Acta Acustica 2022, **6**, 23

https://doi.org/10.1051/aacus/2022018

Topical Issue - Auditory models: from binaural processing to multimodal cognition

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE

OPEN ACCESS

Towards a simplified and generalized monaural and binaural auditory model for psychoacoustics and speech intelligibility

Thomas Biberger* and Stephan D. Ewert

Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, 26111 Oldenburg, Germany

Received 12 May 2021, Accepted 21 April 2022

Abstract – Auditory perception involves cues in the monaural auditory pathways, as well as binaural cues based on interaural differences. So far, auditory models have often focused on either monaural or binaural experiments in isolation. Although binaural models typically build upon stages of (existing) monaural models, only a few attempts have been made to extend a monaural model by a binaural stage using a unified decision stage for monaural and binaural cues. A typical prototype of binaural processing has been the classical equalization-cancelation mechanism, which either involves signal-adaptive delays and provides a single channel output, or can be implemented with tapped delays providing a high-dimensional multichannel output. This contribution extends the (monaural) generalized envelope power spectrum model by a non-adaptive binaural stage with only a few, fixed output channels. The binaural stage resembles features of physiologically motivated hemispheric binaural processing, as simplified signal-processing stages, yielding a 5-channel monaural and binaural matrix feature "decoder" (BMFD). The back end of the existing monaural model is applied to the BMFD output and calculates short-time envelope power and power features. The resulting model accounts for several published psychoacoustic and speech-intelligibility experiments and achieves a prediction performance comparable to existing state-of-the-art models with more complex binaural processing.

Keywords: Auditory Modeling, Psychoacoustic Masking, Binaural Hearing, Speech intelligibility

1 Introduction

Auditory perception is typically binaural, involving signals at both ears. Besides enabling localization based on interaural time and intensity differences, interaural disparities can also be exploited to better detect a target stimulus in spatially separated maskers (spatial release from masking, SRM; e.g., [1, 2]) or an antiphasic tone in diotic noise (binaural masking level difference, BMLD; e.g., [3, 4]). Auditory models have been used to explain and analyze monaural and binaural psychoacoustic phenomena (e.g., [5–9]), and as supportive tools offering instrumental assessment of, e.g., speech intelligibility (SI) and audio quality, that are applicable for the development and control of signal processing (e.g., [10–16]). In such applications, the spectro-temporal composition and interaural differences in the original signal are typically altered [17, 18]. Accordingly, monaural cues relevant for, e.g., spectral and temporal masking [19, 20], and binaural cues relevant for, e.g., sound-source location and apparent source width [15] might be affected. Auditory models, as well as psychoacoustic experiments, have often focused on either monaural or binaural aspects of perception in isolation, and led to a variety of monaural (e.g., [5, 6, 9, 21–26]) and binaural models (e.g., [8, 12, 27–34]). The binaural models typically share "common ground" assumptions of essential monaural preprocessing steps followed by a binaural interaction (BI) stage. In many of these binaural models, the prototype BI is based on the equalization-cancelation mechanism (EC; [28]), providing a "monaural", single-channel output signal after a signal-adaptive binaural noise cancelation. To cancel undesired noise, this single-channel output either uses the optimal internal delay to compensate for external interaural delays in connection with an optimal level compensation (equalization), comparable to an adaptive binaural (or bilateral) beamformer (for an overview, see [35]), or simply selects the better ear (referred to as "better-ear glimpsing" if applied in time-frequency frames, see [1]). Thus, the EC mechanism can easily be applied as the binaural front end to an existing monaural model (for speech intelligibility see, e.g., [12–14, 36, 37]). Providing a monaural or diotic input reverts such models to monaural ones, although they are typically applied to binaural (dichotic) stimuli. Focusing on a large variety of basic binaural psychoacoustic experiments, Breebaart et al. [8, 38, 39] combined a number of internal delays and interaural gains into a matrix of (excitatory-inhibitory) cancelation elements. This avoids using a signal-adaptive mechanism to equalize prior to

*Corresponding author: thomas.biberger@uni-oldenburg.de

cancelation, as in the EC approach; however, a signal-adaptive template mechanism is required to "select" optimal matrix elements by applying weights in the form of a template for a given psychoacoustic experiment. Both the monaural front end and the template-matching procedure used in the Breebaart model have been taken from the (monaural) perception model of [5, 6].

The above-mentioned models use successful concepts for combining monaural and binaural model stages in a combined model. They have, however, been explicitly applied to binaural psychoacoustics or speech intelligibility, whereas their front and back ends without a binaural stage have been explicitly applied to the respective monaural experiments. In addition, the models require a signal-adaptive mechanism in the EC stage and a selection from three output channels (EC approach: Left, EC output: right), or a signal-adaptive template to extract information from the high-dimensional matrix of delay-gain elements.

The question arises whether a simpler, non-adaptive approach is sufficient to model binaural interaction. For speech intelligibility in symmetrically placed interferers, e.g., [2] found that a simple addition of the left and right input channel can explain a large part of the observed spatial release from masking. Such a simplistic binaural interaction has also been suggested by [40] as a midline spatial channel in the human auditory cortex. Additionally, the existence of delay lines as utilized in the EC and Breebaart approach has been questioned in mammals (for a review see [41]). Physiological studies (e.g., [42, 43]) suggest a simpler hemispheric model without delay lines to account for binaural interaction, involving fixed phase delays and excitation as well as inhibition from the contralateral ear. Regarding the development of effective auditory signal-processing models, such a fixed binaural interaction could be beneficial for applications in which computational efficiency is important.

It is thus desirable to evaluate the same model both in monaural and binaural experiments, as well as in basic psychoacoustic tasks and speech intelligibility. The advantage of such a unified modeling approach (see, e.g., [9, 26] for monaural models) is the applicability of the model to a wide variety of stimuli, as well as the potential of the model to directly link performance and cues in basic psychoacoustic tasks, such as detection and discrimination thresholds, to higher-level processes involved in speech intelligibility. In the long run, such a link might help to understand and disentangle peripheral and central deficits in hearing impaired and elderly persons (e.g., [44–48]) and might be useful for a model-driven stimulus design in psychoacoustics and physiology (e.g., [49]).

Here we suggest and examine a combined monaural and binaural model in a variety of "benchmark" psychoacoustic and speech-intelligibility experiments. The combined approach uses the monaural front end and back end of the generalized power spectrum model (GPSM; [26]) which has been successfully applied to monaural psychoacoustics, speech intelligibility, and audio quality assessment [9, 16–18, 26]. A binaural processing stage with five fixed (non-adaptive) output channels is suggested, referred to as binaural matrix feature decoder (BMFD). The output comprises the left (L) and right (R) channels, the L + R channel, and the L–R and R–L channels, incorporating a fixed phase delay and gain. L and R enable better-ear glimpsing in connection with a selection of time-frequency frames across the BMFD output channels in the back end (better ear channels). The three other channels realize a binaural interaction: L + R represents a midline channel, enhancing (frontal) signals that are coherent across ears. In a highly simplified manner, the L–R and R–L channels effectively mimic the outputs expected in hemispheric models of binaural interaction. These channels are comparable to two elements in the delay-gain matrix of the Breebaart model, or to two specific parameter choices in the EC approach. The ability of the current model to account for the monaural and binaural data, and the relevance of the five BMFD output channels are assessed in the following.

2 Model description

The front end of the proposed GPSM with BMFD extension calculates short-time power and envelope power features for each of two better-ear (BE) channels (L: BE_L, R: BE_R) and for the three binaural interaction (BI) channels (L–R: BI_L, L + R: BI_C, R–L: BI_R), comprising the binaural matrix feature decoder. Signal-to-noise ratios (SNRs) based on these features are assessed by a task-dependent decision stage (psychoacoustics or speech intelligibility) in the model back end. The model processes two input stimuli, the target-plus-masker (signal) and masker alone (noise).

The peripheral processing, feature extraction, and decision stages of the GPSM with BMFD extension, illustrated in Figure 1, are similar to those of the monaural mr-GPSM proposed in [26]. In the following, the processing stages related to the envelope power pathway are only roughly described. For a more comprehensive description, the reader is referred to [9, 26].

2.1 Monaural processing stages

The initial *Outer & middle ear filtering* stage (see Fig. 1) weights the input signal with the hearing threshold in quiet [50], followed by the *Auditory Fb*, reflecting basilar membrane filtering by applying a fourth-order Gammatone filterbank with a bandwidth equal to the equivalent rectangular bandwidth of the auditory filter (ERB_N; [51]) and third octave spacing from 63 to 12,500 Hz. In contrast to the Hilbert envelope extraction in [26], each auditory channel is half-wave rectified to simulate the fact that inner hair cells primarily respond to one direction of deflection. To simulate effects of neural adaptation of the auditory system in a simple feed-forward manner, the squared half-wave rectified signals are divided by their first-order low-pass filtered versions (integration time constant of 0.32 ms, cut-off frequency of 500 Hz).

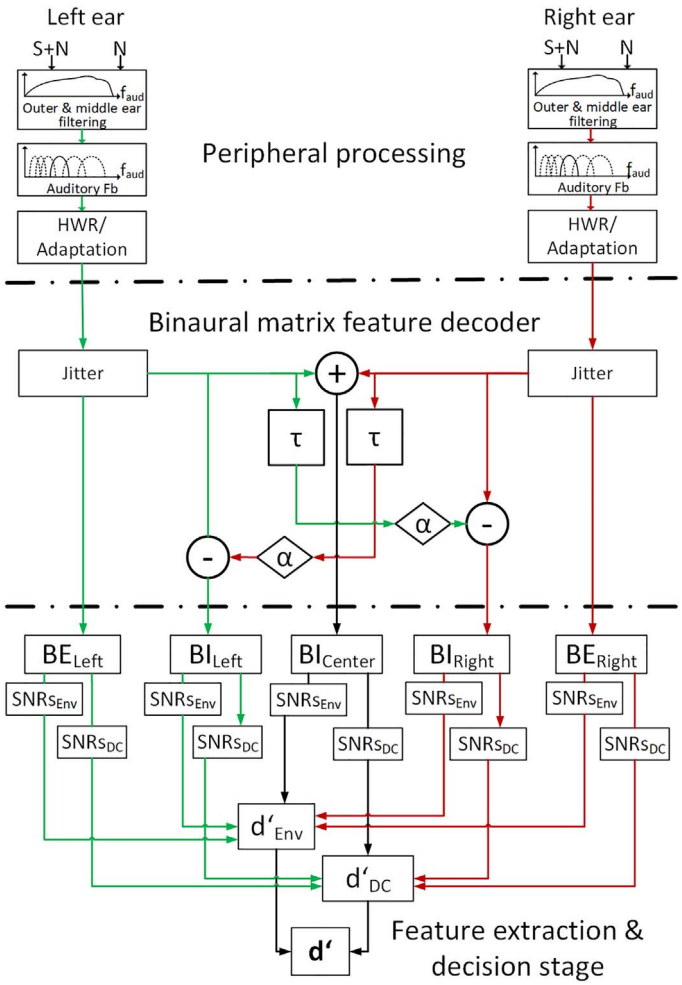

Figure 1. Block diagram of the GPSM with BMFD extension. After peripheral processing, the left and right ear signals are binaurally processed by using the BMFD that provides two better-ear channels BE_L and BE_R and three binaural interaction channels BI_L, BI_C, BI_R. For each of the five BMFD outputs, envelope power and power SNRs are calculated on short-time frames and then combined across the five channels of the BMFD and across auditory and modulation channels, resulting in a sensitivity index d'_{env} based on envelope power SNRs and d'_{DC} based on power SNRs. The final combined d' was then compared to a threshold criterion that assumes that a signal is detected if $d' > (0.5)^{1/2}$.

2.2 Binaural processing stages

The adapted signals from the monaural processing of the left and right ear serve as the input for the binaural processor. First, to limit the performance of the BI, amplitude and time jitter are applied to the input signals independently for each auditory channel. Amplitude and time jitters are generated as zero-mean Gaussian processes with a standard deviation of $\sigma_\epsilon = 0.25$ for amplitude and $\sigma_\delta = 105$ µs for time jitter, as suggested by [28] and also applied by [36] and [37]. Based on the jittered signals, three BI channels BI_L, BI_C, and BI_R are calculated according to equations (1)–(3):

$$BI_L(p,t) = L(p,t) - \alpha R(p, t - \tau(p)), \tag{1}$$

$$BI_C(p,t) = \sqrt{L(p,t)\,R(p,t)}, \tag{2}$$

$$BI_R(p,t) = R(p,t) - \alpha L(p, t - \tau(p)). \tag{3}$$

BI_L results from subtracting the time delayed and amplified right ear channel $\alpha \cdot R(p, t - \tau(p))$ from the left ear channel $L(p, t)$ in each auditory channel p. BI_R is calculated vice versa to BI_L. Based on physiological findings (e.g., [52]) and preliminary tests, a frequency-dependent delay τ equal to a phase shift of $\pi/4$ was chosen, resulting in longer delays for lower frequencies. The amplification factor α equals 3 (see Sect. 3.3 for further details). BI_C accounts for the effect of adding the left and right ear signals prior to auditory processing. Taking the half-wave rectified signal representation into account, this is achieved by the square root of the product $L(p, t)$ and $R(p, t)$, making BI_C a midline channel that is most sensitive to sound images spatially placed in the median plane. In addition to the three BI channels, the (monaural) left and right channel $L(p, t)$ and $R(p, t)$ are passed unaltered as the output of the five-channel BMFD stage. They can be used for better-ear glimpsing in the following feature-extraction stage (referred to as BE_L, BE_R).

2.3 Power and envelope power feature extraction stage

A first-order low-pass filter with a cut-off frequency of 150 Hz [7, 53] is applied to the five output channels of the BMFD. The consecutive processing stages in each of the five BMFD channels are separated into two independent pathways where envelope power SNRs (EPSM; left-hand output of the BE/BI stages of Fig. 1), and power SNRs (PSM; right-hand output of the BE/BI stages of Fig. 1) are calculated. In the following equations, the indices for the BMFD channels are omitted for clarity.

