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

A Computationally Efficient Model for Combined Assessment of Monaural and Binaural Audio Quality

BERNHARD EURICH,* STEPHAN D. EWERT, MATHIAS DIETZ, AND THOMAS BIBERGER
(bernhard.eurich@uol.de) (stephan.ewert@uol.de) (m.dietz@uol.de) (thomas.biberger@uol.de)

Department für Medizinische Physik und Akustik, Universität Oldenburg, Oldenburg, Germany

Audio quality is an important aspect of hearing aids, hearables, and sound reproduction systems because the signal processing of such devices might alter the spectral composition or interaural differences of the original sound and thus might degrade the perceived audio quality. Consequently, an audio quality model applicable to such devices requires accounting for monaural and binaural aspects of audio quality. Fleßner et al. successfully predicted overall audio quality by combining a monaural and binaural audio quality model, which is computationally expensive and thus limits the scope of application. In order to also cover time-critical applications, such as real-time control of algorithms in audio and hearing technology, the authors present a computationally efficient model for overall audio quality in listeners with normal hearing. The suggested model was evaluated with six databases including quality ratings for music and speech signals processed by loudspeakers and algorithms typically applied in modern hearing devices (e.g., acoustic transparency, feedback cancellation or binaural beamforming). The presented model achieved a high prediction performance, indicated by the mean Pearson correlation of 0.9 similar to the more complex model of Fleßner et al., while its calculation time is substantially lower.

0 INTRODUCTION

Audio quality is an important aspect of many signal processing applications ranging from hearing devices to sound reproduction systems. For the evaluation of the perceived audio quality of algorithms or devices, listening tests are considered as the "gold standard." These tests can be carried out as reference-free tests (e.g., [1]), where listeners rate the audio quality of a processed speech or audio signal without any given unprocessed reference signal or as reference-based tests (e.g., [2, 3]), comparing processed and unprocessed (reference) signals. Such listening tests are typically time consuming and expensive and often require expert listeners to gain reliable quality judgements. To overcome these disadvantages, several instrumental audio quality measures have been developed (e.g., [4–7]).

In addition to evaluating signal processing algorithms, instrumental quality measures can also be applied to control algorithms, provided they are computationally efficient. Similarly as for the above-mentioned listening tests, instrumental measures may either predict audio quality without (reference-free or nonintrusive) or with a given reference signal (reference-based or intrusive) as required by the instrumental measures considered in this study. Reference-based instrumental measures typically predict signal fidelity between test and reference signals. They do not capture the listener's preference, which might depend on the type of stimulus and the room acoustics. For example, Kates and colleagues [8] found that a high interaural cross correlation (IACC) corresponded to higher speech clarity ratings, while a low IACC corresponded to greater apparent source width ratings. Thus, depending on the stimuli and context, a listener might prefer acoustically "dry" rooms providing a high IACC for speech (because it facilitates understanding speech), reverberant rooms providing a low IACC for classical music (because it improves the spatial impression [9]), or rooms with intermediate IACC for a jazz club setting. The instrumental quality measures in this study do not provide predictions about the listener's preference ratings.

One relevant field of application for instrumental audio quality measures are wireless and smart headphones, in the following denoted as hearables. These devices have become

*To whom correspondence should be addressed, email: bernhard.eurich@uol.de. Last updated: April 24, 2023 increasingly popular because, in addition to their traditional use for listening to music and streaming audio, they offer signal processing features typically used in hearing aids to restore ambient sound for (hard of hearing) listeners [10]. The signal processing typically involved, such as noise suppression, beamforming, hear-through processing, nonlinear amplification, or attenuation, potentially alters the spectral composition or interaural differences of the original sound. This might be perceived by the listeners as spectral or spatial distortions, degrading the audio quality of signals. Hereby, hear-through processing aims at a natural (ideally acoustically transparent) representation of the external acoustical environment without perceivable distortions, similar to the sound impression with an open ear (without inserted device). This enables perceptually authentic conversations as well as awareness of the acoustic scene, both important in real life but also for augmented, mixed, and virtual-reality applications [11].

Because the human auditory system is limited in its ability to resolve monaural (spectral and temporal) and binaural differences, i.e., interaural level differences (ILDs) and interaural time differences (ITDs), an authentic hear-through processing does not require the exact reproduction of the open-ear signal at the eardrum. A previous study [12] has shown that much of the distortion associated with the hear-through mode in hearables and smart headphones can be attributed to monaural, spectral coloration cues. However, degraded binaural cues play a significant role in standard hearing device algorithms, such as binaural noise reduction and beamforming [13–16]. Given that binaural cues offer substantial advantages for speech intelligibility in realistic, complex acoustic conditions [17–20], for sound localization [21, 22] as well as for listening effort [23], changes in (monaural) spectral coloration alone may not be a sufficient predictor for overall audio quality in such cases.

In the context of hearing aid processing, several recently suggested algorithms for noise reduction or dereverberation were designed not only to improve speech intelligibility but also to preserve binaural cues. Algorithms presented in [15, 14] aimed at finding an optimal trade-off between noise reduction performance and the preservation of the interaural coherence for diffuse noise fields in order to maintain the spatial impression of the acoustical scene. The binaural dereverberation algorithm presented in Jeub et al. [24] was designed to suppress reverberation while maintaining binaural cues. Their listening test showed that for the objective assessment of such binaural-cue-preserving algorithms, instrumental quality measures require accounting for spatial quality aspects, indicating whether the algorithm alters the spatial perception of the original sound. Although quality of spatial and surround sound reproduction was not specifically addressed in this study, outcomes for such systems [25] support the findings from the previously mentioned studies, highlighting the importance of (monaural) spectral and binaural cues for audio quality ratings. Rumsey and colleagues [25] found that spatial fidelity accounted for approximately 30% of the basic audio quality rating of degraded multichannel audio signals. Therefore, they suggested to include both timbral and spatial quality aspects in future perceptual models of audio quality.

Given that this study aims to provide an instrumental measure for assessing the overall audio quality of hearing devices and sound reproduction systems, from the abovementioned studies, it seems reasonable to incorporate spectral coloration and loudness cues as well as binaural cues because they often seem to be the most relevant cues. In the past, several monaural instrumental measures for the assessment of speech and audio quality have been developed [5, 26, 6, 27–31, 4, 32], often designed for different specific applications, such as quality predictions for audio and speech codecs, hearing-aid signal processing, or loudspeaker and headphone distortions. In comparison to those monaural measures, only a few instrumental measures that capture binaural audio quality aspects have been developed [33–37], while such aspects are expected to be important in hearing devices [15, 14, 38–40, 16, 13] and (multichannel) loudspeaker-based sound field reproduction [41, 42, 25].

One publicly available binaural instrumental audio quality measure is the binaural auditory model for audio quality (BAM-Q [33]), an intrusive measure that is based on a perceptually motivated direction of arrival estimation model [43]. It estimates spatial audio quality based on differences between the test and reference signal in ILD, ITD, and an interaural coherence measure called interaural vector strength (IVS).

In order to predict overall audio quality for signals impaired by monaural, binaural, or combined monaural and binaural distortions, Fleßner et al. [7] suggested the instrumental audio quality measure MoBi-Q. Their model combines the outputs of the binaural BAM-Q and the monaural Generalized Power Spectrum Model for quality (GPSM^q [6]) so that overall audio quality is determined by the lower-quality aspect, i.e., either monaural or binaural. The GPSM^q represents an audio quality extension of the psychoacoustic and speech intelligibility model GPSM [44–46]. The combined model MoBi-Q has been shown to account for several artificially introduced monaural and binaural distortions [7] and distortions occurring in hearables [12].

So far, computational efficiency of binaural quality models was not specifically considered. In contrast to combining the outputs of established models, each comprehensive monaural and binaural models, efficient monaural and binaural model for quality (eMoBi-Q) combines previously established relevant cues such as spectral coloration, loudness, and binaural cues within a simpler model structure than in MoBi-Q.

In this regard, three recent findings were utilized:

(1) It has been shown that peripheral filtering in the inner ear limits the bandwidth of both monaural and binaural processing bands in the same way [47–49]. Therefore, the monaural and binaural cues were extracted from the same bandpass signals using a single peripheral filterbank.

(2) Mammalian encoding of ITDs is best described as a two-hemisphere code [22, 50]. Therefore, the complex correlation coefficient γ was used as a sufficient but compact formulation of the two-hemisphere code, reflecting coherence as the magnitude, i.e., the envelope of the correlation

function, and the interaural phase difference (IPD) as the argument. Low coherence, on the one hand, reflects a diffuse intracranial image. This has been associated with the IACC—which is the real part of γ—in, e.g., architectural acoustics [51]. The IPD, on the other hand, reflects laterality. The definition of γ, known from optics [52], thus provides a general description of the relationship between two waves. Applied to the context of binaural perception, it combines physiological plausibility with high predictive power in behavioral data and mathematical efficiency [53, 49].

(3) Because the spectral coloration and loudness cues in combination with the binaural cues have been shown to predict overall audio quality for hearing devices and sound reproduction systems, eMoBi-Q extracts the monaural and binaural features directly from the output of the peripheral filters, without fine-structure and modulation filtering.

