Instrumental Quality Predictions and Analysis of Auditory Cues for Algorithms in Modern Headphone Technology

Thomas Biberger, Henning Schepker, Florian Denk, and Stephan D. Ewert

Abstract
Smart headphones or hearables use different types of algorithms such as noise cancelation, feedback suppression, and sound pressure equalization to eliminate undesired sound sources or to achieve acoustical transparency. Such signal processing strategies might alter the spectral composition or interaural differences of the original sound, which might be perceived by listeners as monaural or binaural distortions and thus degrade audio quality. To evaluate the perceptual impact of these distortions, subjective quality ratings can be used, but these are time consuming and costly. Auditory-inspired instrumental quality measures can be applied with less effort and may also be helpful in identifying whether the distortions impair the auditory representation of monaural or binaural cues. Therefore, the goals of this study were (a) to assess the applicability of various monaural and binaural audio quality models to distortions typically occurring in hearables and (b) to examine the effect of those distortions on the auditory representation of spectral, temporal, and binaural cues. Results showed that the signal processing algorithms considered in this study mainly impaired (monaural) spectral cues. Consequently, monaural audio quality models that capture spectral distortions achieved the best prediction performance. A recent audio quality model that predicts monaural and binaural aspects of quality was revised based on parts of the current data involving binaural audio quality aspects, leading to improved overall performance indicated by a mean Pearson linear correlation of 0.89 between obtained and predicted ratings.

Keywords
audio quality, auditory models, hearables, spatial audio

Modern earphones, in the following denoted as hearables, go far beyond their original application for audio playback, providing many additional features such as medical monitoring, voice assistant systems, active noise control, and hear-through features (Rumsey, 2019; Temme, 2019). It is conceivable that such devices may bridge the gap between a classical hearing aid and a modern HiFi sound reproduction system in the future. Typical applications demand a variety of signal processing algorithms that may either deliberately alter the properties of the signal (e.g., noise suppression, nonlinear amplification, attenuation) or alter signal properties by undesired distortions (e.g., hear-through, feedback suppression). Undesired distortions introduced by hear-through processing and feedback suppression algorithms recently received a lot of interest (e.g., Madsen & Moore, 2014; Marentakis & Liepins, 2014; Maxwell & Zurek, 1995) for applications aiming at faithful reproduction of external sound signals, enabling perceptually authentic conversations as well as awareness of...
the acoustical scene while hearables are inserted in the ear canals of the listeners.

The hear-through mode is one basic feature of many hearables that allows the user to hear the acoustic environment through the device, similar to a hearing aid, and to simultaneously listen to an audio signal from a source device like a smartphone. A natural representation of the acoustical environment, similar to that experienced with an open ear (without inserted device), is desirable. In the optimal case, the listener is not able to distinguish between scenarios where the hearables with activated hear-through mode are inserted in the ear canals and where the ears are open without the inserted device, which is typically referred to as acoustical transparency (Denk et al., 2018; Hoffmann et al., 2013). Because the human auditory system is limited in resolving monaural (e.g., spectral or temporal) and binaural differences (e.g., interaural time differences [ITDs] and interaural level differences [ILDs]), acoustical transparency can be achieved by the hear-through mode without exactly reproducing the open-ear signal at the eardrum. In addition, an important feature in hearables, as in conventional hearing aids, is acoustic feedback suppression to avoid howling or chirping of the device. This may introduce audible distortion.

To assess the perceived audio quality of such hearing devices or algorithms, subjective quality tests can be used. These tests can be carried out as reference-free tests (e.g., ITU-R BS.1534, 2014), where listeners rate the audio quality of a processed audio signal under test without any knowledge of an unprocessed reference signal, and as reference-based tests (e.g., ABX: Munson & Gardner, 1950; Multiple Stimulus with Hidden Reference and Anchor [MUSHRA]: ITU-T. P800, 1996), where listeners are allowed to compare the processed signal with the unprocessed signal to make their quality judgment. Performing subjective listening tests for audio quality evaluation is time consuming, expensive, and often requires qualified (expert) listeners. With the goal of replacing listening tests, several instrumental audio quality measures have been developed over the past few decades. Similar to the reference-free and reference-based listening tests, instrumental measures can be classified into nonintrusive and intrusive measures. Nonintrusive measures do not require an unprocessed reference signal, while intrusive measures explicitly require a reference signal. Given that instrumental quality measures are less time consuming and more cost-effective than listening tests, they lend themselves to application during hearing device and algorithm development, as well as for the development of real-time steering and optimization of signal processing in future devices.

For the assessment of distortions introduced by audio signal processing, intrusive measures based on auditory perception models are commonly used and have been shown to be broadly applicable (Biberger et al., 2018; Harlander et al., 2014). Such measures include the Perceptual Audio Quality Measure (Beerends & Stermerdink, 1992), Perceptual Speech Quality Measure (Beerends & Stermerdink, 1994), Perceptual Evaluation of Speech Quality (Beerends et al., 2002; ITU-T P.862, 2001; Rix et al., 2002), Perceptual Evaluation of Audio Quality (PEAQ; ITU-R BS.1387, 2001; Thiede et al., 2000), Short-Term Partial Loudness Model (Glasberg & Moore, 2005), or the Perceptual Objective Listening Quality Assessment (Beerends et al., 2013a, 2013b).

Moore and colleagues suggested a measure of the perceived naturalness of sounds based on differences in the auditory excitation patterns (D; Moore & Tan, 2004) and a measure for predicting the quality of nonlinearly distorted signals (\(R_{\text{nom}}\); Tan et al., 2004).

An auditory model front end suggested by Kates and Arehart forms the basis for the speech intelligibility model—the Hearing-Aid Speech Perception Index (Kates & Arehart, 2014b)—and the quality models—Hearing-Aid Speech Quality Index (HASQI; Kates & Arehart, 2010), Hearing-Aid Speech Quality Index version 2 (HASQIv2) (Kates & Arehart, 2014a), and Hearing-Aid Audio Quality Index (Kates & Arehart, 2016). Following the idea that a psychoacoustic model that successfully accounts for a large number of relevant psychoacoustic experiments should also be suited as front end for audio quality predictions, Huber and Kollmeier (2006) adapted the Perception Model (PEMO) of Dau et al. (1997a, 1997b), resulting in Perception Model Quality Assessment (PEMO-Q), while the more complex Computational Auditory Signal processing and Perception model (CASP; Jepsen et al., 2008) formed the basis for Computational Auditory Signal processing and Perception model based Quality assessment (CASP-Q) of Harlander et al. (2014). Biberger and Ewert combined the Power Spectrum Model (PSM; Fletcher, 1940; Patterson & Moore, 1986) and Envelope Power Spectrum Model (EPSM; Ewert & Dau, 2000) with multiresolution analysis as suggested by Jørgensen et al. (2013), denoted the Generalized Power Spectrum Model (GPSM), which has been demonstrated to predict the results of several experiments on psychoacoustic masking and speech intelligibility (Biberger & Ewert, 2016, 2017). Recently, Biberger and colleagues suggested the Generalized Power Spectrum Model for quality (GPSM\(^*\); Biberger et al., 2018) that has been shown to predict the perception of a large variety of monaural distortions.

All aforementioned quality models account only for monaural aspects of audio quality, while several applications, including hear-through modes, may also introduce binaural distortions (Denk et al., 2020). Such binaural distortions could arise from processing delays...
and sensitivity differences between the left and right ear channels of the hearing device, resulting in distorted ITDs and ILDs that may alter the perceived spatial image. Thus, the binaural perceptually motivated direction estimation model of Dietz et al. (2011) was adapted by Fleßner et al. (2017) to develop the intrusive binaural auditory model for audio quality (BAM-Q). In their later study (Fleßner et al., 2019), Fleßner and colleagues combined the outputs of the binaural BAM-Q and the monaural GPSM<sup>1</sup>, here denoted MoBi-Q, to predict overall audio quality for monaurally, binaurally, and combined monaurally and binaurally distorted speech, music, and noise signals.

To the knowledge of the authors, it has not yet been tested whether existing instrumental audio quality measures are applicable to the distortions that might occur in modern hearing devices such as hearables. Moreover, it is not clear which auditory cues are mainly impaired by such algorithms. In this context, auditory-inspired instrumental quality measures could help to analyze the contribution of monaural and binaural cues.

To address these two aspects, the current study examined the prediction performance of 13 intrusive monaural and binaural audio quality models for distortions occurring in hearables. Three databases, including sounds processed using algorithms for adaptive feedback cancelation, feedback suppression based on a null-steering beamformer, and hear-through processing, were used to cover a large range of relevant distortions. A comparison of the models’ prediction performance should help to identify models suited for the objective evaluation of algorithms potentially employed in hearables. The best performing quality models are also expected to provide information about auditory cues that are relevant for accounting for quality degradation and potentially help developers to test their algorithms and identify perceptually relevant distortions. Finally, based on these findings, an instrumental measure optimized for the distortions that might occur in hearables is suggested. This instrumental measure will be made publicly available.<sup>1</sup>

**Audio Quality Models**

In total, 13 intrusive auditory-based perceptual measures were examined for their applicability to distortions typically occurring in hearables. First, 11 measures purely based on monaural cues are described, followed by the description of a measure purely based on binaural cues. Third, a perceptual measure combining monaural and binaural cues is explained. There exist only a few approaches (e.g., Schäfer et al., 2013; Seo et al., 2013; Takanen et al., 2014) for predictions of binaural or combined monaural and binaural audio quality that are, to the best knowledge of the authors, not publicly available. Thus, only one binaural and one combined audio quality model were tested in the current study.

Instrumental measures used the original sampling rate given by the input signals of the databases (see the Evaluation section for details). Input signals were up- or downsampled if measures required a certain sampling rate (e.g., PEAQ required a sampling rate of 48 kHz).

