Acta Acustica 2022, **6**, 17

https://doi.org/10.1051/aacus/2022008

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

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

SCIENTIFIC ARTICLE OPEN ACCESS

A comparative study of eight human auditory models of monaural processing

Alejandro Osses Vecchi[1,*], Léo Varnet[1], Laurel H. Carney[2], Torsten Dau[3], Ian C. Bruce[4], Sarah Verhulst[5], and Piotr Majdak[6]

[1] Laboratoire des systèmes perceptifs, Département d'études cognitives, École Normale Supérieure, PSL University, CNRS, 75005 Paris, France
[2] Departments of Biomedical Engineering and Neuroscience, University of Rochester, Rochester, NY 14642, USA
[3] Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[4] Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada
[5] Hearing Technology group, WAVES, Department of Information Technology, Ghent University, 9000 Ghent, Belgium
[6] Acoustics Research Institute, Austrian Academy of Sciences, 1040 Vienna, Austria

Received 4 July 2021, Accepted 28 February 2022

Abstract – A number of auditory models have been developed using diverging approaches, either physiological or perceptual, but they share comparable stages of signal processing, as they are inspired by the same constitutive parts of the auditory system. We compare eight monaural models that are openly accessible in the Auditory Modelling Toolbox. We discuss the considerations required to make the model outputs comparable to each other, as well as the results for the following model processing stages or their equivalents: Outer and middle ear, cochlear filter bank, inner hair cell, auditory nerve synapse, cochlear nucleus, and inferior colliculus. The discussion includes a list of recommendations for future applications of auditory models.

1 Introduction

Computational auditory models reflect our fundamental knowledge about hearing processes and have been accumulated during decades of research (e.g., [1]). Models are used to derive conclusions, reproduce findings, and develop future applications. Usually, models are built in stages that reflect basic parts of the auditory system, such as cochlear filtering, mechanoneural interface, and neural processing, by applying signal-processing methods such as bandpass filtering or envelope processing [2]. Models of monaural processing often form a basis for binaural models (e.g., [3]) and more complex models of auditory-based multimodal cognition (e.g., [4]). For this reason, combined with the increasing popularity of reproducible research [5], it is not surprising that there is an increasing number of auditory models available as software packages (e.g., [6, 7]).

However, models must be used with caution because they approximate auditory processes and are designed and evaluated under a specific configuration for a specific set of input sounds. While the evaluation conditions are selected to test the main properties of the simulated stages, models may provide different predictions when processing unseen sounds. Combined with the wide and low-threshold availability of model implementations, there is a chance of applying a model outside its specific signal or parameter range. Thus, studies comparing models' properties and configurations are important to model users. For example, Saremi *et al.* [12] compared seven models of cochlear filtering with respect to various parameters describing the nonlinear filtering process of an active cochlea, and Lopez-Poveda [13] compared eight models of the auditory periphery based on the reproduction of auditory-nerve properties. Other related studies focus on a specific application (e.g., [10, 11, 14–17]) or provide an introduction to modelling frameworks [18, 19].

In the current study, we compare various monaural auditory models that approximate subcortical neural processing. For this comparison, we use model configurations that reduce the heterogeneity across model outputs, indicating advantages and disadvantages of these configuration choices. These configurations are evaluated using the same set of sound stimuli across models. The selected set of stimuli illustrates critical model properties that can be used as a guideline for the choice of a specific model. These properties include fast and slow temporal aspects, i.e., temporal fine structure and temporal envelope, that are evaluated for a wide range of presentation levels.

We selected a number of auditory models that met two main criteria. First, the selected models describe the auditory path beginning with the acoustic input up to

*Corresponding author: ale.a.osses@gmail.com

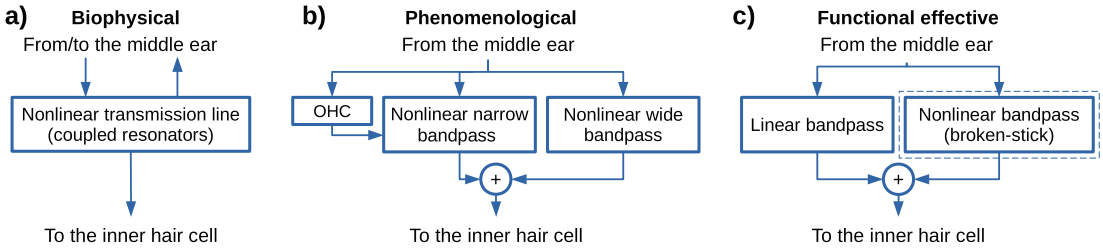

Figure 1. Model families used in this study. The families are defined by the level of detail in simulating the cochlear processing and are sorted by their complexity from left to right. a) Biophysical models using a nonlinear transmission line that contains resonating stages coupled by the cochlear fluid; b) Phenomenological models using nonlinear filters dynamically controlled by an outer-hair-cell (OHC) model; c) Functional effective models using linear filters, optionally combined with static nonlinear filters.

subcortical neural stages, in the cochlear nucleus (brainstem) and the inferior colliculus (midbrain). Consideration of these stages extends previous comparisons of auditory periphery models [12, 13]. Second, the model implementations are publicly available and validated to simulate psychoacoustic performance and/or physiological properties. We use the implementations available in the Auditory Modelling Toolbox (AMT) [8, 9, 20].

Based on our inclusion criteria, some models are excluded from the comparison, e.g., models that have only been evaluated at the level of cochlear filtering, such as models based on Hopf bifurcation [21] and the model of asymmetric resonators with fast-acting compression [22]. Other cochlear models, such as the Gammatone filter bank [23], the dual-resonance nonlinear model [24], the chirp filter bank [25] and the transmission-line model from [26], are included as modules in the selected models. We further excluded models whose structure did not contain one or more of the relevant processing stages that we evaluated in this study (Stages 3–6 in Fig. 2). Two models that entered this category are the power spectrum models, EPSM [27] (no stages 3–5) and GPSM [10, 11] (no stage 5). Lastly, we did not include models focusing on specific psychoacoustic metrics [28–30], despite the fact that such models are often based on comparable auditory stages as those described in this study.

For the sake of simplicity, our analyses are focused on the comparison *across* models rather than on a comparison with experimental data. Nevertheless, we provide experimental references to the simulations that are illustrated throughout this paper. Additionally, to encourage reproducible research in auditory modelling, all our paper figures can be retrieved using AMT 1.1, including (raw) waveform representations of intermediate model outputs.