Therefore, this study aims to provide a simpler and thus computationally more efficient alternative to MoBi-Q. It is specifically targeted at assessing the sound quality of hearing devices and sound reproduction systems and consists of the linear path of GPSM^q combined with γ and ILDs as new binaural features to mimic the perception of binaural cues, replacing the more-complex BAM-Q. At the same time, it will be explored whether γ is suited to assess binaural audio quality. The proposed model is termed "efficient model for combined assessment of monaural and binaural audio quality" (eMoBi-Q), where the "e" symbolizes the computational efficiency obtained by (1) replacing the binaural model BAM-Q by the γ and ILD features, (2) only using the linear part of GPSM^q, (3) using one preprocessing stage with consistent time frames for all model features, and (4) using a very simple backend that associates larger differences between test and reference signals with lower audio quality.

Given that the proposed eMoBi-Q was developed for audio quality predictions of distortions as they may occur in loudspeakers or modern hearing devices, six databases including music, speech, and noise signals processed by loudspeakers, algorithms for hear-through processing, feedback cancellation, binaural beamformer, and artificial binaural distortions were used for the evaluation as they cover a large range of relevant distortions. The performance of eMoBi-Q was compared with the more complex binaural quality measure BAM-Q and the combined quality measure MoBi-Q.

1 MODEL DESCRIPTION

The architecture of the suggested quality measure eMoBi-Q allows for simultaneously analyzing the binaural and monaural features in real time on a unified time scale, providing a frame-by-frame estimate of the binaural and monaural contributions to overall quality.

The relative contribution and perceptual range of the binaural features, γ and ILD, were first calibrated by hand such that eMoBi-Q best replicated ratings in a database of subjective quality ratings of the hear-through mode of hearables [54], following Biberger et al. [12], since the model is aimed at applications in modern hearing and headphone technology. The model was then evaluated with six databases covering a broad range of monaural, binaural, and combined distortions. These databases include audio quality ratings on the acoustical transparency of binaural noise reduction algorithms, binaural magnification and adaptive feedback cancellation (AFC) in hearing devices, loudspeaker distortions, and the acoustical transparency of hearing device prototypes. The considered databases are based on the assessments of normal-hearing (NH) listeners and accordingly the model reflects NH.

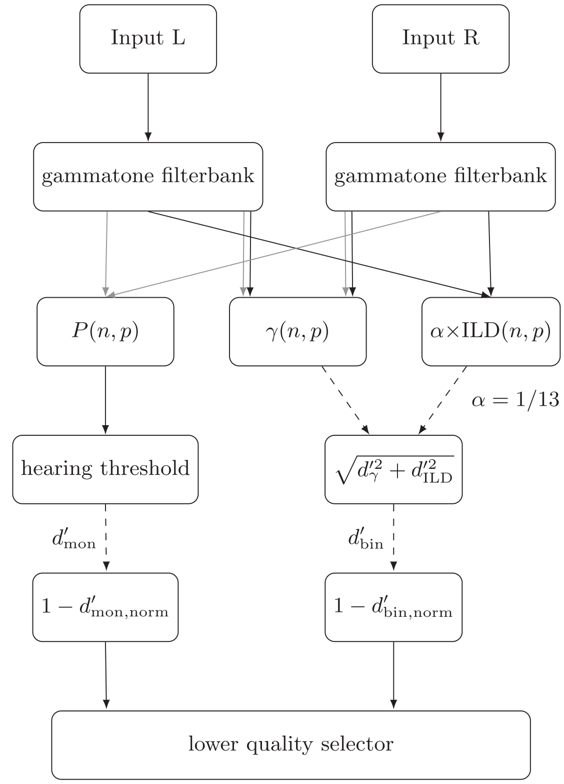


Fig. 1. Block diagram of the proposed model. In both the monaural [local DC power $P(n, p)$] and binaural [$\gamma(p, n)$, ILD(p, n)] paths, the frequency channels p are combined in an optimal manner. The n consecutive 400-ms time frames are combined in an optimal manner for the binaural features and averaged for the spectral coloration feature. Gray lines denote envelope low-pass filtering of the audio signals; dashed lines denote that the discriminability d' was obtained from comparing a test signal with a reference signal.

Fig. 1 shows the block diagram of the suggested model. It requires processed (distorted) test and unprocessed reference signals with either one-channel (monaural) or two-channel (binaural) audio signals as input. In the following, the model frontend with joint preprocessing stages and the calculation of monaural and binaural features are explained, followed by the description of the backend where monaural and binaural features are combined to the final audio quality measure.

1.1 Frontend

1.1.1 Preprocessing

Basilar membrane filtering of the left and right input signals was modeled by a linear fourth-order gammatone filterbank [55, 56], as implemented by Hohmann [57]. This results in 29 bandpass-filtered signals with center frequencies between 315 Hz and 12.5 kHz that have equivalent rectangular bandwidths (ERBs) according to Glasberg and Moore [58]. The covered frequency range has been found to be sufficient to capture the distortions that occur in hearables [12], which are mostly redundant in a wider frequency range, so that a wider frequency filter range does not provide any additional benefit. Since this is also expected for binaural distortions of hearing devices, and in order to provide a unified monaural and binaural frontend, the same bandpass signals are used in both the monaural and binaural model features.

The processing stages in the monaural and binaural path process the bandpass signals in consecutive time frames of 400 ms. The time-frequency signal elements of the left and right ear signals are denoted as $l(n, p)$ and $r(n, p)$ for a time frame n and frequency band p. A first-order low-pass filter with a 150-Hz cutoff frequency was applied to the envelope to model the limited sensitivity to envelope fluctuations [59]. The low-pass–filtered envelope affected the monaural spectral coloration feature, ILD feature, and γ feature for frequency bands above 1,300 Hz. However, no low-pass filtering was applied to the temporal fine-structure processing realized by the γ feature in frequency bands centered below 1,300 Hz.

1.1.2 Monaural Spectral Coloration and Loudness Feature

The monaural feature was calculated by adopting the power spectrum path of the GPSM^q [6]. In case of two-channel (binaural) input signals, left and right channels were concatenated. The Hilbert envelope, calculated for each of the complex-valued gammatone filterbank outputs, was filtered by a first-order low-pass filter with a 150-Hz cutoff frequency to account for the decrease of modulation sensitivity with increasing modulation frequency. The local DC power was extracted from the low-pass–filtered envelope signals. It is half the squared mean of the envelope E across the time frame n of a frequency band p:

$$P(n, p) = \frac{\overline{E(n, p)}^2}{2}. \tag{1}$$

Elements with a local DC power below the hearing threshold in quiet [60] were set to that threshold.

As in Biberger et al. [6], local power increments $\mathrm{SNR_{incr}}$ were computed as[1]

$$\mathrm{SNR}(n, p)_{\mathrm{incr}} = \frac{P_{\mathrm{test}}(n, p) - P_{\mathrm{ref}}(n, p)}{P_{\mathrm{ref}}(n, p)} \tag{2}$$

and local power decrements $\mathrm{SNR_{decr}}$ as

$$\mathrm{SNR}(n, p)_{\mathrm{decr}} = \frac{P_{\mathrm{ref}}(n, p) - P_{\mathrm{test}}(n, p)}{P_{\mathrm{test}}(n, p)}. \tag{3}$$

An upper limit of 13 dB was applied to each time-frequency element of $\mathrm{SNR}(n, p)_{\mathrm{incr}}$ and $\mathrm{SNR}(n, p)_{\mathrm{decr}}$, resulting in a dynamic range of 26 dB in total. Then $\mathrm{SNR}(n, p)_{\mathrm{incr}}$ and $\mathrm{SNR}(n, p)_{\mathrm{decr}}$ are averaged across time segments resulting in $\mathrm{SNR}(p)_{\mathrm{incr}}$ and $\mathrm{SNR}(p)_{\mathrm{decr}}$.

1.1.3 Binaural Features

Two binaural features were extracted for each gammatone-filtered signal:

1.1.3.1 Complex Correlation Coefficient γ The complex-valued correlation coefficient was used because it conveniently combines information about both the interaural coherence $|\gamma|$, reflecting the perceptual compactness of a sound, and the mean IPD as $\arg\{\gamma\}$, reflecting laterality. It is a mathematical formulation of the two-hemisphere channel code underlying neural encoding of interaural differences in mammals [22, 50], capturing temporal fluctuations in the interaural phase. This feature and its assumption on filter bandwidth has been psychoacoustically validated by Encke and Dietz [53], Eurich et al. [49], Eurich and Dietz [61], and Dietz et al. [48].

The gammatone filterbank implementation [57] provides complex-valued outputs signals $l(n, p)$ and $r(n, p)$, utilized for computing the complex correlation coefficient γ:

$$\gamma(n, p) = \frac{\overline{l(n, p)^* r(n, p)}}{\sqrt{\overline{|l(n, p)|^2 |r(n, p)|^2}}}, \tag{4}$$

where $\overline{\bullet}$ denotes the mean over the duration of the time frame. For frequency bands with center frequencies below 1,300 Hz, γ operates on the temporal fine structure of the bandpass signals, while above of 1,300 Hz, it operates on their Hilbert envelopes. This mimics the sensitivity to IPDs in the temporal fine structure at low frequencies as encoded by the human medial superior olive in combination with the sensitivity to IPDs in the envelope at higher frequencies as encoded by the lateral superior olive [62, 63]. As in [49, 61], Fisher's z transform was applied to the coherence (i.e., $|\gamma|$) to normalize the variance and to account for the increasing sensitivity to changes in coherence toward unity. To avoid infinite sensitivity, γ was multiplied by 0.9 [49, 61].

