmeier, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1902–1911 (2006 Nov.). https://doi.org/10.1109/TASL.2006.883259.

[6] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI)," *J. Audio Eng. Soc.*, 58, vol. 5, pp. 363–381 (2010 Jun.).

[7] J.-H. Fleßner, R. Huber, and S. D. Ewert, "Assessment and Prediction of Binaural Aspects of Audio Quality," *J. Audio Eng. Soc.*, vol. 65, no. 11, pp. 929–942 (2017 Nov.). https://doi.org/10.17743/jaes.2017.0037.

[8] T. Biberger, J.-H. Fleßner, R. Huber, and S. D. Ewert, "An Objective Audio Quality Measure Based on Power and Envelope Power Cues," *J. Audio Eng. Soc.*, vol. 66, no. 7/8, pp. 578–593 (2018 Aug.). https://doi.org/10.17743/jaes.2018.0031.

[9] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding," presented at the *112th Convention of the Audio Engineering Society* (2002 May), paper 5553.

[10] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Jan./Feb.).

[11] T. Goossens, S. van de Par, and A. Kohlrausch, "On the Ability to Discriminate Gaussian-Noise Tokens or Random Tone-Burst Complexes," *J. Acoust. Soc. Am.*, vol. 124, no. 4, pp. 2251–2262 (2008 Oct.). https://doi.org/10.1121/1.2973184.

[12] N. Wallaert, B. C. J. Moore, S. D. Ewert, and C. Lorenzi, "Sensorineural Hearing Loss Enhances Auditory Sensitivity and Temporal Integration for Amplitude Modulation," *J. Acoust. Soc. Am.*, vol. 141, no. 2, pp. 971 (2017 Feb.). https://doi.org/10.1121/1.4976080.

[13] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the Authenticity of Individual Dynamic Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795 (2017 Oct.). https://doi.org/10.1121/1.5005606.

[14] H. Kang, T. R. Agus, and D. Pressnitzer, "Auditory Memory for Random Time Patterns," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 2219–2232 (2017 Oct.). https://doi.org/10.1121/1.5007730.

[15] N. Guttman and B. Julesz, "Lower Limits of Auditory Periodicity Analysis," *J. Acoust. Soc. Am.*, vol. 35, no. 4, pp. 610 (1963 Apr.). https://doi.org/10.1121/1.1918551.

[16] V. G. Rajendran, N. S. Harper, K. H. A. Abdel-Latif, and J. W. H. Schnupp, "Rhythm Facilitates the Detection of Repeating Sound Patterns," *Front. Neurosci.*, vol. 10, pp. 464–467 (2016 Jan.). https://doi.org/10.3389/fnins.2016.00009.

[17] P. M. Delgado and J. Herre, "Towards Improved Objective Perceptual Audio Quality Assessment-Part 1: A Novel Data-Driven Cognitive Model," *IEEE/ACM Trans.*

Audio Speech Lang. Process., vol. 32, pp. 4661–4675 (2024 Oct.). https://doi.org/10.1109/TASLP.2024.3477291.

[18] T. Thiede, W. C. Treurniet, R. Bitto, et al., "PEAQ - The ITU-Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29 (2000 Jan./Feb.).

[19] J.-H. Fleßner, T. Biberger, and S. D. Ewert, "Subjective and Objective Assessment of Monaural and Binaural Aspects of Audio Quality," *IEEE Trans. Audio Speech Lang. Process.*, vol. 27, no. 7, pp. 1112–1125 (2019 Jul.). https://doi.org/10.1109/TASLP.2019.2904850.

[20] J. Grosse and S. van de Par, "Perceptually Accurate Reproduction of Recorded Sound Fields in a Reverberant Room Using Spatially Distributed Loudspeakers," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 867–880 (2015 Aug.). https://doi.org/10.1109/JSTSP.2015.2402631.

[21] T. Wendt, *A Computationally Efficient and Perceptually Plausible Room Acoustics Simulation Method*, Ph.D. thesis, Carl-von-Ossietzky Universität Oldenburg, Oldenburg, Germany (2018 Feb.).

[22] T. Biberger and S. D. Ewert, "Envelope and Intensity Based Prediction of Psychoacoustic Masking and Speech Intelligibility," *J. Acoust. Soc. Am.*, vol. 140, no. 2, pp. 1023–1038 (2016 Aug.). https://doi.org/10.1121/1.4960574.

[23] T. Biberger and S. D. Ewert, "The Role of Short-Time Intensity and Envelope Power for Speech Intelligibility and Psychoacoustic Masking," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 1098–1111 (2017 Aug.). https://doi.org/10.1121/1.4999059.