The paper is structured as follows: In Section 2 we provide a brief description of the processing stages in the selected auditory models. Their specific configurations are described in Section 3. The model comparison is presented in Section 4 and contains a description of the set of test stimuli as well as a detailed numerical description of the simulation results. Section 5 starts with a list of applications of the evaluated models, including some general considerations for the application of auditory models in further modelling work. Note that although the detailed analysis in Section 4 is relevant for model users who are interested in a transparent and accurate description of the illustrated model outputs, readers who are only interested in a bigger picture, as it is covered elsewhere (e.g., [1, 2, 13]), may wish to go directly to Sections 5 and 6.

2 Models

We define three model families, classified by their objectives [40], which translate into three different levels of detail in simulating the cochlear processing, as schematised in Figure 1. The selected models are listed in Table 1 and are labelled throughout this paper by the last name of the first author and the year of the corresponding publication. This naming system directly reflects the corresponding model functions implemented in AMT 1.1 [8].

We define the family of *biophysical models* (Fig. 1a) that use a transmission line consisting of many resonant stages coupled by the cochlear fluid. Biophysical models aim at exploring how the properties of the system emerge from biological-level mechanisms, needing a fine-grained description at this level. The biophysical models are represented by `verhulst2015` [34] and its extended version, `verhulst2018` [35] (model version 1.2 [41, 42]). We further define *phenomenological models* which primarily predict physiological properties of the system, using a more abstract level of detail than the biophysical models. The phenomenological models considered here rely on dynamically adapted bandpass-filtering stages combined with nonlinear mappings (Fig. 1b) and are represented by `zilany2014` [32] and its extended version `bruce2018` [36], both combined with the same-frequency inhibition-excitation (SFIE) stages for subcortical processing [43]. Further approximation is given by *functional-effective models* [44], which target the simulation of behavioural (perceptual) performance rather than the direct simulation of neural representations. These models usually approximate the cochlear processing by using static bandpass filtering with an optional nonlinear mapping (Fig. 1c). The linear effective models are represented by `dau1997` [31] and `osses2021` [39] and the nonlinear effective models are represented by `king2019` [37] and `relanoiborra2019` [38]. Given that for each model a similar level of approxima-

Table 1. List of selected models. The model labels used in this study correspond with the model functions in AMT 1.1.

Label	Reference
`dau1997`	Dau *et al.* (1997) [31]
`zilany2014`	Zilany *et al.* (2014) [32] and Carney *et al.* (2015) [33]
`verhulst2015`	Verhulst *et al.* (2015) [34]
`verhulst2018`	Verhulst *et al.* (2018) [35]
`bruce2018`	Bruce *et al.* (2018) [36] and Carney *et al.* (2015) [33]
`king2019`	King *et al.* (2019) [37]
`relanoiborra2019`	Relaño-Iborra *et al.* (2019) [38]
`osses2021`	Osses and Kohlrausch (2021) [39]

tion has been generally used in the design of subsequent model stages, we use the defined categories to reflect the nature of the entire model. The defined categories are not meant to represent a hard boundary for model classification. Hence, we do not discard the existence of other more- or less-detailed models than the selected biophysical and effective models, respectively.

The selected monaural models share common stages of signal processing, as indicated in the schematic diagrams of Figure 2, with some stages even using the same (digital) implementation. Each model stage mimics, with greater or lesser detail, underlying hearing processes along the ascending auditory pathway. The thick vertical lines in Figure 2 indicate the intermediate model outputs which are the basis for our evaluation. Note that these stages are, conceptually speaking, independent of each other. However, because of nonlinear interactions between them, the processing performed by these stages is not commutative and thus requires a step-by-step approach.

2.1 Outer ear

The listener's head, torso, and pinna filter incoming sounds. The ear-canal resonance further emphasises frequencies around 3000 Hz [45]. These effects can be accounted for by filtering the sound with a head-related transfer function (HRTF) (e.g., [46]) or by applying a headphone-to-tympanic-membrane transfer function, as used in `relanoiborra2019` and `osses2021`. The other six selected models that do not include an outer-ear filter, implicitly assume that either the outer ear does not introduce a significant effect in the subsequent sound processing chain, or that the sounds are presented near the tympanic membrane, as is the case for a sound presentation using in-ear earphones.

2.2 Middle ear

Six of the eight evaluated models include a stage of middle-ear filtering. The transfer functions of the middle-ear filters used in these models are shown in Figure 3. The transfer functions in `verhulst2015` and `verhulst2018` approximate the middle-ear forward pressure gain ("M1" in [47]). The humanised `zilany2014` and `bruce2018` models use a linear middle-ear filter [48, 49] that approximates the forward-pressure measurements from [50, 51]. The middle-ear filters in `relanoiborra2019` and `osses2021` are those designed to represent stapes velocity near the oval window of the cochlea [24, 52]. These models use the same filter implementation, but the filter in `osses2021` includes a gain factor to provide a 0-dB amplitude in the frequency range of the passband and a fixed group-delay compensation.

Middle-ear filtering not only introduces a bandpass characteristic to the incoming signal (Fig. 3), but also affects the operating range of cochlear compression in models relying on nonlinear cochlear processing, i.e., `verhulst2015`, `verhulst2018`, `zilany2014`, `bruce2018`, and `relanoiborra2019`. The passband gains of the middle-ear filters are indicated in Table 2 and range between −66.9 dB (`relanoiborra2019`) and +24 dB (`verhulst2015`). In nonlinear models, lower and higher passband gains vary the actual input level to the filter bank, shifting the onset of cochlear compression to higher and lower knee points, respectively.

2.3 Cochlear filtering

A cochlear filter bank performs a spectral analysis of incoming signals to simulate the mechanical oscillations of the basilar membrane (BM) and organ of Corti at different points along the cochlea. The BM and organ of Corti are thought to have multiple modes of vibration which could drive the transduction currents of the IHCs depending on their amplitudes. The cochlear filtering stages in the evaluated models all aim to describe the filtering properties of the main mode(s) of vibration that explain the frequency-tuning curves of AN fibres. Some of the models include further components aimed at capturing the filtering behaviour of additional modes of BM and/or organ of Corti vibration to better explain the level-dependent nonlinear filtering of the cochlea. All the included cochlear filtering stages produce a set of N time-domain signals, for N simulated characteristic frequencies (CFs). Each cochlear section is assumed to either have relatively sharp frequency tuning (Eq. (1), [53]) or broader tuning (Eq. (2), [54]). For CFs expressed in Hz:

$$\mathrm{Q_{ERB}} = 12.7 \cdot (\,\mathrm{CF}\,/1000)^{0.3}, \tag{1}$$

$$\mathrm{Q_{ERB}} = \mathrm{CF}\,/[24.7 \cdot (4.37 \cdot \mathrm{CF}\,/1000 + 1)]. \tag{2}$$