1.1.3.2 Interaural Level Differences ILDs were extracted as the logarithmic power ratio between left and right signals:

$$\mathrm{ILD}(n, p) = 10 \log(\frac{P_l(n, p)}{P_r(n, p)}). \tag{5}$$

The model's sensitivity to binaural distortions was obtained as the difference between the frontend outputs of reference and test signals, denoted as d':

$$d'_{\gamma}(n, p) = |\gamma_{\mathrm{ref}}(n, p) - \gamma_{\mathrm{test}}(n, p)|, \tag{6}$$

$$d'_{\mathrm{ILD}}(n, p) = |\mathrm{ILD}_{\mathrm{ref}}(n, p) - \mathrm{ILD}_{\mathrm{test}}(n, p)|. \tag{7}$$

An upper limit of 10 dB (calibrated to the hear-through-mode database [54]) was applied to $d'_{\mathrm{ILD}}(n, p)$ (cf. BAM-Q

[1] Since the spectral coloration feature was adopted from GPSM^q, the term SNR was also used as historically established in, for example, the underlying GPSM, which predicts psychoacoustic masking and speech intelligibility [44]. However, in the context of audio quality, "signal" and "noise" refer to "processed by device under test" and "unprocessed reference," respectively. Thus, a high SNR means a high local power increment or decrement, which, although unintuitive, means strong distortion.

[33]) to mimic the perceptual saturation of laterality and to avoid disproportionately large ILDs at moments of very low one-sided DC power.

1.2 Backend

For $d'_{\gamma}(n, p)$ and $d'_{\mathrm{ILD}}(n, p)$, information was optimally combined across time frames n and frequency bands p, i.e., assuming a linear, independent combination:

$$d' = \sqrt{\sum_{n} \sum_{p} d'(n, p)^2}. \tag{8}$$

The weighted optimal combination of the two binaural features' sensitivity indices gives the output of the binaural model path:

$$d'_{\mathrm{bin}} = \sqrt{d'^2_{\gamma} + \alpha \, d'^2_{\mathrm{ILD}}}, \tag{9}$$

where the relative weight $\alpha = 1/13$ of the ILD feature was calibrated using the database on the hear-through mode of hearables [54].

Adopted from Biberger et al. [6], the monaural increment and decrement SNRs, $\mathrm{SNR}(p)_{\mathrm{incr}}$ and $\mathrm{SNR}(p)_{\mathrm{decr}}$, were combined by taking the mean for each auditory filter, resulting in $\mathrm{SNR}(p)_{\mathrm{mon}}$. These monaural SNRs were then optimally combined across frequency bands providing the single-valued $\mathrm{SNR}_{\mathrm{mon}}$ to which a logarithmic transformation was applied with lower and upper bounds, similarly as applied in [6] to limit the range of the perceptual submeasure:

$$d'_{\mathrm{mon,\,lim}} = \min(\max(10\log(\mathrm{SNR}_{\mathrm{mon}}) + 10, 0), 26). \tag{10}$$

The dynamic range of the binaural d'_{bin} is limited by a lower bound of zero and upper bound of 23, calibrated to the hear-through-mode database [54]. The perceptual range of both model paths was normalized to $d'_{\mathrm{norm}} \in [0; 1]$.

While the sensitivity indices of the model, d', represent the perceptual distance between reference and test signals, the predicted audio quality was obtained as $1 - d'_{\mathrm{norm}}$.[2] This allows for adopting the linear monaural frontend from [6] and combine it with the new binaural frontend without further calibration.

In psychoacoustic detection tasks, monaural and binaural cues are usually best described by an optimal combination [53]. However, Fleßner et al. [7] concluded from the combination functions tested for MoBi-Q that the overall audio quality is dominated by the lower-quality aspect. For eMoBi-Q, selecting the lower-quality component, i.e., monaural or binaural, also yielded better results than an optimal combination (not shown). Therefore, in order to provide a simple but well-performing combination, the lower-quality component was selected as the overall quality rating. However, the features provided in this model can also be used with other backends (see Sec. 4.3). This combined version of the model was used to predict the subjective ratings of the seven databases described below. Additionally, the performance of the monaural and binaural paths in isolation was compared to previous models, which is discussed in Sec. 4.1.

2 DATABASES

All databases, including subjective quality scores and signals, used in this study were made available from the authors of the respective study. The hear-through mode database of Schepker et al. [54] was used for calibration of the relative weight of the binaural features, i.e., γ and ILDs, as well as upper bounds of ILD cues and of the binaural path, cf. [12]. Six further databases covering a broad variety of monaural, binaural, and combined monaural and binaural distortions, because they typically occur in loudspeakers and hearing technology, were used to evaluate the "calibrated" model.

The hear-through mode database consists of 120 speech (female, male) and music (jazz, piano) items, sampled at 48 kHz. The study aimed to assess the audio quality of various hearables, including six commercial devices and three research devices, in the hear-through mode. To achieve this, recordings were made using a mannequin head equipped with the hearables in a laboratory environment with moderate room reverberation (T60 ≈ 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 mannequin head and adjusted in height to be at ear level with the mannequin head. The mannequin head's open-ear recordings served as the reference signals, ensuring that the sound transmission to the eardrum through the hearable devices matched the acoustic transparency of the open-ear reference. The occluded ear was used as an anchor signal. The subjective evaluation of the hearables was conducted with 17 NH participants by employing a framework like Multiple Stimulus with Hidden Reference and Anchor (MUSHRA).

The following six databases were used for model evaluation. The subjective quality ratings for all these databases were measured in headphone experiments in sound-isolated booths with participants who had NH.

2.1 Binaural Distortions

The database by Fleßner et al. [33] has 114 items, consisting of speech, music, and pink noise signals with a duration of 10 s. The reference signals were diotic and thus perceived in the middle of the head as a narrow spatial image. The test signals were manipulated in ILDs and ITDs to change the perceived apparent source width, listening envelopment, and direction of arrival of the sound source. The listeners rated the perceived difference between a reference

[2] The Weber law suggests that a logarithmic d' axis is more likely to reflect perception than a linear axis [64]. This would suggest associating $log(1/d')$ with audio quality rather than $1 - d'$. However, a backend based on $log(1/d')$ did not give better results and requires a modification of the $d_{\mathrm{mon,\,lim}}$ adopted from [6]. Therefore, in this work, a linear association of d' with audio quality is used, which provides simplicity paired with performance. For future backends, e.g., involving a neural network, a logarithmic association of d' and audio quality may be preferred.

and various test signals on a numerical rating scale ranging from 100 (“no difference”) to 0 (“very strong difference”) by using a procedure similar to the MUSHRA method.

The binaural magnification database, including eight items, sampled at 44.1 kHz, was taken from [33], comprises binaural hearing aid algorithms [65], and magnifies binaural ILD and ITD cues to improve the spatial separation between sound sources. The algorithm was applied to one speaker in a conversation scenario who talks with another (unprocessed) speaker. Such processing shifts the perceived location of the processed speaker, while the spatial position of the other talker does not change. In the unprocessed reference signal, both speakers were perceived in front of the receiver. Different degrees of magnifications were tested, and ten NH listeners rated the overall difference between the reference and test signals by using a procedure similar to MUSHRA.

The database of Gößling et al. [14] contains 32 speech items, sampled at 16 kHz and a duration of about 7 s. In their study, Gößling et al. measured the performance of six noise-reduction algorithms based on the binaural minimum-variance-distortionless-response (MVDR) beamformer, which compromise between noise-reduction performance and preservation of the interaural coherence for diffuse noise fields. An MVDR beamformer with optimal processing strategy, which reduces the SNR between the speech and noise component but perfectly preserves the interaural coherence of the diffuse noise component, was used as the reference signal. The anchor signal was obtained by averaging the left and right output signals of the MVDR with optimal processing strategy algorithm, resulting in a monaural signal. Consequences of such algorithms on the perceived audio quality were assessed for anechoic and echoic (cafeteria) room conditions. Eleven NH listeners rated the perceived audio quality between the test and reference signals by using a MUSHRA-like procedure.

2.2 Monaural Distortions

The loudspeaker database, taken from [6], consists of 336 items (sampled at 44.1 kHz), based on the ratings of ten well-trained NH listeners (“expert listeners”) for the perceived overall sound-quality difference between a high-quality three-way reference loudspeaker and 59 low-to-mid quality three-way and two-way test speaker systems playing 15 music excerpts (20–30 s). All loudspeakers were digitally equalized in order to evaluate quality differences between test loudspeakers with digitally compensated frequency response and a high-quality three-way reference loudspeaker. The played-back music signals were recorded by a mannequin head (Neutric Cortex MK2). The perceived sound-quality differences between reference and test signals were rated by using a quasi-continuous rating scale ranging from 0 (imperceptible differences) to 4 (significant differences).

The AFC database was taken from the study of Nordholm et al. [66]. It consists of 60 diotic 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 mannequin 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). Nordholm et al. 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. Subjective quality ratings from 15 NH subjects were obtained using the MUSHRA method [67].

2.3 Combined Distortions

The acoustic transparency database, taken from the study of Schepker et al. [68], encompasses 140 speech and music items, sampled at 48 kHz. The study aimed to evaluate the audio quality of a real-time hearing device prototype designed for achieving acoustically transparent sound reproduction by applying feedback suppression using a null-steering beamformer and individualized equalization of the sound pressure at the eardrum. The evaluation was conducted under various recording room conditions, including three different reverberation times (T60 $\approx$ 0.35 s, 0.45 s, 1.4 s) and three incoming signal directions (azimuths of 0°, 90°, 225°). For the recording process, a mannequin head equipped with the hearing devices was utilized. The open-ear recordings from the mannequin head served as the reference signals to establish acoustical transparency. A total of 15 NH listeners were involved in the study, and they employed a MUSHRA-like procedure to rate the perceived overall sound quality of each stimulus relative to the reference signal (open-ear).