In the PSM path, the intensity (DC-power) features $P_{DC,j}(p)$ are calculated in short-time windows j by taking the squared mean of the envelope within each auditory channel p

$$P_{\mathrm{DC},j}(p) = \frac{\left[\overline{E}_j(p)\right]^2}{2}. \tag{4}$$

The divisor 2 was applied to make $P_{DC,j}(p)$ comparable to rms-based power values. The duration of the windows depends on the center frequency of the auditory channel, where the lowest center frequency of 63 Hz corresponds to a window length of 45 ms and the highest center frequency of 12.5 kHz provides a window length of 8 ms. As proposed by Rhebergen and Versfeld [11], values for the window duration were taken from [54] and multiplied by 2.5. $P_{\mathrm{DC},j}(p)$ values falling below the hearing threshold are set to 1e–10. Then the $\mathrm{SNR}_{\mathrm{DC},j}(p)$ is calculated between target-plus-masker $P_{\mathrm{DC,targ+mask},j}(p)$ and the masker $P_{\mathrm{DC,mask},j}(p)$ according to

$$\mathrm{SNR}_{\mathrm{DC},j}(p) = \frac{P_{\mathrm{DC,targ+mask},j}(p) - P_{\mathrm{DC,mask},j}(p)}{P_{\mathrm{DC,mask},j}(p)}. \tag{5}$$

For speech intelligibility predictions, optionally a band-importance function (BIF) as used in the ESII, is multiplicatively applied to the power $\mathrm{SNR}_{\mathrm{DC}}(p)$. Note that here, the applied BIF is normalized by its highest value, and thus the $\mathrm{SNR}_{\mathrm{DC}}$ for that channel remains unaffected, while all other channels become attenuated.

In the EPSM path, the envelopes are initially processed by a modulation filter bank consisting of bandpass filters ranging from 2 to 256 Hz with a Q-value of 1, and a third-order low-pass filter with a cut-off frequency of 1 Hz. Based on [55], only modulation filter center frequencies up to one fourth of the corresponding auditory channel center frequency are considered. Consecutively, the AC-coupled envelope power $P_{\mathrm{env},j}(p,n)$ is calculated for each auditory channel p, modulation channel n, and time window i, as proposed in [25], by applying a lower limit of −27 dB for the envelope power, reflecting the limitation in human sensitivity to amplitude modulation (AM) [22, 53]. The envelope power based signal-to-noise ratio $\mathrm{SNR}_{\mathrm{env},i}(p,n)$ between the target-plus-masker and masker envelope power is calculated according to [25], and a logarithmic weighting of envelope power SNRs is applied for auditory channels with a target-plus-masker level below 35 dB, while envelope power SNRs above that level are unaffected by the weighting.

Taken together, the output of the model front end consists of level-weighted envelope power SNRs, $\mathrm{SNR}_{\mathrm{envW},i}(p,n)$, and power SNRs, $\mathrm{SNR}_{\mathrm{DC},j}(p)$, for each of the five BMFD output channels.

2.4 Decision stage

The envelope power and power based SNRs are subjected to a task-specific decision stage for predicting psychoacoustic detection or discrimination thresholds and SI data.

2.4.1 Psychoacoustics

In the first step, $\mathrm{SNR}_{\mathrm{envW},i}(p,n)$ in each of the five front end output channels are combined by taking the largest value for each time frame within each auditory and modulation channel, resulting in $\mathrm{SNR}_{\mathrm{envWC},i}(p,n)$. $\mathrm{SNR}_{\mathrm{envWC},i}(p,n)$ is then averaged across temporal segments i per modulation filter, resulting in a two-dimensional representation of envelope power $\mathrm{SNR}_{\mathrm{env}}(p,n)$. The same procedure is applied to combine $\mathrm{SNR}_{\mathrm{DC},j}(p)$ across the five channels, resulting in the $\mathrm{SNR}_{\mathrm{DCC},j}(p)$ which is then averaged across temporal segments j, resulting in a one-dimensional representation of power SNRs over auditory channels denoted as $\mathrm{SNR}_{\mathrm{DC}}(p)$

Finally, the envelope power and power SNRs $[\mathrm{SNR}_{\mathrm{env}}(p,n), \mathrm{SNR}_{\mathrm{DC}}(p)]$ are combined in the same manner as proposed in [26]:

$$\mathrm{SNR} = \max\left[\beta \cdot \left(\sum_{p=1}^{M}\sum_{n=1}^{N}\mathrm{SNR}_{\mathrm{env}}(p,n)^2\right)^{\frac{1}{2}}, \gamma \cdot \left(\sum_{p=1}^{M}\mathrm{SNR}_{\mathrm{DC}}(p)^2\right)^{\frac{1}{2}}\right]. \tag{6}$$

Initially, envelope power and power SNRs are combined across auditory and modulation channels (in case of envelope power) and auditory channels [inner brackets in Eq. (6)] and multiplied with the empirically determined correction factors $\beta = 0.21$ and $\gamma = 0.45$. Both correction factors are identical to those proposed in [9, 26] and are used due to violation of the assumption of independent observations in the auditory and modulation channels, due to the use of an overlapping bandpass filter. Finally, the domain (envelope or power), providing the highest SNR-value is chosen, representing the overall SNR.

As in [9, 26] the decision criterion used in this study is based on [7] assuming that a signal is detected if the SNR > −6 dB (equivalent to a power ratio of 0.25), which, according to [56], can also be expressed as the sensitivity index $d' = (2 \cdot \mathrm{SNR})^{1/2} \approx (0.5)^{1/2}$.

2.4.2 Speech intelligibility

Following [26], the overall SNR (Eq. (6)) is obtained at a given input target-to-masker ratio as described for psychoacoustic predictions, however without applying any weights ($\beta = \gamma = 1$). The overall SNR is converted to the sensitivity index d' and then transformed into percent correct by using equations (6) and (7) from [26].

The two parameters k and q convert the overall SNR into d' by applying $d' = k(\mathrm{SNR})^q$, while the two parameters m and σ_s refer to the response-set size and the redundancy of the speech material. Based on the response size m, both parameters σ_N and μ_N are established according to equations (A1) and (A2) in [57], required to calculate the percent correct responses $P_{\mathrm{correct}}(d') = \Phi\left(\frac{d'-\mu_N}{\sqrt{\sigma_S^2+\sigma_N^2}}\right)$.

2.5 Model configurations

All model versions with binaural extension tested in this study had the same settings as the monaural GPSM-versions in [9, 26]: For psychoacoustic experiments, auditory filters had a third-octave spacing ranging from 63 to 12,500 Hz, while for SI experiments, auditory filters ranged from 63 to 8000 Hz. For SI predictions, the band-importance weighting, as proposed by Table 3 of [58] was exclusively applied to the power SNRs. Each of the models used exactly the same set of parameters for all experiments.

3 Psychoacoustic evaluation

3.1 Monaural experiments

In this study, the same set of headphone-based monaural psychoacoustic experiments were used for model evaluation as in [9, 26]. Thus, these experiments are only briefly explained in the following. For more detailed information, the reader is referred to [9] or the respective original publications.

In Experiment 1 (Intensity discrimination and hearing thresholds), just noticeable level differences (JNDs) as a function of the reference level (20, 30, 40, 50, 60, 70 dB) were measured for a 1-kHz pure tone (in quiet) and broadband noise ranging from 0.1 to 8 kHz [59]. The target interval contained an increased level $L_t = L_0 + \Delta L$, where L_0 corresponded to the reference level and ΔL corresponded to the JND, which can be rewritten in terms of intensities as $\Delta L = 10\log_{10}\frac{I_t}{I_o} = \frac{\Delta I + I_0}{I_o}$. Hearing thresholds ranging from 50 Hz to 10 kHz were taken from [50].

Experiment 2 (Spectral masking with narrow-band and pure-tone maskers) measured masking patterns for four different signal-masker combinations of noise-in-tone (NT), noise-in-noise (NN), tone-in-tone (TT) and tone-in-noise (TN) [60]. The noise corresponds to a Gaussian noise with a bandwidth of 80 Hz, while the tone refers to a sinusoidal stimulus. The masker had a fixed center frequency of 1 kHz, while the signal had frequencies of 0.25, 0.5, 0.75, 0.9, 1.0, 1.1, 1.25, 1.5, 2, 3, and 4 kHz. All signal-masker combinations had random phases, with the exception of the TT condition, where each stimulus had a fixed phase of 90°. Data for the masker levels of 45 and 85 dB are described here.

Experiment 3 (Tone-in-noise masker) was patterned on [24] and reflects detection thresholds of a 2-kHz pure tone signal in the presence of a band-limited (0.02–5 kHz) Gaussian-noise masker with signal durations from 5 to 200 ms. The masker had a duration of 500 ms and the signal was temporally centered in the masker. The presentation level of the masker was 65 dB SPL.

Experiment 4 (AM-depth discrimination) is based on the study from [61], in which an AM-depth discrimination function for a 16-Hz sinusoidal AM with respect to fixed reference AM depths was measured for sinusoidally modulated broadband noise (1.952–4 kHz) and pure-tone carriers (4 kHz) at an overall presentation level of 65 dB SPL. The AM depth of the (standard) reference signal m_s ranged, in 5-dB steps, from −28 to −3 dB. The increased AM depth of the target signal is given by $m_c = m_s\sqrt{1+m_{\mathrm{inc}}}$. Within the measurement, the fractional increment $m_{\mathrm{inc}} = (m_c^2 - m_s^2)/m_s^2$ was varied in dB (10 log m_{inc}).

In Experiment 5 (AM detection), temporal-modulation transfer functions (TMTF) for three narrow-band noise carriers with bandwidths of 3, 31, and 314 Hz [5] and for broadband noise carriers [22] were assessed. The narrow-band noise carriers were centered at 5 kHz and a sinusoidal AM of 3, 5, 10, 20, 30, 50, and 100 Hz was used. The narrow-band carrier level was 65 dB SPL, and the stimuli were adjusted to have equal power after AM. The broadband noise carriers ranged from 0.001 to 6 kHz, and a sinusoidal AM of 4, 8, 16, 32, 64, 128, 256, 512, and 1024 Hz was applied. The level of the broadband carriers was 77 dB SPL.

Experiment 6 (Amplitude modulation masking) was taken from [9] and measured AM masking and detection thresholds for a sinusoidal target modulation in the presence of a sinusoidal or square-wave masker modulation. The effect of varying the carrier type (broadband and pure-tone carriers), masker waveform (sinusoidal or square-wave), and modulation rate of the target (4 and 16 Hz) and masker (16 and 64 Hz) were examined in four different stimulus configurations that are shown in Table 1 of [9].

3.2 Binaural experiments

Seven published binaural headphone experiments were used for the model evaluation. In Experiments 1–5, the effects of spectral and temporal parameters on binaural processing were examined. This comprised experiments for interaural level- and time differences (ILDs, ITDs), BMLDs depending on signal frequency and interaural phase relationships in wideband masker configurations, forward- and backward masking, and detection thresholds depending on signal duration. Experiment 6 examined the interaction between timing and level cues. Experiment 7 extended the rather artificial stimuli of the previous experiment to a more realistic condition, examining spatial unmasking of pure-tones by simulating spatial source positions with head-related transfer functions (HRTFs). This experiment involved natural ILDs (head-shadow effect) and IPDs in combination.

The maskers used in the binaural experiments had a duration of 400 ms unless otherwise stated. In several binaural experiments target and masker signals comprised interaural manipulations indicated by subscripts: The subscript 0 indicates no interaural phase shift (in phase), the subscript π indicates an interaural phase shift of π (out of phase), and the subscript m indicates that the

corresponding signal was presented monaurally. Accordingly, a N_0S_π stimulus indicates that the noise signal N_0 is interaurally in phase, while the target signal S_π is interaurally out of phase. The experiments are only briefly described in the following and the reader is referred to [38, 39] for Experiments 1–5, or the original literature for further details.

Experiment 1 (ITD discrimination) was based on the ITD experiments from [62–64], where discrimination thresholds for ITDs were measured for pure-tone stimuli at various frequencies. The reference stimuli were presented diotically at a level of 65 dB SPL, while the target stimuli were presented at the same level but had an ITD. The frequencies tested ranged from 90 to 1500 Hz.

Experiment 2 (IID discrimination) was based on the IID experiments from [65–67], where thresholds for IID in [65, 66] were measured for pure tones at various frequencies ranging from 62.5 to 4000 Hz. Threshold IIDs in [67] were measured with 10-Hz narrow-band noises with center frequencies ranging from 300 Hz to 7 kHz. The reference stimuli were presented diotically at a level of 65 dB SPL. The target stimuli had an IID, resulting in an overall level of (65 + IID/2) dB SPL for the one ear and (65 − IID/2) dB SPL for the other ear. Consequently the target stimuli were lateralized either to the left or right side.

Experiment 3 (Frequency and interaural phase relationships in wideband conditions) was based on experiments of [3, 4, 68, 69], in which the thresholds of the four binaural conditions N_0S_π, $N_\pi S_0$, N_0S_m, and $N_\pi S_m$, were measured as a function of the frequency of the pure-tone signal (125, 250, 500, 1000, 2000, and 4000 Hz). The masker was a low-pass noise with a cutoff frequency of 8 kHz and a spectral level of 40 dB/Hz.

Experiment 4 (N_0S_π depending on signal duration) was based on experiments of [70–73], in which N_0S_π detection thresholds were measured as a function of the target signal (S_π) duration. The masker signal (N_0) was a 500-ms wideband noise with a spectral density of 36.2 dB/Hz. The target signal was a pure tone at either 500 Hz or 4 kHz, with signal durations ranging from 2 to 256 ms.

Experiment 5 (Temporal phase transition) was based on the experiments of [74], in which to estimate the temporal resolution of binaural auditory system, $N_0N_\pi S_\pi$, $N_\pi N_0 S_\pi$, $N_\pi N_{\pi,\text{-15dB}} S_\pi$, $N_{\pi,\text{-15dB}} N_\pi S_\pi$, thresholds were measured as a function of the temporal position of the target signal (S_π) relative to the masker-phase transition ($N_\pi N_0$ or $N_0 N_\pi$). The broadband noise maskers with a duration of 750 ms were bandpass filtered from 100 to 2000 Hz and had a spectral level of 40 dB/Hz. The N_0N_π masker started with an interaural phase of N_0 that switched to N_π after 375 ms. Accordingly, $N_\pi N_0$ started with a 375 ms interaurally out-of-phase segment followed by a 375 ms in-phase segment. The interaurally out-of-phase masker $N_\pi N_{\pi,\text{-15dB}}$ was attenuated by 15 dB 375 ms after its onset. The interaurally out-of-phase masker $N_{\pi,\text{-15dB}} N_\pi$ was amplified by 15 dB 375 ms after its onset. S_π was an interaurally out-of-phase pure tone of 500 Hz, with a duration of 20 ms. The masked threshold was measured as a function of the delay time between the transition of the noise segments and the signal offset.