**Monaural Models**

The ITU standardized PEAQ (ITU-R BS.1387, 2001; Thiede et al., 2000) was developed to predict the audio quality of low-bit-rate coded audio signals. PEAQ incorporates two different ear models from which a large variety of features, such as envelope modulation, partial noise loudness, audible linear distortion, noise-to-mask ratio, or signal bandwidth, are calculated for the processed and unprocessed signals. Based on such features, model output variables (MOVs) are derived, which are assumed to represent relevant quality-degrading aspects, for example averaged temporal envelope differences or partial loudness of additive distortions. A trained multilayer perceptron neural network is used to map a selected set of MOVs to a single measure of audio quality. The training data set resulted from audio quality ratings of normal-hearing (NH) listeners for music and speech signals processed by low-bit-rate audio codecs. Such algorithms mainly introduced nonlinear distortions.

The linear distortion measure D of Moore and Tan (2004) is based on peripheral preprocessing in which the excitation patterns of the reference and the test signals are calculated on an equivalent rectangular bandwidth (ERB)-number scale (Moore & Glasberg, 1983), from which excitation differences are derived. A combination of the standard deviation of the spectrally weighted excitation differences (first-order excitation differences) and the standard deviation of the slopes of the excitation differences (second-order excitation differences) provides the output measure D. A curvilinear relationship between D and subjective ratings was observed by Moore and Tan (2004)<sup>3</sup>. To obtain a more linear relationship, a transformation was applied that was also used in this study (see appendix in Biberger et al., 2018). The linear measure D was developed using the data set provided by Moore and Tan (2003), where NH listeners rated the audio quality (perceived naturalness) of music and speech signals impaired by linear filtering. D was separately developed with music and

<sup>1</sup>In-text citation was updated after the paper was first published.
speech signals, from which two sets of optimized model parameters were derived.

The nonlinear distortion measure $R_{\text{nonlin}}$ of Tan et al. (2004) analyzes the reference and test signals using simulated auditory filters that are uniformly spaced on an ERB-number scale. A correlation analysis between the reference and the test signals is performed in short time frames of 30 ms, where each frame is weighted according to its level. The weighted frames are summed across auditory filters and averaged across frames to obtain the output measure $R_{\text{nonlin}}$. Since Tan et al. (2004) observed a curvilinear behavior between $R_{\text{nonlin}}$ and subjective ratings, they used a nonlinear transformation to obtain a more linear relationship between predicted and subjective ratings. Such a transformation was also applied in this study (see appendix in Biberger et al., 2018). $R_{\text{nonlin}}$ was developed using three data sets provided by Tan et al. (2003), where NH listeners rated the audio quality of music and speech signals impaired by artificial nonlinear processing (e.g., hard symmetrical/asymmetrical clipping or center clipping) and nonlinearities from transducers. The nonlinear distortion measure was separately optimized with music and speech signals, resulting in two sets of model parameters for each of the three databases. Because no generalized parameter set was provided, the user has to either use one of the existing optimized parameter sets from the study of Tan et al. (2004) or optimize the fitting parameters for the database under test, as was suggested by the authors. However, the latter procedure makes it difficult to compare $R_{\text{nonlin}}$ to other instrumental measures as they do not have such a priori knowledge. Thus, in this study, the same fitting parameters as used in Biberger et al. (2018), derived by averaging the fitting parameters for speech and music given in Figure 2 of Tan et al. (2004), were applied to $R_{\text{nonlin}}$.

The combined quality model $S_{\text{overall}}$ of Moore et al. (2004) is based on the linear component $S_{\text{lin}}$ and the nonlinear component $S_{\text{nonlin}}$ derived from $D$ and $R_{\text{nonlin}}$, respectively, as described in the appendix of Biberger et al. (2018). The combined measure is obtained as $S_{\text{overall}} = \tau \cdot S_{\text{lin}} + (1 - \tau) \cdot S_{\text{nonlin}}$, where $\tau$ is 0.3. $S_{\text{overall}}$ was optimized using data sets, where NH listeners rated the audio quality of music and speech signals impaired by linear filtering and nonlinear processing. Because the measure was separately optimized for music and speech, for each data set two optimized parameter sets were provided.

The HASQI (Kates & Arehart, 2010) is based on a cochlear model including a middle-ear filter, a linear gammatone filterbank used to extract the envelope for each auditory channel, instantaneous compression, attenuation, dB conversion (providing an approximate conversion of signal intensity into a perceptually motivated scale linked to just-noticeable differences in intensity and loudness perception), and low-pass filtering. From the cochlear model output, a nonlinear quality index, based on cepstrum correlation, and a linear quality index, adopted from Moore and Tan (2004)*, accounting for (long-term) spectral differences between the reference and test signal, are calculated and combined to give the final overall quality index. The main difference between HASQI and HASQIv2 (Kates & Arehart, 2014a) is an additional analysis of temporal fine structure (TFS) to account for nonlinear distortions in HASQIv2, while the nonlinear and linear distortion indices proposed in the original HASQI are maintained. HASQIv2 also uses a modified version of the model of the auditory periphery used in HASQI that includes the following aspects: a conversion of all input signals to a 24-kHz sampling rate, broader auditory filters with increasing signal intensity, different outer hair cell dynamic-range compression rule, and inner hair cell firing-rate adaption. Quality judgments made by NH and hearing-impaired (HI) listeners for speech stimuli containing noise and nonlinear processing were used to optimize the nonlinear part of HASQI, while judgments of speech stimuli with linear filtering were used to optimize the linear part. For validation of the entire HASQI (combination of linear and nonlinear parts), quality judgments made by NH and HI listeners for speech containing combinations of noise and nonlinear processing with linear filtering were used. HASQIv2 was optimized using the same data as was used for HASQI.

The PEMO-Q (Huber & Kollmeier, 2006) front end was adopted from the psychoacoustic model PEMO (Dau et al., 1997a, 1997b), which includes the following auditory processing stages: linear (gammatone) basilar membrane filtering, hair cell transduction, adaptation, and a modulation filterbank. The front-end outputs of PEMO and PEMO-Q, also denoted the internal representations (IR), provide cues mainly based on amplitude modulation (AM). Based on the IR, three quality measures, the Perceptual Similarity Measure (PSM), the time-dependent PSM (PSM), and the Objective Difference Grade (ODG), were calculated with the back end of PEMO-Q. Harlander et al. (2014) demonstrated that the ODG had better overall prediction performance than PSM and PSM. Thus, this study considers only the ODG measure, which calculates the (power-weighted) linear cross-correlation between the IR of the test and reference signals in successive time frames of 10 ms, from which the 5th percentile gives the final PSM. The ODG is derived by mapping the PSM to a Subjective Difference Grade (ITU-R BS.1116-1, 1997)-like scale by applying a nonlinear regression function. As in Harlander et al. (2014), we used the original PEMO-Q as described in Huber and Kollmeier (2006) and a modified version, PEMO-QISO (Harlander et al., 2014), that additionally includes a hearing threshold based on ISO
226 (2003). Similarly as for PEAQ, the data set used to optimize PEMO-Q was derived from audio quality ratings of NH listeners for music and speech signals processed by low-bit-rate audio codecs.

CASP-Q (Harlander et al., 2014) can be considered as an updated version of PEMO-Q with a front end adopted from the CASP model described by Jepsen et al. (2008). The substantial changes compared with the PEMO-Q front end are outer- and middle-ear transformations, nonlinear basilar membrane filtering, and a squaring expansion after hair cell transduction. The IR of CASP-Q mainly represents AM-based features. CASP-Q uses the same back-end processing as PEMO-Q, thus providing the similar quality measures PSM, PSMt, and ODG. Here, for the same reason as stated earlier for PEMO-Q, only the ODG measure is considered. It should be mentioned that the CASP-Q front end has some general modifications compared with the original CASP, such as an adjustment of the amplification following hair cell processing and modified adaptation loops, which are important for audio quality predictions (for more details, see Harlander et al., 2014). Both CASP-Q versions, CASP-QISO and CASP-QnoExp as suggested by Harlander et al. (2014), were used in this study. CASP-QISO additionally applies a hearing threshold based on ISO 226 (2003), while CASP-QnoExp includes a hearing threshold but does not have the expansion stage. As mentioned by Jepsen et al. (2008), the expansion stage transforms the half-wave rectified and low-pass filtered signal into an intensity-like representation, which was motivated by physiological findings of Yates et al. (1990) and Muller et al. (1991). CASP-Q was optimized with a database derived by Hu and Loizou (2007), where NH listeners rated the audio quality of speech signals processed by noise reduction algorithms.

The GPSMq (Biberger et al., 2018) represents an audio quality extension of the GPSM, which has been demonstrated to predict the results of many psychoacoustic and speech intelligibility experiments (Biberger & Ewert, 2016, 2017). GPSMq applies a linear, fourth-order gammatone filterbank with bandwidths equal to one ERB\textsubscript{N}, and cochlear compression. The mechano-electrical transduction process in the inner hair cells is modeled by half-wave rectification followed by a 770-Hz fifth-order low-pass filter. These steps are followed by further processing of TFS for auditory filters tuned at or below 1.4 kHz and temporal envelope processing for auditory filters tuned to higher frequencies. The binaural feature extraction stage uses complex outputs from the left and right channels to calculate the interaural transfer function, from which interaural phase differences and ITDs are derived. The interaural vector strength (IVS), which is similar to the interaural coherence (IC), is also derived from the interaural transfer function. In addition, ILDs are calculated from the energy ratio between the right and left filters. The back-end processing combines the submeasures ILD, ITD, and IVS that are calculated in consecutive time frames of 400 ms. The ILD and ITD submeasures can be used to predict changes in perceived source location and changes in the apparent source width (ASW). Perceived diffusiveness and ASW are often related to IC (e.g., Ando & Kurihara, 1986; Blauert & Lindemann, 1986; Damaske & Ando, 1972; Kendall, 1995), where perceived diffusiveness and ASW increase as IC decreases, and thus the IVS submeasure is assumed to predict differences for both perceptual attributes. The submeasures for each frame are averaged across time and auditory bands, and then combined by a nonlinear regression method, providing the final output measure binQ. BAM-Q was optimized with a data set for which 10 NH listeners rated spatial audio quality degradations for manipulations of static and dynamic binaural properties of music, noise, and speech signals.