3 RESULTS

Prediction performance is characterized by the Pearson linear correlation coefficient r_{Pearson} (accuracy), Spearman rank coefficient r_{rank} (monotonicity), RMS error (RMSE), and epsilon-insensitive RMSE (RMSE* [69]) between subjective and objective ratings. RMSE* is based on the 95% confidence-interval–weighted RMSE and includes a first-order mapping of the objective scores. Given that RMSE* calculations require the standard deviations of the subjective scores, it was only calculated for databases where this information was available. Results are summarized in Table 1. There, r_{Pearson} and r_{rank} for each database are given as predicted by eMoBi-Q. Additionally, scores in parantheses denote the performance obtained by the binaural features in isolation (for binaural distortion databases) or the spectral coloration feature in isolation (for monaural distortion databases).

Fig. 2 shows subjective quality scores and objective scores for eMoBi-Q for the binaural distortions in three databases [33, 14]. In Figs. 2–4, subjective and objective, i.e., instrumentally assessed quality scores, are given on the abscissa and on the ordinate, respectively. Black circles

Table 1. Performance in terms of r_{Pearson}, r_{rank}, RMSE, and RMSE* between subjective and objective ratings for the seven databases predicted by the proposed eMoBi-Q.

Distortion	Database	Study	r_{Pearson}	r_{rank}	RMSE	RMSE*
Binaural	Artificial distortions	Fleßner et al. [33]	0.85 (0.85)	0.85 (0.85)	10.4	⋯
	Binaural magnification	Fleßner et al. [33]	0.96 (0.96)	0.95 (0.95)	5.7	⋯
	MVDR-based algorithms	Gößling et al. [14]	0.89 (0.98)	0.80 (0.95)	13.7	2.5
Monaural	Loudspeakers	Biberger et al. [6]	0.86 (0.90)	0.88 (0.87)	14.8	3.3
	AFC	Nordholm et al. [66]	0.99 (0.99)	0.98 (0.98)	5.7	1.2
Combined	Acoustic transparency	Schepker et al. [68]	0.86	0.85	13.6	2.1
	Calibration: hear-through mode	Schepker et al. [54]	0.90	0.90	9.9	1.5

Note. The correlation coefficients given in parentheses are those obtained with the binaural model path in isolation (for the databases on binaural distortions) or monaural model path in isolation (for the databases on monaural distortions). RMSE* is only provided for those databases where the standard deviations of the subjective scores were available.

and blue diamonds denote predictions determined by either spectral or binaural features, respectively.

For the calibration database, eMoBi-Q achieved $r_{\text{Pearson}} = 0.9$ and $r_{\text{rank}} = 0.9$ (RMSE* = 2.2). eMoBi-Q performed well the for the artificial binaural distortions in the database by Fleßner et al. ($r_{\text{Pearson}} = 0.85$, $r_{\text{rank}} = 0.85$) and gave accurate predictions for the magnification hearing aid algorithm, indicated by $r_{\text{Pearson}} = 0.96$ and $r_{\text{rank}} = 0.95$, as well as for the MVDR-based algorithms ($r_{\text{Pearson}} = 0.89$, $r_{\text{rank}} = 0.8$, RMSE* = 2.46). Table 1 shows that for these databases, prediction performance increases when only the binaural path of eMoBi-Q is used.

In Fig. 3, eMoBi-Q scores are plotted over subjective quality scores for the monaural distortions in the loudspeaker and AFC databases. The eMoBi-Q provided good-quality predictions for the loudspeaker database ($r_{\text{Pearson}} = 0.86$, $r_{\text{rank}} = 0.88$) and very accurate predictions for the AFC database ($r_{\text{Pearson}} = 0.99$, $r_{\text{rank}} = 0.98$, RMSE* = 1.2), respectively. The prediction performance of eMoBi-Q for combined monaural and binaural distortions are shown in Fig. 4. Next to the hear-through-mode database used for calibration of the binaural path, eMoBi-Q also replicated the ratings on the acoustic transparency of hearing aid prototypes [68] very well ($r_{\text{Pearson}} = 0.86$, $r_{\text{rank}} = 0.85$). Without further optimization and parameter adjustment procedures, the presented combined model eMoBi-Q achieved average r_{Pearson} and r_{rank} coefficients between subjective ratings and objective model ratings of 0.9 and 0.89, respectively, for seven databases.

4 DISCUSSION

The presented eMoBi-Q was shown to predict a range of monaural, binaural, and combined distortions well. The involved features are the complex correlation coefficient γ, which incorporates interaural coherence ($|\gamma|$) characterizing compactness and the IPD ($\arg\{\gamma\}$), ILD representing laterality, and a simplistic monaural representation for spectral coloration and loudness. With the model structure being transparent and simple, developers can incorporate the features into their analyses according to their own requirements.

4.1 Comparison to Other Instrumental Quality Measures

An instrumental quality measure intended to predict overall audio quality requires capturing aspects that degrade monaural and binaural audio quality. To assess the power of the auditory cues analyzed in eMoBi-Q, the prediction

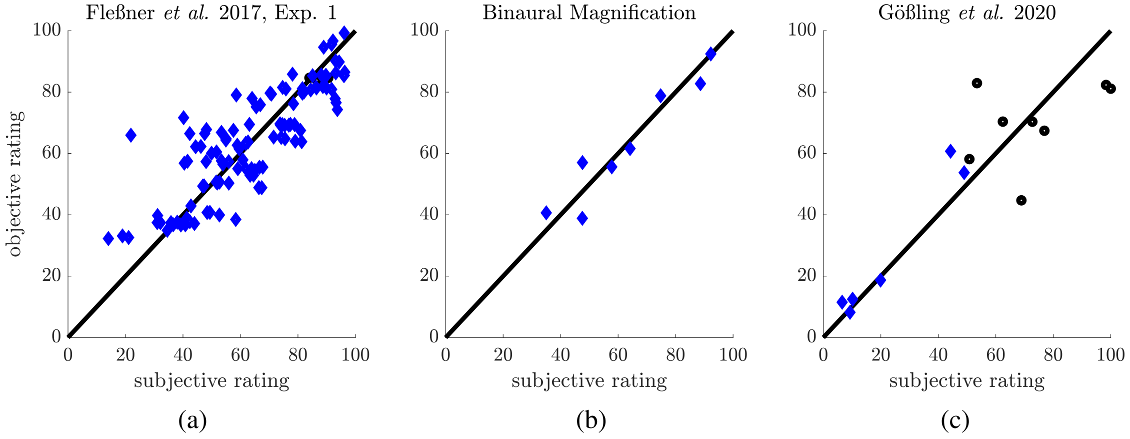
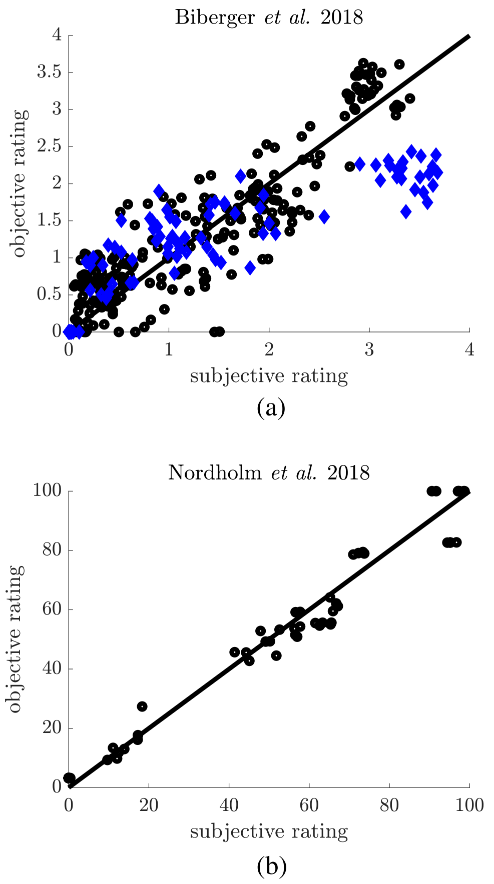


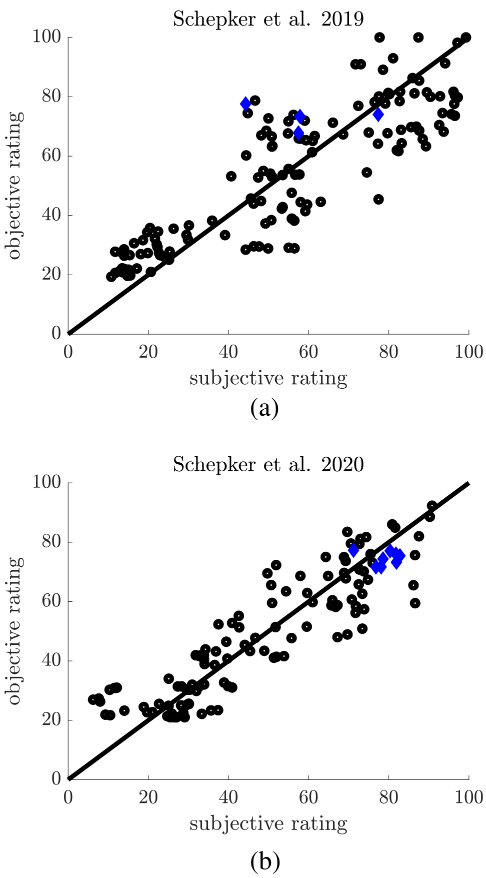
Fig. 2. Subjective and objective quality scores for the databases on binaural distortions. Black circles denote conditions determined by the monaural path of eMoBi-Q, i.e., spectral coloration based measure, being lower than the binaural distortion measure; blue/lighter diamonds (color online) denote those determined by a lower binaural path. (a) Database by [33] on artificial distortions in ITDs, ILDs, and head-related transfer functions; (b) binaural magnification database [65, 33]; (c) database on noise reduction algorithms based on the binaural MVDR beamformer by [14].



