

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

Trends in Hearing
Volume 25: 1–22
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165211001219
journals.sagepub.com/home/tia



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

Received 10 February 2021; Revised 10 February 2021; accepted 18 February 2021

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

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Oldenburg, Germany

Corresponding author:

Thomas Biberger, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Straße 11, Oldenburg, Lower Saxony 26129, Germany.
Email: thomas.biberger@uni-oldenburg.de



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_{nonlin} ; 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^q; 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 GFSM^q, 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.¹

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 multi-layer 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 (2004b). 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 speech signals, from which two sets of optimized model parameters were derived.

The nonlinear distortion measure R_{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_{nonlin} . Since Tan et al. (2004) observed a curvilinear behavior between R_{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_{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_{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_{nonlin} .

The combined quality model S_{overall} of Moore et al. (2004) is based on the linear component S_{lin} and the nonlinear component S_{nonlin} derived from D and R_{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 τ is 0.3. S_{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 (2004b), 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 adaptation. 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_t), 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_t . 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_t . The ODG is derived by mapping the PSM_t 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-Q_{ISO} (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, PSM_t, 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-Q_{ISO} and CASP-Q_{noExp} as suggested by Harlander et al. (2014), were used in this study. CASP-Q_{ISO} additionally applies a hearing threshold based on ISO 226 (2003), while CASP-Q_{noExp} 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 GPSM^q (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). GPSM^q applies a linear, fourth-order gammatone filterbank with bandwidth equal to the equivalent rectangular bandwidth of the auditory filter (ERB_N; Glasberg & Moore, 1990; Moore & Glasberg, 1983) that simulates the behavior of the basilar membrane, followed by calculating the low-pass filtered Hilbert envelope (cutoff frequency of 150 Hz) to account for decreased modulation sensitivity at high modulation frequencies. The low-pass filtered Hilbert envelopes form the basis for calculating the local envelope power and the local DC power. The local DC power is calculated in rectangular windows with a fixed duration of 375 ms for each auditory filter. After modulation filterbank processing of the Hilbert envelopes, the local envelope power is calculated in rectangular windows, where the window duration is related to the inverse of the center frequency of the corresponding modulation bandpass filter. These calculation steps are performed for the reference and the test signals, from which local envelope-power signal-to-noise ratios (SNRs) and power SNRs are derived, averaged over time and combined across audio and modulation channels. The resulting single-valued envelope-power SNR and power SNR

are additively combined and transformed by a logarithmic function to give the objective perceptual measure (OPM). The data sets used for optimizing GPSM^q include speech and music signals processed by audio codecs, audio source separation, noise reduction algorithms, and loudspeaker and their subjective quality ratings from NH listeners.

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_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^q and the binaural BAM-Q. The GPSM^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^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^q (OPM_{dual}) and the binaural BAM-Q (binQ) by selecting the (transformed) output that showed the largest quality degradation (minimum operation):

$$\text{Overall quality} = \min(\log_{10}(0.0528 * \text{OPM}_{\text{dual}}), 0.0078 * \text{binQ}) \quad (1)$$

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 Cancellation. The *adaptive feedback cancellation* (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 cancellation algorithm (with perfect a priori knowledge about the feedback path) served as reference signals, while signals processed without feedback cancellation 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.

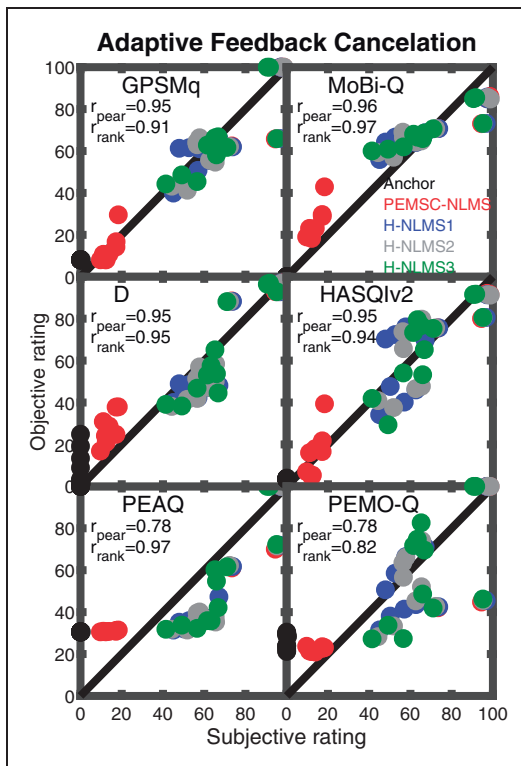


Figure 1. Audio Quality Predictions of GPSM^q, MoBi-Q, PEAQ, D, HASQiv2, and PEMO-Q for the *Adaptive Feedback Cancellation* (AFC) Database. Algorithms are represented by different colors. The individual prediction performance is given in each panel by Accuracy (r_{pear}) and Monotonicity (r_{rank}). The names of the algorithms are indicated in the right top panel.

GPSM^q = 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.