Experiment 6 (Time-intensity-trading) was based on experiments of [75], in which d' was measured for several combinations of fixed ITDs (0, +10, +20, +30, and +40 µs; a positive sign indicates left-ear leading) and varying IIDs (ranging from 0 to −3 dB; a negative sign indicates that the level in the right ear was higher than in the left ear) to examine the extent to which time differences can be traded against level differences. The reference signal was a diotic pure tone of 500 Hz (centered sound image). The test signal had an ITD promoting lateralization to the left side, and an IID promoting lateralization to the right side. The lowest d' measured for a given IID at a fixed ITD indicated that the test signal was most similar to a centered image.

Experiment 7 (Spatial unmasking) was based on [76], in which detection thresholds for different spatial configurations of pure-tone targets and broadband maskers were measured. Both target and masker sources had a distance of 1 m to the receiver. The masker was placed either at 0, 45° or 90°. For each of the three masker locations, detection thresholds were measured for target locations of −90°, −45°, 0°, 45° or 90°. The masker was a low-pass-filtered Gaussian noise with a cutoff frequency of 5 kHz. The target signal was a 165-ms-long pure tone of 500 Hz including 30-ms raised-cosine ramps. The masker presentation level was 64 dB SPL. Spatial locations of targets and maskers were simulated in [76] by using individualized HRTFs; it was, however, also demonstrated that both spectral and binaural effects were largely maintained in HRTFs simulated by KEMAR or a spherical-head model. Here, HRTFs from the CIPIC database [77] were used.

3.3 Results and discussion

To disentangle the contribution of the binaural interaction (BI_L, BI_C, BI_R) and better-ear (BE_L, BE_R) BMFD channels, predictions from three model versions were compared. Model predictions based on all five channels are abbreviated as BMFD and represented by open diamonds. Model predictions based on the three binaural interaction channels are abbreviated as $BI_{L,C,R}$ (open squares), while predictions based on only the left and right BI channel are abbreviated as $BI_{L,R}$ (open circles).

In [9, 26], the monaural GPSM model was calibrated to match data from amplitude modulation and intensity JND experiments. In the current study, the suggested model with binaural extension applied identical calibration parameters $\beta = 0.21$ and $\gamma = 0.45$ [see Eq. (6)] to those used in [9, 26].

3.3.1 Monaural Experiments

The upper part of Table 1 reports root-mean-squared errors (RMSEs) and the coefficient of determination (R^2) between experimental data and predictions based on BMFD, $BI_{L,R}$, and the monaural mr-GPSM [26]. For the monaural experiments, stimuli were only provided to the left-ear input channel of the BMFD, and the right-ear input channel was set to zero. As obvious from the RMSE- and R^2-values, BMFD predictions largely agree with those from the monaural mr-GPSM. Given the similarity of both

Table 1. Root-mean-square errors (RMSE) and coefficients of determination (R^2; squared cross-correlation coefficient) between data and model predictions for the monaural and binaural psychoacoustic experiments. The RMSE values of the time-intensity trading experiment have no units, as they are based on d'.

Monaural experiments						
Experiments	BMFD		$BI_{L,R}$		mr-GPSM [26]	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
1. Hearing threshold	3.3 dB	0.99	3.3 dB	0.99	1.7 dB	0.99
Intensity JNDs	0.2 dB	0.66	0.2 dB	0.64	0.3 dB	0.57
2. Spectral masking	9.5 dB	0.82	9.5 dB	0.8	7.9 dB	0.9
3. Tone in noise	1.3 dB	0.99	1.3 dB	0.99	2.1 dB	0.99
4. AM detection	4.0 dB	0.71	4 dB	0.78	4.5 dB	0.68
5. AM discrimination	2.4 dB	0.94	2.4 dB	0.92	1.6 dB	0.94
6 AM masking	4.6 dB	0.77	4.7 dB	0.79	6.2 dB	0.73
Binaural experiments						
Experiments	BMFD		$BI_{L,C,R}$		$BI_{L,R}$	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
1. ITD discrimination	0.042 ms	0.04	0.042 ms	0.1	0.041 ms	0.12
2. IID discrimination	0.4 dB	0.14	0.4 dB	0.08	0.4 dB	0.15
3. Frequency and interaural phase relationships in wideband conditions	9.1 dB	0.86	8.5 dB	0.85	6.7 dB	0.88
4. N_0S_π dependence on signal duration	2.9 dB	0.92	3.0 dB	0.92	3.2 dB	0.9
5. Temporal phase transition	2.6 dB	0.80	2.7 dB	0.8	2.7 dB	0.81
6. Time-intensity-trading	0.5	0.38	0.6	0.58	0.6	0.61
7. Spatial unmasking	2.3 dB	0.91	2.0 dB	0.94	2.0 dB	0.98

models for the monaural data, detailed figures to compare the measured and predicted data are not shown here. The similarity is expected, as the BMFD has only a few modifications that potentially influence monaural prediction performance. As shown in Table 1, prediction performance was not degraded when instead of all five BMFD outputs, only BI_L and BI_R ($BI_{L,R}$) were used. This result was also expected, because when the right input channel is set to zero, BI_L only depends on the left ear channel, and in such monaural conditions, BI_L is equal to BE_L. Accordingly, reducing the number of output channels of the BMFD would be sufficient to capture important monaural psychoacoustic effects, but may not suffice to account for all the binaural aspects that are assumed to be important to explain a variety of data from binaural psychoacoustic and SI experiments. To summarize, for monaural experiments tested in this study, the GPSM with binaural BMFD extension largely maintained the prediction performance of the monaural mr-GPSM.

3.3.2 Binaural Experiments

In Figures 2–7, subjective and predicted data for the binaural experiments are represented by filled and open symbols, respectively. The error bars shown in the figures refer to ± one standard deviation. The lower part of Table 1 reports root-mean-square errors (RMSE) and the coefficient of determination (R^2) between experimental data and predictions based on BMFD, $BI_{L,C,R}$, and $BI_{L,R}$.

As illustrated in the upper panel of Figure 2, data of [62–64] showed that ITD thresholds decrease with increasing target tone frequency, where the smallest ITD threshold of about 0.012 ms was found at 1 kHz and steeply increased above about 1200 Hz [64]. These decreasing threshold ITDs represent a more or less constant IPD of about 0.05 rad (~3°). For frequencies above 1 kHz, measured ITD thresholds increase, which is due to a reduced phase-locking ability of the IHCs for higher frequencies. Please note, that the authors of [64] classified their measured ITD discrimination thresholds using data from their *most-sensitive listeners* and *less-sensitive listeners*. The data shown in Figure 2 are based on the average across the two *most-sensitive listeners* (L1, L2) from the original dataset of [64] following the notion that auditory models for normal-hearing listeners typically mimic the limits of human auditory perception. For all three model versions, predicted ITD thresholds are higher than observed in the data, particularly at low frequencies. Here, a nearly constant IPD of about 0.07–0.08 rad (~4°–5°) was predicted, which is higher than the measured IPD. In agreement with the data, predicted ITD thresholds decreased with increasing frequency and showed their lowest threshold of 0.023 µs at about 700 Hz. For frequencies above 900 Hz $BI_{L,R}$ predictions showed increased ITD thresholds, while predictions based on $BI_{L,C,R}$ and BMFD showed increased thresholds up to about 1200 Hz, followed by slightly decreased thresholds up to 1500 Hz. For all three model versions, ITD thresholds slightly decrease (while IPD thresholds slightly increases) for frequencies above 1.5 kHz, and do not show the steep increase observed in [64]. This departure from the data is caused by an identical temporal jitter applied in the model for the target and reference signal per ear channel, resulting in short-time envelope power cues. Such cues were not present or used in listeners. It should be noted

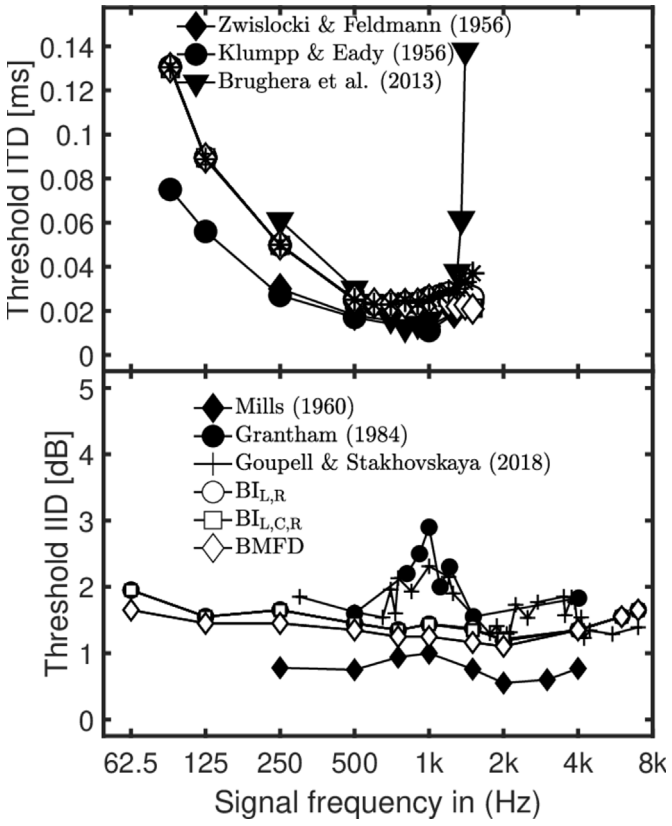

Figure 2. Empirical data (filled symbols) and model predictions (open symbols) for ITD thresholds in ms (upper panel) and IID thresholds in dB (lower panel). The asterisks in the upper panel refer to BMFD-based predictions using short-time power SNRs in combination with long-time envelope power SNRs, instead of short-time power and envelope power SNRs used in the BMFD-based predictions represented by open diamonds.

that predictions based only on short-time power SNRs follow the steep increase of the ITD thresholds. Thus, a combination of short-time power SNRs in combination with long time envelope power SNRs (shown as asterisks in Fig. 2) predicted increasing thresholds for frequencies above 1.2 kHz. The consequences of such combinations of psychoacoustic and speech intelligibility predictions are addressed in Section 5.

The lower panel of Figure 2 shows measured IID thresholds adopted from the studies of [65–67]. In [65] rather similar IID thresholds (average threshold of about 0.8 dB) were measured across frequencies ranging from 250 Hz to 4 kHz, where a maximum of about 1 dB was reached at 1 kHz. In [66] about 1.3 dB higher overall IID thresholds, with substantially increased thresholds around 1 kHz, were measured. Such an increased threshold at 1 kHz for narrow band noises with center frequencies between 300 Hz and 7 kHz is also observed in the data of [67]. Predicted IID thresholds for the three model versions decreased slightly from about 2 dB at 62.5 Hz to about 1.1 dB at 2 kHz, and increased again for higher frequencies. The predicted IID pattern agrees well with the average of the three data sets. Predicted thresholds for $BI_{L,R}$, and $BI_{L,C,R}$ between frequencies from 62.5 Hz to 2 kHz are on average 0.2 dB higher than those from BMFD.

The upper four panels of Figure 3 show measured N_0S_m, $N_\pi S_m$ N_0S_π, $N_\pi S_0$, thresholds adopted from the studies of [3, 4, 68, 69]. All threshold patterns show a V shape, with a minimum at 250 Hz. For the monaural target (S_m), thresholds are lower for N_0S_m than for $N_\pi S_m$, while for the binaural target (S_π or S_0), thresholds are lower for N_0S_π than for $N_\pi S_0$. The resulting threshold differences of $N_\pi S_m - N_0S_m$ and $N_\pi S_0 - N_0S_\pi$ are shown in both lower panels of Figure 3. The largest differences, up to about 9.5 dB, occur for signal frequencies below 500 Hz. $BI_{L,R}$ predictions (open circles) show an overall pattern similar to the data, and accordingly, the predicted $N_\pi S_m - N_0S_m$ and $N_\pi S_0 - N_0S_\pi$ patterns largely agree with the data. For $N_\pi S_m$ and $N_\pi S_0$, both middle panels in Figure 3 show larger deviations between the data and the $BI_{L,C,R}$ and BMFD predictions at 250 Hz and 500 Hz. This deviation is based on the contribution of the BI_C channel that overestimates human performance for the $N_\pi S_m$ and $N_\pi S_0$ conditions. Accordingly, large deviations between data and predictions are observed in the difference patterns in the lower two panels for $BI_{L,C,R}$ and BMFD at 250 Hz.

Measured N_0S_π thresholds as a function of signal duration adopted from [70–73] are shown in Figure 4. For the target signal with a frequency of 500 Hz, for durations up to about 60 ms, thresholds decrease with a slope of about 4.5 dB per duration doubling, while for longer signal durations a slope of about 1.5 dB per duration doubling is observed. For the 4 kHz target signal, the data shows a slope of about 3 dB per duration doubling. For all three model versions, nearly identical thresholds were observed with, on average, higher thresholds than observed in the data. For both signal frequencies, the predicted thresholds decreased with about a 3 dB per doubling of the signal duration, as the signal's energy increases by 3 dB per doubling of the duration. Such an increase in signal duration means that more short-time frames of the model provide an SNR-advantage, which effectively lowers the threshold.