(a)

(b)

Fig. 3. Subjective and objective quality assessments for the databases on monaural distortions. (a) Loudspeaker database [6]; (b) AFC database [66].



(a)

(b)

Fig. 4. Subjective and objective quality assessments for the databases on combined monaural and binaural distortions. (a) Acoustic transparency database [68]; (b) database on the quality of the hear-through mode of hearables [54], used for calibration of the binaural path.

performance of the isolated monaural and binaural paths of eMoBi-Q are compared with existing monaural and binaural instrumental quality measures in the following[3]. Besides an adequate representation of monaural and binaural cues, the combination of such cues is also important to gain reasonable overall quality outcomes. Therefore, eMoBi-Q is additionally compared to an existing instrumental measure for overall audio quality.

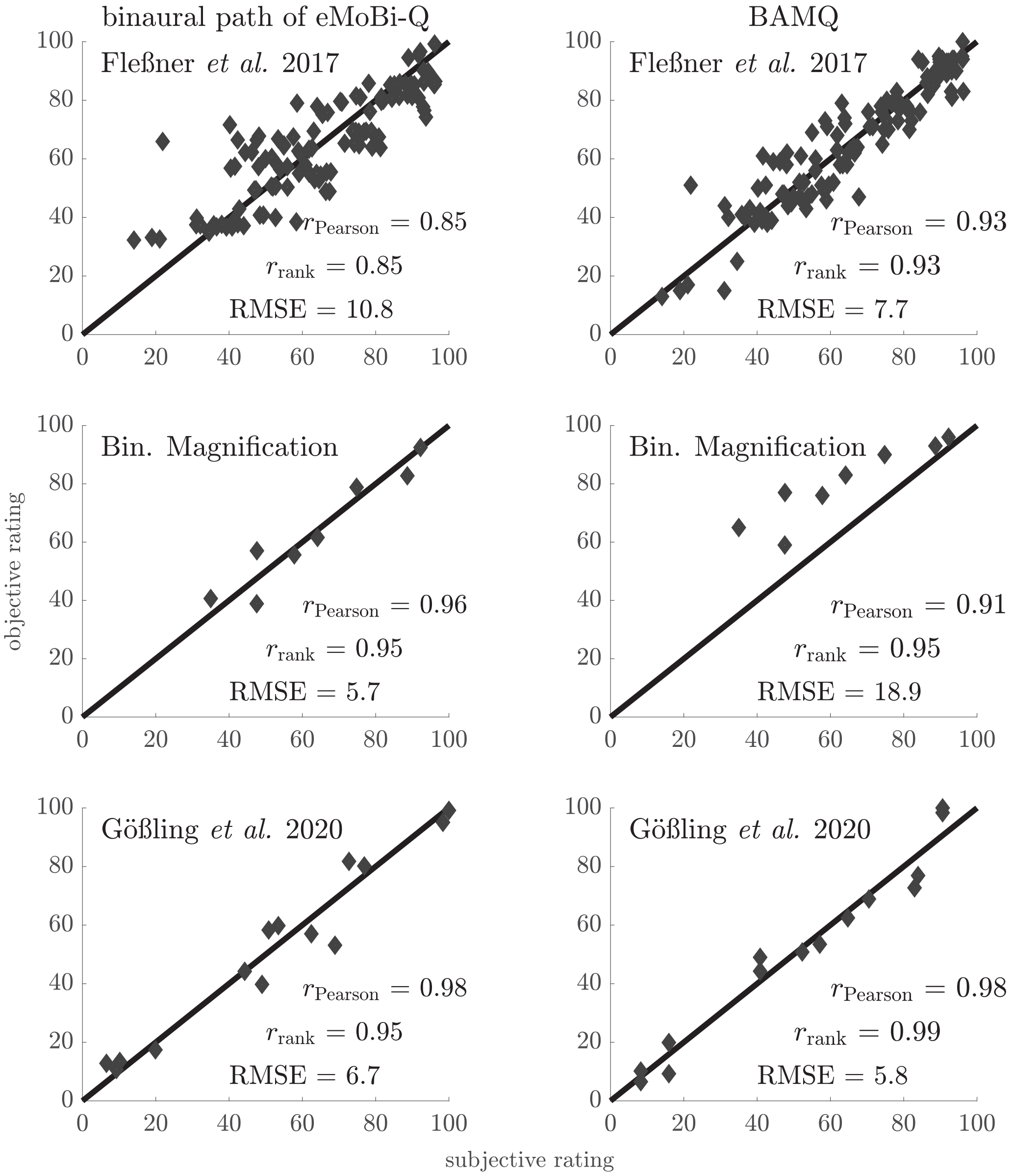
4.1.1 Binaural Measures

One goal in this study was to assess whether the simplistic and computationally efficient binaural auditory model of Eurich et al. [49] is suitable to predict binaural audio quality. For that reason, the prediction performance of the binaural path of eMoBi-Q was compared to that of the established binaural audio quality model BAM-Q [33] for the three binaural databases in this study. As shown in Fig. 5 for the databases for binaural magnification and MVDR beamformers, the binaural path of eMoBi-Q has a prediction performance comparable to BAM-Q, which is also indicated by similar r_{Pearson} and r_{rank} above 0.9 for both models, as well as by comparable RMSEs. However, for the database of Fleßner et al. [33], the prediction performance of the binaural path of eMoBi-Q ($r_{\text{Pearson}} = 0.85$, $r_{\text{rank}} = 0.85$, RMSE = 10.8) is lower than that of BAM-Q ($r_{\text{Pearson}} = 0.93$, $r_{\text{rank}} = 0.93$, RMSE = 7.7). Given that BAM-Q has been trained on the Fleßner database, it is not surprising that BAM-Q outperforms eMoBi-Q for that database.

The features extracted by BAM-Q—ITD, ILD, and IVS—are related to the features of eMoBi-Q, γ, and ILD. The backend of BAM-Q, however, involves the "multivariate adaptive regression splines" [70, 71] consisting of forward and backward passes to fit the relative importance of the three features to the data as well as further computations to obtain the quality ratings. In the present binaural model, however, $1 - d'_{\text{norm}}$ is directly used as binaural quality rating. The proposed model can serve as a basis for potentially more elaborate backends to further optimize prediction accuracy. However, when the relative contribution of the ILD feature is increased to $\alpha = 1/8$, performance of the binaural path of eMoBi-Q for the Fleßner database becomes closer to BAM-Q ($r_{\text{Pearson}} = 0.88$, $r_{\text{rank}} = 0.89$, RMSE = 9.4).

To test for significant RMSE differences between the binaural path of eMoBi-Q and BAM-Q, a Wilcoxon signed-

[3] MATLAB implementations of eMoBi-Q, MoBi-Q, BAM-Q, and GPSM^{q} are publicly available and can be found under www.faame4u.com. eMoBiQ can be downloaded from Zenodo: B. Eurich, S. D. Ewert, M. Dietz, and T. Biberger, "Efficient Monaural and Binaural Model for Audio Quality (eMoBi-Q)," https://doi.org/10.5281/zenodo.12671474.



(a) (b)

Fig. 5. Performance comparison of the binaural path of the present eMoBi-Q in isolation (a) with the established binaural quality model BAM-Q (b). The databases used on binaural distortions are the data from experiment 1 in Fleßner et al. [33]; the binaural magnification database [65, 33] and the database of binaural cue preservation in binaural MVDR beamformers are from Gößling et al. [14].

rank test was performed. This nonparametric test was chosen because the RMSE did not follow a normal distribution. For the seven databases, the RMSEs were not significantly different ($p = 1.00$).

In a nutshell, the similar overall performance of the two models highlights the strength of the simplistic binaural path of eMoBi-Q and the suitability of the complex correlation coefficient γ for binaural quality assessment. Therefore, the binaural path of eMoBi-Q could also provide a useful binaural extension for other monaural audio quality models.

4.1.2 Monaural Measures

The subjective quality ratings in the databases on loudspeakers and AFC were well replicated by the current eMoBi-Q model as indicated by r_{Pearson} values of 0.86 and 0.98, respectively. eMoBi-Q and the isolated monaural path of eMoBi-Q achieved the same prediction performance for the database on AFC (compare results without and with parentheses in Table 1), while for the (dichotic) loudspeaker database, eMoBi-Q performed slightly worse than the isolated monaural path of eMoBi-Q. This is due to the cue redundancy in the monaural and binaural features (see below). Specifically, interaural coherence cues (i.e., $|\gamma|$) are present because the loudspeaker database compares recordings in rooms.

The studies of Biberger et al. [6, 12] demonstrated that for the loudspeaker and AFC databases, accurate predictions of the perceptual effects of spectral distortions are important. Therefore, the naturalness model [4], Hearing-Aid

Table 2. Performance comparison of eMoBi-Q (bold text) and MoBi-Q [7] (Roman text) in terms of r_{Pearson}, r_{rank}, RMSE, and RMSE* between subjective and objective ratings.