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 ($T_{60} \approx 0.35$ s, 0.45 s, 1.4 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

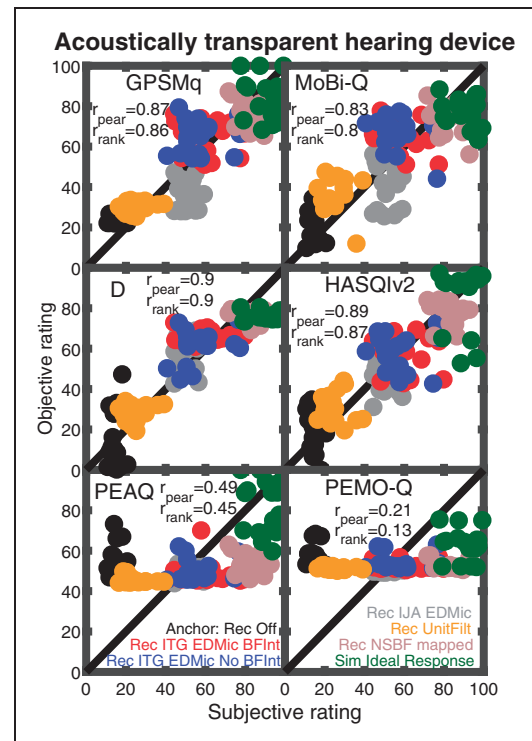


Figure 2. Audio Quality Predictions of GPSM^q, MoBi-Q, PEAQ, D, HASQiv2, and PEMO-Q for the *Acoustically Transparent Hearing Device* (ATHD) Database. Hearing device settings are indicated by different colors. Prediction performance is given in each panel by Accuracy (r_{pear}) and Monotonicity (r_{rank}). The algorithm names are indicated in the two bottom panels.

GPSM^q = 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.

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 framework² (Völker et al., 2018). Participants were instructed in writing to rate the perceived overall sound quality of each stimulus relative to 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, the reader is referred to Schepker et al. (2020). Subjective results are based on data for 17 NH subjects using a MUSHRA-like framework² (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 \pm 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

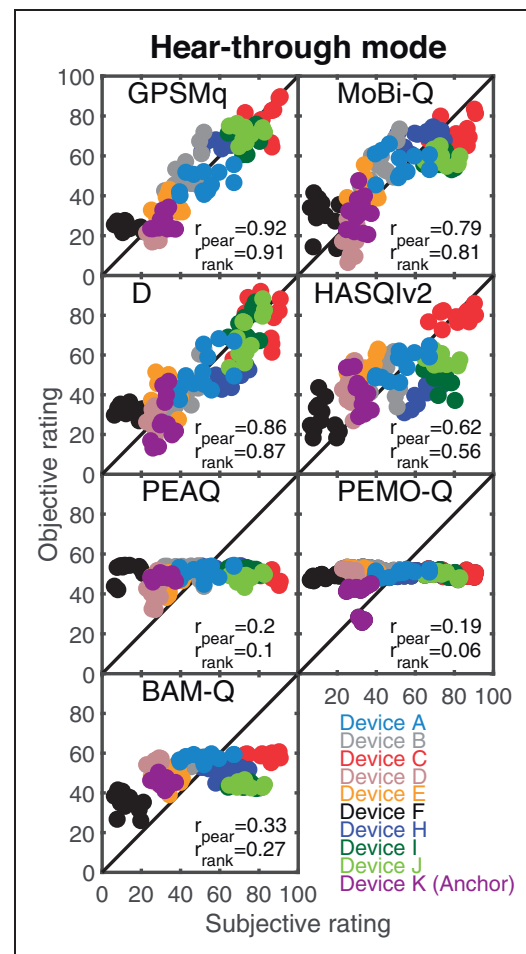


Figure 3. Audio Quality Predictions for GPSM^q, MoBi-Q, PEAQ, BAM-Q, D, HASQlv2, and PEMO-Q for the *Hear-Through Mode* (HTM) Database. Hearing devices are indicated by different colors. Prediction performance is given in each panel by *Accuracy* (r_{pear}) and *Monotonicity* (r_{rank}). For better visualization of the relationship between subjective and objective scores, the upper left panel shows GPSM^q 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 GPSM^q, which incorporates such perceptual limit by default. For further details about the devices, refer to Schepker et al. (2020). GPSM^q = Generalized Power Spectrum Model for quality; HASQlv2 = 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.

measures can be found in Emiya et al. (2011) and Harlander et al. (2014).

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 Cancellation

Figure 1 shows subjective scores and objective scores for GPSM^q, 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_{pear} 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). R_{nonlin} 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 R_{nonlin} 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 R_{nonlin} , S_{overall} , achieved values for *Accuracy* and *Monotonicity* of 0.92 and 0.94, respectively. The *Consistency* value of S_{overall} (0.48) fell between the *Consistency* values for D (0.67) and R_{nonlin} (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-Q_{ISO}, showed good prediction performance, with similar *Accuracy* (PEMO-Q: 0.78; PEMO-Q_{ISO}: 0.8) and *Monotonicity* (PEMO-Q: 0.82; PEMO-Q_{ISO}: 0.86). CASP-Q_{ISO} and CASP-Q_{noExp} gave poor predictions for the AFC database with *Accuracy* and *Monotonicity* values < 0.6 . GPSM^q predictions agreed well with subjective ratings, indicated by an *Accuracy* value of 0.95 and a *Monotonicity* value of 0.91 (see Figure 1). GPSM^q 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 GPSM^q, 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. R_{nonlin} showed poor prediction performance (*Accuracy* value of 0.17, *Monotonicity* value of 0.01) for the ATHD database. The combined measure S_{overall} based on D and R_{nonlin} , 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-Q_{ISO}, but also the more complex CASP-Q_{ISO}, and CASP-Q_{noExp}, gave poor predictions, with *Accuracy* values ≤ 0.41 and *Monotonicity* values ≤ 0.22 . GPSM^q 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^q, 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 R_{nonlin} gave poor prediction performance (*Accuracy* of 0.18, *Monotonicity* of 0.14). S_{overall} 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-Q_{ISO}, but also CASP-Q_{ISO}, and CASP-Q_{noExp}, gave poor predictions with *Accuracy* values ≤ 0.19 and *Monotonicity* values ≤ 0.23 . GPSM^q 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^q, 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^q and D achieved the highest mean *Accuracy*, *Monotonicity*, and *Consistency* and the lowest *RMSE**, indicating the best overall prediction performance. The performance of MoBi-Q and HASQIv2 was somewhat lower than for GPSM^q and D. As shown in Table 2, GPSM^q and D show the same standard deviation for *Accuracy*, *Monotonicity*, and *Consistency*, while the *RMSE** values of GPSM^q 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^q, D, MoBi-Q, and HASQIv2. For each of the models, GPSM^q, D, and MoBi-Q, a significant *RMSE** difference to R_{nonlin} , PEAQ, PEMO-Q, PEMO-Q_{ISO}, CASP-Q_{ISO}, and CASP-Q_{noExp} was observed, while there was no significant *RMSE** difference to HASQI and S_{overall} . 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^q, 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^q 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^q 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^q 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.