Figure 5 shows masked thresholds measured by [74]. In $N_0N_\pi S_\pi$ and $N_\pi N_0 S_\pi$ conditions, lower thresholds (large BMLD) were measured for target signals (S_π) in the interaurally in-phase masker segments (N_0) than for S_π in interaurally out-of-phase masker segments (N_π). Similarly, for the corresponding "monaural" $N_\pi N_{\pi,\text{-15dB}} S_\pi$ and $N_{\pi,\text{-15dB}} N_\pi S_\pi$ conditions, S_π in attenuated N_π segments resulted in lower thresholds compared to S_π in not attenuated N_π segments. While a gradual release from masking was observed when shifting S_π from the N_π segment into the N_0 segment (upper-left panel), a very steep release from masking was observed for the corresponding "monaural" $N_\pi N_{\pi,\text{-15dB}} S_\pi$ condition (lower-left-panel). A similar behavior was found for the $N_0N_\pi S_\pi$ and the $N_{\pi,\text{-15dB}} N_\pi S_\pi$ conditions. Similar predicted masked thresholds were observed for the three model versions and the predicted steepness of the transition was the same for all four conditions. The predicted BMLD in $N_\pi N_0 S_\pi$ (upper-left panel) and the predicted masking effect in $N_0N_\pi S_\pi$ (upper-right panel) are somewhat smaller than observed in the data. Overall,

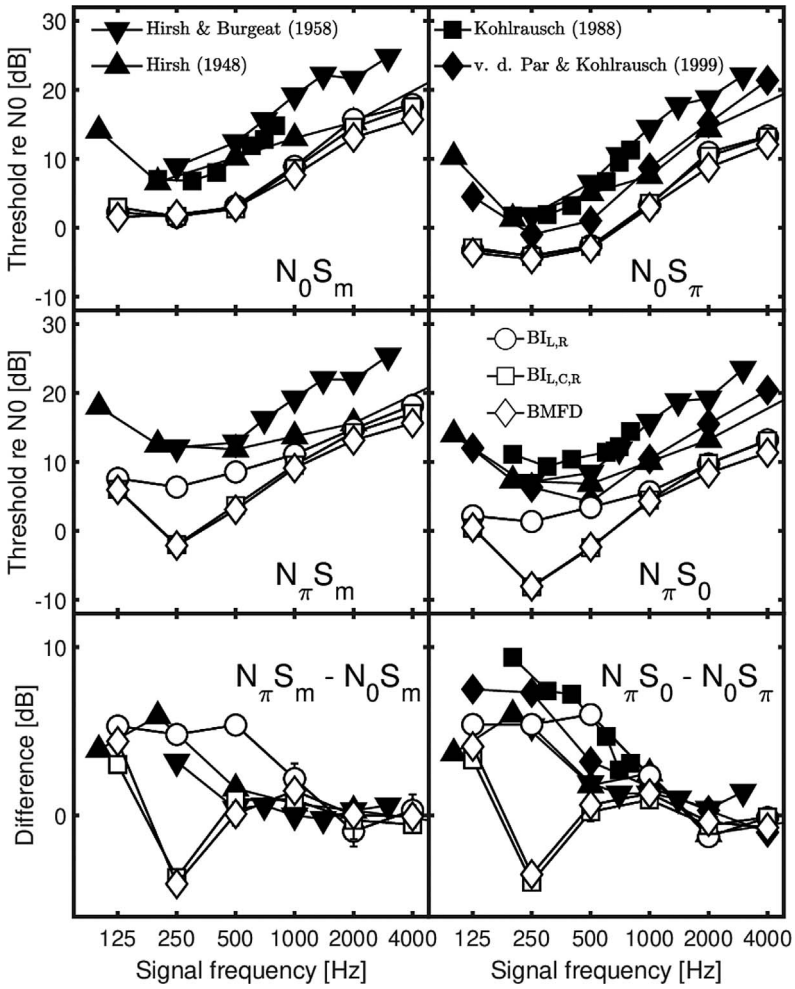

Figure 3. Empirical data (filled symbols) and model predictions (open symbols) for masked thresholds for wideband N_0S_m (upper-left panel), N_0S_π (upper-right panel), $N_\pi S_m$ (middle-left panel), and $N_\pi S_0$ (middle-right panel) conditions as a function of the frequency of the signal. Differences in thresholds between the $N_\pi S_m$ and N_0S_m are shown in the lower-left panel, while the lower-right panel represents differences in threshold between $N_\pi S_0$ and N_0S_π.

the predictions largely agree with the experimental data, which is also indicated by reasonable RMSE and R^2 values of about 2.7 dB and 0.8, respectively.

The upper and lower panel of Figure 6 show measured d's from the time-intensity-trading experiment of subjects S1 and S4 from [75], respectively (see their Fig. 1). For clarity, only these two subjects – with the largest difference in performance – are shown. Likewise, the model predictions for the BI channels and all five channels are split across the two panels for better visibility. Both subjects show that for increasing ITD of 0, 10, 20, 30, and 40 µs, a larger opposing ILD was required for "trading", yielding the lowest sensitivity index d' for discrimination of the trading stimulus from the diotic reference signal. It is obvious that the model based only on the BI channels (upper panel of Fig. 6) can only mimic the general pattern, while there are large differences in the sensitivity and the ILD required for trading as a function of ITD. Moreover, the model with all five BMFD output channels (lower panel of Fig. 7) shows even larger deviations from the data, and fails to predict a clear dependency of ILD on ITD. Overall, the better-performing BI model versions are closer to the performance of subject S4 to that S1. To estimate the best fit of the model, the RMSE and R^2 values reported in Table 1 for Experiment 6 are based only on the data of subject S4. Nevertheless, there are severe deviations.

The three panels in Figure 7 show the spatial unmasking benefit for masker azimuth locations of 0°, 45°, and 90° measured in [76]. Spatial unmasking is calculated as the difference between detection thresholds of the spatially co-located target and masker and the target location indicated at the abscissa. In each of the three panels, the measured spatial unmasking increased as the target moved away from the masker location. The largest separation between target and masker resulted in about 15 dB of unmasking for spatial configurations for which the masker was placed at 0° (upper-most panel) and 45° (middle panel), while the target was either placed at −90° or at +90°. The predicted spatial unmasking pattern of all three model versions largely agreed with the measured data, except for the lower maximum unmasking of about 12 dB. All three model versions predicted similar spatial unmasking for

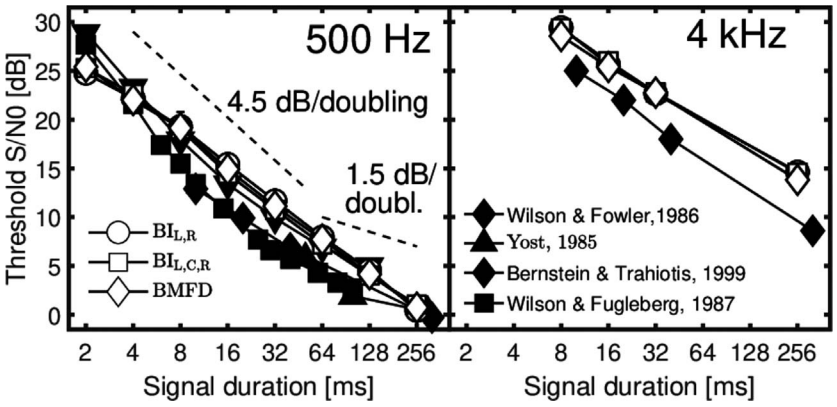

Figure 4. Empirical data (filled symbols) and model predictions (open symbols) for N_0S_π thresholds as a function of the signal duration. Data and predictions are shown for signal frequencies of 500 Hz (left panel) and 4 kHz (right panel).

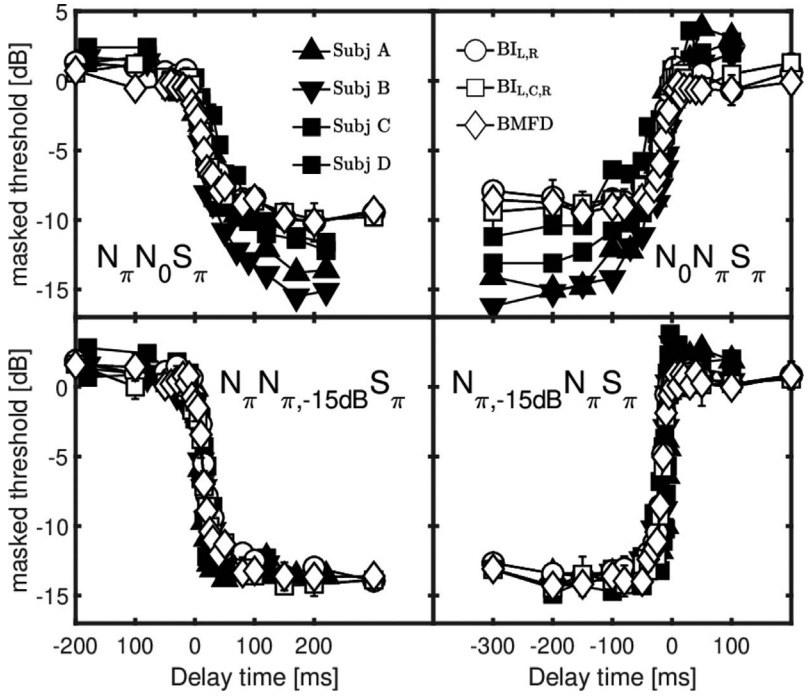

Figure 5. Empirical data (filled symbols) and model predictions (open symbols) for $N_\pi N_0 S_\pi$ (upper-left panel) and $N_0 N_\pi S_\pi$ (upper-right panel) thresholds as a function of the temporal position of the signal center relative to the masker-phase transition. Monaural thresholds for $N_\pi N_{\pi,\text{-15dB}} S_\pi$ and $N_{\pi,\text{-15dB}} N_\pi S_\pi$ are shown in the lower-left and lower-right panels. Filled symbols represent four subjects measured by [74].

maskers at 0° and 45° azimuth location, while for the masker location of 90°, $BI_{L,R}$ predicted on average about 2.4 dB lower spatial unmasking than $BI_{L,C,R}$ and BMFD. It should be noted that the CIPIC HRTF database used here misses azimuth angles in the range from above 80° to below 110°, and thus, ±80° were used instead of ±90° azimuth with, however, very similar expected results.

The lower part of Table 1 summarizes RMSE and R^2 between experimental data and predictions for the three model versions. For most binaural experiments, the three model versions BMFD, $BI_{L,C,R}$, and $BI_{L,R}$ achieve a comparable prediction performance. In Experiment 3 (Frequency and interaural phase relationships in wideband conditions), $BI_{L,R}$ achieved a substantially better performance compared to the other two versions, while in Experiment 7 (Time-intensity trading), $BI_{L,C,R}$ and $BI_{L,R}$ gave more accurate predictions than BMFD. Therefore, BI_L and BI_R are sufficient to explain most of the data of the binaural psychoacoustic experiments used in this study.

Overall, Table 1 showed that the GPSM with binaural BMFD extension, accounts for several monaural and binaural psychoacoustic experiments, while using the $BI_{L,R}$ extension based only on binaural interaction channels gave most accurate predictions.

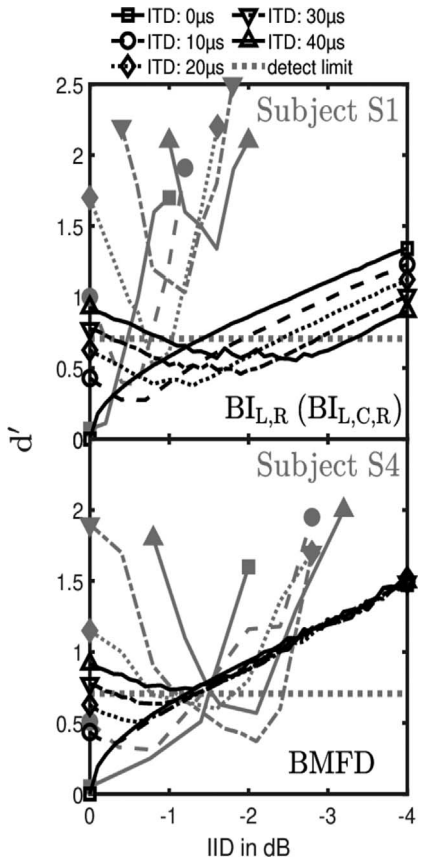

Figure 6. Empirical data (grey lines, filled symbols) and model predictions (black lines, open symbols) for the time-intensity trading experiment of [75] with ITDs of 0, 10, 20, 30, and 40 µs. The ordinate represents d', while the abscissa represents the ILD in dB. Since $BI_{L,R}$ and $BI_{L,C,R}$ predicts nearly identical d', only $BI_{L,R}$ predictions are shown in the upper panel for improved clarity. The lower panel represents predictions from BMFD. The dashed horizontal lines indicate the decision criterion of the models, e.g. differences between test and reference signals resulting in d' values below the criterion are assumed not to be detectable.

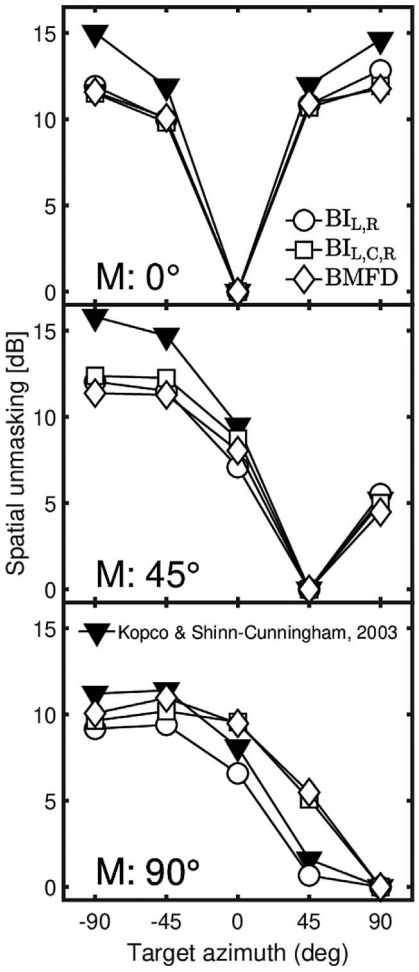

Figure 7. Empirical data (filled symbols) and model predictions (open symbols) for the spatial unmasking experiment of [76] with a 500-Hz pure tone in the presence of a broadband noise masker. The abscissa represents the azimuth angle (in degrees) of the target signal, and the ordinate represents spatial unmasking in dB. Spatial unmasking was calculated as the difference between detection thresholds of the spatially co-located target and masker and the target location indicated on the abscissa. The three panels represent different masker locations of 0°, 45° and 90°.