The models `verhulst2015` and `verhulst2018` use a transmission-line model fitted to otoacoustic estimates of human cochlear filtering [26], whose outputs represent BM velocity [34, 35], which provides the input to a model stage that maps BM velocity to IHC stereociliar deflection. In `zilany2014` and `bruce2018`, the filtering is based on a chirp filter bank [25, 55] tuned to a human cochlea [48, 49, 56], that contains two parallel processing paths of static and outer-hair-cell (OHC) controlled filters (C2 and C1 in [32]), representing BM and organ of Corti motion that drive two separate IHC transduction functions

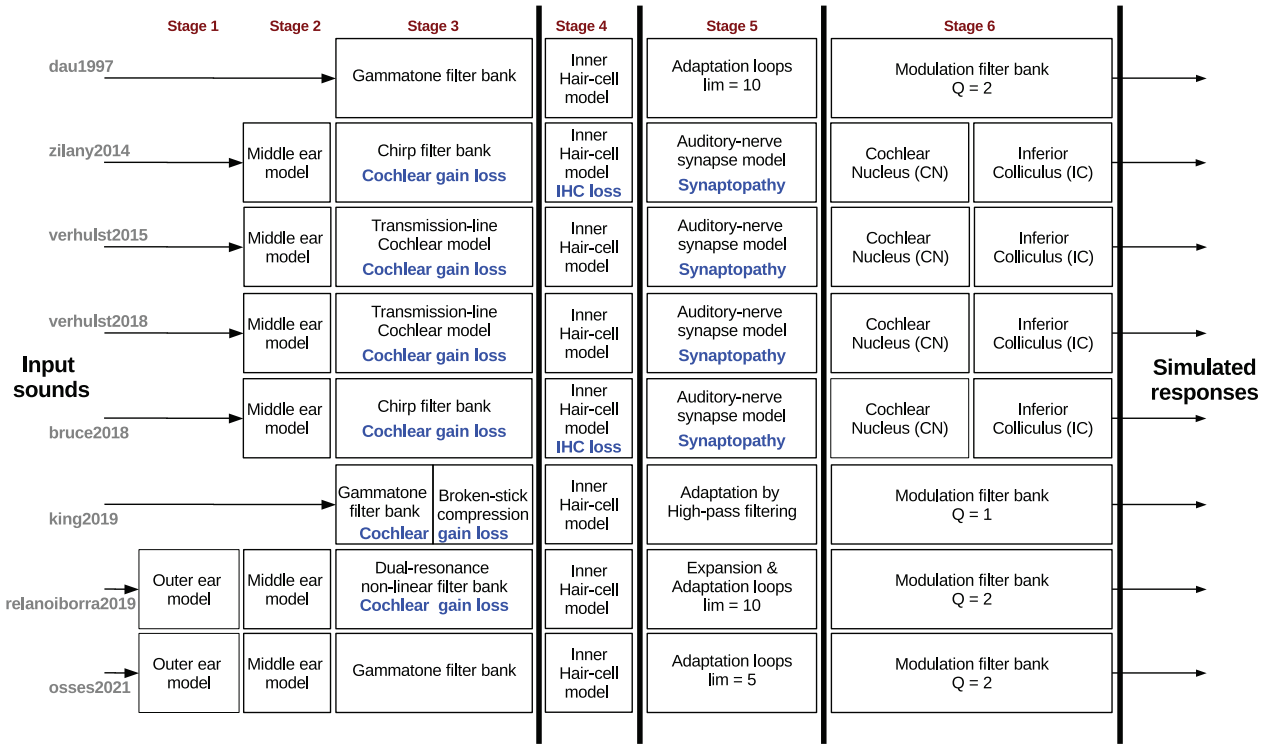

Figure 2. Block diagrams of the selected auditory models. Vertical lines: Intermediate model outputs as the basis for the evaluation in the corresponding sections. Blue: Type of hearing impairment that can be accounted for in the corresponding stage (see a brief overview in Sect. 5.1).

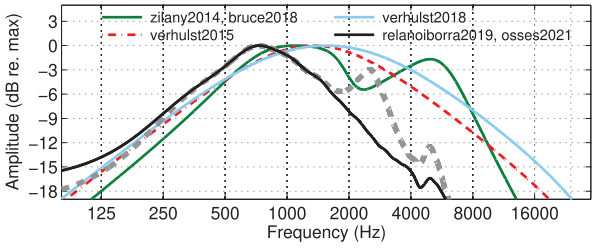

Figure 3. Amplitude spectra of the four middle-ear filters used in six of the evaluated models. The lines were shifted vertically to display their individual maximum at 0 dB. For `relanoiborra2019` and `osses2021`, the grey dashed line shows the combined response of the outer- and middle-ear filters. *Literature*: Figure 1A from [47] and Figure 3 from [51].

[55, 57]. The cochlear filters in these models are assumed to be tuned according to equation (1).[1]

In `dau1997` and `osses2021`, the linear Gammatone filter bank from [23] is used. `King2019` uses the Gammatone filter bank from [23] followed by a compressive stage acting above a given knee point. In `relanoiborra2019`, the cochlear processing is simulated by the dual-resonance nonlinear (DRNL) filter bank [24]. The cochlear filters of these models are assumed to be tuned according to equation (2).

2.4 Inner hair cell

The inner hair cells (IHCs) transform the mechanical BM and organ of Corti oscillations into receptor potentials, subsequently initiating neuronal discharges in the auditory nerve (AN) [58]. In the most simple approach, the IHC processing can be simulated as an envelope extractor that removes phase information for high stimulus frequencies, implemented as a half-wave rectification followed by a lowpass (LP) filter. This approach is used in `dau1997`, `king2019`, `relanoiborra2019`, and `osses2021`, in which the LP filters have −3-dB cut-off frequencies ($f_{\text{cut-off}}$) between 1000 and 2000 Hz. In `zilany2014`, `bruce2018`, and `verhulst2015`, a nonlinear transformation is applied to the output of the cochlear filter bank, followed by a cascade of LP filters with $f_{\text{cut-off}}$ of 3000 Hz (`zilany2014` and `bruce2018`) and 1000 Hz (`verhulst2015`). The resulting $f_{\text{cut-off}}$ of each model ranges between 642 Hz (`verhulst2015`) and 1000 Hz (`dau1997`, `relanoiborra2019`, `king2019`), as indicated in Table 2. In `verhulst2018`, a more sophisticated IHC model is used [59], that is implemented as a three-channel non-spiking Hodgkin–Huxley type model, with each of the channels representing mechanoelectrical and (fast and slow) potassium-gated processing [35, 59].