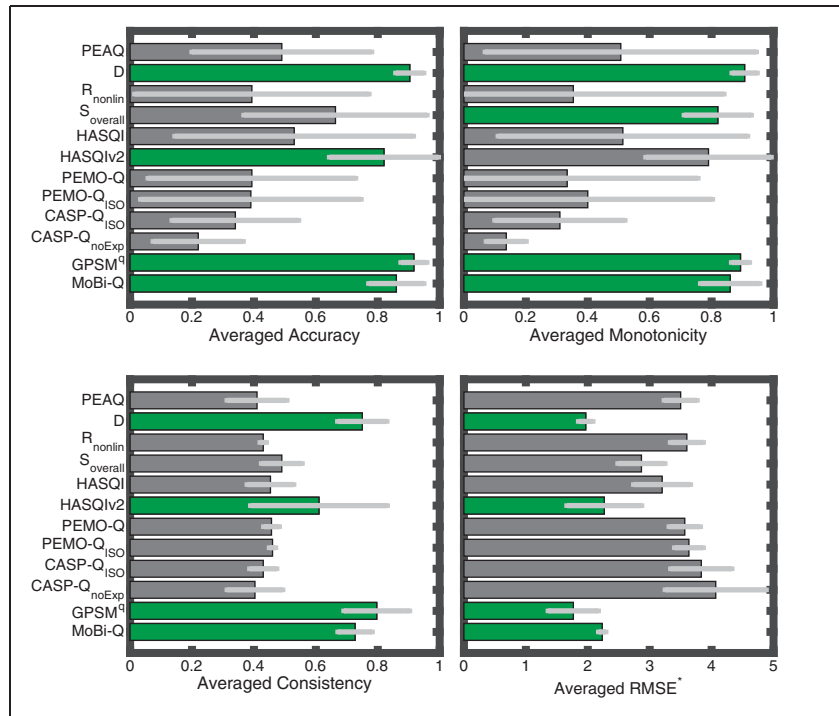


Figure 4. The Bars Show Mean Accuracy (Upper Left Panel), Mean Monotonicity (Upper Right Panel), Mean Consistency (Lower Left Panel), and Mean RMSE* (Lower Right Panel) for the Instrumental Measures Across the AFC, ATHD, and HTM Databases. The error bars indicate \pm one standard deviation for Accuracy, Monotonicity, Consistency, and RMSE* across the three databases. In each panel, the four best performing instrumental measures are highlighted in green.

PEAQ = Perceptual Evaluation of Audio Quality; HASQI = Hearing-Aid Speech Quality Index; HASQIv2 = Hearing-Aid Speech Quality Index version 2; PEMO-Q = Perception Model based Quality assessment; CASP-Q = Computational Auditory Signal processing and Perception model based Quality assessment; GPSM^q = Generalized Power Spectrum Model for quality; RMSE* = epsilon-insensitive root mean square error.

The spectral cue analysis of GPSM^q 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 GPSM^q 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, GPSM^q (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

GPSM^q_{DC} 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 GPSM^q_{DC} 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 GPSM^q predictions comparable to those for the linear measure D, GPSM^q predictions are here based on local power-based SNRs, and thus named GPSM^q_{DC} in the following. This GPSM^q_{DC} accounts only for spectral (linear) distortions, while modulation-based features are not considered. Because the GPSM^q_{DC} output was not transformed by a logarithmic function, as is done for the overall GPSM^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

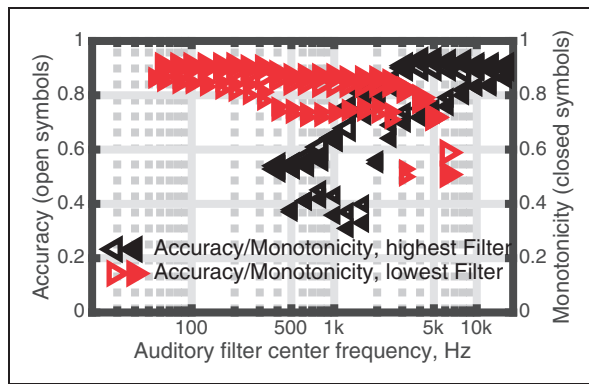


Figure 5. Effect of the Range of Auditory Filter Center Frequencies on Accuracy (Open Symbols) and Monotonicity (Closed Symbols) for $\text{GPSM}_{\text{DC}}^{\text{q}}$ (Large Symbols), Purely Based on Local Power-Based SNRs, and D (Small Symbols). Right-pointing triangles show the effect of varying the center frequency of the lowest auditory channel, while the center frequency of the highest auditory channel is fixed at 16 kHz for $\text{GPSM}_{\text{DC}}^{\text{q}}$ 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 $\text{GPSM}_{\text{DC}}^{\text{q}}$ and 55 Hz for D.