Binaural Model

The binaural audio quality model (BAM-Q; Fleßner et al., 2017) is based on the binaural psychoacoustic model front end of Dietz et al. (2011). The peripheral processing stages include outer and middle-ear filtering followed by a linear, fourth-order gammatone filterbank with bandwidths equal to one ERB\textsubscript{N}, and cochlear compression. The mechano-electrical transduction process in the inner hair cells is modeled by half-wave rectification followed by a 770-Hz fifth-order low-pass filter. These steps are followed by further processing of TFS for auditory filters tuned at or below 1.4 kHz and temporal envelope processing for auditory filters tuned to higher frequencies. The binaural feature extraction stage uses complex outputs from the left and right channels to calculate the interaural transfer function, from which interaural phase differences and ITDs are derived. The interaural vector strength (IVS), which is similar to the interaural coherence (IC), is also derived from the interaural transfer function. In addition, ILDs are calculated from the energy ratio between the right and left filters. The back-end processing combines the submeasures ILD, ITD, and IVS that are calculated in consecutive time frames of 400 ms. The ILD and ITD submeasures can be used to predict changes in perceived source location and changes in the apparent source width (ASW). Perceived diffusiveness and ASW are often related to IC (e.g., Ando & Kurihara, 1986; Blauert & Lindemann, 1986; Damaske & Ando, 1972; Kendall, 1995), where perceived diffusiveness and ASW increase as IC decreases, and thus the IVS submeasure is assumed to predict differences for both perceptual attributes. The submeasures for each frame are averaged across time and auditory bands, and then combined by a nonlinear regression method, providing the final output measure binQ. BAM-Q was optimized with a data set for which 10 NH listeners rated spatial audio quality degradations for manipulations of static and dynamic binaural properties of music, noise, and speech signals.
Combined Monaural and Binaural Model

The MoBi-Q model of Fleßner et al. (2019) combines the quality outputs of a modified version of the monaural GPSM\textsuperscript{q} and the binaural BAM-Q. The GPSM\textsuperscript{q} modification was necessary to reduce sensitivity of the model to binaural distortions such as ILDs and ITDs based on monaural features and artifacts such as level differences and phase distortions. In the combined MoBi-Q model, it was thus ensured that the binaural features were exclusively captured by BAM-Q. Because the monaural GPSM\textsuperscript{q} processes the left and right channels of a binaural input signal separately, and then averages the output measures from the left and the right channels, it is sensitive to certain binaural cues without applying such a binaural modification. Fleßner et al. (2019) demonstrated that overall quality is dominated by whichever aspect is lower in quality, either monaural or binaural. Accordingly, Fleßner and colleagues suggested a combination of the outputs of the monaural GPSM\textsuperscript{q} (OPM\textsubscript{dual}) and the binaural BAM-Q (binQ) by selecting the (transformed) output that showed the largest quality degradation (minimum operation):

$$\text{Overall quality} = \min \left( \log_{10}(0.0528+\text{OPM\textsubscript{dual}}), \ 0.0078+\text{binQ} \right)$$  \hspace{1cm} (1)

In the study of Fleßner et al. (2019), 16 NH listeners evaluated 119 items containing music, speech, and noise that had either monaural (50 items) and binaural (24 items) distortions in isolation or combined monaural and binaural distortions (45 items). For monaural distortions in isolation, overall quality was dominated by the monaural pathway, while for binaural distortions, overall quality was dominated by the binaural pathway. For combined monaural and binaural distortions, overall quality was dominated by the monaural pathway for 35 items and by the binaural pathway for 10 items.

Evaluation

Three databases with different types of distortions were chosen that covered a broad range of distortions affecting quality in hearables. The first database was based on monaural (right ear) dummy head recordings and thus included only monaural distortions. Data were taken from the study of Nordholm et al. (2018). The second and third databases were based on binaural dummy head recordings and were taken from the studies of Schepker et al. (2019, 2020). These databases potentially included monaural and binaural distortions. Denk et al. (2020) demonstrated that the devices examined by Schepker et al. (2020) impaired the representation of monaural and binaural cues, and thus monaural and binaural distortions were expected to contribute to quality degradations for the third database of Schepker et al. (2020). The second database (Schepker et al., 2019) was not technically evaluated, but test signals mainly showed monaural spectral differences compared with the test signal, while spatial differences such as changes in ASW and source location could be perceived as well (as indicated by informal listening tests).

The databases taken from Schepker et al. (2019, 2020) are balanced in the sense that the impairments range from intermediate to small, and distortions are homogeneously distributed over time. Depending on the tested segment, the database from Nordholm et al. (2018) includes some signals where the distortions are concentrated in a short segment. Thus, distortions for the stimuli of Nordholm et al. (2018) are perceptually different to the distortions for the stimuli of Schepker et al. (2019, 2020).

Databases

Adaptive Feedback Cancelation. The adaptive feedback cancelation (AFC) database was taken from the study of Nordholm et al. (2018). It consists of 60 monaural items, based on speech and music material, sampled at 16 kHz. All signals were recorded using a microphone placed in the right ear of a dummy head in an anechoic chamber for two different sound source positions (azimuths of 0° and 90°), resulting in four audio signals (2 × speech and 2 × music). In hearing devices such as hearing aids and hearables, acoustic coupling between the loudspeakers and the microphones generates feedback loops that can be suppressed by AFC algorithms. Nordholm et al. (2018) examined four AFC algorithms using four signals and three signal segments (initial and reconvergence phase, steady-state phase). Signals processed with an ideal feedback cancelation algorithm (with perfect a priori knowledge about the feedback path) served as reference signals, while signals processed without feedback cancelation served as anchor signals. This resulted in 48 items based on the AFC algorithms plus 12 items based on the anchor algorithm. The 12 items based on the reference algorithm were used by the reference-based instrumental measures as reference. For convenience, the algorithm names referring to Nordholm et al. (2018) are provided in Figure 1; their exact function is beyond the scope of the current article. Subjective quality ratings from 15 NH subjects were obtained using the MUSHRA method (ITU-R BS.1534-1, 2003). Besides the instruction to rate the perceived overall audio quality of the test signals compared with the reference signal, the listeners were instructed to rate at least one of the signals with a score of 100 (no perceptible difference) and at least one signal with a score of 0 (very strong difference). The listening test
was carried out in a quiet office room, and signals were presented via headphones. Averaged MUSHRA scores for distorted test signals (including the anchor) ranged from 0 to 100.

**Acoustically Transparent Hearing Device.** The *acoustically transparent hearing device* (ATHD) database was taken from the study of Schepker et al. (2019). The database consists of 140 speech (female, male) and music (piano, jazz) items, sampled at 48 kHz. Schepker et al. evaluated the audio quality of a real-time hearing device prototype intended to achieve acoustically transparent sound presentation. This device applies feedback suppression based on a null-steering beamformer and individualized equalization of the sound pressure at the eardrum. Six signal processing conditions representing feedback suppression in combination with different equalization strategies and their effect on perceived audio quality were assessed for three recording room reverberation times \( T_{60} \approx 0.35 \text{s}, 0.45 \text{s}, 1.4 \text{s} \) and three incoming signal directions (azimuths of 0°, 90°, 225°). A dummy head with inserted hearing devices was used for recordings. The loudspeakers used for signal playback were placed at a distance of approximately 2 m from the dummy head and adjusted in height to be at ear level with the dummy head. The dummy head open-ear recordings served as the reference signals for acoustical transparency. A low-quality anchor signal (denoted as Rec Off in Figure 2) was obtained using dummy head occluded-ear recordings, with hearing devices inserted but without signal processing. The algorithm names are provided in Figure 2, and the reader is referred to Schepker et al. (2019) for further details. The subjective evaluation was carried out by 15 NH subjects via a MUSHRA-like framework (Völker et al., 2018). Participants were instructed in writing to rate the perceived overall sound quality of each stimulus relative to

---

**Figure 1.** Audio Quality Predictions of GPSMq, MoBi-Q, PEAQ, D, HASQv2, and PEMO-Q for the *Adaptive Feedback Cancelation* (AFC) Database. Algorithms are represented by different colors. The individual prediction performance is given in each panel by Accuracy \( (r_{\text{pear}}) \) and Monotonicity \( (r_{\text{rank}}) \). The names of the algorithms are indicated in the right top panel.

GPSMq = Generalized Power Spectrum Model for quality; HASQv2 = Hearing-Aid Speech Quality Index version 2; PEAQ = Perceptual Evaluation of Audio Quality; PEMO-Q = Perception Model based Quality assessment.

**Figure 2.** Audio Quality Predictions of GPSMq, MoBi-Q, PEAQ, D, HASQv2, and PEMO-Q for the *Acoustically Transparent Hearing Device* (ATHD) Database. Hearing device settings are indicated by different colors. Prediction performance is given in each panel by Accuracy \( (r_{\text{pear}}) \) and Monotonicity \( (r_{\text{rank}}) \). The algorithm names are indicated in the two bottom panels.

GPSMq = Generalized Power Spectrum Model for quality; HASQv2 = Hearing-Aid Speech Quality Index version 2; PEAQ = Perceptual Evaluation of Audio Quality; PEMO-Q = Perception Model based Quality assessment.
the (open-ear) reference. The listening test was carried out in a sound-isolated cabin, and signals were presented over headphones. Averaged MUSHRA scores for distorted test signals (including the anchor) ranged from about 8 to 95.