[1] The models `verhulst2015` and `verhulst2018` use an updated version of equation (1): $Q_{\text{ERB}} = 11.46 \cdot (\text{CF}/1000)^{0.25}$ [34, 35]. This equation produces sharp tuning curves but with slightly lower Q-factors compared to those given by equation (1), with values between 6.8 (CF = 125 Hz) and 19.3 (CF = 8000 Hz).

Table 2. Model configurations and numeric results. *Middle ear*: Details of frequency response of the middle-ear filters. *Cochlear filter bank*: 40 dB and 100 dB refer to filter characteristics between 160 and 8000 Hz in response to white noises of 40 and 100 dB SPL, respectively. *IHC*: Parameters and frequency response of the LP filter structures. *Subcortical processing*: Theoretical and estimated BMF of the modulation filter with the closest BMF to 100 Hz and peak-to-peak amplitude of the simulated IC outputs in response to 70-dB-peSPL clicks of positive (A) and negative ($-A$) (see Sect. 4.4 for more details). *Performance*: Time required to process a 1-s long stimulus using each of the selected auditory models, between Stages 1 and 6 (Fig. 2). All $f_{\text{cut-off}}$ in this table were measured at the −3 dB points of the amplitude spectrum.

Stage	Model							
Parameter	`dau1997`	`zilany2014`	`verhulst2015`	`verhulst2018`	`bruce2018`	`king2019`	`relanoiborra2019`	`osses2021`
Middle ear								
Passband gain in Fig. 3 (dB)	–	−6.0	24.0	18.0	−6.0	–	−66.9	0.00
Lower $f_{\text{cut-off}}$ (Hz)	–	577.6	601.0	601.1	577.6	–	474.8	474.8
Higher $f_{\text{cut-off}}$ (Hz)	–	6061.9	2995.3	3993.1	6061.9	–	1230.2	1230.2
Cochlear filter bank								
Reference gain in Figs. 4–5 (dB)	−0.6	−44.1	−77.9	−101.9	−44.1	−43.9	26.9	−1.9
40 dB: Number of filters	34	51	59	52	51	34	36	34
40 dB: Average filter bandwidth (ERB)	0.903	0.588	0.505	0.579	0.588	0.904	0.848	0.905
100 dB: Number of filters	34	20	12	15	20	33	27	34
100 dB: Average filter bandwidth (ERB)	0.903	1.57	3.046	2.299	1.57	0.925	1.147	0.905
IHC								
Number of filter sections	1	7	1	–	7	1	1	5
Order of each filter section	1	1	2	–	1	1	2	1
$f_{\text{cut-off}}$ of each filter section (Hz)	1000	3000	1000	–	3000	1000	1000	2000
$f_{\text{cut-off}}$ of the total filter structure (Hz)	1000	966	642	–	966	1000	1000	771
Subcortical processing								
Theoretical BMF (Hz)	77.2	85.4	82.4	82.4	85.4	94.0	77.2	77.2
Estimated BMF from Fig. 14a (Hz)	70	35	45	60	45	65	70	70
Unit of the amplitude	MU	spikes/s	µV	µV	spikes/s	a.u.	MU	MU
Click of amplitude A	407.3	116.5	0.245	0.401	117.6	$1.64 \cdot 10^{-4}$	423.1	155.9
Click of amplitude −A	437.4	115.8	0.237	0.407	117.3	$1.67 \cdot 10^{-4}$	408.9	157.3
Performance								
Total time (s)	0.80	86.6	122.9	319.5	163.1	1.86	7.70	0.622
Number of cochlear channels	31	66	401	401	66	31	60	31
Time per channel (s)	0.026	1.31	0.306	0.797	2.47	0.060	0.128	0.020

2.5 Auditory nerve

The transduction from IHC receptor potentials into patterns of neural activity can be derived from the interaction between the IHC and AN. Several AN synapse models have been inspired by the three-store diffusion model [58], assuming that the release of synaptic material is managed in three storage compartments. For steady-sound inputs, this model predicts a rapid neural firing shortly after the sound onset with a decreasing rate towards a plateau discharge rate, a phenomenon called adaptation (e.g., [60]).

The AN synapse models in `verhulst2015`, `verhulst2018`, and `zilany2014` are based on [58], but `zilany2014` further incorporates a power-law adaptation following the diffusion model from [32]. The synapse model in `bruce2018` uses a diffusion model based on [61] to: (1) have limited release sites, and (2) come after the power-law adaptation instead of before it [36]. The outputs of these models simulate the firing of neurons having a specific spontaneous rate of high-, medium-, and/or low-spontaneous rates.

The effective models, on the other hand, rely on a more coarse AN simulation, expressed in arbitrary units (a.u.). In `king2019`, adaptation is simulated by applying a highpass filter with a cut-off frequency of 3 Hz [37]. In `dau1997`, `relanoiborra2019`, and `osses2021`, adaptation is simulated by so-called adaptation loops [44] that introduce a nearly logarithmic compression to stationary input signals and a linear transformation for fast signal fluctuations (Appendix B in [39]). The arbitrary units of these transformed outputs are named model units (MUs).

2.6 Subcortical neural processing

AN firing patterns propagate to higher stages along the auditory pathway, first through the auditory brainstem, then towards more cortical regions [62]. On its way, AN spiking is mapped onto fluctuation patterns by neurons that are sensitive to the amplitude of low-frequency fluctuations [63]. This fluctuation sensitivity has been approximated using various approaches. Our analyses focus on model approximations of the modulation processing circuits of the ventral cochlear nucleus (CN) and inferior colliculus (IC) [43], as well as on different modulation-filter-bank variants [27, 31]. As a result, we exclude the analysis of other subcortical structures such as those that play a particular role in the binaural interaction between ears (e.g., the dorsal cochlear nucleus and superior olive) [62, 64].

The modulation processing in the ventral CN and IC can be simulated using the same-frequency inhibition-excitation (SFIE) model, resulting in a widely tuned modulation filter (Q-factor $\approx$ 1) with a best-modulation frequency (BMF) depending on the parameters of the model [33, 43]. The SFIE model has already been used in combination with the biophysical and phenomenological models described here. For example, `zilany2014` has been combined with the SFIE model using between one and three modulation filters (e.g., [65]). Or, `verhulst2015` and `verhulst2018` have used the SFIE model with one modulation filter centred at a BMF of 82.4 Hz (see Tab. 2) [35, 41]. Further, `bruce2018` can be combined with the SFIE model in the UR EAR 2020b toolbox [66]. Note that `zilany2014`, `verhulst2015`, and `verhulst2018` have used the output of their mean firing rate generator—an output that can be conceptualised as peri-stimulus time histograms (PSTHs) [67]—as an input to the SFIE model. In `bruce2018`, because of the stochastic processes in its spike generator, repeated processing of the same stimulus is recommended to obtain a faithful PSTH that can appropriately account for power-law adaptation properties (see Sect. 3 in [36]).