correlation coefficient) values are considered, as they are not affected by such a curvilinear relationship. $\text{GPSM}_{\text{DC}}^{\text{q}}$ 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. $\text{GPSM}_{\text{DC}}^{\text{q}}$ 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 $\text{GPSM}_{\text{DC}}^{\text{q}}$ 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, $\text{GPSM}_{\text{DC}}^{\text{q}}$ 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. A similar result was observed for the original GPSM^{q} , where predictions suggest greatest accuracy for 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 $\text{GPSM}_{\text{DC}}^{\text{q}}$ (*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 $\text{GPSM}_{\text{DC}}^{\text{q}}$: 0.91) as it is used by D, would not degrade the prediction performance of $\text{GPSM}_{\text{DC}}^{\text{q}}$, 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 $\text{GPSM}_{\text{DC}}^{\text{q}}$. This implies that there might be more redundant information in GPSM^{q} than in D, which allows reduction of the frequency range of GPSM^{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^{q} and D, but rather differences in the postprocessing of auditory excitation patterns. GPSM^{q} and MoBi-Q assess (first-order) excitation differences between the reference and the target signals. HASQiv2 and D evaluate a weighted sum of the standard deviation of the first-order and second-order excitation differences. A comparison of the mean *Accuracy* ($\text{GPSM}_{\text{DC}}^{\text{q}}$: 0.88; D: 0.90) and mean *Monotonicity* ($\text{GPSM}_{\text{DC}}^{\text{q}}$: 0.91; D: 0.91) for $\text{GPSM}_{\text{DC}}^{\text{q}}$ 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 R_{nonlin} 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^q 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^q 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 measures, 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 R_{nonlin} . 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 R_{nonlin} to the AFC database improved *Accuracy* from 0.83 to 0.96, while *Monotonicity* values were not affected. Accordingly, prediction performance of S_{overall} , representing the combination of the linear measure D and the nonlinear measure R_{nonlin} , was also improved (*Accuracy*: 0.98; *Monotonicity*: 0.95; *Consistency*: 0.82; *RMSE**: 1.4) by applying the optimized R_{nonlin} . However, optimizing R_{nonlin} to the ATHD and HTM databases did not substantially improve the prediction performance of R_{nonlin} and S_{overall} . Distortions occurring in those databases might not be sufficiently captured by R_{nonlin} .

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.

GPSM^q 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 the MUSHRA evaluation, listeners compared the audio quality of a reverberant reference signal with that of a

Table 1. Accuracy (Acc), Monotonicity (Mon), Consistency (Con), and RMSE* Results for Different Instrumental Measures (Rows) for the Adaptive Feedback Cancellation (AFC), Acoustically Transparent Hearing Device (ATHD), and Hear-Through Mode (HTM) Databases (Columns).

Measure \ DB	AFC				ATHD				HTM			
	Acc	Mon	Con	RMSE*	Acc	Mon	Con	RMSE*	Acc	Mon	Con	RMSE*
PEAQ	0.78	0.97	0.30	3.8	0.49	0.45	0.48	3.4	0.20	0.10	0.45	3.3
D	0.95	0.95	0.66	2.1	0.90	0.90	0.78	1.9	0.86	0.87	0.81	1.9
R _{nonlin}	0.83	0.91	0.42	3.4	0.17	0.01	0.43	3.9	0.18	0.14	0.44	3.5
S _{overall}	0.92	0.94	0.48	2.6	0.73	0.79	0.56	2.7	0.34	0.73	0.44	3.3
HASQI	0.92	0.94	0.45	2.7	0.52	0.46	0.53	3.3	0.15	0.14	0.38	3.6
HASQIv2	0.95	0.94	0.40	2.2	0.89	0.87	0.84	1.7	0.62	0.56	0.59	2.9
PEMO-Q	0.78	0.82	0.48	3.6	0.21	0.12	0.46	3.8	0.19	0.06	0.43	3.3
PEMO-Q _{ISO}	0.80	0.86	0.45	3.5	0.21	0.11	0.46	3.9	0.16	0.23	0.47	3.5
CASP-Q _{ISO}	0.50	0.55	0.38	4.4	0.41	0.22	0.46	3.6	0.11	0.16	0.45	3.5
CASP-Q _{noExp}	0.21	0.11	0.30	5.0	0.37	0.21	0.46	3.7	0.08	0.09	0.45	3.5
GPSM ^q	0.95	0.91	0.70	2.0	0.87	0.86	0.78	2.0	0.93	0.91	0.91	1.3
BAM-Q	–	–	–	–	–	–	–	–	0.33	0.27	0.48	3.4
MoBi-Q	0.96	0.97	0.67	2.2	0.83	0.8	0.73	2.3	0.79	0.81	0.78	2.2

Note. Bold font indicates the best performing measure for Accuracy, Monotonicity, Consistency, and RMSE* for each database. ATHD = acoustically transparent hearing device; CASP-Q = Computational Auditory Signal processing and Perception model based Quality assessment; GPSM^q = Generalized Power Spectrum Model for quality; BAM-Q = Binaural Auditory Model for audio Quality; AFC = adaptive feedback cancellation; HTM = hear-through mode; PEAQ = Perceptual Evaluation of Audio Quality; HASQI = Hearing-Aid Speech Quality Index; HASQIv2 = Hearing-Aid Speech Quality Index version 2; PEMO-Q = Perception Model based Quality assessment.