Distortion	Database	r_{Pearson}		r_{rank}		RMSE		RMSE*	
Binaural	Artificial distortions	**0.85**	0.93	**0.85**	0.93	**10.4**	7.7	···	···
	Binaural magnification	**0.96**	0.98	**0.95**	1.00	**5.7**	3.9	···	···
	MVDR-based algorithms	**0.89**	0.98	**0.80**	0.97	**13.7**	5.9	**2.5**	1.1
Monaural	Loudspeakers	**0.86**	0.67	**0.88**	0.62	**14.8**	21.8	**2.2**	3.3
	AFC	**0.99**	0.83	**0.98**	0.80	**5.7**	10.4	**1.2**	3.3
Combined	Acoustic transparency	**0.86**	0.83	**0.85**	0.80	**13.6**	15.0	**2.1**	2.3
	Calibration: hear-through mode	**0.90**	0.79	**0.90**	0.81	**9.9**	13.9	**1.5**	2.2

Speech Quality Index version 2 [28], and GPSM$^{\mathrm{q}}$ [6], each explicitly accounting for spectral differences between reference and test signals, were used as monaural comparison models. For the loudspeaker and AFC databases, eMoBi-Q performs similarly to the naturalness model, Hearing-Aid Speech Quality Index version 2, and GPSM$^{\mathrm{q}}$ (loudspeaker database: r_{Pearson} values of 0.85, 0.8, and 0.9; AFC database: $r_{\mathrm{Pearson}} = 0.95$ for all three comparison models).

4.1.3 Combined Measures

Evaluating the binaural and monaural paths of eMoBi-Q in isolation has shown that binaural and monaural cues in hearing devices and loudspeakers are generally well predicted. This gives developers the choice of using the paths in isolation or in combination.

Biberger et al. [12] tested the combination of GPSM$^{\mathrm{q}}$ and BAM-Q [33] with the acoustic transparency database [68] (see Sec. 2.3). While GPSM$^{\mathrm{q}}$ alone performed well ($r_{\mathrm{Pearson}} = 0.87$, $r_{\mathrm{rank}} = 0.86$), performance was slightly reduced when it was combined with BAM-Q in MoBi-Q ($r_{\mathrm{Pearson}} = 0.83$, $r_{\mathrm{rank}} = 0.80$). This is not the case for eMoBi-Q, which performed as equally well as GPSM$^{\mathrm{q}}$ ($r_{\mathrm{Pearson}} = 0.88$, $r_{\mathrm{rank}} = 0.87$).

An even more substantial detrimental impact of BAM-Q combined with GPSM$^{\mathrm{q}}$ was observed for the hear-through mode database [54] (MoBi-Q: $r_{\mathrm{Pearson}} = 0.79$, $r_{\mathrm{rank}} = 0.81$; GPSM$^{\mathrm{q}}$: $r_{\mathrm{Pearson}} = 0.92$, $r_{\mathrm{rank}} = 0.91$). Given that eMoBi-Q was calibrated on the hear-through mode database [54], it seems plausible that it achieved a better performance ($r_{\mathrm{Pearson}} = 0.90$, $r_{\mathrm{rank}} = 0.90$) than MoBi-Q without any a priori knowledge about that database.

The reduced prediction performance of the combined model compared with the monaural or binaural path in isolation can be explained by binaural distortions also being reflected in spectral distortions. Because ILDs are extracted as the logarithmic power ratio between the left and right bandpass signals, interaural differences in DC power are detected by both the ILD and DC-power feature of eMoBi-Q. Furthermore, as discussed by [7] and [12], the way of combining monaural and binaural paths has a major impact on the overall predicted quality and carries the risk of obtaining a large number of degrees of freedom, overfitting, and significant degradation of prediction performance. For this reason and for the sake of simplicity, no specific weighting of the monaural or binaural paths was used in eMoBi-Q.

Because one result of [7] was that the lower-quality component determines the overall quality, this was applied to eMoBi-Q. The result is a lean combined monaural and binaural instrumental quality measure with fewer degrees of freedom, which, at the same time, achieves a slightly higher performance than the more complex MoBi-Q on combined distortions considered in this study. Over the seven considered databases, however, a Wilcoxon signed-rank test showed that the RMSEs of eMoBi-Q and MoBi-Q were not significantly different ($p = 0.81$).

4.2 Computational Efficiency

The goal was to provide a computationally efficient audio quality model that can serve as both a real-time hearing device control and a development tool. This was achieved by including only those processing steps necessary to evaluate the overall audio quality of hearing instruments and sound reproduction systems. This resulted in a simpler model structure of eMoBi-Q compared with MoBi-Q and, thus, in a potential reduction of computational load. As shown in Table 3, for a 1-s two-channel audio signal, eMoBi-Q's signal processing takes about 257 ms, while the current implementation, MoBi-Q [7], needs 17 s.[4] Given that the tested implementation of MoBi-Q was not designed with a focus on computational efficiency, the current performance difference should, however, be considered with caution. While eMoBi-Q is expected to be considerably faster than MoBi-Q, the authors assume that runtime differences with an optimized implementation of MoBi-Q are substantially smaller than currently reported.

An obvious redundancy in MoBi-Q are two separate peripheral filter stages for the binaural (BAM-Q) and monaural model (GPSM$^{\mathrm{q}}$), and consequently, eMoBi-Q uses a joint filterbank for the monaural and binaural model pathways. More importantly, the frontend processing in MoBi-Q includes several further processing stages like fine-structure and modulation filters from which IVS, IPD, and amplitude-modulation–based features are calculated, which has been dispensed with in eMoBi-Q.

Because of the low computational complexity of the monaural and binaural feature calculation in eMoBi-Q, the peripheral filterbank requires 48% of the runtime, and the

[4] The models were run in MATLAB on an Intel(R) Core(TM) i7-8565U CPU at 1.80 GHz machine, using a single thread.

Table 3. Comparison of computation times for a 1-s two-channel audio signal between eMoBi-Q and the (currently unoptimized) implementation of MoBi-Q. For each of the two models, the model stages that require the most processing times are shown.

Model	Model stages	Computation time
eMoBi-Q	**Entire model**	**0.257 s**
	Peripheral filtering	0.123 s (48%)
	Envelope low-pass filtering	0.059 s (23%)
MoBi-Q	**Entire model**	**17 s**
	Binaural BAM-Q	11.9 s (70%)
	Monaural GPSM^q	5.1 s (30%)
	BAM-Q: peripheral, fine-structure, and modulation filtering	9.35 s (55%)
	GPSM^q: peripheral filtering, cue normalization	4.76 s (28%)
	BAM-Q: IVS and IPD feature calculation	0.85 s (4%)

subsequent envelope low-pass filtering another 23%, meaning that approximately 71% of the total runtime is spent in this initial stage. Reducing the number of frequency bands can therefore further reduce the computational load. To explore this potential, the model was evaluated with a reduced number of frequency bands. The lowest and highest center frequencies were kept constant at 315 and 12,500 Hz, respectively, while the density of the frequency bands in between was reduced from 1 filter per ERB (default) to 0.8, 0.5, and, as an extreme case, 0.2 filters per ERB. With the filter bandwidth unadjusted, this led to a reduction in filter overlap and, in extreme cases, to the neglect of frequency ranges between the filters.

With 0.5 filters per ERB, the runtime was reduced from 257 to 121 ms, which means the runtime is approximately proportional to the number of frequency bands. The resulting performance in terms of their r_{rank} between subjective data and model predictions is shown in Fig. 6. Depending on the individual database, low to moderate performance losses were observed for 0.8 and 0.5 filters per ERB. Only for the database of Gößling et al. [14], a more significant loss was observed at 0.5 filters per ERB. For the extreme case of 0.2 filters per ERB, however, substantial performance losses were observed for three of the seven databases (binaural calibration, MVDR beamformers, and loudspeaker database).

The authors hypothesize that, based on the used set of seven databases, distortions that occur in one frequency band are likely to also occur in at least one neighboring frequency band. Thus, even if the sensitivity of the model is not constant over the entire frequency range (it is constant for the standard density of one filter per ERB, where transfer functions cross at their 3-dB–down points), a large part of the distortions that determine the subjective ratings are captured. Compensating the lower density of frequency bands with larger filter bandwidths led to more substantial performance losses (not shown). The more significant loss for the database by Gößling et al. on binaural cue preservation in MVDR beamformers, however, shows that binaural audio quality in such applications relies on cues that are not necessarily represented in adjacent frequency regions.

The authors conclude that for some time-critical applications, such as real-time evaluation, it may be useful to use the model with a reduced number of frequency bands. However, in order to maintain generalizability to different stimuli with different bandwidths, it is recommended that the center frequencies of the remaining filters cover a wide range, such as 315–12,500 Hz.

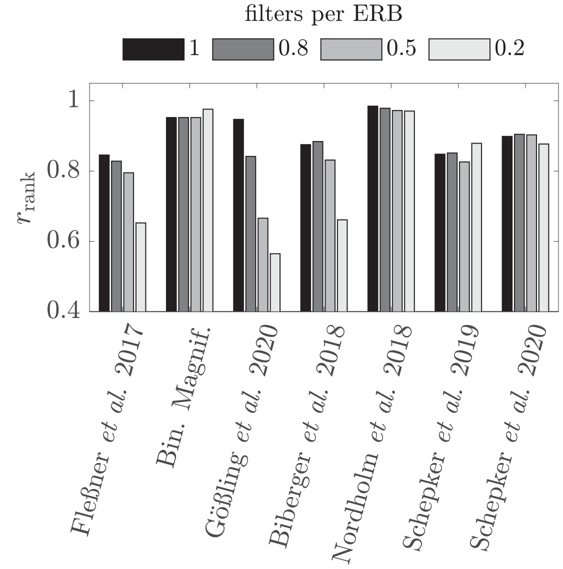


Fig. 6. Performance of the presented eMoBi-Q in terms of prediction monotonicity (r_{rank}) for the seven considered databases for different spacings of the frequency bands. While the lower and uppermost center frequencies are kept constant, the distance between center frequencies is increased. A lower number of frequency bands reduces computational load. Results given in Figs. 2–5 use one filter per ERB.