The effective models, on the other hand, approximate the subcortical neural processing based on the modulation-filter-bank concept [27, 31]. In `dau1997`, `king2019`, `relanoiborra2019`, and `osses2021`, linear modulation filter banks are used, covering a range of BMFs up to 1000 Hz. In `dau1997`, twelve modulation filters with a Q-factor of 2 and overlapped at their −3 dB points are used. The same modulation filters are used in `relanoiborra2019` and `osses2021`, but an additional 150-Hz LP filter is applied [27, 68] and the number of filters is limited so that the highest BMF is less than a quarter of the corresponding CF [69]. In `king2019`, the filter bank is used with a wider tuning (Q = 1, as suggested in [27, 70]), using ten 50%-overlapped filters having a maximum BMF of 120 Hz [37].

3 Model configuration

We evaluated the intermediate model outputs that are indicated by thick vertical black lines in Figure 2. The evaluation points are located after the cochlear filter bank (Stage 3), the IHC processing stage (Stage 4), the AN synapse stage or equivalent (Stage 5), and after the IC processing stage or equivalent (Stage 6). Starting with the default parameters of each model, we introduced small adjustments to obtain the most comparable model outputs. All comparisons can be reproduced with the function `exp_osses2022` from AMT 1.1 [8].

3.1 Level scaling

The same set of sound stimuli was used as input to all models. The waveform amplitudes were assumed to represent sound pressure expressed in Pascals (Pa). The models `zilany2014`, `verhulst2015`, `verhulst2018`, `bruce2018`, and `relanoiborra2019` use this level convention and did not require further level scaling. The models `dau1997`, `king2019`, and `osses2021` interpret sound pressures between −1 and 1 Pa as amplitudes in the range ±0.5, thus a factor of 0.5 (attenuation by 6 dB) was applied to the generated stimuli to meet the level convention of these models. For these latter models, which include mostly level-independent stages, such calibration is relevant because the adaptation loops (used in `dau1997`, `osses2021`, also extensible to `relanoiborra2019`) include level-dependent scaling (Eqs. (B1)–(B3) in [39]).

In `king2019`, a calibrated knee point (default of 30 dB) is used in its cochlear compression stage (Stage 3). All signal levels are reported as root-mean-square (rms) values referenced to 20 µPa, in dB sound pressure level (dB SPL).

3.2 Cochlear filtering

The phenomenological and effective models can be set to simulate responses at any CF. However, because of the nature of the transmission-line structure, the selected biophysical models have a discrete tonotopy that translates into a discrete set of available CFs.

The models `verhulst2015` and `verhulst2018` were set to 401 cochlear sections spaced at $\Delta x = 0.068$ mm with tonotopic distances x_n ranging between $x_1 = 3.74$ mm and $x_{401} = 30.9$ mm, that are related to CFs between $\mathrm{CF}_1 = 12010$ Hz and $\mathrm{CF}_{401} = 113$ Hz, according to the base-to-apex mapping [71],

$$\mathrm{CF}_n = A_0 \cdot \left(10^{-a \cdot x_n/1000}\right) - A \cdot k, \tag{3}$$

where x_n (in mm) can be obtained as $x_1 + \Delta x \cdot (n-1)$, and $A = 165.4188$ Hz, $a = 61.765$ 1/m, $k = 0.85$, and $A_0 = 20682$ Hz. Note that when reporting results, we indicate the cochlear section number n and its corresponding CF_n.

The cochlear-filtering parameters of `zilany2014` and `bruce2018` were those adapted to a human cochlea [48, 49]. Moreover, in order to analyse separately the effects of cochlear filtering and IHC processing in `zilany2014`, the C1- and C2-path outputs from the chirp filter bank were added and analysed before the IHC nonlinear mapping was applied. This analysis follows a similar rationale as analysing the main output of the DRNL filter bank in `relanoiborra2019` (see Fig. 3a from [24]).

Finally, in `king2019`, we used a compression factor of 0.3 for all simulated CFs, which is different from the one-channel (on-CF) compression used in [37, 72].

3.3 Inner hair cell and auditory nerve

Default parameters were used for the IHC and AN stages of the evaluated effective models. However, the biophysical and phenomenological models require the choice of parameters to simulate a population of AN fibres. For each CF we simulated 20 fibres, having either high- (HSR), medium- (MSR), or low-spontaneous rates (LSR), distributed in a 0.6–0.2–0.2 ratio [73, 74], resulting in a 12–4–4 configuration (HSR–MSR–LSR). Note that for `verhulst2015` and `verhulst2018`, this deviates from the standard 13–3–3 configuration [34, 35]. For `verhulst2015` and `verhulst2018`, the spontaneous rates of each fibre type were 68.5, 10, and 1 spikes/s for HSR, MSR, and LSR, respectively, as used in human-tuned simulations [35]. For `zilany2014`, the spontaneous rates of each fibre type were 100, 4, and 0.1 spikes/s and for `bruce2018` were 70, 4, and 0.1 spikes/s for HSR, MSR, and LSR, respectively. We further disabled the random fractional noise generators in `zilany2014` and `bruce2018` [75], and the random spontaneous rates in `bruce2018` ("std" from Tab. I in [36] was set to zero). With this configuration, the mean-rate synapse outputs of `verhulst2015`, `verhulst2018`, `zilany2014`, and `bruce2018` are deterministic. For this reason, to obtain population responses, we simulated the AN processing of each type of neuron only once and then weighted them by factors of 0.6, 0.2, and 0.2 for HSR, MSR, and LSR fibres, respectively. In contrast, the PSTH outputs that are reported for `zilany2014` and `bruce2018` are not deterministic, requiring the simulation of each AN fibre for each CF. Therefore, PSTH population responses were obtained by counting the average number of spikes in time windows of 0.5 ms across 100 repetitions of the corresponding stimuli.