**Hear-Through Mode.** The hear-through mode (HTM) database was taken from the study of Schepker et al. (2020). The database consists of 120 speech (female, male) and music (jazz, piano) items, sampled at 48 kHz. The study examined the audio quality of the hear-through mode of six commercial hearables (referred to as Devices A, B, C, D, E, and F) and three research devices (Devices H, I, and J). A dummy head with inserted hearables was used for recordings in a laboratory with moderate room reverberation ($T_{60} \approx 0.45$ s) to assess the devices in realistic but controlled acoustic conditions. Four audio signals were recorded for three playback directions (azimuths of 0°, 90°, and 225°) with loudspeakers placed at a distance of approximately 2 m from the dummy head and adjusted in height to be at ear level with the dummy head. The dummy head’s open-ear recordings served as reference signals, and thus the sound transmission to the eardrum with the hearable should be equivalent to the open-ear reference signal to achieve acoustic transparency. The occluded ear, using Device J turned off, was used as anchor signal (denoted Device K in Figure 3). For further details about the devices, refer to Schepker et al. (2020). Subjective results are based on data for 17 NH subjects using a MUSHRA-like framework (Völker et al., 2018). Participants were instructed in writing to rate the perceived overall sound quality of the stimuli recorded with the different devices relative to the (open-ear) reference. The test was carried out in a sound-isolated cabin, and signals were presented over headphones. Averaged MUSHRA scores for distorted test signals (including the anchor) ranged from about 10 to 82.

**Objective Performance Measures for Model Predictions**

As suggested by Emiya et al. (2011) and applied in Harlander et al. (2014) and Biberger et al. (2018), prediction performance was individually calculated for each database on the basis of three measures: **Accuracy**, **Monotonicity**, and **Consistency**. Accuracy was quantified by the Pearson linear correlation coefficient, Monotonicity by the Spearman rank correlation coefficient, and Consistency was based on the number of discrepancies in quality prediction using an interval of ± one standard deviation of the subjective quality ratings, instead of the two-standard-deviation interval used by Emiya et al. (2011) and Harlander et al. (2014). The calculation of Consistency requires relating objective scores to the subjective results, and this was done using linear regression. After this transformation, subjective and objective model scores were expressed on a 100-point scale. More detailed explanations of these measures can be found in Emiya et al. (2011) and Harlander et al. (2014).

---

**Figure 3.** Audio Quality Predictions for GPSMq, MoBi-Q, PEAQ, BAM-Q, D, HASQIv2, and PEMO-Q for the Hear-Through Mode (HTM) Database. Hearing devices are indicated by different colors. Prediction performance is given in each panel by Accuracy ($r_{pear}$), Monotonicity ($r_{rank}$), and Consistency ($\rho$). For better visualization of the relationship between subjective and objective scores, the upper left panel shows GPSMq predictions without lower and upper perceptual limits (see equation 6 in Biberger et al., 2018). Disregarding the perceptual limits results in a slightly lower Accuracy of 0.92 compared with the Accuracy of 0.93 for the standard GPSMq, which incorporates such perceptual limit by default. For further details about the devices, refer to Schepker et al. (2020). GPSMq = Generalized Power Spectrum Model for quality; HASQIv2 = Hearing-Aid Speech Quality Index version 2; PEAQ = Perceptual Evaluation of Audio Quality; PEMO-Q = Perception Model based Quality assessment; BAM-Q = Binaural Auditory Model for audio Quality.
In addition to Accuracy, Monotonicity, and Consistency, the widely used objective performance measure epsilon-insensitive root mean square error (also denoted as RMSE*; ITU-T Rec P.1401, 2012) was calculated, including first-order mapping of the objective scores, for cross-validation. RMSE* is based on the 95% confidence-interval-weighted RMSE.

Accuracy, Monotonicity, and Consistency values of one represent the best achievable prediction performance, while values of zero represent the worst prediction performance. Small RMSE* values indicate accurate predictions, while large values represent discrepancies between subjective and objective scores.

Results

Table 1 compares the prediction performance of all audio quality models (rows) across the three databases (columns). BAM-Q predictions are only shown for the HTM database. Pretests indicated that binaural distortions played a minor role for the ATHD, while the AFC consists of monaural recordings. The bold values in Table 1 indicate the best performing instrumental measure for Accuracy, Monotonicity, Consistency, and RMSE* for each database, respectively.

For clarity, the relationships between subjective and objective scores for the three databases are shown in Figures 1–3 only for the four best performing models of this study and the widely used PEAQ and PEMO-Q.

Adaptive Feedback Cancelation

Figure 1 shows subjective scores and objective scores for GPSM4, MoBi-Q, PEAQ, D, HASQIv2, and PEMO-Q for the AFC database. The abscissa of each panel in Figure 1 represents subjective scores, while the ordinate represents objective scores. Each panel gives the Accuracy and Monotonicity, abbreviated as r_per and r_rank, of the corresponding instrumental measure. The PEAQ predictions agreed well with subjective ratings from the AFC database and gave, together with MoBi-Q, the highest Monotonicity value of 0.97. Large differences between the Accuracy and Monotonicity values indicate a curvilinear relationship between subjective ratings and PEAQ scores, which is also represented in Figure 1. The naturalness measure D performed very well and showed high values for Accuracy and Monotonicity of 0.95 (see Figure 1). Rnonlin also achieved reasonably good prediction performance for the AFC database represented by Accuracy and Monotonicity values of 0.83 and 0.91. However, predicted quality scores of Rnonlin were often lower than subjective scores, resulting in a small Consistency value of 0.42 and a high RMSE* value of 3.4. The combination of D and Rnonlin, Soverall, achieved values for Accuracy and Monotonicity of 0.92 and 0.94, respectively. The Consistency value of Soverall (0.48) fell between the Consistency values for D (0.67) and Rnonlin (0.42).

Both HASQI versions performed very well for the AFC database with Accuracy and Monotonicity values > 0.9, and small RMSE* values (HASQI: 2.7; HASQIv2: 2.2).

PEMO-Q and PEMO-QISO, showed good prediction performance, with similar Accuracy (PEMO-Q: 0.78; PEMO-QISO: 0.8) and Monotonicity (PEMO-Q: 0.82; PEMO-QISO: 0.86). CASP-QISO and CASP-QnoExp gave poor predictions for the AFC database with Accuracy and Monotonicity values < 0.6. GPSM4 predictions agreed well with subjective ratings, indicated by an Accuracy value of 0.95 and a Monotonicity value of 0.91 (see Figure 1). GPSM4 predictions produced the fewest outliers as indicated by the highest Consistency value of 0.7 and the lowest RMSE* value of 2.0. MoBi-Q achieved the highest Accuracy and Monotonicity of 0.96 and 0.97, respectively. Accurate predictions of MoBi-Q are also represented by Consistency and RMSE* of 0.67 and 2.2, respectively. As signals in this database are monaural, MoBi-Q predictions are purely based on the monaural pathway.

Acoustically Transparent Hearing Devices

Figure 2 shows subjective scores and objective scores for GPSM4, MoBi-Q, PEAQ, D, HASQIv2, and PEMO-Q for the ATHD database. The abscissa and ordinate of each panel are the same as in Figure 1. PEAQ gave poor prediction performance with Accuracy and Monotonicity values of 0.49 and 0.45. The naturalness measure D gave the best prediction performance (see Figure 2), indicated by the highest Accuracy and Monotonicity values of 0.9 and 0.9, and the lowest RMSE* value of 1.9. Rnonlin showed poor prediction performance (Accuracy value of 0.17, Monotonicity value of 0.01) for the ATHD database. The combined measure Soverall based on D and Rnonlin, achieved moderate prediction performance (Accuracy value of 0.73, Monotonicity value of 0.79).

HASQI performed rather poorly (Accuracy of 0.52, Monotonicity of 0.46), while HASQIv2, gave very accurate predictions, indicated by high values for Accuracy and Monotonicity of 0.89 and 0.87, and the highest Consistency value 0.84, and the lowest RMSE* value of 1.7.

PEMO-Q, and PEMO-QISO, but also the more complex CASP-QISO, and CASP-QnoExp, gave poor predictions, with Accuracy values ≤ 0.41 and Monotonicity values ≤ 0.22. GPSM4 predictions agreed well with subjective ratings, indicated by Accuracy and Monotonicity values of 0.87 and 0.86, and a high Consistency value of 0.78. MoBi-Q showed good prediction performance, indicated by Accuracy, Monotonicity, and Consistency...
values of 0.83, 0.8, and 0.73, respectively. For 32 test items (out of 140 test items), the binaural pathway predicted greater quality degradations than the monaural pathway.

**Hear-Through Mode**

Figure 3 shows subjective scores and objective scores for GPSM, MoBi-Q, PEAQ, BAM-Q, D, HASQIv2, and PEMO-Q for the HTM database. Figure 3 shows that PEAQ predictions differed substantially from subjective ratings (Accuracy of 0.2, Monotonicity of 0.1). The naturalness measure D gave accurate predictions (see Figure 3) indicated by Accuracy and Monotonicity of 0.86 and 0.87, while Rnonlin gave poor prediction performance (Accuracy of 0.18, Monotonicity of 0.14). Soverall achieved moderate performance, indicated by a Monotonicity value of 0.73. However, low values for Accuracy and Consistency of 0.34 and 0.44 and a high value of RMSE* of 3.3 represent the nonlinear relationship between objective and subjective scores as well as some large deviations between objective and subjective scores of up to 60 points on the 0 to 100 point scale used by the MUSHRA protocol.

HASQI performed rather poorly (Accuracy of 0.15, Monotonicity of 0.14) for the HTM database, while HASQIv2 showed moderate prediction performance, indicated by Accuracy, Monotonicity, and Consistency values of 0.62, 0.56, and 0.59, respectively.