4 Speech intelligibility evaluation

4.1 Experiment S1: Symmetric maskers

The binaural model extension was tested for the headphone-based binaural (dichotic) speech intelligibility experiments of [2]. In these, speech reception thresholds (SRTs; SNRs at which 50% of target speech was understood), were measured for frontal target speech [German Oldenburger Satztest (OLSA), [78]] in the presence of two co-located or spatially separated maskers with different spectro-temporal characteristics, but identical long-term spectrum.

Four stationary speech-shaped noise (SSN) based maskers and two speech maskers were used in [2]: The SAM masker was obtained by applying an 8-Hz sinusoidal amplitude modulation with 100% modulation depth to the SSN masker, yielding regular temporal modulations coherent across all auditory channels (co-modulation). For the BB masker, the SSN was multiplied with the Hilbert envelope of a broadband speech signal (ten randomly selected OLSA sentences), introducing temporal gaps that reflect the modulations of intact speech. Temporal irregularities of the speech envelope were coherent across all auditory channels. For the across-frequency shifted (AFS) masker, the speech envelope was randomly shifted in eight groups (each consisting of four adjacent auditory frequency channels), resulting in incoherent AMs across auditory channels. As speech maskers, a male version of the International Speech Test Signal (ISTS; [79]), composed of intact continuous speech uttered by six different female talkers in different languages, was used as "nonsense" speech. A single talker (ST) masker used randomly selected parts of ten concatenated OLSA sentences spoken by a male speaker differing from the target OLSA material.

Two spatial target-masker configurations were measured for each masker: In the co-located configuration, target and masker sources were placed in front of the receiver (0°). In the spatially separated configuration, the masker positions were changed to both sides at ±60° relative to the frontal direction. Speech intelligibility improvements, depending on the spatial separation between target and masker, are expressed as SRM. A single masker had a level of 65 dB SPL, and accordingly the presentation of two statistically independent masker sequences resulting in an overall masker level of 68 dB SPL. A detailed description of the experiment can be found in [2].

4.2 Experiment S2: Asymmetric masker

The binaural model extension was further tested for the headphone-based binaural (dichotic) speech intelligibility experiments of [80]. The SRTs were measured for a frontal target speech (OLSA) as a function of the masker azimuth angle (−180°, −145°, −100°, −45°, 0°, 45°, 80°, 125°, 180°). HRTFs from the CIPIC database [77] were used for spatialization of the stimuli. The stationary noise masker had a long-term spectrum similar to that of the target speech. The masker was presented at a level of 65 dB SPL.

4.2.1 Results and discussion

4.2.1.1 Experiment S1

Measured and predicted SRTs shown in Figure 8 are represented by gray and black symbols, respectively. Co-located maskers are indicated by filled symbols, and separated maskers by open symbols. Predicted SRTs are averaged over 5 repeated simulations each based on 20 OLSA sentences. Each model version was calibrated to the speech material as proposed in [25] by adjusting the parameters k to match the co-located SSN data. The values of the other parameters q, m, σ_s shown in Table 2 are identical to those used for OLSA predictions in [26].

For noise maskers (SSN, SAM, AFS, and BB) presented co-located with target speech, the highest SRTs were measured for stationary SSN and fluctuating AFS maskers, and listeners only took advantage of listening to dips when speech was presented in fluctuating SAM and BB maskers. The highest SRT was measured when speech was masked by the single talker (ST), resulting in about 5.5 dB higher thresholds compared to the SSN masker. A spatial separation of target speech and maskers resulted in SRM values ranging between 4.3 and 13.5 dB. The smallest SRM of about 4.3 dB was observed for the SSN maskers, while the largest SRM values of 10.1 and 13.5 dB were observed for ISTS and ST maskers.

BMFD predictions (filled circles in the upper panel of Fig. 8) for the co-located BB and AFS maskers largely agree with data, while BMFD predictions for co-located SAM maskers are about 3 dB higher than data. For BMFD, the largest differences between predicted and measured SRTs of up to 13 dB can be observed for co-located ISTS and ST maskers. The similarity between the ST masker and the target sentence makes it difficult for the listener to separate the target from the interfering speech masker (informational masking, e.g. [81]), which results in high SRTs and high variability across listeners. In contrast to human listeners, the current model, as other intrusive SI models, has a priori knowledge about the target speech and the masker signals, and is only limited by aspects of amplitude modulation and energetic masking (and not informational masking), yielding substantially lower thresholds for the speech-like maskers. For the spatially separated conditions (open circles in the upper panel of Fig. 8), BMFD predictions fit well for SSN and AFS, while they underestimate human performance for SAM and BB, and again overestimate human performance for the speech-like maskers ISTS and ST, as can be expected (see above). Regarding the SRM (lower panel of Fig. 8), BMFD predictions show a good agreement with the data for SSN, SAM (about 2 dB reduced SRM) and AFS. For BB, the predicted SRM is about 3 dB lower and for ISTS and ST up to 5 dB lower than the measured SRM. For ISTS and ST, these differences are partly caused by larger discrepancies between predicted and measured SRTs in co-located conditions.

In a further step, each of the five BMFD outputs was analyzed to identify the channel contributing most. Here, $\mathrm{BI_C}$ with highest sensitivity to the hemispheric midline, represented as downward-pointing triangles in Figure 8, gave the largest contribution to SI predictions; this is clearly shown by very similar predictions of BMFD and $\mathrm{BI_C}$ in Figure 8. This agrees well with the findings of [2], where a simple binaural summation of the left and right ear signals (prior to the model) showed similar results for predictions using the binaural speech intelligibility model (BSIM; [12]). For this summed diotic input, BSIM effectively reduces to similar processing as suggested in the monaural ESII [11] model, using a short-time assessment of power-based SNRs. In contrast, the current $\mathrm{BI_C}$ predictions are based on both short-time envelope power and power SNRs. It should be noted that although predictions of both of the power pathways of BMFD and BSIM are based on power SNRs, substantial differences exist, such as the SNR combination across time frames and auditory channels, which could have an influence on predicted SRTs.

Analyzing the contribution of envelope power and power SNRs revealed that AM cues are mostly dominant. Predictions only based on envelope power SNRs provided by the center binaural interaction channel are denoted as $\mathrm{BI_C^{AC}}$ and shown as diamonds in Figure 8. With the exception of the BB masker condition, $\mathrm{BI_C^{AC}}$-based predictions already explain most of the SRM observed in the data.

For spatially co-located conditions, predictions only based on power SNRs provided by the center binaural interaction channel $\mathrm{BI_C^{DC}}$ (not shown in Fig. 8 for clarity) are about 12 dB lower than the measured data, indicating that the benefit from dip listening based on short-time power SNRs is substantially larger than that observed in the listeners. The SRM predicted by $\mathrm{BI_C^{DC}}$ (RMSE: 5 dB) for fluctuating maskers follows a similar pattern as observed for $\mathrm{BI_C^{AC}}$ (RMSE: 2.8 dB), however, with larger deviations

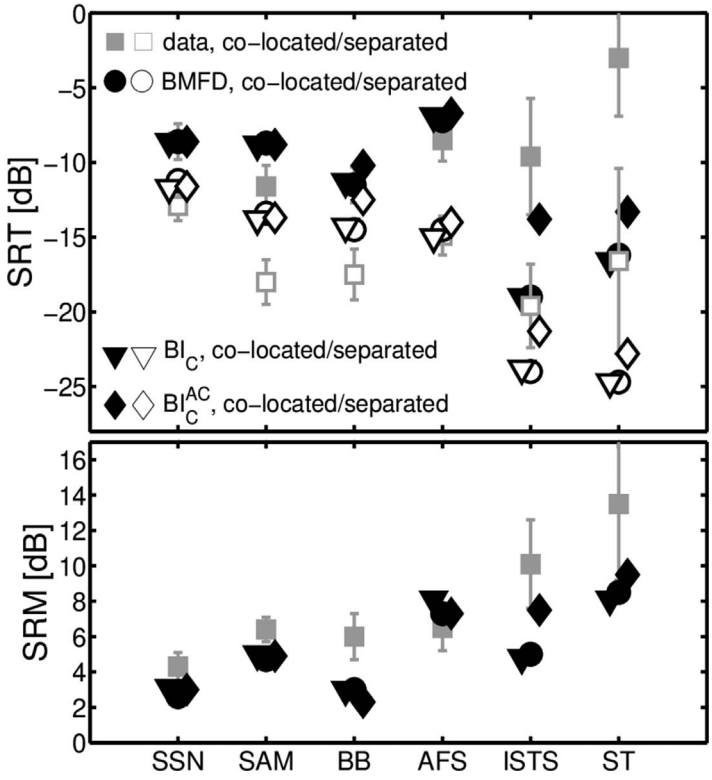

Figure 8. The upper panel shows SRT results of Experiment S1, while the lower panel shows the respective SRM. Data is represented by squares, while predictions are given by circles, triangles, and diamonds, respectively. The spatially co-located and separated masker conditions are indicated by filled and open symbols, respectively.

Table 2. Parameter settings of the three model versions to match the co-located SSN data. In Experiment S1, the $\overline{k}$ value results from averaging the individual k values from five repeated simulations, each based on 20 sentences, while the $\overline{k}$ value applied to Experiment S2 was determined from 20 sentences.

	$\overline{k}$	q	m	σ_s
Experiment S1				
BMFD	0.6	0.5	50	0.6
BI_C	0.72	0.5	50	0.6
BI_C^{AC}	0.72	0.5	50	0.6
Experiment S2				
BMFD	0.265	1.2	8000	0.6
BI_C	0.42	1.2	8000	0.6
BI_C^{AC}	0.42	1.2	8000	0.6

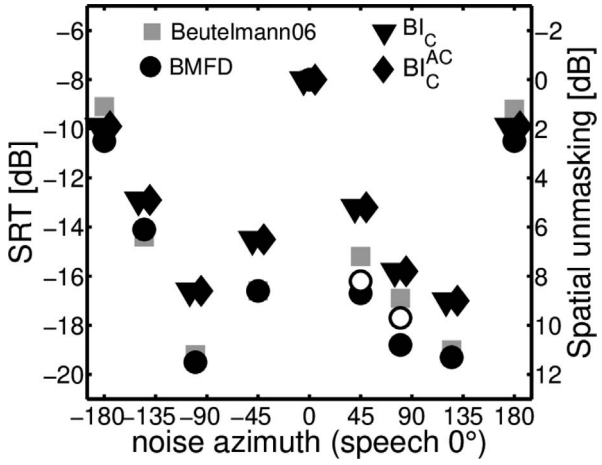

Figure 9. Measured and predicted SRTs from Experiment S2 as a function of noise masker azimuth angles are represented by grey and black symbols, respectively. For comparison, BMFD-based predictions for spatial unmasking of a 500-Hz pure tone in the presence of a broadband noise masker from experiment 7 of the psychoacoustic evaluation are represented as black open circles.

to the measured SRM as indicated by RMSE values. Such deviations can mostly be explained by the lower thresholds in the co-located conditions based on dip listening, effectively reducing the predicted SRM.

Although BI_C does not play an important role for the binaural psychoacoustic experiments in this study, it can successfully account for a large part of the SRM in the speech intelligibility experiments.

4.2.1.2 Experiment S2

In Figure 9, the same three model versions BMFD, BI_C and BI_C^{AC} as in Experiment S1 were used and the same symbols and style is used for plotting data and predictions as in Figure 8. Predicted SRTs are based on 20 OLSA sentences, which was sufficient to reach stable results with the stationary maskers used here. The same calibration

procedure as described in Experiment S1 was applied to adjust predicted SRTs to match the co-located (0° azimuth) data. The calibration parameters k, q, m, σ_s of each model version are shown in Table 2.

The abscissa of Figure 9 represents the noise azimuth angle, while the left and right ordinates show SRTs and spatial unmasking, respectively. Spatial unmasking was calculated as the difference between the configurations in which the target and masker are co-located at 0° azimuth and the masker location indicated on the abscissa. The highest measured SRT of −8 dB is observed for the co-located target and masker at 0° azimuth. Speech intelligibility improved when masker and target were spatially separated, with the lowest SRTs of −19.2 dB and −19 dB for azimuth angles of −100° and 125°. The noise masker presented at ±180° resulted in slightly lower SRTs of about −9 dB than those observed for the azimuth angle of 0°.

Predicted SRTs of the three model versions for the azimuth angle of 0° were adjusted to perfectly match the measured SRT. The BMFD (filled circles) agreed largely with measured SRTs, as indicated by RMSE and R^2 values of 1 dB and 0.97. The largest deviation between BMFD predictions and data was 1.9 dB and was observed for the azimuth angle of 80°. Figure 8 clearly shows that predictions of $\mathrm{BI_C}$ (downward-pointing triangles) and $\mathrm{BI_C^{AC}}$ (diamonds), both based only on the center channel of the BMFD, are less accurate than predictions based on all five BMFD channels, resulting in a larger RMSE value of 1.6 dB for $\mathrm{BI_C}$ and $\mathrm{BI_C^{AC}}$ (and the same R^2 of 0.97). Since $\mathrm{BI_C}$ and $\mathrm{BI_C^{AC}}$ predicted identical SRTs, it can be expected that envelope power SNRs were dominant in the center binaural interaction channel. Predictions based on $\mathrm{BI_C^{DC}}$ (not shown in Fig. 9 for clarity) reached a lower RMSE of 0.6 dB and a slightly higher R^2 of 0.98. Taken together, for both speech intelligibility experiments, the center binaural interaction channel appears to capture the most relevant cues required for predictions of spatial unmasking.

For direct comparison of spatial unmasking (or SRM) in psychoacoustics and speech intelligibility, BMFD predictions for Experiment 7 (spatial unmasking of tone in noise) are shown in Figure 9 as black open circles for the comparable spatial arrangement of target and masker. The BMFD predictions show a slightly larger spatial unmasking for speech intelligibility than for tone in noise; this is likely related to the spectral composition of the target and masker signals.

5 General discussion

The proposed model explores the ability of a highly simplified, fixed (non-adaptive) binaural interaction stage to account for key aspects of binaural psychoacoustics and speech intelligibility using spatially separated interferers. The investigated 5-channel BMFD stage was incorporated into an existing monaural model using power and envelope power SNR cues. The suggested model maintains the ability of the former monaural approach to account for monaural psychoacoustic key phenomena. Binaural psychoacoustics was well covered, except for larger discrepancies in time-intensity trading. For speech intelligibility, the key aspects where also predicted, with larger discrepancies for speech-like interferers. Here, aspects of informational masking that are generally not covered by signal-processing models play a role, as has been previously shown in other speech intelligibility models.