3.4 Subcortical neural processing

The default configuration of the model stages of subcortical processing (Stage 6, Fig. 2) differs in the number of modulation filters (from 1 to 12) and in their tuning across models. In our study, we use only one modulation filter targeting a BMF of approximately 80 Hz (see "Theoretical BMF" in Tab. 2) using a Q-factor of approximately 1 for `zilany2014`, `verhulst2015`, `verhulst2018`, `bruce2018`, and `king2019`, and a Q-factor of 2 for `daul997`, `relanoiborra2019`, and `osses2021`.

For the biophysical and phenomenological models, we used the SFIE model [33, 43] using two different configurations. The SFIE model [43] integrated in `verhulst2015` and `verhulst2018` has CN parameters with excitatory and inhibitory time constants of $\tau_{\mathrm{exc}} = 0.5$ ms and $\tau_{\mathrm{inh}} = 2$ ms, a delay $D = 1$ ms, and a strength of inhibition of $S = 0.6$. The IC stage uses $\tau_{\mathrm{exc}} = 0.5$ ms, $\tau_{\mathrm{inh}} = 2$ ms [35], $D = 2$ ms, and $S = 1.5$ [43], achieving a BMF of 82.4 Hz.

For `zilany2014` and `bruce2018`, the SFIE model is a separate stage [33], implemented as `carney2015` in AMT 1.1, where either the mean-rate (`zilany2014`) or the PSTH outputs (`bruce2018`) are used as inputs. In our analysis, we only used the output of the band-enhanced IC cell, which corresponds to the SFIE model from [43]. The CN parameters were identical to those for the biophysical models. The IC parameters were $\tau_{\mathrm{exc}} = 1.11$ ms, $\tau_{\mathrm{inh}} = 1.67$ ms, $D = 1.1$ ms, and $S = 0.9$, achieving a BMF of 83.9 Hz [33]. Note the different inhibition strength S between models. In the biophysical models, the IC output is dominated by inhibitory responses ($S > 1$) whereas in the phenomenological models the IC output is dominated by excitatory responses ($S < 1$).

4 Evaluation

In this section we analyse the outputs of the eight selected auditory models in a number of test conditions, whose results are presented in Figures 4–15. We aimed at a comparison across models and thus, for the sake of clarity, we refrained from a direct comparison to ground-truth

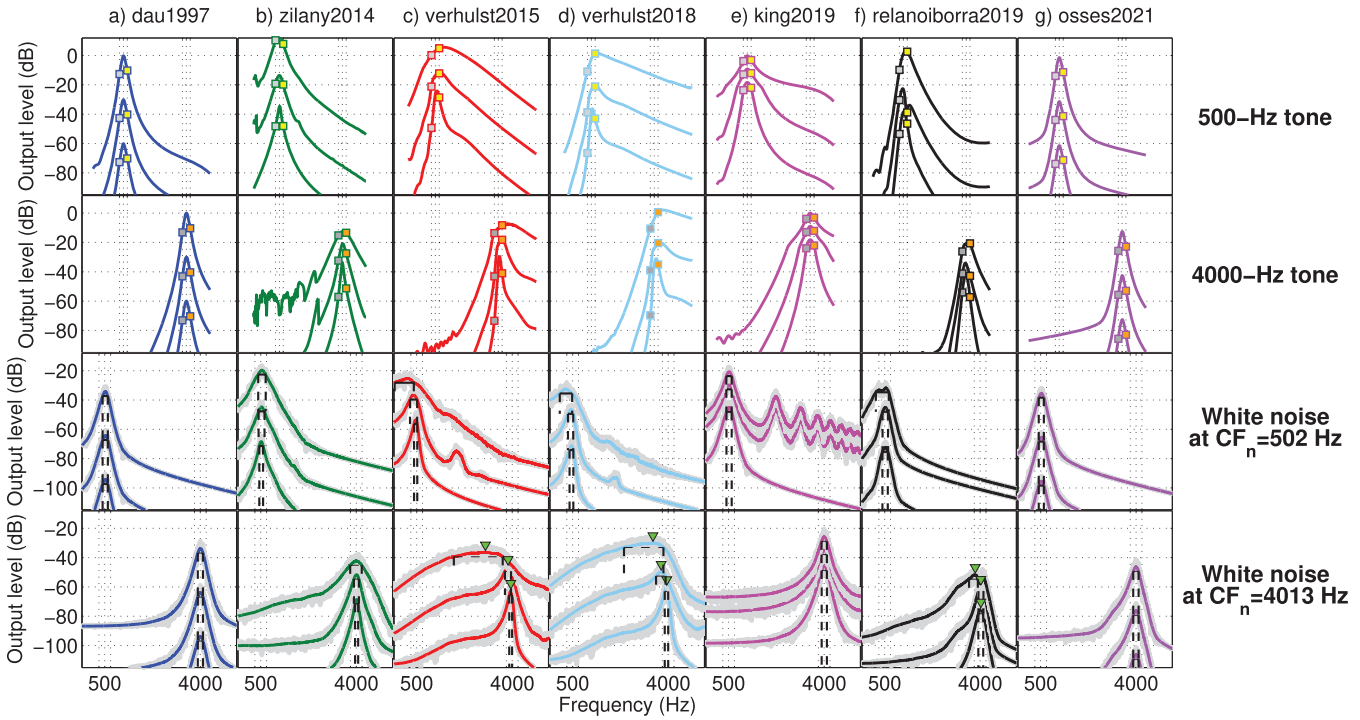

Figure 4. Filter bank responses to pure tones at 500 Hz and 4000 Hz (top two rows) and spectral magnitudes of single-channel responses to white noises (bottom two rows) for sounds of 40-, 70-, or 100-dB SPL (bottom-to-top coloured curves, respectively). In the first two rows, the responses represent excitation patterns at the simulated CFs. The coloured markers indicate the amplitudes at the (off-frequency) CFs one ERB_N below (grey) and one ERB_N above (yellow or orange) the on-frequency CF (502 or 4013 Hz). These markers are highlighted using the same colours in Figure 5. In the third and fourth rows, the average FFT response of two cochlear-filters (CFs of 502 Hz or 4013 Hz) in response to six 500-ms white noise sections are shown in grey, and the corresponding smoothed responses are shown in colour. This type of responses was used to assess the quality factors of Figure 6. The dashed black lines indicate the corresponding −3-dB filter bandwidths. All responses were shifted vertically by the reference gains given in Table 2 (see the text for details). *Literature*: Figure 1A from [76], Figure 2C–E from [77], and Figure 2 from [12].

references from physiological data. However, such a comparison is important and interesting. For this reason, we provide references where similar experimental and/or simulation analyses have been presented. These references are indicated as "Literature" in the caption of the corresponding figure. Alternatively, the outputs of the biophysical and phenomenological models may be considered as referential because they have been primarily developed to reflect physiological (human or animal) responses to sounds. This latter assumption always requires a careful consideration, especially when translating findings from animal to human physiology.