PEMO-Q, and PEMO-QISO, and also CASP-QISO, and CASP-QnoExp gave poor predictions with Accuracy values ≤ 0.19 and Monotonicity values ≤ 0.23. GPSM4 predictions achieved the best performance, indicated by Accuracy, Monotonicity, and Consistency values of 0.93, 0.91, and 0.91 and a low value for RMSE* of 1.3. BAM-Q predictions, purely based on binaural distortions, were poor (Accuracy of 0.33, Monotonicity of 0.27). MoBi-Q gave good prediction performance (Accuracy of 0.79, Monotonicity of 0.81). For 32 test items (out of 120 test items), the binaural pathway predicted greater quality degradations than the monaural pathway.

**Overall Performance**

To compare the overall performance of the models, the average Accuracy, Monotonicity, Consistency, and RMSE* and their standard deviation across databases are summarized in Figure 4, where the four overall best performing instrumental measures are highlighted in green. Because distortions in the AFC, ATHD, and HTM databases were assessed by different listener groups and in different comparison contexts (including different anchor signals), performance measures were calculated for each database and then compared across the three databases.

In each of the panels in Figure 4, GPSM, D, MoBi-Q, and HASQIv2 (with exception of Monotonicity) achieved the best average prediction performance of all instrumental measures examined in this study. The mean Accuracy, Monotonicity, Consistency, and RMSE* and the standard deviation for the four best performing measures across the AFC, ATHD, and HTM databases are given in Table 2. This table shows that GPSM achieved D and HASQIv2 was somewhat lower than for GPSM and D. As shown in Table 2, GPSM and D show the same standard deviation for Accuracy, Monotonicity, and Consistency, while the RMSE* values of GPSM showed slightly larger variability than those of D. A one-way repeated-measures analysis of variance showed a significant main effect of measure, \( F(1.7, 3.5) = 10.5, p < 0.05 \) on RMSE*. A post hoc pairwise comparison using the least significant difference test (protected \( t \) test) showed no significant RMSE* differences between GPSM, D, MoBi-Q, and HASQIv2. For each of the models, GPSM, D, and MoBi-Q, a significant RMSE* difference to Rnonlin, PEAQ, PEMO-Q, PEMO-QISO, CASP-QISO, and CASP-QnoExp was observed, while there was no significant RMSE* difference to HASQI and Soverall. Further, there were no significant RMSE* differences between HASQIv2 and the other instrumental measures.

**Discussion**

**Comparison of the Instrumental Measures**

**Analysis of Auditory Cues Used by the Top Four Measures.** The best performing audio quality models in this study, GPSM, MoBi-Q, HASQIv2, and D, explicitly account for spectral distortions by evaluating auditory excitation patterns. Therefore, it can be concluded that (monaural) spectral cues are highly relevant for the hearable algorithms assessed in this study. However, this also raises the question of what differences between these four models are responsible for the differences in their prediction performance.

The underlying procedure for predicting quality degradations produced by spectral distortions is identical for GPSM and MoBi-Q. Prediction differences can be explained by (a) an additional effect of binaural distortions accounted for by MoBi-Q, (b) a slightly modified GPSM front end (see Fleßner et al., 2019) in MoBi-Q that is largely insensitive to binaural cues, and (c) to obtain the overall quality measure of MoBi-Q, the GPSM output measure OPM was modified for combination with the binaural output measure binQ.
HASQIv2 applies a slightly modified version of D to account for spectral distortions. The main differences between these models are additional auditory processing stages applied by HASQIv2 for the analysis of TFS and spectral envelope differences.

The spectral cue analysis of GPSMq and MoBi-Q on one hand, and HASQIv2 and D on the other hand, mainly differs in auditory frequency range and in the postprocessing of auditory excitation patterns. Both GPSMq and MoBi-Q calculate auditory excitation patterns, for filter center frequencies from 315 to 12 500 Hz, while for D the center frequencies range from 55 to 16 800 Hz (Moore & Tan, 2004, suggested that speech is evaluated with a filter range from 123 to 10 900 Hz). The lowest auditory filter center frequency of 55 Hz for D agrees with recent findings of Jurado and Moore (2010) suggesting that there are no auditory filters with center frequencies below about 50 Hz. For HASQIv2, intended to predict speech quality, center frequencies range from 80 to 8000 Hz.

To examine the effect of auditory filter center frequency range on prediction performance, GPSMq (large symbols) and D predictions (small symbols) for the HTM database were calculated for different frequency ranges, as shown in Figure 5. Right-pointing triangles (gray; red online) are for the lowest auditory filter center frequency, as indicated on the x axis, while the highest auditory filter was always centered at 16 kHz for GPSMq and 16.8 kHz for D. The leftmost, small, closed right-pointing triangle represents the Monotonicity value of D for a filter range from 55 to 16 800 Hz. Left-pointing triangles (black) are for the highest auditory filter center frequency indicated on the x axis, while the center frequency of the lowest auditory filter was always set to 63 Hz for GPSMq and 55 Hz for D. The rightmost, small, closed left-pointing triangle represents the Monotonicity value of D for a filter range from 55 to 16 800 Hz. The left abscissa indicates Accuracy (open symbols), while the right abscissa indicates Monotonicity (closed symbols). To make GPSMq predictions comparable to those for the linear measure D, GPSMq predictions are here based on local power-based SNRs, and thus named GPSMqDC in the following. This GPSMqDC accounts only for spectral (linear) distortions, while modulation-based features are not considered. Because the GPSMqDC output was
Figure 5. Effect of the Range of Auditory Filter Center Frequencies on Accuracy (Open Symbols) and Monotonicity (Closed Symbols) for GPSM\textsubscript{DC} (Large Symbols), Purely Based on Local Power-Based SNRs, and D (Small Symbols). Right-pointing triangles show the effect of varying the center frequency of the lowest auditory channel, while the center frequency of the highest auditory channel is fixed at 16 kHz for GPSM\textsubscript{DC} and 16.8 kHz for D. The left-pointing triangles are for a variation of the highest auditory channel, while the lowest auditory channel has a fixed center frequency of 63 Hz for GPSM\textsubscript{DC} and 55 Hz for D.

not transformed by a logarithmic function, as is done for the overall GPSM\textsuperscript{q} measure (see Equation 6 in Biberger et al., 2018), subjective and objective quality scores show a curvilinear relationship. Therefore, in the following mainly the Monotonicity (Spearman rank correlation coefficient) values are considered, as they are not affected by such a curvilinear relationship. GPSM\textsubscript{DC} reached the highest Monotonicity values from 0.90 to 0.91 (Accuracy: 0.87 – 0.90) when the center frequency of the lowest auditory filter was ≤ 315 Hz as shown by large diamonds in Figure 5. The Monotonicity dropped to 0.88 when the center frequency of the lowest auditory filter was 400 Hz and dropped from 0.79 to 0.72 when the center frequency of the lowest auditory filter was changed from 4000 to 5000 Hz. GPSM\textsubscript{DC} reached the highest Monotonicity values from 0.91 to 0.93 (Accuracy: 0.89 – 0.90) when the center frequency of the highest auditory filter was ≥ 3150 Hz (see large, closed left-pointing triangles in Figure 5). The Monotonicity of GPSM\textsubscript{DC} was reduced from 0.91 to 0.85 when the center frequency of the highest auditory filter was changed from 3150 to 2500 Hz and dropped further from 0.75 to 0.64 when the center frequency was changed from 1250 to 1000 Hz. To summarize, GPSM\textsubscript{DC} predictions suggest that best performance is achieved with a center frequency ≤ 315 Hz for the lowest auditory filter and a center frequency ≥ 4 kHz for the highest auditory filter. As the original center frequencies of the lowest (315 Hz) and highest (12,500 Hz) auditory filters of GPSM\textsubscript{DC} (Monotonicity: 0.9) lie within the suggested range of auditory filter bandwidths, a wider bandwidth, for example, from 63 to 16 000 Hz (Monotonicity of GPSM\textsubscript{DC}: 0.91) as it is used by D, would not degrade the prediction performance of GPSM\textsubscript{DC}, at least for the HTM database.

D reached the highest Monotonicity value of 0.87 (Accuracy: 0.85 – 0.86) when the center frequency of the lowest auditory filter was ≤ 63 Hz, as shown by the small, closed right-pointing triangles in Figure 5. Monotonicity dropped to 0.84 when the center frequency of the lowest auditory filter was 200 Hz and dropped from 0.74 to 0.53 when the center frequency of the lowest auditory filter was changed from 2500 to 3150 Hz. D reached the highest Monotonicity values from 0.83 to 0.87 (Accuracy: 0.83 – 0.86) when the center frequency of the highest auditory filter was ≥ 8000 Hz (see small, closed left-pointing triangles in Figure 5). The Monotonicity of D was reduced from 0.80 to 0.76 when the center frequency of the highest auditory filter was changed from 8000 to 6300 Hz and dropped further from 0.7 to 0.64 when the center frequency was changed from 3150 to 2500 Hz. Figure 5 shows that predictions for D are more sensitive to frequency range variations than to predictions for GPSM\textsubscript{DC}. This implies that there might be more redundant information in GPSM\textsuperscript{q} than in D, which allows reduction of the frequency range of GPSM\textsuperscript{q} without a significant degradation of prediction performance. Further, the results in Figure 5 confirm that a frequency range from 55 to 16 800 Hz (2 to 40 ERB), as suggested by Moore and Tan (2004), gave the highest Accuracy and Monotonicity values for D.

The earlier analysis implies that differences in frequency range are probably not the reason for performance differences between GPSM\textsuperscript{q} and D, but rather differences in the postprocessing of auditory excitation patterns. GPSM\textsuperscript{q} and MoBi-Q assess (first-order) excitation differences. A comparison of the mean Accuracy (GPSM\textsubscript{DC}: 0.88; D: 0.90) and mean Monotonicity (GPSM\textsubscript{DC}: 0.91; D: 0.91) for GPSM\textsubscript{DC} and D across the AFC, ATHD, and HTM databases indicates that both first-order and second-order auditory excitation differences are suitable for capturing the spectral distortions in the databases used in this study.