Table 2. Mean Accuracy (\overline{Acc}), Monotonicity (\overline{Mon}), Consistency (\overline{Con}), and $\overline{RMSE^*}$ and the Corresponding Standard Deviation for the Four Best Performing Instrumental Measures Calculated Across the AFC, ATHD, and HTM Databases.

	\overline{Acc}	\overline{Mon}	\overline{Con}	$\overline{RMSE^*}$
GPSM ^q	0.92 ± 0.04	0.89 ± 0.03	0.80 ± 0.10	1.8 ± 0.40
D	0.90 ± 0.05	0.91 ± 0.04	0.75 ± 0.08	2.0 ± 0.12
MoBi-Q	0.86 ± 0.09	0.86 ± 0.10	0.73 ± 0.09	2.2 ± 0.09
HASQIv2	0.82 ± 0.18	0.79 ± 0.20	0.61 ± 0.22	2.3 ± 0.60

Note. Bold font indicates the best performing instrumental measure for Accuracy, Monotonicity, Consistency, and RMSE*. GPSM^q = Generalized Power Spectrum Model for quality; HASQIv2 = Hearing-Aid Speech Quality Index version 2; RMSE* = epsilon-insensitive root mean square error.

reverberant signal processed by algorithms or hearables. To assess the influence of T_{60} on the four best performing instrumental measures, the female speech and jazz music samples recorded in rooms with T_{60} of about 0.35 s, 0.45, and 1.4 s from the ATHD databases were used.

GPSM^q, D, MoBi-Q, and HASQIv2 were not explicitly trained with reverberant signals, and it was stated in the original publications that effects of reverberation were not considered during model development. Nevertheless, the prediction performance of GPSM^q, D, and MoBi-Q for distortions in the ATHD database was not degraded by increasing reverberation, as shown by Accuracy and Monotonicity in Table 3, while the performance of HASQIv2 dropped for the longest reverberation time of about 1.4 s. Considering the importance of

spectral cues for GPSM^q, 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 $T_{60} \approx 1.4$ s, prediction performance of this nonlinear part of HASQIv2 (HASQIv2_{nonlin}) 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 HASQIv2_{TFS} and HASQIv2_{CepCorr} for $T_{60} \approx 1.4$ 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 for audio quality prediction of sounds in reverberant conditions than cues based on TFS or 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

Table 3. Accuracy (Acc) and Monotonicity (Mon) Results for GPSM^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 ^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 _{lin}	0.89	0.86	0.87	0.89	0.82	0.86
HASQIv2 _{nonlin}	0.89	0.87	0.86	0.79	0.67	0.64
HASQIv2 _{TFS}	0.75	0.70	0.78	0.80	0.55	0.54
HASQIv2 _{CepCorr}	0.80	0.82	0.75	0.74	0.56	0.54

Note. Predictions based on the linear part of HASQIv2, denoted HASQIv2_{lin}, and predictions based on the nonlinear part denoted HASQIv2_{nonlin} are shown. Predictions of the nonlinear part of HASQIv2 are based on TFS and cepstrum correlation features. To disentangle their contribution to HASQIv2_{nonlin}, prediction results of HASQIv2_{TFS} and HASQIv2_{CepCorr} are also provided. Bold values indicate the best performing objective measure for Accuracy and Monotonicity. GPSM^q = Generalized Power Spectrum Model for quality; HASQIv2 = Hearing-Aid Speech Quality Index version 2.

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^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 the binaural quality model BAM-Q. When only monaural aspects were considered, Devices F, D, and K (occluded ear, served as anchor signal) were given low scores by the monaural audio quality models GPSM^q and D. These models did not correctly rank the scores for Devices F, D, and K. This indicates either that monaural distortions were insufficiently represented by the monaural models or that binaural distortions significantly contributed to perceived audio quality. If the latter is true, audio quality predictions of the combined monaural and binaural model MoBi-Q should reflect the subjective ranking of Devices F, D, and K, but it did not. This can be explained by the method of combining monaural and binaural quality predictions, where only the domain (either monaural or binaural) with dominant quality differences was taken into account. As is shown in the following section, an additive combination

of monaural and binaural objective ratings gave more accurate quality predictions.

Device C achieved the highest subjective quality rating. The technical evaluation of Denk et al. (2020) revealed no significant delay differences and only slightly distorted ITDs and ILDs for Device C compared with the open-ear reference. The measurements of Denk et al. further showed very good binaural cue conservation for Devices A and B. These measurements agree well with quality predictions of the binaural BAM-Q, where 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^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_{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-filter effect for frequencies < 1 kHz (see Figure 6 in

Denk et al., 2020), it achieved fairly high subjective quality ratings, as predicted by GPSM^q 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 GPSM^q (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.

$$\text{Overall quality} = \frac{1}{1 + e^{-0.1188 \cdot (\text{OPM}_{\text{dual}} - 38.9817)}} + \frac{1}{1 + e^{-0.0192 \cdot (\text{binQ} + 20.8263)}} \quad (2)$$

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.