It is conceivable that the current simplified approach might not reach the performance of other "specialist", dedicated monaural and binaural models for psychoacoustics and speech intelligibility for all of the experiments considered here. Hence, a comparison with other state-of-the-art comparison models is given in Section 5.3. However, as the current approach is based on former work [9, 16–18, 26], it is assumed to generalize well to other unknown data, which makes the model interesting in the context of instrumental (spatial) audio quality predictions. Further, the straightforward processing in the BMFD stage is generally advantageous for real-time applications, e.g., for control of signal processing algorithms in hearing supportive devices or as a hearing-aid processing stage itself. Finally, the current approach demonstrated that the physiologically motivated hemispheric interaural interaction in mammals (e.g., [42, 43]), as realized here in the two binaural interaction channels $\mathrm{BI_L}$ and $\mathrm{BI_R}$, predicts the results of a wide variety of perception experiments.

5.1 Contribution of binaural interaction and better-ear channels

For the binaural psychoacoustic experiments used in this study, the two $\mathrm{BI_L}$ and $\mathrm{BI_R}$ channels appear sufficient to account for the data. $\mathrm{BI_C}$ has only a negligible effect on the predicted data as also indicated by very similar RMSE and R^2 values shown in Table 1 for the model versions including $\mathrm{BI_c}$ ($\mathrm{BI_{L,C,R}}$) and excluding $\mathrm{BI_c}$ ($\mathrm{BI_{L,R}}$), except for the binaural Experiment 3 on interaural phase effects under wideband conditions. There, the predicted thresholds based on $\mathrm{BI_C}$ were significantly better than human performance in $\mathrm{N_\pi S_\mathit{m}}$ and $\mathrm{N_\pi S_0}$ conditions (see middle panels in Fig. 7). Accordingly the predicted difference patterns for $\mathrm{N_\pi S_\mathit{m}} - \mathrm{N_0 S_\mathit{m}}$ and $\mathrm{N_\pi S_0} - \mathrm{N_0 S_\pi}$ show a large deviation from measured data of up to 10 dB at 250 Hz. In general, both better-ear channels $\mathrm{BE_L}$ and $\mathrm{BE_R}$ make no substantial contribution to the binaural psychoacoustic experiments.

For speech intelligibility, the importance of the five BMFD channels is different, and $\mathrm{BI_C}$ accounted for a large part of the data (see Figs. 8 and 9). Under the current SI conditions, a frontal target was presented in either co-located or spatially separated maskers. In view of the psychoacoustic conditions, the co-located condition can be regarded as $\mathrm{N_0 S_0}$, while the separated condition can be considered as $\mathrm{S_0}$ plus noise with a frequency-dependent interaural phase difference. Accordingly, the correlation coefficient for the co-located maskers in Experiment S1 of the speech intelligibility evaluation is about 0.99, while it is strongly reduced for spatially separated masker configurations, indicated by values ranging between −0.03 and −0.11. In the

separated conditions, the BI_C channel amplifies the coherent frontal target speaker (S_0), while spatially separated maskers with IPDs $\neq 0$ are incoherently added or might be partially cancelled.

The role of the five BMFD channels for speech intelligibility can be further assessed by analyzing the distribution of envelope power and power SNRs contributing the most across frequency and over the five binaural processing channels (not shown): For all spatially separated conditions, BI_C shows the greatest contribution (in agreement with the additive approach in [2]). For the co-located conditions, no large differences in the contributions of all channels were observed. BI_L and BI_R contribute slightly more, resulting in about 1 dB lower SRTs for BI_L and BI_R than for the other three channels. Regarding the SRM, in line with the psychoacoustic experiments, the two better-ear channels $BE_{L,R}$ contributed less, resulting in consistently lower predicted SRM than the three binaural interaction channels. Although BI_L and BI_R might be less important in the current spatial configuration with a frontal target, where BI_C was most beneficial, they can be assumed to be more important when the target is placed to either side of the head. Moreover, both BI_L and BI_R are also assumed to be important for the evaluation of spatial audio quality, as inaccuracies in the audio rendering of sound reproduction systems may alter the spatial properties, e.g., location, or apparent source width, of an auditory object.

As shown in Figure 2, measured ITD thresholds steeply increased above 1.2 kHz, which was not covered by the predictions. In the model, short-time windows of higher modulation filters, applied in the envelope power pathway, captured phase differences between jittered target and reference signals. Such undesired sensitivity, which is related to the use of the identical jitter for the target and reference signal, can be avoided by using long-term envelope power SNRs, while power SNRs are still analyzed in short-time windows, as also used earlier in [26]. With this modification, the prediction performance for the current monaural and binaural psychoacoustic experiments is comparable (not shown) to that of the suggested model, in agreement with the ability of the similar monaural approach in [26] to account for a large variety of monaural psychoacoustic and speech intelligibility experiments. Conversely, in [16], an analysis of envelope-power SNRs in short-time frames was required for audio quality predictions to sufficiently capture (monaural) cues related to nonlinear distortions.

5.2 Comparison of the binaural stage to other literature models

The outputs of the suggested BMFD stage can be considered as a simplification of the delay-gain matrix and the left/right channel in Breebaart et al. [8], or as specific fixed states of the EC model [28]. Given the conceptual similarity of these two models and the widespread use of the EC approach as binaural processing stage in numerous auditory models (e.g., [12, 36, 37]), the current results might be interesting for other literature models.

The three $BI_{L,C,R}$ channels are comparable to elements in the matrix of the Breebaart model with a specific delay and gain in the respective auditory frequency channel. The $BE_{L,R}$ channels are directly comparable to the individual ear signals passed to the detector stage in the Breebaart model, in parallel to outputs of the delay-gain matrix. In the Breebaart model, internal delays of up to 5 ms (π phase shift at 100 Hz) and a gain difference up to 10 dB between both ears are realized. These parameters broadly cover the current choice in the $BI_{L,R}$ channels. Thus the difference between the suggested model and the Breebaart model is the reduction of degrees of freedom in the binaural interaction stage to parameters that are directly motivated by the physiology of mammals (e.g., [52]).

Similarly, each of the five BMFD outputs represents a specific state of the EC approach. Again the difference is that the EC stage can realize arbitrary delays and gains (for the equalization of the noise in the left and right channel) to optimally cancel the noise at the output, while $BI_{L,C,R}$ represents a fixed, potentially suboptimal, realization of the EC process. Alternatively, the left or right ear input can be directly routed to the EC output, comparable to the better-ear channels $BE_{L,R}$ in the current BMFD stage.

Based on the five BMFD outputs, envelope power and power SNRs are calculated and combined to give an overall d'. In contrast to other models like the B-sEPSM [37] and BSIM [12] where SI prediction are either based on envelope power SNRs or power SNRs, the current approach combines both types of SNRs. As shown in Figure 7, envelope power SNRs capture most of the measured SRM. It should be noted that predictions based only on power SNRs also agree with the measured SRM pattern. For fluctuating maskers, however, SRTs predicted by power SNRs are often substantially lower than measured SRTs, which was also observed in [26]. As suggested in [26], a forward-masking function or SNR limitation could be applied to counteract that effect.

The envelope power $SNR_{envW,i}(p,n)$ and $SNR_{DC,j}(p)$ are combined across the five BMFD outputs by taking the largest value for each time frame within each auditory and modulation channel. Such a procedure allows fast switching between the five BMFD outputs, in line with findings of [82]. However, psychophysical studies (e.g., [74], also considered here, see Fig 5) and a recent SI study [83] implied some limitations of the binaural auditory system in following temporal changes of ITDs (or IPDs). This is often referred to as binaural sluggishness, and suggests a binaural temporal window with a time constant of up to about 200 ms. Such binaural sluggishness was not included in the current model. This enables a fast switching between the monaural (better-ear) and binaural interaction channels, resulting in the same slope of the transition as shown in Figure 5, representing data of [74]. Thus, for some conditions, prediction performance could be improved if aspects of (task dependent) binaural sluggishness were integrated into the current model by using a temporal window as suggested in [8].

Table 3. Root-mean-square errors (RMSE) and coefficients of determination (R^2; squared cross-correlation coefficient) between data and model predictions for the Breebaart model [8].

Experiments	Breebaart	
	RMSE	R^2
1. ITD discrimination	0.035 ms	0.23
2. IID discrimination	0.6 dB	0.015
3. Frequency and interaural phase relationships in wideband conditions	2.8 dB	0.87
4. N_0S_π depending on signal duration	2.2 dB	0.94
5. Temporal phase transition	3.6 dB	0.77

5.3 Performance comparison to literature models

An important question is how the prediction performance of the current simplified binaural model compares to existing state-of-the-art approaches using more sophisticated binaural stages.

Regarding monaural psychoacoustics, the monaural GPSM [26] and the current model with binaural extension achieved a comparable performance (see Tab. 1). In turn, the monaural GPSM and the widely used perception model of [5, 6], with identical model parameters for all tested monaural experiments, showed a similar performance [9, 26], suggesting comparable results for the current binaural extension.

Regarding binaural psychoacoustics, the Breebaart model [8], as a binaural extension of [5, 6], has been chosen as the comparison model. In Table 3, the prediction performance of the Breebaart model is given by RMSE and R^2 as estimated from their predictions for the five binaural experiments taken from original publications [38, 39]. Since the model was not available to us, no predictions were made for experiments 6 (Time-intensity trading) and 7 (Spatial unmasking). Comparing Table 1 and Table 3, the Breebaart model and the BMFD-based models achieved a comparable performance in predicting the general data pattern represented by R^2, while the Breebaart model showed overall smaller RMSE values. Particularly in Experiment 3 (Frequency and interaural phase relationships in wideband conditions), the Breebaart model had a 6.8 dB lower RMSE than the BMFD model. Low R^2 values in Tables 1 and 2 for the experiments 1 and 2 indicate that both the BMFD-based models and the Breebaart model only provided poor predictions for ITD and IID discrimination. For experiment 1, this is because none of these models sufficiently predict the steeply increasing ITD thresholds at higher frequencies as shown in the data of [64] in Figure 2. BMFD-based predictions founded only on power SNRs follow the increase of ITD thresholds for frequencies above 1 kHz better, as indicated by R^2 of 0.15. As shown in Figure 3, measured IID thresholds show a peak at around 1 kHz, which is neither predicted by the BMFD-based models nor by the Breebaart model, and accordingly resulted in low R^2 values. Nevertheless, low RMSE values of 0.4 dB and 0.6 dB between measured and predicted IID thresholds from the BMFD-based models and the Breebaart model indicate reasonably small absolute errors between the measured and predicted thresholds. It should be noted that the Breebaart model has been tested for a large variety of other binaural psychoacoustic experiments that were not considered in the current study. Thus, for a more conclusive picture about the prediction performance of both models, the suggested BMFD approach has to be tested with other experiments. Further, it should be mentioned that the Breebaart model used an experiment-dependent template, which is an advantage, and is expected to improve prediction accuracy compared to the suggested BMFD approach that used identical parameter settings for all psychoacoustic experiments.

For comparison of SI prediction performance, the (short-time) BSIM [12], which combines the ESII [11] with the EC-mechanism, was used as an established model. For the first SI experiment (see Fig. 8), BSIM and the current model obtained a similar prediction performance for SRTs ($RMSE_{BSIM}$: 6.6 dB, $RMSE_{BMFD}$: 5.7 dB) and SRM ($RMSE_{BSIM}$: 3.1 dB, $RMSE_{BMFD}$: 3.3 dB), while the most accurate predictions for the current model were obtained by using only envelope power SNRs from the center channel BI_C^{AC} ($RMSE_{SRTs}$: 4.3 dB, $RMSE_{SRM}$: 2.6 dB). In the second SI experiment (see Fig. 9), BSIM predictions (RMSE: 0.7 dB; R^2: 0.98) agreed somewhat better with the data than did BMFD predictions (RMSE: 1.0 dB, R^2: 0.97). Predictions of the current model based on power SNRs (as also used in BSIM) and using only the center binaural interaction channel (BI_C^{DC}) achieved a similar performance (RMSE: 0.6 dB; R^2: 0.98) to BSIM predictions. Thus, despite its strongly simplified binaural processing stages, in the two SI experiments the suggested BMFD approach obtained a performance comparable to BSIM using the signal-adaptive binaural EC processing. For spatial configurations in which the target is in front of the listeners, as in the SI experiments considered in the current study, the current results suggest that the most relevant spatial cues are captured by the center binaural interaction channel, which would allow a further simplification of binaural processing.

5.4 Model limitations and simplification of physiological processes

The peripheral processing component of the suggested modeling approach does not incorporate nonlinear cochlear processing as observed in normal-hearing listeners and, accordingly, it cannot account for level-dependent effects, e.g., in the upward spread of masking [19] and in forward masking [20]. Despite such limitations, several models for psychoacoustics [5, 6, 8, 21, 31], speech intelligibility [11–14], and audio quality [10, 15–17] successfully used linear cochlear processing for the prediction of normal-hearing

data, often in combination with nonlinearities in feature calculation or in the decision stage, justifying the current linear approach for a simplified and computationally efficient auditory model.

The L-R and R-L processing after delay and amplification in the current $BI_{L,R}$ channels represents a strongly simplified realization of the hemispheric processing suggested in more detailed models (e.g., [41, 84]) that were based on (simulated) neuronal responses. A key feature of these approaches is the characteristic (hemispheric) net neural activation as a function of ITD for high frequencies in the lateral superior olive (LSO) and for low frequencies in the medial superior olive (MSO), see, e.g., the bottom row in Figure 5 of [41].