Prediction Performance of the Instrumental Measures. PEMO-Q, CASP-Q, and \textit{Rnonlin} by design do not explicitly account for spectral distortions and thus gave on
average poor to moderate prediction performance for the three databases of this study. They have been demonstrated to give accurate predictions for nonlinear distortions (see Harlander et al., 2014; Huber & Kollmeier, 2006; Tan et al., 2004), but this is less relevant for audio quality predictions of the AFC, ATHD, and HTM databases.

PEAQ gave a very high Monotonicity value of 0.97 but a substantially lower Accuracy value of 0.78 for the AFC database. This indicates a curvilinear relationship between subjective and objective quality ratings (see Figure 1), while the selected MOVs captured most of the signal degrading aspects. The audio quality of the stimuli in the ATHD and HTM databases was often rated as intermediate, while PEAQ was intended to predict quality for small signal degradations of audio codecs, which could explain the poor prediction performance. Creusere et al. (2007) demonstrated that the prediction performance of PEAQ substantially increased when the MOVs were individually weighted for audio sequences with small (sequences compressed at 32 to 64 kb/s, resulting in good to excellent subjective quality ratings) and large (sequences compressed at 8 to 16 kb/s, resulting in poor to fair subjective quality ratings) distortions. Applying such a recalculated weighting of the MOVs might also help to increase the prediction performance of PEAQ.

HASQI (mean Accuracy: 0.53) achieved substantially lower average prediction performance than HASQIv2 (mean Accuracy: 0.82). This is surprising because the two models apply the same concept, while using different models of auditory periphery, to account for linear distortions, which are dominant for the three databases used in this study. A comparison of the mean Accuracy across the three databases only based on the linear parts clearly demonstrates the advantage of the revised peripheral stages in HASQIv2 (mean Accuracy: 0.64; mean Monotonicity: 0.67) compared with HASQI (mean Accuracy: 0.5; mean Monotonicity: 0.54) but does not explain such big differences in the overall performance. Motivated by the work of Tan et al. (2004), HASQIv2 includes a short-time correlation-based analysis of the TFS that is absent in HASQI. The combined TFS and cepstrum correlation analysis, representing the nonlinear analysis of HASQIv2, captured the effects of a large number of distortions in this study, as the mean Accuracy across the three databases was larger for the nonlinear part (mean Accuracy: 0.79; mean Monotonicity: 0.77) than for the linear part (mean Accuracy: 0.64; mean Monotonicity: 0.67). For each of the three databases, HASQIv2 predictions based on cepstrum correlation (mean Accuracy: 0.76) achieved higher Accuracy values than predictions based on TFS analysis (mean Accuracy: 0.62), while the highest Accuracy was obtained by combining the two features. Moreover, cepstrum correlation based predictions of HASQIv2 gave considerably better performance than cepstrum correlation based predictions of HASQI, which underlines the importance of the revised peripheral stages in HASQIv2. Therefore, the joint cepstrum correlation and TFS analysis in HASQIv2, in combination with the revised model of auditory periphery, explain the differences in prediction performance between HASQI and HASQIv2. While the linear part of HASQI (a modified version of D) has moderate prediction performance, the original version of D (mean Accuracy: 0.90, mean Monotonicity: 0.91) gave very accurate predictions. The comparison of these two measures implies a nonoptimal modification of D in HASQI for the databases tested in this study.

BAM-Q predictions were calculated only for the HTM database, for which some hearables had substantial interaural distortions, while the other databases do not have or have only slight binaural distortions. The poor prediction performance of BAM-Q indicates that binaural distortions are not the dominant factor for audio quality degradations in the HTM database.

The combined monaural and binaural model MoBi-Q was one of the four best performing instrumental measures, as shown in Figure 4. The accurate predictions of MoBi-Q for the AFC (Accuracy: 0.96; Monotonicity: 0.97) and ATHD (Accuracy: 0.83; Monotonicity: 0.8) databases with no or limited binaural distortions indicate that the modified monaural GPSM captures most of the relevant distortions. The technical evaluation of Denk et al. (2020) for the hearables of the HTM database revealed large interaural differences for some devices, which, however, were subject to large monaural distortions as well. An audio quality model that combines monaural and binaural aspects of audio quality may give more accurate quality predictions for a database containing both monaural and binaural distortions than a quality model that considers either monaural or binaural distortions. The monaural GPSM provided the most accurate prediction performance (Accuracy: 0.93) for the HTM database. The reason why MoBi-Q (Accuracy: 0.83) predictions were less accurate for that database is further examined in the two last subsections within the discussion.

Despite the success of purely monaural audio quality models in this study, it should be mentioned that such approaches are not expected to give sufficiently reliable quality ratings for applications such as spatial sound reproduction or binaural algorithms in hearing aids, where signal processing strategies might introduce stronger interaural differences than in the current study. Only instrumental measures that additionally capture binaural quality aspects are expected to accurately predict audio quality for such applications as listeners also use monaural and binaural information to make their
quality judgment. Thus, an instrumental quality measure combining monaural and binaural cues is in principle more powerful as a purely monaural quality measure, as it covers additional quality aspects. On the other hand, purely monaural quality measures can give accurate quality predictions when monaural distortions are dominant as shown in this study.

**Influence of Training Data Sets**

One goal of this study was to assess the applicability of instrumental quality measures to distortions typically occurring in hearables for both music and speech signals. Besides aiming at accurate predictions for certain types of distortion, many instrumental measures are designed to predict aspects of either speech or audio (music) quality. Accordingly, different data sets have been used by the developers to optimize their instrumental quality measures, as described earlier.

As reported by Kates and Arehart (2010, 2014a), HASQI and HASQIV2 were trained with speech stimuli to predict effects on speech quality. Hearing-Aid Audio Quality Index (not included in this study) is an adapted instrumental measure, closely related to the HASQI measures, but intended to predict audio quality. Because all three databases used in this study contain music and speech stimuli, while the HASQI measures were originally designed for speech quality, better overall performance can be expected when only speech stimuli are considered. Indeed, HASQIV2 applied to the speech signals only achieved higher Accuracy values (AFC: 0.97; ATHD: 0.94; HTM: 0.66) than shown in Table 1. Interestingly, applying HASQIV2 exclusively to music signals resulted in only slightly lower Accuracy values (AFC: 0.97; ATHD: 0.88; HTM: 0.63) than for the speech quality predictions. Thus, HASQIV2 is able to capture relevant signal degrading aspects for distortions occurring in this study for music and speech signals, while the most critical point for performance when applied to both types of signals seems to be the joint representation of predictions for speech and music quality.

Predictions of the linear measure D are based on the final fitting parameters suggested in Table 1 of Moore and Tan (2004), which can be expected to give reasonable predictions for speech and music signals with linear distortions. However, in the current study, D was always based on center frequencies from 55 to 16 800 Hz, which agrees with the findings of Moore and Tan (2003) for music signals, while according to that study frequencies below about 123 Hz and above 10 900 Hz do not contribute much to quality ratings for linearly distorted speech signal. Thus, a narrower frequency range in combination with speech-optimized fitting parameters given in Table 1 of Moore and Tan (2004) might further improve the prediction performance of D.

In this study, the same fitting parameters as used in Biberger et al. (2018), derived by averaging the fitting parameters for speech and music given in Figure 2 of Tan et al. (2004), were applied to Rnonlin. Such averaged parameters were used to enable a fair comparison to other out-of-the-box models, while in Tan et al. (2004), the fitting was done separately for the speech and music signals for each database, to linearize the relationship between subjective and objective ratings. As can be expected, optimization of Rnonlin to the AFC database improved Accuracy from 0.83 to 0.96, while Monotonicity values were not affected. Accordingly, prediction performance of Soverall, representing the combination of the linear measure D and the nonlinear measure Rnonlin, was also improved (Accuracy: 0.98; Monotonicity: 0.95; Consistency: 0.82; RMSE*: 1.4) by applying the optimized Rnonlin. However, optimizing Rnonlin to the ATHD and HTM databases did not substantially improve the prediction performance of Rnonlin and Soverall. Distortions occurring in those databases might not be sufficiently captured by Rnonlin.

As reported by Huber and Kollmeier (2006) and Thiede et al. (2000), PEMO-Q and PEAQ were both mainly optimized for music signals and for low-bit rate audio codecs, which often introduced smaller signal degradations than the algorithms and devices considered in the current study. Although PEMO-Q does not explicitly account for spectral cues, which are important here, it can be expected that PEMO-Q and PEAQ would benefit from recalibration with a data set containing similar distortions as used in this study.

GPSMq was trained with music and speech signals and a large variety of distortions. This might explain why it accounts well for the variety of distortions occurring in the current study. As for the other instrumental measures, adjusting model parameters for speech and music signals or optimizing the combination of auditory features according to the distortions in the current data sets is also expected to improve prediction results.

A more detailed assessment of the influence of the databases used for optimizing the models is beyond the scope of this article. However, it can be concluded that besides the auditory feature representation in the models, the data sets used for model calibration have a strong impact on prediction performance, as they define stimulus properties and the perceptual range of signal impairments, where both aspects influence the fitting or learning procedure used to derive an optimal feature combination.

**Effects of Room Reflections on Instrumental Quality Ratings**

To represent realistic room situations, the recording rooms of the ATHD and HTM databases had reverberation times ($T_{60}$) ranging from about 0.35 s to 1.4 s. In
were used. About 0.35 s, 0.45, and 1.4 s from the ATHD databases.

To assess the influence of reverberant signal processed by algorithms or hearables.