[24] H. Fletcher, "Auditory Patterns," *Rev. Mod. Phys.*, vol. 12, no., pp. 47–65 (1940 Jan.). https://doi.org/10.1103/RevModPhys.12.47.

[25] R. D. Patterson and B. C. J. Moore, "Auditory Filters and Excitation Patterns as Representations of Frequency Resolution," in *Frequency Selectivity in Hearing*, pp. 123–177 (Academic Press, New York, NY, 1986).

[26] S. D. Ewert and T. Dau, "Characterizing Frequency Selectivity for Envelope Fluctuations," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1181–1196 (2000 Sep.). https://doi.org/10.1121/1.1288665.

[27] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An Efficient Auditory Filterbank Based on the Gammatone Function," presented at the *IOC Speech Group on Auditory Modelling at Royal Signals and Radar Establishment (RSRE)* (Malvern, UK) (1987 Dec.).

[28] V. Hohmann, "Frequency Analysis and Synthesis Using a Gammatone Filterbank," *Acta Acust. united Acust.*, vol. 88, no. 3, pp. 433–442 (2002 May).

[29] B. C. J. Moore and B. R. Glasberg, "Suggested Formulae for Calculating Auditory Filter Bandwidths and Excitation Patterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753 (1983 Sep.). https://doi.org/10.1121/1.389861.

[30] ISO, "Acoustics — Reference Zero for the Calibration of Audiometric Equipment — Part 7: Reference Threshold of Hearing Under Free-Field and Diffuse-Field Listening Conditions," *ISO Standard 389-7:2005SO Standard 389-7:2005* (2005 Nov.).

[31] K. S. Rhebergen, N. J. Versfeld and W. A. Dreschler, "Extended Speech Intelligibility Index for the Prediction of the Speech Reception Threshold in Fluctuating Noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997 (2006 Dec.). https://doi.org/10.1121/1.2358008.

[32] J. L. Verhey, T. Dau, and B. Kollmeier, "Within-Channel Cues in Comodulation Masking Release (CMR): Experiments and Model Predictions Using a Modulation-Filterbank Model," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2733–2745 (1999 Nov.). https://doi.org/10.1121/1.428101.

[33] B. C. J. Moore and C.-T. Tan, "Development and Validation of a Method for Predicting the Perceived Naturalness of Sounds Subjected to Spectral Distortion," *J. Audio Eng. Soc.*, vol. 52, no. 9, pp. 900–914 (2004 Sep.).

[34] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, "Predicting the Perceived Quality of Nonlinearly Distorted Music and Speech Signals," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 699–711 (2004 Jul.).

[35] T. Dau, D. Püschel, and A. Kohlrausch, "A Quantitative Model of the 'Effective' Signal Processing in the Auditory System: I. Model Structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622 (1996 Jun.). https://doi.org/10.1121/1.414959.

[36] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling Auditory Processing of Amplitude Modulation. I. Detection and Masking With Narrow-Band Carriers," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2892–2905 (1997 Nov.). https://doi.org/10.1121/1.420344.

[37] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling Auditory Processing of Amplitude Modulation. II. Spectral and Temporal Integration," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2906–2919 (1997 Nov.). https://doi.org/10.1121/1.420345.

[38] C. J. Darwin, M. T. Turvey, and R. G. Crowder, "An Auditory Analogue of the Sperling Partial Report Procedure: Evidence for Brief Auditory Storage," *Cognitive Psychol.*, vol. 3, no. 2, pp. 255–267 (1972 Apr.). https://doi.org/10.1016/0010-0285(72)90007-2.

[39] T. Wendt, S. van de Par, and S. D. Ewert, "A Computationally-Efficient and Perceptually Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. Audio Eng. Soc.*, vol. 62, no. 11, pp. 748–766 (2014 Nov.). https://doi.org/10.17743/jaes.2014.0042.

[40] C. Kirsch, T. Wendt, S. van de Par, H. Hu, and S. D. Ewert, "Computationally-Efficient Simulation of Late Reverberation for Inhomogeneous Boundary Conditions and Coupled Rooms," *J. Audio Eng. Soc.*, vol. 71, no. 4, pp. 186–201 (2023 Apr.). https://doi.org/10.17743/jaes.2022.0053.

[41] A. Lindau, V. Erbes, S. Lepa, et al., "A Spatial Audio Quality Inventory (SAQI)," *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 984–994442 (2002 May). https://doi.org/10.3813/AAA.918778.

[42] O. J. Dunn and V. Clark, "Comparison of Tests of the Equality of Dependent Correlation Coefficients," *J. Am. Statist. Assoc.*, vol. 66, no. 336, pp. 904–908 (1971 Dec.).