4.3 Limitations and Reasonable Model Extensions

Besides the shown range of distortion types that are well captured by the presented model, there are also distortion types the model is not expected to be accounted for: The presented model does not include a feature to capture nonlinear distortions, which makes the model unsuitable to evaluate the audio quality of, e.g., audio codecs. Furthermore, distortions such as spectral subtraction, introducing musical tones, are not expected to be accurately detected by the current version of the model without modulation filters.

The frame length of 400 ms was chosen because the focus was on detecting realistic binaural distortions [33, 7, 12] and computational efficiency. Fast dynamic binaural distortions, such as phasewarp (i.e., a binaural beat created by an interaural spectrum shift), are, however, not detected. A future version could possibly include fine-structure–based feature extraction as used in Eurich and Dietz [61], which would increase sensitivity to such mostly artificial distortions at the cost of higher computational load.

To address the redundancy of monaural and binaural cues, which partially results in performance degradation when the monaural DC power path is added to the binaural path and vice versa, a unified monaural and binaural path could be developed for a future version. Alternatively, ILDs and ITDs could be canceled out in the DC power path, as in MoBi-Q. Also, a sophisticated procedure to fit the relative weighting of the model features could potentially slightly improve performance. In the study of Qiao et al. [72], a simple neural network was trained to map the monaural and binaural features of MoBi-Q for timbral, spatial, and overall quality. For their test databases, containing signals processed by binaural rendering algorithms and ambisonics reproduction, such mapping provided more accurate predictions than the original feature combination suggested in MoBi-Q. Thus, replacing the straightforward combination of monaural and binaural features in eMoBi-Q by a carefully trained neural network might also improve the prediction performance. However, the focus of this model was on efficiency, simplicity, and, considering the few degrees of freedom, generalizability.

5 CONCLUSION

A computationally efficient and lean instrumental measure for combined monaural and binaural audio quality assessment was presented. While a number of monaural instrumental quality measures have been established in the past, tools for assessing binaural aspects of audio quality are limited, although spatial cue preservation is important for, e.g., binaural hearing aids and sound-field reproduction. The presented model is a simplified version of MoBi-Q, providing a lean structure and a new, compact binaural path.

The predictive power of the presented model was shown to be comparable with more computationally complex quality models for seven databases involving a range of monaural, binaural, and combined distortions.

Due to the simple structure, the resulting computational efficiency, and the unified analysis timescales of the monaural and binaural paths, the model is suitable for a range of applications. It has the potential for real-time control of algorithms, e.g., in hearing aids, but can also be used as an analysis tool for developers to monitor perceptually relevant distortions. The model will be publicly available from the University of Oldenburg and will be part of the Auditory Modeling Toolbox [73].

6 ACKNOWLEDGMENT

Thomas Biberger and Stephan Ewert were funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) Project-ID 352015383–SFB 1330 A2. Bernhard Eurich and Mathias Dietz were funded by Deutsche Forschungsgemeinschaft Project-ID 352015383–SFB 1330 B4 and the European Research Council under the European Union's Horizon 2020 research and innovation programme (Grant No. 716800).

7 REFERENCES

[1] ITU-T, "Methods for Subjective Determination of Transmission Quality," *Recommendation ITU-T P.800* (1996 Aug.).

[2] W. A. Munson and M. B. Gardner, "Standardizing Auditory Tests," *J. Acoust. Soc. Am.*, vol. 22, no. 5 (Supplement), p. 675 (1950 Sep.). https://doi.org/10.1121/1.1917190.

[3] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).

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

[5] N. Harlander, R. Huber, and S. D. Ewert, "Sound Quality Assessment Using Auditory Models," *J. Audio Eng. Soc.*, vol. 62, no. 5, pp. 324–336 (2014 Jun.).

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

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

[8] J. M. Kates, K. H. Arehart, and L. O. Harvey, "Integrating a Remote Microphone With Hearing-Aid Processing," *J. Acoust. Soc. Am.*, vol. 145, no. 6, pp. 3551–3566 (2019 Jun.).

[9] M. Barron and A. H. Marshall, "Spatial Impression due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure," *J. Sound Vibr.*, vol. 77, no. 2, pp. 211–232 (1981 Jul.).

[10] S. F. Temme, "Testing Audio Performance of Hearables," in *Proceedings of the AES International Conference on Headphone Technology* (2019 Aug.), paper 6.

[11] R. Gupta, R. Ranjan, J. He, W.-S. Gan, and S. Peksi, "Acoustic Transparency in Hearables for Augmented Reality Audio: Hear-through Techniques Review and Challenges," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2020 Aug.), paper 3-7.

[12] T. Biberger, H. Schepker, F. Denk, and S. D. Ewert, "Instrumental Quality Predictions and Analysis of Auditory

Cues for Algorithms in Modern Headphone Technology," *Trends Hear.*, vol. 25, paper 233121652110012 (2021 Jan.). https://doi.org/10.1177/23312165211001219.

[13] P. Derleth, E. Georganti, M. Latzel, et al., "Binaural Signal Processing in Hearing Aids," *Semin. Hear.*, vol. 42, no. 3, pp. 206–223 (2021 Aug.). https://doi.org/10.1055/s-0041-1735176.

[14] N. Gößling, D. Marquardt, and S. Doclo, "Performance Analysis of the Extended Binaural MVDR Beamformer With Partial Noise Estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 462–476 (2020 Dec.). https://doi.org/10.1109/TASLP.2020.3043674.

[15] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural Coherence Preservation in Multi-Channel Wiener Filtering-Based Noise Reduction for Binaural Hearing Aids," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2162–2176 (2015 Dec.). https://doi.org/10.1109/TASLP.2015.2471096.

[16] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic Beamforming for Hearing Aid Applications," in S. Haykin and K. J. R. Liu (Eds.), *Handbook on Array Processing and Sensor Networks*, pp. 269–302 (Wiley, Hoboken, NJ, 2010). https://doi.org/10.1002/9780470487068.ch9.

[17] A. W. Bronkhorst and R. Plomp, "The Effect of Head-Induced Interaural Time and Level Differences on Speech Intelligibility in Noise," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1508–1516 (1988 Apr.). https://doi.org/10.1121/1.395906.

[18] A. W. Bronkhorst and R. Plomp, "Effect of Multiple Speechlike Maskers on Binaural Speech Recognition in Normal and Impaired Hearing," *J. Acoust. Soc. Am.*, vol. 92, no. 6, pp. 3132–3139 (1992 Dec.). https://doi.org/10.1121/1.404209.

[19] A. W. Bronkhorst, "The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions," *Acta Acust. united Acust.*, vol. 86, no. 1, pp. 117–128 (2000 Jan.).

[20] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The Benefit of Binaural Hearing in a Cocktail Party: Effect of Location and Type of Interferer," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843 (2004 Jan.). https://doi.org/10.1121/1.1639908.

[21] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1996), revised ed. https://doi.org/10.7551/mitpress/6391.001.0001.

[22] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of Sound Localization in Mammals," *Physiol. Rev.*, vol. 90, no. 3, pp. 983–1012 (2010 Jul.). https://doi.org/10.1152/physrev.00026.2009.

[23] J. Rennies and Jr. G. Kidd, "Benefit of Binaural Listening as Revealed by Speech Intelligibility and Listening Effort," *J. Acoust. Soc. Am.*, vol. 144, no. 4, pp. 2147–2159 (2018 Oct.). https://doi.org/10.1121/1.5057114.

[24] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-Based Dereverberation Preserving Binaural Cues," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1732–1745 (2010 Sep.). https://doi.org/10.1109/TASL.2010.2052156.

[25] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 968–976 (2005 Aug.). https://doi.org/10.1121/1.1945368.

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

[27] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI)," *J. Audio Eng. Soc.*, vol. 58, no. 5, pp. 363–381 (2010 Jun.).

[28] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117 (2014 Mar.).

[29] J. M. Kates and K. H. Arehart, "The Hearing-Aid Audio Quality Index (HAAQI)," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 2, pp. 354–365 (2016 Feb.). https://doi.org/10.1109/TASLP.2015.2507858.

[30] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, et al., "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I—Temporal Alignment," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384 (2013 Jul.).

[31] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) – A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proceedings of theIEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752 (Salt Lake City, UT) (2001 May). https://doi.org/10.1109/ICASSP.2001.941023.

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

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

[34] M. Schäfer, M. Bahram, and P. Vary, "An Extension of the PEAQ Measure by a Binaural Hearing Model," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8164–8168 (Vancouver, Canada) (2013 May). https://doi.org/10.1109/ICASSP.2013.6639256.

[35] J.-H. Seo, S. B. Chon, K.-M. Sung, and I. Choi, "Perceptual Objective Quality Evaluation Method for High Quality Multichannel Audio Codecs," *J. Audio Eng. Soc.*, vol. 61, no. 7/8, pp. 535–545 (2013 Aug.).

[36] M. Takanen, O. Santala, and V. Pulkki, "Visualization of Functional Count-Comparison-Based Binaural Au-

ditory Model Output," *Hear. Res.*, vol. 309, pp. 147–163 (2014 Mar.). https://doi.org/10.1016/j.heares.2013.10.004.