The prediction performance of HASQIv2 dropped for the longest reverberation time of about 1.4 s. Considering the importance of spectral cues for GPSMq, D, and MoBi-Q predictions in this study, it seems that spectral cues are hardly affected by reverberation. As already mentioned for HASQIv2, the nonlinear part appears to be more important for quality predictions of distortions occurring in this study than the linear part. For the echoic recording room with $T_{60} \approx 1.4 \text{s}$, prediction performance of this nonlinear part of HASQIv2 (HASQIv2nonlin) was clearly degraded, as shown by the Accuracy and Monotonicity values in Table 3, while the performance of the linear part was barely degraded by reverberation. The severely degraded prediction performances of HASQIv2TF and HASQIv2CepCorr for $T_{60} \approx 1.4 \text{s}$ indicate that both TFS and cepstral correlation features may be unreliable predictors of audio quality in rooms with moderate reverberation. This implies that, at least for the ATHD database, spectral cues are more reliable than TFS/cepstrum correlation.

Comparison of the HTM Database With Technical Measures

Denk et al. (2020) technically evaluated the hear-through mode of the hearing devices from the HTM database to identify artefacts that potentially impair audio quality. Hear-through impulse responses, the hear-through frequency response at the eardrum, conservation of binaural cues, and self-noise were measured. These measures...
revealed large differences between the HTMs of the devices and the open ear, potentially affecting perceived acoustic transparency.

In the following, auditory-model-based quality predictions are compared with the technical measures used by Denk et al. (2020) and corresponding subjective data of Schepker et al. (2020), to assess whether the measured differences between the hearing devices are reflected in the predictions of the audio quality models. For this comparison, the focus lay on the monaural models GPSM\textsuperscript{q} and D, as they provided accurate predictions for the HTM database, and the MoBi-Q that accounts for monaural and binaural distortions.

Device F gave the lowest subjective quality scores, which could be explained by large delay differences (left: 0.8 ms, right: 10.4 ms) between the left and right devices, poor conservation of ILDs and ITDs, and spectral ripples in the hear-through (diffuse-field) frequency response of the right channel (see Figures 4, 6, and 7 in Denk et al., 2020). Device F gave the lowest scores for Devices A to C showed on average the highest and similar quality scores (see Figure 3). The differences in the subjective quality ratings for Devices A to C must be explained by monaural differences, which can be observed in the middle panel of Figure 6 in Denk et al. (2020), showing HRTFs at the eardrum for the hear-through case. The hear-through response of Device C matched the open-ear response over a large frequency range. Large deviations from the open-ear response only occurred at frequencies above 10 kHz. The hear-through response of Device A showed spectral ripples below 1 kHz, while the response of Device B showed a large attenuation below 0.5 kHz and above 10 kHz compared with the open-ear response. All of the three best performing monaural models GPSM\textsuperscript{q}, D, and HASQIv2, which correctly predicted higher scores for Device C than for Devices A and B, explicitly account for (monaural) spectral cues. Other audio quality models that do not explicitly represent (monaural) spectral cues, such as PEMO-Q, CASP-Q and R\textsubscript{nonlin}, failed to account for such distortions.

Device J (denoted as UOL Tr. Earpiece in Denk et al., 2020) included some distortions of ITDs and ILDs. Despite the fact that the hear-through response of the right channel of Device J showed a substantial comb-

---

**Table 3.** Accuracy (Acc) and Monotonicity (Mon) Results for GPSM\textsuperscript{q}, D, MoBi-Q, and HASQIv2 Across Female Speech and Jazz Music Samples for Different Reverberation Times ($T_{60}$).

|                  | $T_{60} \approx 0.35$ s | $T_{60} \approx 0.45$ s | $T_{60} \approx 1.4$ s |
|------------------|-------------------------|-------------------------|------------------------|
|                  | Acc | Mon | Acc | Mon | Acc | Mon |
| GPSM\textsuperscript{q} | 0.86 | 0.81 | 0.93 | 0.96 | 0.88 | 0.91 |
| D                | 0.92 | 0.89 | 0.95 | 0.94 | 0.90 | 0.92 |
| MoBi-Q           | 0.83 | 0.79 | 0.91 | 0.93 | 0.83 | 0.84 |
| HASQIv2          | 0.94 | 0.89 | 0.94 | 0.93 | 0.81 | 0.76 |
| HASQIv2\textsubscript{lin} | 0.89 | 0.86 | 0.87 | 0.89 | 0.82 | 0.86 |
| HASQIv2\textsubscript{nonlin} | 0.89 | 0.87 | 0.86 | 0.79 | 0.67 | 0.64 |
| HASQIv2\textsubscript{TFS} | 0.75 | 0.70 | 0.78 | 0.80 | 0.55 | 0.54 |
| HASQIv2\textsubscript{CepCorr} | 0.80 | 0.82 | 0.75 | 0.74 | 0.56 | 0.54 |

*Note.* Predictions based on the linear part of HASQIv2, denoted HASQIv2\textsubscript{lin}, and predictions based on the nonlinear part denoted HASQIv2\textsubscript{nonlin} are shown. Predictions of the nonlinear part of HASQIv2 are based on TFS and cepstrum correlation features. To disentangle their contribution to HASQIv2\textsubscript{nonlin}, prediction results of HASQIv2\textsubscript{TFS} and HASQIv2\textsubscript{CepCorr} are also provided. Bold values indicate the best performing objective measure for Accuracy and Monotonicity. GPSM\textsuperscript{q} = Generalized Power Spectrum Model for quality; HASQIv2 = Hearing-Aid Speech Quality Index version 2.
filter effect for frequencies < 1 kHz (see Figure 6 in Denk et al., 2020), it achieved fairly high subjective quality ratings, as predicted by GPSM⁴ and D. This demonstrates the importance of using auditory-based quality models, as technical measures can show large differences between the processed (Device J) and the unprocessed (open-ear) signals, which might have only a minor impact on subjective quality ratings.

Self-noise was another factor measured by Denk et al. (2020) to characterize hearing device performance. This measure makes sense in a quiet environment, where self-noise generated by the devices might disturb listeners and thus reduce audio quality. However, quality rating scores were obtained for speech and music signals in rooms with mild reverberation ($T_{60} \approx 0.45$ s), where it was not expected that self-noise would influence quality ratings. This is supported by the fact that Device C received the highest subjective quality score but had the highest self-noise. It is not expected that self-noise had an effect on the audio quality predictions in this study. Further, it should be mentioned that no self-noise normalization was carried out in the study of Schepker et al. (2020), to preserve potential effects of self-noise in realistic situations.

As shown in this section, it may be beneficial for algorithm developers as well as for the developers of auditory models to jointly compare measures that technically describe system properties with predictions from (reliable) audio quality models.

**Implication for Joint Predictions of Monaural and Binaural Distortions Occurring in Hearing Devices**

MoBi-Q predictions for the HTM database showed some deviations from subjective scores, which might be explained by the method of combining the outputs of the monaural GPSM⁴ (OPM$_{dual}$ with binaural modification to reduce its sensitivity to ILD and ITD differences) and the binaural BAM-Q (binQ), as mentioned in the previous section.

Here, two modifications of MoBi-Q are suggested. First, sigmoid functions were applied to OPM$_{dual}$ and binQ, where the slope and the sigmoid’s midpoint were fitting parameters. The values of these parameter as shown in the denominators of Equation 2, resulted from a least-squares fitting procedure to the subjective quality ratings for the HTM database.

$$\text{Overall quality} = \frac{1}{1 + e^{-0.1188(\text{OPM}_{dual} - 38.9817)}} + \frac{1}{1 + e^{-0.0192(\text{binQ} + 20.8263)}}$$

Second, the transformed OPM$_{dual}$ and binQ values were added to predict overall quality. It should be noted that Equation 2 is used for all stimulus types. The sigmoid function of Equation 2, shown in Figure 7, allows overall quality to range from 0.6 (very strong differences between reference and test signals) to about 1.9 (no perceptible differences). A rescaling was applied to bound the MoBi-Q$_{add}$ quality scores between 0 and 1.

The sigmoid transformation adapts the OPM$_{dual}$ and binQ scores, which were originally calibrated to exclusively monaural and binaural distortions in the database of Fleßner et al. (2019), to the current HTM database. The fitted sigmoid function parameters allow assessment of the contribution of monaural and binaural quality aspects. The revised MoBi-Q version is denoted MoBi-Q$_{add}$ in the following.

MoBi-Q$_{add}$ achieved better prediction performance for the HTM database ($\text{Accuracy}: 0.92; \text{Monotonicity}: 0.9, \text{Consistency}: 0.93; \text{RMSE}^*: 1.3$ dB) than MoBi-Q ($\text{Accuracy}: 0.79; \text{Monotonicity}: 0.81; \text{Consistency}: 0.78; \text{RMSE}^*: 2.2$ dB).

As shown in Figure 6, MoBi-Q$_{add}$ also gave very good prediction performance for the AFC ($\text{Accuracy}: 0.95; \text{Monotonicity}: 0.96; \text{Consistency}: 0.62; \text{RMSE}^*: 2.1$ dB) and ATHD ($\text{Accuracy}: 0.85; \text{Monotonicity}: 0.82; \text{Consistency}: 0.75; \text{RMSE}^*: 2.2$ dB) databases, resulting in better overall performance ($\text{ACC}: 0.91; \text{Mon}: 0.89; \text{Con}: 0.77; \text{RMSE}^*: 1.9$ dB) of MoBi-Q$_{add}$ than for MoBi-Q (see Table 2). The sigmoid functions of Equation 2, shown in Figure 7, indicate that overall quality was...
mainly driven by the monaural GPSM. Further, the contribution of the binaural BAM-Q to overall quality was strongest for strong signal degradations.