The (hemispheric) net neural activation is only partly modeled with the current subtraction process of the half-wave rectified continuous time signal, as illustrated in Figure 10, and is reminiscent of that observed in the LSO (first two rows in Figure 5 of [41]). The left panel of Figure 10 shows the linear response of BI_L (red lines) and BI_R (black lines), normalized to the response at 0° IPD, as a function of the IPD (negative sign indicates left ear leading, no ILD) for τ (delay) of $\pi/4$ and α of 3. The strongest contralateral inhibition occurs when the contralateral ear is leading with an IPD of τ. The least inhibition occurs when the ipsilateral ear is leading with an IPD of π–τ, resulting in internal phase differences of π between the excitatory and inhibitory channels. The current τ value of $\pi/4$ provides a sufficiently steep slope around zero IPD to ensure a sufficient sensitivity for small interaural phase differences, and is consistent with physiological findings. Smaller values would further increase IPD sensitivity and would improve predictions for data of the ITD experiment shown in Figure 2. The α factor of 3 was selected empirically and leads to a complete inhibition by a contralaterally leading ear with a level that is up to 10 dB lower. Larger values would widen the troughs in the response pattern in the left panel of Figure 10, while smaller values would result in narrower troughs. Additional simulations for α values ranging between 3 and 5 resulted in similar prediction performance (not shown). The current α agrees well with the range of interaural gain differences applied in the Breebaart model. The right panel of Figure 10 represents the linear response as a function of the ILD (negative sign indicates that the level in the right ear was higher than in the left ear, no IPD). The response of the ipsilateral ear increases for higher levels in the ipsilateral than in the contralateral ear, while inhibition occurs for higher levels in the contralateral ear.

In more detailed neural models (e.g., [41, 42]), the hypothesis of timed inhibition is that the contralateral inhibitory post-synaptic potential (PSP) precedes the contralateral excitatory PSP for low-frequency processing in the MSO, resulting in a delay of the contralaterally evoked net excitation and the observed hemispheric excitation as a function of ITD. The delayed excitatory interaction, as well as the temporal smearing of excitatory and inhibitory effects represented in the PSPs, are not covered by the current (over) simplified model. Moreover, processing in the

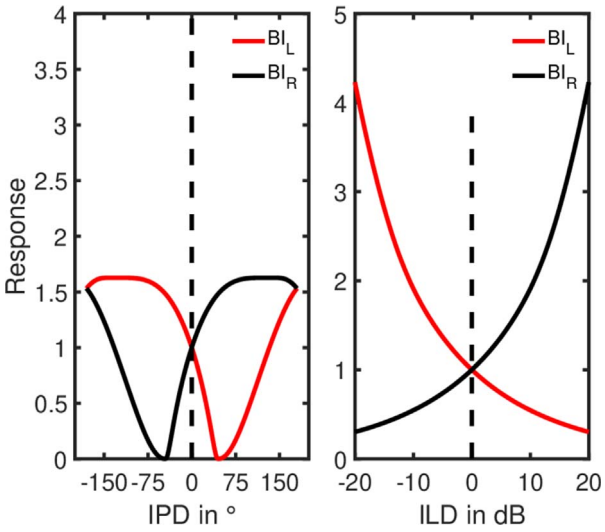

Figure 10. Response of the BI_L and BI_R channels as a function of IPD (left panel) and ILD (right panel) for a 500 Hz pure tone. Negative IPDs indicate left ear leading, while negative ILDs indicate that the level in the left ear was higher than in the right ear. Response shown in both panels are based on the same τ and α values of $\pi/4$ and 3 as they were used for all simulations in this study. Note that for clarity, amplitude and phase jitter were turned off.

LSO and MSO for low and high frequencies, respectively, is observed in the physiology. Conversely, the current model only uses subtraction of the waveforms, disregarding details of PSP simulation, resembling (envelope) ITD processing assumed in the LSO for high frequencies (see center panel of Figure 5 in [41]). This inhibitory processing is used for all frequencies, involving interaural temporal fine structure (TFS) differences at low frequencies and temporal envelope differences at high frequencies. An improvement of the current model can be expected if it incorporated both excitatory and inhibitory effects more faithfully, at however, the cost of simplicity.

To compare inhibitory vs. excitatory interaction in the context of the current model, we replaced the current subtractive (inhibitory) processing by an additive (excitatory) processing. Such additive processing is closely linked to binaural models that are based on the concept of a neurally-based cross-correlation (e.g., [32]) or coincidence detection (e.g., [8]), that mimics the behavior of excitatory-excitatory cells in the MSO (e.g., [43]). Physiological findings in mammals, however, imply that the MSO receives both excitatory and inhibitory inputs (e.g., [43]) as was applied in more comprehensive models (e.g., [64, 85]). The additive processing results in an overall similar prediction performance for the psychoacoustic experiments (not shown). However, large τ values above about $3\pi/4$ had to be used to ensure sufficiently large response differences between stimuli with and without interaural phase shifts. Although the additive processing also explained most of data from the binaural psychoacoustic experiments used in this study, the SRM predictions in SI experiments were often substantially lower than observed in data (not shown).

Accordingly, the RMSE between predicted and measured SRM was higher for the additive processing (RMSE of 5.5 dB) than for the current subtractive processing (RMSE of 3.3 dB).

5.5 Relation to binaural signal processing algorithms

The five outputs of the suggested fixed BMFD stage can be translated to binaural signal processing, and is potentially applicable in hearing supportive devices. The difference between the model stage and audio-signal processing is that the model operates on a half-wave rectified internal representation, whereas audio-signal processing operates on the input waveform at the ears. This difference is important for the binaural interaction channels in which the ear signals are combined after nonlinear processing in the model. As outlined in the introduction, the processing of BI_C was designed to resemble the effect of summation of the waveform in the ears. For BI_L and BI_R, the subtraction of the unipolar (half-wave rectified) signals is followed by another half-wave rectification (setting all negative values to zero), which makes the result more comparable to that of a subtraction of the waveforms. Thus, as a signal-processing algorithm, BI_C represents a (spatially broadly tuned) fixed broadside beamformer (tuning to front and back). Taking the phase delays and subtraction into account, BI_L and BI_R conceptually represent fixed (non-adaptive) first-order differential microphone beamformers with a (frequency-dependent) steering vector. Finally, taking the head shadow effect into account, BE_L and BE_R can be interpreted as beamformers pointing to the left and the right. Thus, the findings with the current modeling may suggest that the auditory system selects the favorable outputs of a discrete number of beamformers (in our model five) in the time-frequency frames, depending on the task and spatial configuration of the input.

In comparison to the adaptive EC model, the current approach cannot optimize parameters to specifically cancel certain signal parts (or directions), as is possible in the adaptive differential microphone. Further simplifying the current selection of the optimal BMFD channel in time-frequency frames by the selection of a single broadband channel, the BMFD might be applicable in hearing-aid processing as five spatially broadly tuned binaural beamformers from which the optimal output is selected, e.g., based on the direction of arrival of the intended target. As is the case with other beamformers (e.g., [86, 87]), advantages in complex multitalker environments can be achieved by increasing the signal-to-noise ratio for a (known) target direction. However, depending on the algorithm used, binaural cues might not be preserved. To partly preserve binaural cues at the BMFD outputs, parts of the unprocessed signal can be added, further reducing directivity. Such simplistic "mild" beamformers might also be suited in ecologically valid situations involving head movements, in which the additional benefit of more elaborated processing might be limited (e.g., [88]). As indicated by the current speech-intelligibility results for a frontal (speech) target and symmetrically or asymmetrically spatially separated maskers, humans appear to just use a simple broadside binaural beamformer (BI_C).

6 Summary and conclusions

The main goal of this study was to examine how well a modeling approach with strongly simplified assumptions about fixed (non-adaptive) binaural interaction processing can predict data from both binaural psychoacoustic- and speech-intelligibility experiments. For this, the generalized power spectrum model [26] was extended by a five-channel binaural matrix feature decoder, comprising two better-ear and three binaural interaction channels, to account for monaural and binaural aspects in psychoacoustic and speech intelligibility experiments. The binaural processing comprises the left (L) and right (R) better-ear channels, the L + R channel (BI_C) and L-R (BI_L) and R-L (BI_R) channels, incorporating a fixed phase delay ($\pi/4$) and amplification factor (3). The model was tested in a monaural and binaural "benchmark" of overall 13 psychoacoustic experiments and 2 speech intelligibility experiments. The following conclusions can be drawn:

- The suggested binaural model accounts for several key temporal and spectral aspects in classical binaural experiments and also explains a large amount of spatial release from masking in speech-intelligibility experiments. The model maintains the predictive power of the earlier monaural approach for monaural psychoacoustics.
- For predicting the psychoacoustic experiments of this study, the L–R and R–L binaural interaction channels, physiologically motivated by hemispheric processing, were most important, as the target signal often contained an interaural phase shift (S_π). The L + R "midline" channel played no important role.
- For predicting the speech-intelligibility experiments, with a frontal target and symmetrically spatially separated maskers (somewhat similar to a S_0 plus noise with frequency-dependent interaural phase difference condition in psychoacoustics), the L + R channel was most important to account for SRT and the spatial release from masking.
- Overall, the results indicate that human performance in binaural tasks might be based on a smart selection of spectro-temporal segments at the output of only a few, fixed binaural interaction channels.
- For the current set of psychoacoustic and speech intelligibility experiments, the suggested modeling approach with a strongly simplified binaural processing achieved a prediction performance that is comparable to state-of-the-art models with more complex binaural processing.

Acknowledgments

We would like to thank M. Dietz, B. Eurich, and J. Encke for helpful remarks. We would also like to thank the members of the Medizinische Physik and Birger Kollmeier for continued support. This work was supported

by the Deutsche Forschungsgemeinschaft (DFG – 352015383 – SFB1330 A2 and DFG – 390895286 – EXC 2177/1). English language services were provided by https://stels-ol.de.

Data availability statement

MATLAB implementations of the psychoacoustic models used in this study are provided under: www.faame4u.com

References

1. D.S. Brungart, N. Iyer: Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. Journal of the Acoustical Society of America 132 (2012) 2545–2556. https://doi.org/10.1121/1.4747005.
2. S.D. Ewert, W. Schubotz, T. Brand, B. Kollmeier: Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers. Journal of the Acoustical Society of America 142 (2017) 12–28. https://doi.org/10.1121/1.381578.
3. I. Hirsh: The influence of interaural phase on interaural summation and inhibition. Journal of the Acoustical Society of America 20 (1948) 536–544. https://doi.org/10.1121/1.1916992.
4. S. van de Par, A. Kohlrausch: Dependence of binaural masking level differences on center frequency, masker bandwidth and interaural parameters. Journal of the Acoustical Society of America 106 (1999) 1940–1947. https://doi.org/10.1121/1.427942.
5. T. Dau, B. Kollmeier, A. Kohlrausch: Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. Journal of the Acoustical Society of America 102 (1997) 2892–2905. https://doi.org/10.1121/1.420344.
6. T. Dau, B. Kollmeier, A. Kohlrausch: Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. Journal of the Acoustical Society of America 102 (1997) 2906–2919. https://doi.org/10.1121/1.420345.
7. S.D. Ewert, T. Dau: Characterizing frequency selectivity for envelope fluctuations. Journal of the Acoustical Society of America 108 (2000) 1181–1196. https://doi.org/10.1121/1.1288665.
8. J. Breebaart, S. van de Par, A. Kohlrausch: Binaural processing model based on contralateral inhibition. I. Model setup. Journal of the Acoustical Society of America 110 (2001) 1074–1088. https://doi.org/10.1121/1.1383297.
9. T. Biberger, S.D. Ewert: Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility. Journal of the Acoustical Society of America 140 (2016) 1023–1038. https://doi.org/10.1121/1.4960574.
10. B.C.J. Moore, C.-T. Tan: Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion. Journal of the Audio Engineering Society 52 (2004) 900–914.
11. K.S. Rhebergen, N.J. Versfeld: A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. Journal of the Acoustical Society of America 117 (2005) 2181–2192. https://doi.org/10.1121/1.1861713.
12. R. Beutelmann, T. Brand, B. Kollmeier: Revision, extension and evaluation of a binaural speech intelligibility model. Journal of the Acoustical Society of America 127 (2010) 2479–2497. https://doi.org/10.1121/1.3295575.
13. M. Lavandier, J.F. Culling: Prediction of binaural speech intelligibility against noise in rooms. Journal of the Acoustical Society of America 127 (2010) 387–399. https://doi.org/10.1121/1.3268612.
14. A.H. Andersen, J.M. de Haan, Z.-H. Tan, J. Jensen: Predicting the intelligibility of noisy and non-linearly processed binaural speech. IEEE/ACM Transactions on speech, Audio and Language Processing 24 (2016) 1908–1920. https://doi.org/10.1109/TASLP.2016.2588002.
15. J.-H. Fleßner, R. Huber, S.D. Ewert: Assessment and prediction of binaural aspects of audio quality. Journal of the Audio Engineering Society 65 (2017) 929–942. https://doi.org/10.17743/jaes.2017.0037.
16. T. Biberger, J.-H. Fleßner, R. Huber, S.D. Ewert: An objective audio quality measure based on power and envelope power cues. Journal of the Audio Engineering Society 66 (2018) 578–593. https://doi.org/10.17743/jaes.2018.0031.
17. J.-H. Fleßner, T. Biberger, S.D. Ewert: Subjective and objective assessment of monaural and binaural aspects of audio quality. IEEE Transactions on Audio, Speech and Language Processing 27 (2019) 1112–1125. https://doi.org/10.1109/TASLP.2019.2904850.
18. T. Biberger, H. Schepker, F. Denk, S.D. Ewert: Instrumental quality predictions and analysis of auditory cues for algorithms in modern headphone technology. Trends in Hearing 25 (2021) 1–22. https://doi.org/10.1177/23312165211001219.
19. R.D. Patterson, B.C.J. Moore: Auditory filters and excitation patterns as representations of frequency resolution. In: B.C.J. Moore, Ed. Frequency selectivity in hearing. London: Academic Press, 1986.
20. C.J. Plack, A.J. Oxenham: Basilar-membrane nonlinearity and the growth of forward masking. Journal of the Acoustical Society of America 103 (1998) 1598–1608. https://doi.org/10.1121/1.421294.
21. H. Fletcher: Auditory patterns. Reviews of Modern Physics 12 (1940) 47–65. https://doi.org/10.1103/RevModPhys.12.47.
22. N.F. Viemeister: Temporal modulation transfer functions based upon modulation thresholds. Journal of the Acoustical Society of America 66 (1979) 1364–1380. https://doi.org/10.1121/1.383531.
23. B.R. Glasberg, B.C.J. Moore: 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 (2005) 906–918.
24. M.L. Jepsen, S.D. Ewert, T. Dau: A computational model of human auditory signal processing and perception. Journal of the Acoustical Society of America 124 (2008) 422–438. https://doi.org/10.1121/1.2924135.
25. S. Jørgensen, S.D. Ewert, T. Dau: A multi-resolution envelope-power based model for speech intelligibility. Journal of the Acoustical Society of America 134 (2013) 436–446. https://doi.org/10.1121/1.4807563.
26. T. Biberger, S.D. Ewert: The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking. Journal of the Acoustical Society of America 142 (2017) 1098–1111. https://doi.org/10.1121/1.4999059.
27. L.A. Jeffress: A place theory of sound localization. Journal of Comparative and Physiological Psychology 41 (1948) 35–39. https://doi.org/10.1037/h0061495.
28. N.I. Durlach: Equalization and cancellation theory of binaural masking-level differences. Journal of the Acoustical Society of America 35 (1963) 1206–1218. https://doi.org/10.1121/1.1918675.
29. W. Lindemann: Extension of a binaural cross-correlation model by contralateral inhibition. Journal of the Acoustical Society of America 80 (1986) 1608–1622. https://doi.org/10.1121/1.394325.
30. R.M. Stern, G.D. Shear: Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay. Journal of the Acoustical Society of America 100 (1996) 2278–2288. https://doi.org/10.1121/1.417937.