[37] P. Manocha, A. Kumar, B. Xu, et al., "SAQAM: Spatial Audio Quality Assessment Metric," in *Proceedings of INTERSPEECH*, pp. 649–653 (Incheon, South Korea) (2022 Sep.). https://doi.org/10.21437/Interspeech.2022-406.

[38] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech Enhancement for Multimicrophone Binaural Hearing Aids Aiming to Preserve the Spatial Auditory Scene," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, paper 12 (2016 Feb.). https://doi.org/10.1186/s13634-016-0314-6.

[39] N. Yousefian, P. C. Loizou, and J. H. L. Hansen, "A Coherence-Based Noise Reduction Algorithm for Binaural Hearing Aids," *Speech Commun.*, vol. 58, pp. 101–110 (2014 Mar.). https://doi.org/10.1016/j.specom.2013.11.003.

[40] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Robustness Analysis of Binaural Hearing Aid Beamformer Algorithms by Means of Objective Perceptual Quality Measures," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 315–318 (New Paltz, NY) (2007 Oct.). https://doi.org/10.1109/ASPAA.2007.4393016.

[41] F. E. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *J. Audio Eng. Soc.*, vol. 33, no. 1/2, pp. 2–32 (1985 Feb.).

[42] A. Gabrielsson and B. Lindström, "Perceived Sound Quality of High-Fidelity Loudspeakers," *J. Audio Eng. Soc.*, vol. 33, no. 1/2, pp. 33–53 (1985 Feb.).

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

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

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

[46] T. Biberger and S. D. Ewert, "Towards a Simplified and Generalized Monaural and Binaural Auditory Model for Psychoacoustics and Speech Intelligibility," *Acta Acust.*, vol. 6, paper 23 (2022 Jun.).

[47] M. Mc Laughlin, T. P. Franken, M. van der Heijden, and P. X. Joris, "The Interaural Time Difference Pathway: A Comparison of Spectral Bandwidth and Correlation Sensitivity at Three Anatomical Levels," *J. Assoc. Res. Otolaryngol.*, vol. 15, no. 2, pp. 203–218 (2014 Apr.). https://doi.org/10.1007/s10162-013-0436-6.

[48] M. Dietz, J. Encke, K. I. Bracklo, and S. D. Ewert, "Tone Detection Thresholds in Interaurally Delayed Noise of Different Bandwidths," *Acta Acust.*, vol. 5, paper 60 (2021 Dec.). https://doi.org/10.1051/aacus/2021054.

[49] B. Eurich, J. Encke, S. D. Ewert, and M. Dietz, "Lower Interaural Coherence in Off-Signal Bands Impairs Binaural Detection," *J. Acoust. Soc. Am.*, vol. 151, no. 6, pp. 3927–3936 (2022 Jun.). https://doi.org/10.1121/10.0011673.

[50] D. McAlpine, D. Jiang, and A. R. Palmer, "A Neural Code for Low-Frequency Sound Localization in Mammals," *Nature Neurosci.*, vol. 4, no. 4, pp. 396–401 (2001 Apr.). https://doi.org/10.1038/86049.

[51] T. Okano, L. L. Beranek, and T. Hidaka, "Relations Among Interaural Cross-Correlation Coefficient (IACC_E), Lateral Fraction (LF_E), and Apparent Source Width (ASW) in Concert Halls," *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 255–265 (1998 Jul.).

[52] D. Just and R. Bamler, "Phase Statistics of Interferograms With Applications to Synthetic Aperture Radar," *Appl. Opt.*, vol. 33, no. 20, pp. 4361–4368 (1994 Jul.). https://doi.org/10.1364/AO.33.004361.

[53] J. Encke and M. Dietz, "A Hemispheric Two-Channel Code Accounts for Binaural Unmasking in Humans," *Commun. Biol.*, vol. 5, no. 1, paper 1122 (2022 Oct.). https://doi.org/10.1038/s42003-022-04098-x.

[54] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo, "Acoustic Transparency in Hearables—Perceptual Sound Quality Evaluations," *J. Audio Eng. Soc.*, vol. 68, no. 7/8, pp. 495–507 (2020 Jul.).

[55] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An Efficient Auditory Filterbank Based on the Gammatone Function," presented at the *Meeting of the IOC Speech Group on Auditory Modelling at RSRE* (Malvern, UK) (1987 Dec.), 1–33.

[56] J. Holdsworth, R. Patterson, I. Nimmo-Smith, and P. Rice, "Implementing a Gammatone Filter Bank," *Annex C of the SVOS Final Report (Part A: The Auditory Filterbank)* (1988 Feb.).

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

[58] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes From Notched-Noise Data," *Hear. Res.*, vol. 47, no. 1-2, pp. 103–138 (1990 Aug.). https://doi.org/10.1016/0378-5955(90)90170-T.

[59] A. Kohlrausch, R. Fassel, and T. Dau, "The Influence of Carrier Level and Frequency on Modulation and Beat-Detection Thresholds for Sinusoidal Carriers," *J. Acoust. Soc. Am.*, vol. 108, no. 2, pp. 723–734 (2000 Aug.). https://doi.org/10.1121/1.429605.

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

[61] B. Eurich and M. Dietz, "Fast Binaural Processing but Sluggish Masker Representation Reconfiguration," *J. Acoust. Soc. Am.*, vol. 154, no. 3, pp. 1862–1870 (2023 Sep.). https://doi.org/10.1121/10.0021072.

[62] M. W. H. Remme, R. Donato, J. Mikiel-Hunter, et al., "Subthreshold Resonance Properties Contribute to

the Efficient Coding of Auditory Spatial Cues," *Proc. Nat. Acad. Sci.*, vol. 111, no. 22, pp. E2339–E2348 (2014 May). https://doi.org/10.1073/pnas.1316216111.

[63] J. Klug and M. Dietz, "Frequency Dependence of Sensitivity to Interaural Phase Differences in Pure Tones," *J. Acoust. Soc. Am.*, vol. 152, no. 6, pp. 3130–3141 (2022 Dec.). https://doi.org/10.1121/10.0015246.

[64] P. Heil and B. Friedrich, "How to Define Thresholds for Level and Interaural-Level-Difference Discrimination: Insights From Scedasticities and Distributions," *Hear. Res.*, vol. 436, paper 108837 (2023 Sep.). https://doi.org/10.1016/j.heares.2023.108837.

[65] B. Kollmeier and J. Peissig, "Speech Intelligibility Enhancement by Interaural Magnification," *Acta Oto-Laryngol.*, vol. 109, no. sup469, pp. 215–223 (1990 Jan.).

[66] S. Nordholm, H. Schepker, L. T. T. Tran, and S. Doclo, "Stability-Controlled Hybrid Adaptive Feedback Cancellation Scheme for Hearing Aids," *J. Acoust. Soc. Am.*, vol. 143, no. 1, pp. 150–166 (2018 Jan.).https://doi.org/10.1121/1.5020269.

[67] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," *Recommendation ITU-R BS.1534-1* (2003 Mar.).

[68] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo, "Subjective Sound Quality Evaluation of an Acoustically Transparent Hearing Device," in *Proceedings of the AES International Conference on Headphone Technology* (2019 Aug.), paper 18.https://doi.org/10.1080/00016489.1990.12088432.

[69] ITU-T, "Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models," *Recommendation ITU-T P.1401* (2012 Jul.).

[70] J. H. Friedman, "Multivariate Adaptive Regression Splines," *Ann. Stat.*, vol. 19, no. 1, pp. 1–67 (1991 Mar.). https://doi.org/10.1214/aos/1176347963.

[71] G. Jekabsons, "ARESLab: Adaptive Regression Splines Toolbox for Matlab/Octave," User's Manual, version 1.5.1 (2011 Jun.).

[72] Y. Qiao, N. Zacharov, and P. F. Hoffmann, "Prediction of Timbral, Spatial, and Overall Audio Quality With Independent Auditory Feature Mapping," presented at the *153rd Convention of the Audio Engineering Society* (2022 Oct.), paper 48.

[73] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A Toolbox for Reproducible Research in Auditory Modeling," *Acta Acust.*, vol. 6, paper 19 (2022 May). https://doi.org/10.1051/aacus/2022011.

THE AUTHORS

Bernhard Eurich

Stephan D. Ewert

Mathias Dietz

Thomas Biberger

Bernhard Eurich has a joint education in music, technology, and hearing science. He received his B.Eng. in Audio and Video Engineering at Robert Schumann Hochschule and University for Applied Sciences Düsseldorf in 2017. In 2019, he obtained his M.Sc. in Hearing Technology at University of Oldenburg. He finalized his doctorate in the division for Physiology and Modeling of Auditory Perception in 2023, working on physiologically plausible and computationally efficient models of binaural perception.

•

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

•

Mathias Dietz received his Diploma in physics in Münster, Germany, in 2006 and his doctorate in hearing research from the Carl von Ossietzky Universität Oldenburg (Germany) in 2009. He worked as a postdoctoral researcher in Oldenburg in 2009–2011, followed by a year at the University College London Ear Institute, London, UK. In 2012–2015, he was a junior research group leader again in Oldenburg. In 2016–2018, he was Associate Professor and Canada Research Chair for Binaural Hearing at the National Centre for Audiology, at Western University, London, Ontario, Canada. In 2018, he was appointed professor, leading the division Physiology and Modeling of Auditory Perception at the Carl von Ossietzky Universität Oldenburg and awarded a European Research Council Starting Grant "Individualized Binaural Diagnostics and Technology."

•

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