To assess MoBi-Q add for other signal distortions, the database of Fleßner et al. (2019) was used, which introduced a variety of monaural and binaural distortions to music, noise, and speech signals. Here, MoBi-Q add (Accuracy: 0.83; Monotonicity: 0.79; Consistency: 0.92; RMSE*: 1.4 dB) gave slightly lower prediction performance than MoBi-Q (Accuracy: 0.86; Monotonicity: 0.8; Consistency: 0.95; RMSE*: 1.3 dB). This can be explained by differences in the contribution of monaural and binaural distortions between the two databases: Fleßner et al. (2019) used artificial monaural and binaural distortions that gave comparable subjective quality ratings and thus similar perceptual salience of monaural

and binaural distortions. The model predictions in this study indicate that the devices in the HTM database mainly introduced monaural distortions, while binaural distortions were of minor importance. Because MoBi-Q add was optimized using the HTM database, a strong contribution of monaural distortions is also represented in the combination of the monaural and binaural model pathways in MoBi-Q add. For that reason, MoBi-Q add slightly underestimated quality degradations from binaural distortions in the database of Fleßner et al. (2019).

A further relevant aspect concerns differences of the monaural component of MoBi-Q from the monaural GPSM (Biberger et al., 2018). The monaural GPSM in MoBi-Q is applied to the reference and the test signals from the left and the right ears and thus potentially shows some sensitivity to interaural differences mediated by monaural signal features and artifacts. A modified GPSM was used in MoBi-Q to reduce its sensitivity to ILDs and ITDs and to ensure that the binaural features ILDs and ITDs were only captured by the binaural component of the model BAM-Q (see Figure 4 in Fleßner et al., 2019). Because GPSM provided very accurate predictions for the three databases used here, the question arises whether the modified GPSM in MoBi-Q can be replaced by the original GPSM without impairing the prediction performance of MoBi-Q. To assess this, an additive combination applying the sigmoid functions of Equation 2, but using different fitting parameters to combine the outputs of the original GPSM (without binaural modification) and BAM-Q, was tested (MoBi-Q add, origGPSM). The results given in Table 4 indicate slightly better prediction performance for the current AFC, ATHD, and HTM databases. However, across all four databases (AFC, ATHD, HTM, and Fleßner et al., 2019), MoBi-Q add provided consistently high prediction performance achieving Accuracy values ≥ 0.83 and a mean Accuracy of 0.89 for all databases, and so this appears to be the best

Figure 7. Sigmoid Functions (see Equation 2) Applied to the Monaural GPSM With Binaural Modification and the Binaural BAM-Q in MoBi-Q add. The x axis represents the objective quality score from the original model, while the abscissa represents the transformed quality score.

Table 4. Accuracy (Acc), Monotonicity (Mon), Consistency (Con), and RMSE Results for the Suggested Additive Combination of the Outputs of GPSM (With Binaural Modification (Fleßner et al., 2019) and BAM-Q (Denoted as MoBi-Q add) and for an Alternative Approach That Also Applies an Additive Combination, but Using the Outputs of the Original GPSM (Biberger et al., 2018) Without Binaural Modification and BAM-Q (Denoted MoBi-Q add, origGPSM).

| Measure | MoBi-Q add | MoBi-Q add, origGPSM | MoBi-Q |
|---------|------------|----------------------|---------|
| Acc     | 0.95       | 0.96                 | 0.96    |
| Mon     | 0.96       | 0.94                 | 0.97    |
| Con     | 0.62       | 0.70                 | 0.67    |
| RMSE*   | 2.1        | 1.8                  | 2.2     |
| DB      |            |                      |         |
| AFC     | 0.85       | 0.87                 | 0.83    |
| ATHD    |            | 0.74                 | 0.80    |
| HTM     | 0.92       | 0.94                 | 0.95    |
| Fleßner et al. (2019) | 0.83 | 0.75 | 0.86 |

Note. For comparison, results for the original MoBi-Q (Fleßner et al., 2019), which were presented in Table 1 and in the text, are reproduced in this table. Bold values indicate the best performing objective measure for Accuracy, Monotonicity, Consistency, and RMSE*. ATHD = acoustically transparent hearing device; AFC = adaptive feedback cancelation; HTM = hear-through mode.
broadly applicable model version. It should be mentioned that the other instrumental measures used in this study, as far as they provide a proper feature representation, are expected to improve their prediction performance as well, when they are optimized for the distorted signals of the HTM database. However, it cannot be expected that monaural instrumental measures achieve better performance than MoBi-Q_add for the distortions used for the database of Fleßner et al. (2019). This was confirmed for GPSM^q, D, and HASQIv2 which obtained Accuracy values of 0.75, 0.64, and 0.18, respectively. Consequently, MoBi-Q_add shows the highest mean Accuracy across the AFC, ATHD, HTM and Fleßner et al. (2019) databases.

This analysis demonstrated why it is important to test instrumental measures with artificial signals, as well as with real algorithms or devices, to achieve a large variety of distortions with different monaural and binaural contributions. Although the proposed instrumental measure MoBi-Q_add was evaluated using four databases with different monaural and binaural distortions occurring in music, speech, and noise signals, further databases with other types of distortions (e.g., from noise reduction algorithms) related to hearables should be assessed in the future, to draw a more conclusive picture about its predictive power and limitations.

**Summary and Conclusions**

Thirteen auditory-based instrumental audio quality measures were evaluated using three databases including music, noise, and speech signals impaired by distortions that typically occur in smart headphones or hearables. The following conclusions can be drawn:

- The monaural GPSM^q (Biberger et al., 2018) and the measure of perceived naturalness D (Moore & Tan, 2004) achieved better average prediction performance across a large variety of signal distortions related to hearables than the other auditory-based quality models tested in this study. Two other quality measures, MoBi-Q (Fleßner et al., 2019) and HASQIv2 (Kates & Arehart, 2014a), also achieved high prediction performance for the distortions considered in this study.
- Accurate predictions of the perceptual effects of spectral distortions in instrumental quality measures are important for application to algorithms in smart headphones or hearables. Binaural distortions made lower contribution to perceived overall audio quality than monaural distortions.
- Audio quality predictions for distorted signals recorded in rooms with different reverberation times implied that spectral cues are more reliable for quality prediction in reverberation than cues based on TFS or cepstrum correlation.
- A modified and additive combination of the monaural and binaural quality components (GPSM^q and BAM-Q outputs) in MoBi-Q_add based on Fleßner et al. (2019) is suggested. MoBi-Q_add provided the best, consistently and homogeneously high prediction performance, achieving Pearson linear correlation coefficient values \( \geq 0.83 \) (a mean Pearson linear correlation coefficient value of 0.89) for the current three databases and the database of Fleßner et al. (2019). The suggested MoBi-Q_add will be made publicly available.¹

**Acknowledgments**

The authors would like to thank the members of Medizinische Physik and Birger Kollmeier for continued support. The authors would also like to thank James Kates and Kathryn Arehart for providing the HASQI and HASQIv2 code, Rainer Huber for providing his implementations of D and Rnonlin, and J.-H. Fleßner for helpful discussions regarding predictions of BAM-Q and MoBi-Q. Further, the authors would like to thank Brian C. J. Moore and the two anonymous reviewers for their helpful comments on an earlier version of the article.

**Declaration of Conflicting Interests**

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Deutsche Forschungsgemeinschaft (DFG – 352015383 – SFB1330 A2 and additionally, A4, and C1).

**ORCID iDs**

Thomas Biberger ¹ https://orcid.org/0000-0002-6314-1914
Florian Denk ¹ https://orcid.org/0000-0003-3490-123X

**Notes**

1. A MATLAB implementation of the MoBi-Q_add with revised back end is provided under: www.faame4u.com
2. The MUSHRA drag and drop (Völker et al., 2018) was designed to maximize the accessibility of MUSHRA for elderly and technically nonexperienced listeners, who constitute the typical target group in hearing aid evaluation. As shown in Figure 3 in Völker et al. (2018), the buttons representing the test items are placed via drag and drop within a rating field ranging from bad to excellent.
Schäfer, M., Bahram, M., & Vary, P. (2013, May). An extension of the PEAQ measure by a binaural hearing model. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada. https://doi.org/10.1109/icassp.2013.6639256

Schepker, H., Denk, F., Kollmeier, B., & Doclo, S. (2019, August). Subjective sound quality evaluation of an acoustically transparent hearing device. Proceedings of the 2nd AES Conference on Headphone Technology, San Francisco, USA.

Schepker, H., Denk, F., Kollmeier, B., & Doclo, S. (2020). Subjective quality evaluation of commercial hearing assistive devices with transparency features. Journal of the Audio Engineering Society, 68(7/8), 495–507. https://doi.org/10.17743/jaes.2020.0045

Seo, J.-H., Chon, S. B., Sung, K.-M., & Choi, I. (2013). Perceptual objective quality evaluation method for high quality multichannel audio codecs. Journal of the Audio Engineering Society, 61(7/8), 535–545.

Takanen, M., Wierstorf, H., Pulkki, V., & Raake, A. (2014, August). Evaluation of sound field synthesis techniques with a binaural auditory model. Proceedings of the 55th AES Conference, Helsinki, Finland.

Tan, C.-T., Moore, B. C. J., Zacharov, N., & Mattila, V.-V. (2004). Predicting the perceived quality of nonlinearly distorted music and speech signals. Journal of the Audio Engineering Society, 52(7/8), 699–711.

Temme, S. F. (2019, November). Testing audio performance of hearables. Proceedings of the 2nd AES Conference on Headphone Technology, San Francisco, USA.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K., & Feiten, B. (2000). PEAQ - The ITU standard for objective measurement of perceived audio quality. Journal of the Audio Engineering Society, 48(1/2), 3–29.

Völker, C., Bisitz, T., Huber, R., Kollmeier, B., & Ernst, S. M. A. (2018). Modifications of the MUlti Stimulus Test with Hidden Reference and Anchor (MUSHRA) for use in audiology. International Journal of Audiology, 57, 92–104. https://doi.org/10.1080/14992027.2016.1220680

Yates, G. K., Winter, I. M., & Robertson, D. (1990). Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. Hearing Research, 43(3), 203–220. https://doi.org/10.1016/0378-5955(90)90121-5