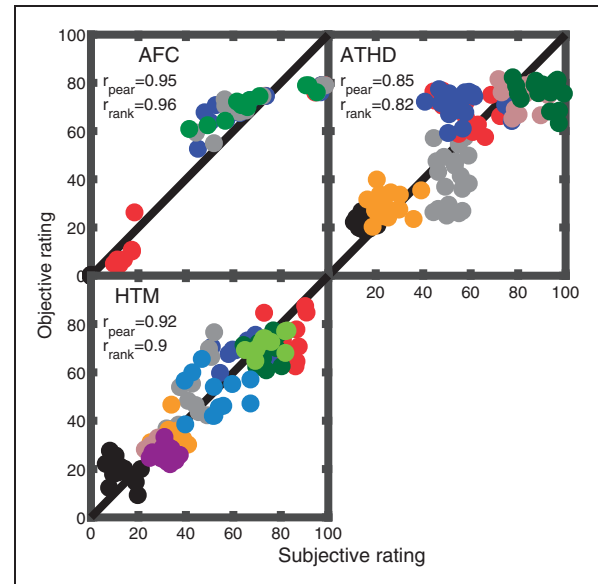


Figure 6. Predictions From MoBi-Q_{add} for the AFC, ATHD, and HTM Databases. The x axis represents subjective ratings, while the abscissa represents model predictions.

AFC = adaptive feedback cancelation; ATHD = acoustically transparent hearing device; HTM = hear-through mode.

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 (*Accuracy*: 0.92; *Monotonicity*: 0.9; *Consistency*: 0.93, *RMSE**: 1.3 dB) than MoBi-Q (*Accuracy*: 0.79; *Monotonicity*: 0.81; *Consistency*: 0.78; *RMSE**: 2.2 dB).

As shown in Figure 6, MoBi-Q_{add} also gave very good prediction performance for the AFC (*Accuracy*: 0.95; *Monotonicity*: 0.96; *Consistency*: 0.62; *RMSE**: 2.1 dB) and ATHD (*Accuracy*: 0.85; *Monotonicity*: 0.82; *Consistency*: 0.75; *RMSE**: 2.2 dB) databases, resulting in better overall performance (*Acc*: 0.91; *Mon*: 0.89; *Con*: 0.77; *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 GSPM^q. 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

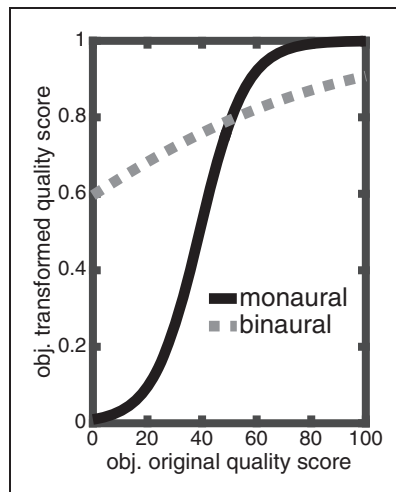


Figure 7. Sigmoid Functions (see Equation 2) Applied to the Monaural GPSM^q With Binaural Modification and the Binaural BAM-Q in MoBi-Q_{add}. The x axis represents the objective quality score from the original model, while the abscissa represents the transformed quality score.

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^q (Biberger et al., 2018). The monaural GPSM^q 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^q 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^q provided very accurate predictions for the three databases used here, the question arises whether the modified GPSM^q in MoBi-Q can be replaced by the original GPSM^q 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^q (without binaural modification) and BAM-Q, was tested (MoBi-Q_{add,origGPSM^q}). 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 broadly applicable model version. It should be

Table 4. *Accuracy* (Acc), *Monotonicity* (Mon), *Consistency* (Con), and *RMSE** Results for the Suggested Additive Combination of the Outputs of GPSM^q 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^q (Biberger et al., 2018) Without Binaural Modification and BAM-Q (Denoted MoBi-Q_{add,origGPSM^q}).

DB	Measure	MoBi-Q _{add}				MoBi-Q _{add,origGPSM^q}				MoBi-Q			
		Acc	Mon	Con	RMSE*	Acc	Mon	Con	RMSE*	Acc	Mon	Con	RMSE*
AFC		0.95	0.96	0.62	2.1	0.96	0.94	0.70	1.8	0.96	0.97	0.67	2.2
ATHD		0.85	0.82	0.75	2.2	0.86	0.87	0.74	2.0	0.83	0.80	0.73	2.3
HTM		0.92	0.90	0.93	1.3	0.93	0.92	0.94	1.2	0.79	0.81	0.78	2.2
Fleßner et al. (2019)		0.83	0.79	0.92	1.4	0.74	0.75	0.87	1.8	0.86	0.80	0.95	1.3

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 cancellation; HTM = hear-through mode.

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 ≥ 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 R_{nonlin}, 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  <https://orcid.org/0000-0002-6314-1914>
 Florian Denk  <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.