4.1 Cochlear filtering

Sound processing in the cochlea depends not only on the frequency but also on the level of the input stimulus [78]. The amplitude of the BM vibration displacement increases for higher levels, following an amplitude growth that comprises linear and compressive regimes [79]. We illustrate this level dependency for a set of pure tones and white noises. The pure tones had frequencies of 500 or 4000 Hz, with a duration of 100 ms. The white noise was generated as a fixed 3-s long waveform with a flat spectrum between 20 and 20000 Hz. All sounds were gated on and off with a 10-ms raised-cosine ramp and had levels of 40, 70, and 100 dB SPL. The obtained responses are shown in Figure 4, which were vertically shifted by the gains indicated in Table 2. These gains were derived for each model using the 1000-Hz pure tone of 100 dB SPL as a reference. The frequency responses were level-dependent for `zilany2014`, `verhulst2015`, `verhulst2018`, `king2019`, and `relanoiborra2019`. For `verhulst2015`, `verhulst2018`, and `relanoiborra2019`, we further observed a change in the location of their maximum amplitude (green triangles in Fig. 4, fourth row). Although a shift of the responses with increasing the stimulus level is supported by experimental data [76, 78, 80], this argument was later challenged [81] and rather attributed to shallower upper tails in the responses. A final observation is that `king2019`, `zilany2014`, `verhulst2015` and `verhulst2018`, showed a certain amount of distortion in the frequency responses of 70- and 100-dB SPL sounds (Fig. 4, second and third rows). For the responses to white noises at CF = 502 Hz, the distortions occur in the upper tail of the cochlear responses as a side effect of the (amplitude) compression stage. This frequency-response distortion is most prominent in `king2019` due to its

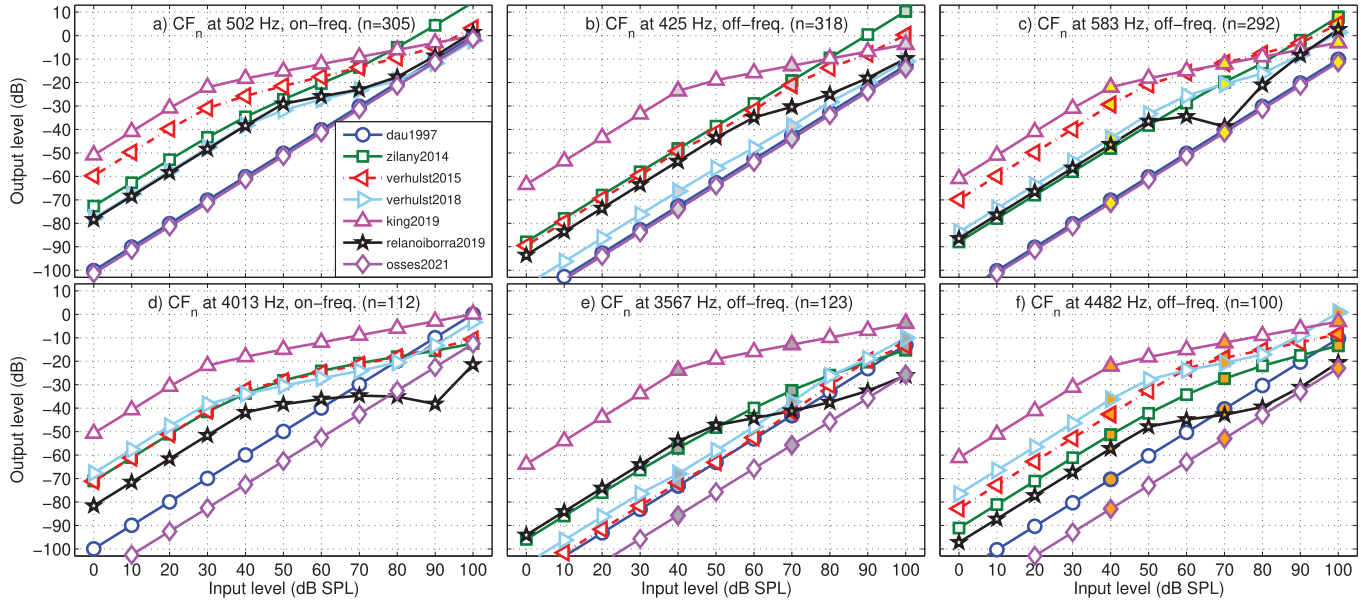

Figure 5. Input–output (I/O) curves for pure tones at 500 Hz (panels a–c) and 4000 Hz (panels d–f), at six model CF$_n$ frequencies (see Eq. (3)). Left (a,d): On-frequency simulations, i.e., output of the cochlear filter with the CF tuned to that of the stimulus frequency. Middle (b,e), right (c,f): Off-frequency simulations, one ERB below and above the on-frequency, respectively. The exact simulated on- and off-frequency CFs are indicated in the title of each panel. The filled markers indicate off-CF amplitudes that are also highlighted in the corresponding frequency responses of Figure 4. All I/O curves were shifted vertically by the reference gains given in Table 2 (see the text for details). *Literature*: Figures 1–3 from [79] and Figure 3 from [12].

broken-stick compression (rise of the amplitudes to the power of 0.3).[2]

4.1.1 Compressive growth

The set of pure tones was further used to assess the curves relating the input stimulus levels with levels at the output of the cochlear filter banks, known as input–output (I/O) curves. For this analysis we included stimulus levels between 0 and 100 dB SPL in steps of 10 dB. The I/O curves were obtained for (1) the on-frequency CF tuned to the frequency of the input stimulus, and (2) the off-frequency responses of cochlear filters tuned to one equivalent rectangular bandwidth number (ERB$_N$) [54] below and above the stimulus frequency.

The obtained I/O curves are shown in Figure 5 for on-frequency (left panels) and off-frequency simulations (± 1 ERB$_N$, middle and right panels). The I/O curves were vertically shifted by the reference gains indicated in Table 2 (as in Fig. 4). As expected for the level-independent Gammatone filters used in `dau1997` and `osses2021`, the curves were linear in all panels of Figure 5. For the remaining models, more compressive behaviour was observed for on-frequency curves (left panels) while more linear curves were obtained for off-frequency CFs (middle and right panels), except for `relanoiborra2019` and `king2019`, that had on- and off-frequency compression.