[43] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory Model Based Direction Estimation of Concurrent Speakers From Binaural Signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605 (2011 May/Jun.). https://doi.org/10.1016/j.specom.2010.05.006.

[44] B. Eurich, S. D. Ewert, M. Dietz, and T. Biberger, "A Computationally Efficient Model for Combined Assessment of Monaural and Binaural Audio Quality," *J. Audio Eng. Soc.*, vol. 72, no. 9, pp. 536–551 (2024 Sep.). https://doi.org/10.17743/jaes.2022.0161.

[45] T. Biberger, H. Schepker, F. Denk, and S. D. Ewert, "Instrumental Quality Predictions and Analysis of Auditory Cues for Algorithms in Modern Headphone Technology," *Trends Hear.*, vol. 25, pp. 1–22 (2021 Mar.). https://doi.org/10.1177/23312165211001219.

[46] A. Chabot-Leclerc, S. Jørgensen, and T. Dau, "The Role of Auditory Spectro-Temporal Modulation Filtering and the Decision Metric for Speech Intelligibility Prediction," *J. Acoust. Soc. Am.*, vol. 135, no. 6, pp. 3502–3512 (2014 Jun.). https://doi.org/10.1121/1.4873517.

[47] J. H. McDermott and E. P. Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence From Sound Synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940 (2011 Sep.). https://doi.org/10.1016/j.neuron.2011.06.032.

[48] T. Dau, T. Piechowiak, and S. D. Ewert, "Modeling Within- and Across-Channel Processes in Comodulation Masking Release," *J. Acoust. Soc. Am.*, vol. 133, no. 1, pp. 350–364 (2013 Jan.). https://doi.org/10.1121/1.4768882.

[49] R. McWalter and J. H. McDermott, "Adaptive and Selective Time Averaging of Auditory Scenes," *Current Biology*, vol. 28, no. 9, pp. 1405–1418 (2018 May). https://doi.org/10.1016/j.cub.2018.03.049.

[50] R. McWalter and T. Dau, "Cascaded Amplitude Modulations in Sound Texture Perception," *Front. Neurosci.*, vol. 11, no. 485, pp. 1–12 (2017 Sep.). https://doi.org/10.3389/fnins.2017.00485.

[51] T. D. Griffiths and J. D. Warren, "What Is an Auditory Object?" *Nature Rev. Neurosci.*, vol. 5, pp. 887–892 (2004 Nov.). https://doi.org/10.1038/nrn1538.

[52] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117 (2014 Mar.). https://doi.org/10.17743/jaes.2014.0006.

THE AUTHORS



Thomas Biberger



Jan-Hendrik Fleßner



Stephan D. Ewert

Thomas Biberger received a Diploma degree in media technology with a focus on audio/video signal processing from Hochschule für Angewandte Wissenschaften Hamburg, Hamburg, Germany, in 2009, and M.Sc. degree in hearing technology and audiology from Universität Oldenburg, Oldenburg, Germany, in 2012. After working as a research associate in the area of applied psychoacoustics at the Acoustics Group of the Universität Oldenburg, he joined the Medical Physics Group of the Universität Oldenburg and received a Ph.D. degree in the field of auditory modeling in 2018. His current research focus is on auditory object-based perception modeling for complex scenes.

•

Jan-Hendrik Fleßner received a B.Eng. (FH) degree in hearing technology and audiology from the Oldenburg University of Applied Science, Oldenburg, Germany, in 2011, and M.Sc. degree in hearing technology and audiology from Universität Oldenburg, Oldenburg, Germany, in 2013. He received his Ph.D. degree in 2018, on the topic of assessment and prediction of binaural audio quality.

•

Stephan D. Ewert studied physics and received a Ph.D. degree from the Carl von Ossietzky Universität Oldenburg, Germany, in 2002. During his Ph.D. project, he spent a 3-month stay as a visiting scientist with the Research Lab of Electronics, Massachusetts Institute of Technology, Cambridge, MA. From 2003 to 2005, he was an assistant professor with the Centre of Applied Hearing Research, Technical University of Denmark, Lyngby, Denmark. In 2005, he rejoined Medizinische Physik at the Universität Oldenburg, where he has been the head of the Psychoacoustic and Auditory Modeling Group since 2008. His field of expertise is in psychoacoustics and acoustics with a strong emphasis on perceptual models of hearing and virtual acoustics. He has authored various papers on spectro-temporal processing, binaural hearing, and speech intelligibility. He also focused on perceptual consequences of hearing loss, hearing-aid algorithms, instrumental audio quality prediction, and room acoustics simulation.