References

- Ando, Y., & Kurihara, Y. (1986). Nonlinear response in evaluating the subjective diffuseness of sound fields. *Journal of*

- the Acoustical Society of America*, 80(3), 833–836. <https://doi.org/10.1121/1.393906>
- Beerends, J. G., Hekstra, A. P., Rix, A. W., & Hollier, M. P. (2002). Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II - Psychoacoustic model. *Journal of the Audio Engineering Society*, 50(10), 765–778.
- Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., & Keyhl, M. (2013a). Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I - temporal alignment. *Journal of the Audio Engineering Society*, 61(6), 366–384.
- Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., & Keyhl, M. (2013b). Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part II - perceptual model. *Journal of the Audio Engineering Society*, 61(6), 385–402.
- Beerends, J. G., & Stemerink, J. A. (1992). A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40(12), 963–978.
- Beerends, J. G., & Stemerink, J. A. (1994). A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 42(3), 115–123.
- Biberger, T., & Ewert, S. D. (2016). Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility. *Journal of the Acoustical Society of America*, 140(2), 1023–1038. <http://dx.doi.org/10.1121/1.4960574>
- Biberger, T., & Ewert, S. D. (2017). The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking. *Journal of the Acoustical Society of America*, 142(2), 1098–1111. <http://dx.doi.org/10.1121/1.4999059>
- Biberger, T., Fleßner, J.-H., Huber, R., & Ewert, S. D. (2018). An objective audio quality measure based on power and envelope power cues. *Journal of the Audio Engineering Society*, 66(7/8), 578–593. <https://doi.org/10.17743/jaes.2018.0031>
- Blauert, J., & Lindemann, W. (1986). Spatial mapping of intracranial auditory events for various degrees of interaural coherence. *Journal of the Acoustical Society of America*, 79(3), 806–813. <https://doi.org/10.1121/1.393471>
- Creusere, C. D., Kallakuri, K. D., & Vanam, R. (2007). An objective metric of human subjective audio quality optimized for a wide range of audio fidelities. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1), 129–136. <https://doi.org/10.1109/TASL.2007.907571>
- Damaske, P., & Ando, Y. (1972). Interaural crosscorrelation for multichannel loudspeaker reproduction. *Acustica*, 27(4), 232–238.
- Dau, T., Kollmeier, B., & Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *Journal of the Acoustical Society of America*, 102(5), 2892–2905. <http://dx.doi.org/10.1121/1.420344>
- Dau, T., Kollmeier, B., & Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *Journal of the Acoustical Society of America*, 102(5), 2906–2919. <http://dx.doi.org/10.1121/1.420345>
- Denk, F., Hiipakka, M., Kollmeier, B., & Ernst, S. M. A. (2018). An individualised acoustically transparent earpiece for hearing devices. *International Journal of Audiology*, 57, 62–70. <http://dx.doi.org/10.1080/14992027.2017.1294768>
- Denk, F., Schepker, H., Doclo, S., & Kollmeier, B. (2020). Acoustic transparency in concurrent hearables – Technical evaluation. *Journal of the Audio Engineering Society*, 68(7/8), 508–521. <https://doi.org/10.17743/jaes.2020.0042>
- Dietz, M., Ewert, S. D., & Hohmann, V. (2011). Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5), 592–605. <https://doi.org/10.1016/j.specom.2010.05.006>
- Emiya, V., Vincent, E., Harlander, N., & Hohmann, V. (2011). Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7), 2046–2057. <https://doi.org/10.1109/TASL.2011.2109381>
- Ewert, S. D., & Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *Journal of the Acoustical Society of America*, 108(3), 1181–1196. <http://dx.doi.org/10.1121/1.1288665>
- Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, 12(47), 47–65. <https://doi.org/10.1103/RevModPhys.12.47>
- Fleßner, J.-H., Biberger, T., & Ewert, S. D. (2019). Subjective and objective assessment of monaural and binaural aspects of audio quality. *IEEE Transactions on Audio, Speech and Language Processing*, 27(7), 1112–1125. <https://doi.org/10.1109/TASLP.2019.2904850>
- Fleßner, J.-H., Huber, R., & Ewert, S. D. (2017). Assessment and prediction of binaural aspects of audio quality. *Journal of the Audio Engineering Society*, 65(11), 929–942. <https://doi.org/10.17743/jaes.2017.0037>
- Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1–2), 103–138. [https://doi.org/10.1016/0378-5955\(90\)90170-T](https://doi.org/10.1016/0378-5955(90)90170-T)
- Glasberg, B. R., & Moore, B. C. J. (2005). Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds. *Journal of the Audio Engineering Society*, 53(10), 906–918.
- Harlander, N., Huber, R., & Ewert, S. D. (2014). Sound quality assessment using auditory models. *Journal of the Audio Engineering Society*, 62(5), 324–336. <https://doi.org/10.17743/jaes.2014.0020>
- Hoffmann, P., Christensen, F., & Hammershøi D. (2013, August). *Insert earphone calibration for hear-through options*. Proceedings of the Audio Engineering Society Conference 51: Loudspeakers and Headphones, Helsinki, Finland.
- Hu, Y., & Loizou, P. C. (2007). A comparative intelligibility study of single-microphone noise reduction algorithms. *Journal of the Acoustical Society of America*, 122(3), 1777–1786. <https://doi.org/10.1121/1.2766778>