31. L.R. Bernstein, C. Trahiotis: Enhancing interaural-delay-based extents of laterality at high frequencies by using "transposed stimuli". Journal of the Acoustical Society of America 113 (2003) 3335–3347. https://doi.org/10.1121/1.1570431.
32. L.R. Bernstein, C. Trahiotis: Lateralization produced by interaural temporal and intensitive disparities of high-frequency, raised-sine stimuli: Data and modeling. Journal of the Acoustical Society of America 131 (2012) 409–415. https://doi.org/10.1121/1.3662056.
33. M. Dietz, S.D. Ewert, V. Hohmann, B. Kollmeier: Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences. Brain Research 1220 (2008) 234–245. https://doi.org/10.1016/j.brainres.2007.09.026.
34. J. Klug, L. Schmors, G. Ashida, M. Dietz: Neural rate difference model can account for lateralization of high frequency stimuli. Journal of the Acoustical Society of America 148 (2020) 678–691. https://doi.org/10.1121/10.0001602.
35. S. Doclo, S. Gannot, D. Marquardt, E. Hadad: Binaural speech processing with application to hearing devices. In: E. Vincent, T. Virtanen, S. Gannot, Eds. Audio source separation and speech enhancement, Wiley, 2018. https://doi.org/10.1002/9781119279860.ch18.
36. R. Wan, N.I. Durlach, H.S. Colburn: Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. Journal of the Acoustical Society of America 136 (2014) 768–776. https://doi.org/10.1121/1.4884767.
37. A. Chabot-Leclerc, E.N. MacDonald, T. Dau: Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain. Journal of the Acoustical Society of America 140 (2016) 192–205. https://doi.org/10.1121/1.4954254.
38. J. Breebaart, S. van de Par, A. Kohlrausch: Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters. Journal of the Acoustical Society of America 110 (2001) 1089–1104. https://doi.org/10.1121/1.1383298.
39. J. Breebaart, S. van de Par, A. Kohlrausch: Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters. Journal of the Acoustical Society of America 110 (2001) 1105–1117. https://doi.org/10.1121/1.1383299.
40. P.M. Briley, A.M. Goman, A.Q. Summerfield: Physiological evidence for a midline spatial channel in human auditory cortex. JARO: Journal of the Association for Research in Otolaryngology 17 (2016) 331–340. https://doi.org/10.1007/s10162-016-0571-y.
41. B. Grothe, M. Pecka: The natural history of sound localization in mammals – a story of neuronal inhibition. Frontiers in Neural Circuits 8 (2014) 116. https://doi.org/10.3389/fncir.2014.00116.
42. M. Pecka, A. Brand, O. Behrend, B. Grothe: Interaural time difference processing in the mammalian medial superior olive: The role of glycinergic inhibition. Journal of Neuroscience 28 (2008) 6914–6925. https://doi.org/10.1523/JNEUROSCI.1660-08.2008.
43. B. Grothe, M. Pecka, D. McAlpine: Mechanisms of sound localization in mammals. Physiological Reviews 90 (2010) 983–1012. https://doi.org/10.1152/physrev.00026.2009.
44. S. Kortlang, M. Mauermann, S.D. Ewert: Suprathreshold auditory processing deficits in noise: Effects of hearing loss and age. Hearing Research 331 (2016) 27–40. https://doi.org/10.1016/j.heares.2015.10.004.
45. N. Paraouty, S.D. Ewert, N. Wallaert, C. Lorenzi: Interactions between amplitude modulation and frequency modulation processing: Effects of age and hearing loss. Journal of the Acoustical Society of America 140 (2016) 121–131. https://doi.org/10.1121/1.4955078.
46. N. Wallaert, B.C.J. Moore, C. Lorenzi: Comparing the effects of age on amplitude modulation detection. Journal of the Acoustical Society of America 139 (2016) 3088–3096. https://doi.org/10.1121/1.4953019.
47. N. Wallaert, B.C.J. Moore, S.D. Ewert, C. Lorenzi: Sensorineural hearing loss enhances auditory sensitivity and temporal integration for amplitude modulation. Journal of the Acoustical Society of America 141 (2017) 971–980. https://doi.org/10.1121/1.4976080.
48. S.D. Ewert, N. Paraouty, C. Lorenzi: A two-path model of auditory modulation detection using temporal fine structure and envelope cues. European Journal of Neuroscience 51 (2018) 1265–1278. https://doi.org/10.1111/ejn.13846.
49. S.D. Ewert: Defining the proper stimulus and its ecology – mammals. In: B. Fritzsch (Ed.), The senses: A comprehensive reference, Elsevier, 2020. https://doi.org/10.1016/B978-0-12-809324-5.24238-7.
50. ISO 389-7: Acoustics-reference zero for the calibration of audiometric equipment. Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions. International Organization for Standardization, Geneva, Switzerland, 2005.
51. B.C.J. Moore, B.R. Glasberg: Suggested formulae for calculating auditory filter bandwidth and excitation patterns. Journal of the Acoustical Society of America 74 (1983) 750–753. https://doi.org/10.1121/1.389861.
52. T. Marquardt, D. McAlpine: A π-limit for coding ITDs: Implications for binaural models. In: B. Kollmeier, Ed. Hearing – From sensory processing to perception, Springer, 2007. https://doi.org/10.1007/978-3-540-73009-5_44.
53. A. Kohlrausch, R. Fassel, T. Dau: The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. Journal of the Acoustical Society of America 108 (2000) 723–734. https://doi.org/10.1121/1.429605.
54. B.C.J. Moore: An Introduction to the Psychology of Hearing, 4th ed., Academic, London, 1997.
55. J.L. Verhey, T. Dau, B. Kollmeier: Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model. Journal of the Acoustical Society of America 106 (1999) 2733–2745. https://doi.org/10.1121/1.428101.
56. W.P. Tanner, R.D. Sorkin: The theory of signal detectability. In: J.V. Tobias, Ed. Foundation of modern auditory function, Academic, New York, 1972.
57. S. Jørgensen, T. Dau: Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. Journal of the Acoustical Society of America 130 (2011) 1475–1487. https://doi.org/10.1121/1.3621502.
58. ANSI S3.5: Methods for calculation of the speech intelligibility index (Standards Secreteriat). Acoustical Society of America, New York, 1997.
59. A.J.M. Houtsma, N.I. Durlach, L.D. Braida: Intensity perception. XI. Experimental results on the relation of intensity resolution to loudness matching. Journal of the Acoustical Society of America 68 (1998) 807–813. https://doi.org/10.1121/1.384819.
60. B.C.J. Moore, J.I. Alcántara, T. Dau: Masking patterns for sinusoidal and narrow-band noise maskers. Journal of the Acoustical Society of America 104 (1998) 1023–1038. https://doi.org/10.1121/1.423321.
61. S.D. Ewert, T. Dau: External and internal limitations in amplitude-modulation processing. Journal of the Acoustical Society of America 116 (2004) 478–490. https://doi.org/10.1121/1.1737399.

62. R.G. Klumpp, H.R. Eady: Some measurements of interaural time difference thresholds. Journal of the Acoustical Society of America 28 (1956) 859–860. https://doi.org/10.1121/1.1908493.
63. J. Zwislocki, R.S. Feldman: Just noticeable differences in dichotic phase. Journal of the Acoustical Society of America 28 (1956) 860–864. https://doi.org/10.1121/1.1908495.
64. A. Brughera, L. Dunai, W.M. Hartmann: Human interaural time differences thresholds for sine tones: The high-frequency limit. Journal of the Acoustical Society of America 133 (2013) 2839–2855. https://doi.org/10.1121/1.4795778.
65. A. Mills: Lateralization of high-frequency tones. Journal of the Acoustical Society of America 32 (1960) 132–134. https://doi.org/10.1121/1.1907864.
66. D.W. Grantham: Interaural intensity discrimination: Insensitivity at 1000 Hz. Journal of the Acoustical Society of America 75 (1984) 1191–1194. https://doi.org/10.1121/1.390769.
67. M.J. Goupell, O.A. Stakhovskaya: Across-channel interaural-level-difference processing demonstrates frequency dependence. Journal of the Acoustical Society of America 143 (2018) 645–658. https://doi.org/10.1121/1.5021552.
68. I. Hirsh, M. Burgeat: Binaural effects in remote masking. Journal of the Acoustical Society of America 30 (1958) 827–832. https://doi.org/10.1121/1.1930084.
69. A. Kohlrausch: Auditory filter shape derived from binaural masking experiments. Journal of the Acoustical Society of America 84 (1988) 573–583. https://doi.org/10.1121/1.396835.
70. W.A. Yost: Prior stimulation and the masking-level difference. Journal of the Acoustical Society of America 78 (1985) 901–906. https://doi.org/10.1121/1.392920.
71. R. Wilson, C. Fowler: Effects of signal duration on the 500-Hz masking-level difference. Scandinavian Audiology 15 (1986) 209–215. https://doi.org/10.3109/01050398609042145.
72. R. Wilson, R. Fugleberg: Influence of signal duration on the masking-level difference. Journal of Speech, Language, and Hearing Research 30 (1987) 330–334. https://doi.org/10.1044/jshr.3003.330.
73. L.R. Bernstein, C. Trahiotis: The effects of signal duration on N_0S_0 and N_0S_π thresholds at 500 Hz and 4 kHz. Journal of the Acoustical Society of America 105 (1999) 1776–1783. https://doi.org/10.1121/1.426715.
74. B. Kollmeier, R.H. Gilkey: Binaural forward and backward masking: Evidence for sluggishness in binaural detection. Journal of the Acoustical Society of America 87 (1990) 1709–1719. https://doi.org/10.1121/1.399419.
75. E.R. Hafter, S.C. Carrier: Binaural interaction in low-frequency stimuli: The inability to trade time and intensity completely. Journal of the Acoustical Society of America 51 (1972) 1852–1862. https://doi.org/10.1121/1.1913044.
76. N. Kopčo, B.G. Shinn-Cunningham: Spatial unmasking of nearby pure-tone targets in a simulated anechoic environment. Journal of the Acoustical Society of America 87 (2003) 2856–2870. https://doi.org/10.1121/1.1616577.
77. V.R. Algazi, R.O. Duda, D.M. Thompson, C. Avendano: The CIPIC HRTF database, in: Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, 4–24 October 2001, New Platz, NY, USA, pp. 99–102. https://doi.org/10.1109/ASPAA.2001.969552.
78. K.C. Wagner, T. Brand, B. Kollmeier: Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests [Development and evaluation of a sentence test for German language III: Design, optimization and evaluation of the Oldenburger sentence test]. Zeitschrift für Audiologie 38 (1999) 86–95.
79. I. Holube, S. Fredelake, M. Vlaming, B. Kollmeier: Development and analysis of an International Speech Test Signal (ISTS). Int. J. Audiol. 49 (2010) 891–903. https://doi.org/10.3109/14992027.2010.506889.
80. R. Beutelmann, T. Brand: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. Journal of the Acoustical Society of America 120 (2006) 331–342. https://doi.org/10.1121/1.2202888.
81. D.S. Brungart: Informational and energetic masking effects in the perception of two simultaneous talkers. Journal of the Acoustical Society of America 109 (2001) 1101–1109. https://doi.org/10.1121/1.1345696.
82. I. Siveke, S.D. Ewert, B. Grothe, L. Wiegrebe: Psychophysical and physiological evidence for fast binaural processing. Journal of Neuroscience 28 (2008) 2043–2052. https://doi.org/10.1523/JNEUROSCI.4488-07.2008.
83. C.F. Hauth, T. Brand: Modelling sluggishness in binaural unmasking of speech for maskers with time-vaying interaural phase differences. Trends in Hearing 22 (2018) 1–10. https://doi.org/10.1177/2331216517753547.
84. J. Encke, W. Hemmert: Extraction of inter-aural time differences using a spiking neuron network model of the medial superior olive. Frontiers in Neuroscience 12 (2018) 140. https://doi.org/10.3389/fnins.2018.00140.
85. J. Bouse, V. Vencovský, F. Rund: Functional rate-code models of the auditory brainstem for predicting lateralization and discrimination data of human binaural perception. Journal of the Acoustical Society of America 145 (2019) 1–15. https://doi.org/10.1121/1.5084264.
86. V. Best, J. Mejia, K. Freeston, R.J. van Hoesel, H. Dillon: An evaluation of the performance of two binaural beamformers in complex and dynamic multitalker environments. International Journal of Audiology 54 (2015) 727–735. https://doi.org/10.3109/14992027.2015.1059502.
87. N. Gößling, D. Marquardt, S. Doclo: Performance Analysis of the extended binaural MVDR beamformer with partial noise estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 462–476. https://doi.org/10.1109/TASLP.2020.3043674.
88. M.M.E. Hendrikse, G. Grimm, V. Hohmann: Evaluation of the influence of head movement on hearing aid algorithm performance using acoustic simulations. Trends in Hearing 24 (2020) 1–20. https://doi.org/10.1177/2331216520916682.

Cite this article as: Biberger T. & Ewert SD. 2022. Towards a simplified and generalized monaural and binaural auditory model for psychoacoustics and speech intelligibility. Acta Acustica, **6**, 23.