- Huber, R., & Kollmeier, B. (2006). PEMO-Q - A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6), 1902–1911. <https://doi.org/10.1109/TASL.2006.883259>
- ISO 226. (2003). *Acoustics – Normal equal-loudness-level contours*. Geneva, Switzerland: International Organization for Standardization.
- ITU-R BS.1116-1. (1997). *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. Geneva, Switzerland: International Telecommunications Union.
- ITU-R BS.1387. (2001). *Method for objective measurements of perceived audio quality*. Geneva, Switzerland: International Telecommunications Union.
- ITU-R BS.1534. (2014). *Method for the subjective assessment of intermediate quality levels of coding systems*. Geneva, Switzerland: International Telecommunications Union.
- ITU-R BS.1534-1. (2003). *Method for the subjective assessment of intermediate quality levels of coding systems*. Geneva, Switzerland: International Telecommunications Union.
- ITU-T. P800. (1996). *Methods for subjective determination of transmission quality*. Geneva, Switzerland: International Telecommunications Union.
- ITU-T P.862. (2001). *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. Geneva, Switzerland: International Telecommunications Union.
- ITU-T Rec P.1401. (2012). *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*. Geneva, Switzerland: International Telecommunications Union.
- Jepsen, M. L., Ewert, S. D., & Dau, T. (2008). A computational model of human auditory signal processing and perception. *Journal of the Acoustical Society of America*, 124(1), 422–438. <http://dx.doi.org/10.1121/1.2924135>
- Jørgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *Journal of the Acoustical Society of America*, 134(1), 436–446. <http://dx.doi.org/10.1121/1.4807563>
- Jurado, C., & Moore, B. C. J. (2010). Frequency selectivity for frequencies below 100 Hz: Comparisons with mid-frequencies. *Journal of the Acoustical Society of America*, 128(6), 3585–3596. <https://doi.org/10.1121/1.3504657>
- Kates, J. M., & Arehart, K. H. (2010). The hearing-aid speech quality index (HASQI). *Journal of the Audio Engineering Society*, 58(5), 363–381.
- Kates, J. M., & Arehart, K. H. (2014a). The hearing-aid speech quality index (HASQI) version 2. *Journal of the Audio Engineering Society*, 62(3), 99–116. <https://doi.org/10.17743/jaes.2014.0006>
- Kates, J. M., & Arehart, K. H. (2014b). The hearing-aid speech perception index (HASPI). *Speech Communication*, 65, 75–93. <https://doi.org/10.1016/j.specom.2014.06.002>
- Kates, J. M., & Arehart, K. H. (2016). The hearing-aid audio quality index (HAAQI). *IEEE Transactions on Audio, Speech and Language Processing*, 24(2), 354–365. <https://doi.org/10.1109/TASLP.2015.2507858>
- Kendall, G. S. (1995). The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19(4), 71–87. <https://doi.org/10.2307/3680992>
- Madsen, S. M. K., & Moore, B. C. J. (2014). Music and hearing aids. *Trends in Hearing*, 18, 1–29. <https://doi.org/10.1177/2331216514558271>
- Marentakis, G., & Liepins, R. (2014). Evaluation of hear-through sound localization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems ACM, Toronto, Canada* (pp. 267–270). Association for Computing Machinery.
- Maxwell, J. A., & Zurek, P. M. (1995). Reducing acoustic feedback in hearing aids. *IEEE Transactions on Speech and Audio Processing*, 3(4), 304–313. <https://doi.org/10.1109/89.397095>
- Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3), 750–753. <http://dx.doi.org/10.1121/1.389861>
- Moore, B. C. J., & Tan, C.-T. (2003). Perceived naturalness of spectrally distorted speech and music. *Journal of the Acoustical Society of America*, 114(1), 408–419. <https://doi.org/10.1121/1.1577552>
- Moore, B. C. J., & Tan, C.-T. (2004). Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion. *Journal of the Audio Engineering Society*, 52(9), 900–914.
- Moore, B. C. J., Tan, C.-T., Zacharov, N., & Mattila, V.-V. (2004). Measuring and predicting the perceived quality of music and speech subjected to combined linear and non-linear distortion. *Journal of the Audio Engineering Society*, 52(12), 1228–1244.
- Muller, M., Robertson, D., & Yates, G. K. (1991). Rate-versus-level functions of primary auditory nerve fibres: Evidence for square law behaviour of all fibre categories in the guinea pig. *Hearing Research*, 55(1), 50–56. [https://doi.org/10.1016/0378-5955\(91\)90091-M](https://doi.org/10.1016/0378-5955(91)90091-M)
- Munson, W. A., & Gardner, M. B. (1950). Standardizing auditory tests. *Journal of the Acoustical Society of America*, 22, 675. <http://dx.doi.org/10.1121/1.1917190>
- Nordholm, S., Schepker, H., Tran, L. T. T., & Doclo, S. (2018). Stability-controlled hybrid adaptive feedback cancellation scheme for hearing aids. *Journal of the Acoustical Society of America*, 143(1), 150–166. <https://doi.org/10.1121/1.5020269>
- Patterson, R. D., & Moore, B. C. J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In B. C. J. Moore (Ed.), *Frequency selectivity in hearing* (pp. 123–177). New York: Academic.
- Rix, A.W., Hollier, M. P., Hekstra, A. P., & Beerends, J. G. (2002). Perceptual evaluation of speech quality (PESQ). The new ITU standard for end-to-end speech quality assessment part I - Time-delay compensation. *Journal of the Audio Engineering Society*, 50(10), 755–765.
- Rumsey, F. (2019). Headphone technology: Hear-through, bone conduction, noise canceling. *Journal of the Audio Engineering Society*, 67(11), 914–919.
- 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. (2003). The effect of nonlinear distortion on the perceived quality of music and speech signals. *Journal of the Audio Engineering Society*, 51(11), 1012–1031.
- 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*, 45(3), 203–220. [https://doi.org/10.1016/0378-5955\(90\)90121-5](https://doi.org/10.1016/0378-5955(90)90121-5)