For `zilany2014/bruce2018`, the I/O curves were fairly linear in response to 500-Hz tones (top panels) for both on- and off-frequency CFs. For 4000-Hz tones, a prominent compressive behaviour was observed in the on-frequency curves (Fig. 5d) where, additionally, the curve for `verhulst2018` turned from a compressive to a linear regime for signal levels above 80 dB. The off-frequency I/O curves obtained for `verhulst2018` were similar to those for `verhulst2015` but had overall lower and higher amplitudes for the pure tones of 500 Hz (Fig. 5b–c) and 4000 Hz (Fig. 5e–f)), respectively, as a consequence of the differences in their middle-ear filters (see Fig. 3). The tendency to a more linear regime in off-frequency CFs has been shown previously [79]. This is in fact the basis for having compression only applied to the on-frequency channel in `king2019` [37, 72]. However, the default compression rate of 0.3 for the on-frequency channel with no compression for off-frequency channels leads to an unrealistic level balance between on- and off-frequency channels.

4.1.2 Frequency selectivity: Filter tuning

The frequency selectivity of each filter bank was computed in response to the described frozen noise waveform, presented at 40, 70, and 100 dB SPL. The estimates of frequency selectivity were obtained from FFT responses averaged across 500-ms non-overlapped analysis windows,

[2] Note that `king2019` was primarily developed to simulate responses to pure tones or narrow-band signals using cochlear channels with CFs that are either on-frequency (one cochlear channel) [110] or spanning ± 2 ERB$_N$ around the on-frequency CF (five cochlear channels) [37, 72]. The prominent frequency distortions shown in Figure 4e (third row) are thus outside the tested CF ranges for this model.

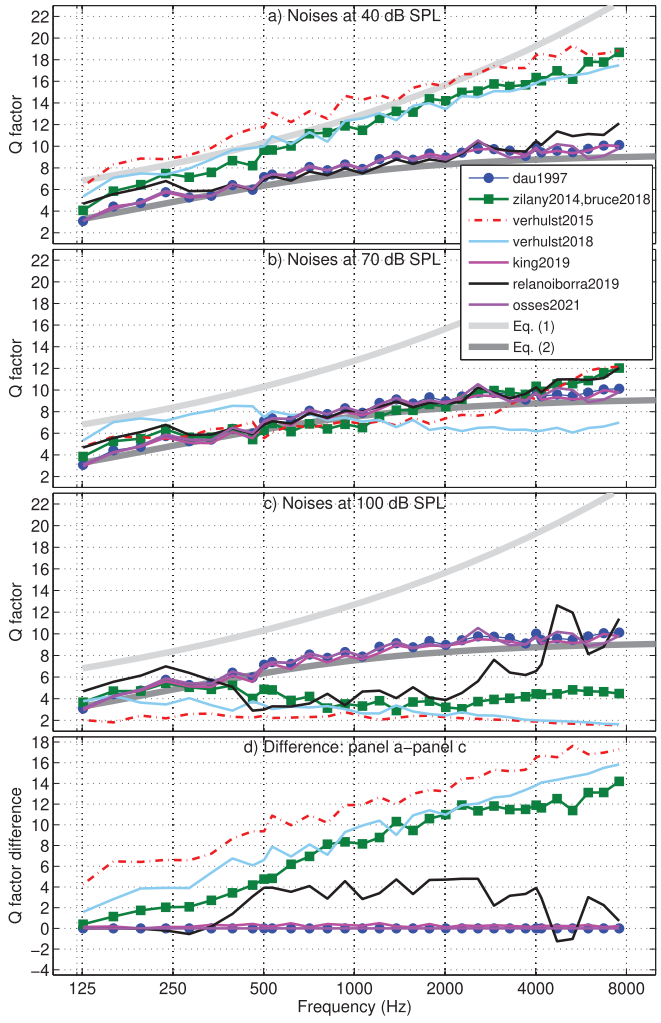

Figure 6. Filter tuning expressed as quality factors Q for noises of 40, 70, and 100 dB SPL (panels a–c), and Q-factor difference obtained from the results of 40- and 100-dB noises (panel d). *Literature*: Figure 4 from [53] and Figure 4B from [12].

meaning that the estimates were obtained from six statistically-independent noise sections.

The frequency response of thirty-two filters with CFs between 126 Hz (n = 396 in Eq. (3)) and 9587 Hz (n = 24 in Eq. (3)) at steps of n = 12 bins was obtained. For illustration purposes, we also included in this analysis the on-frequency CFs used in Figure 5 (CF = 502 Hz, n = 305; CF = 4013 Hz, n = 112), whose obtained responses are shown in Figure 4. For each filter response, a quality factor $Q_{-3\ dB}$ = CF/BW was obtained, where BW is the bandwidth defined by the lower and upper 3-dB down points of each filter transfer function (black dashed lines in Fig. 4).

The frequency-selectivity simulations for each of the filter banks are shown in Figure 6 for noises of 40 (panel a), 70 (panel b), and 100 dB SPL (panel c). The analytical filter tuning curves given by equations (1) and (2) are indicated as light and dark grey traces in Figure 6. Note that with this comparison, we assume that the Q factors within one ERB are similar to $Q_{-3\ dB}$ values. The results for 40-dB noises show that the frequency selectivity follows either the analytical tuning of equation (1) (`zilany2014`, `bruce2018`, `verhust2015`, and `verhulst2018`) or the tuning of equation (2) (`dau1997`, `relanoiborra2019`, `king2019`, and `osses2021`). When looking at the results for higher levels (Fig. 6b–c), no change in tuning was observed for `dau1997` and `osses2021`, as expected for linear models. For the nonlinear models, the results for 70-dB noises in Figure 6b showed overall lower Q factors, but with only a small change for `king2019` and `relanoiborra2019`. The results for 100-dB noises in Figure 6c showed a further lowering of Q factors in the biophysical and phenomenological models, reaching values as low as Q ≈ 2 in `verhulst2015`, lower Q factors for frequencies up to about 4000 Hz in `relanoiborra2019`, and virtually unaffected Q factors in `king2019`. A closer inspection to the outputs of `king2019` revealed that there was a filter broadening as a consequence of its broken-stick nonlinearity stage, but this broadening predominantly affected the frequency responses outside the range defined by the 3-dB bandwidth used to derive the Q factors (see Fig. 4e). To illustrate the Q-factor transition when increasing the signal level in each model, the difference between Q factors obtained from 40- to 100-dB noises is shown in Figure 6d, where a decrease in Q factor with increasing signal level is represented by a positive Q-factor difference.

Additionally, we observed that `relanoiborra2019` and `king2019` introduce a change in selectivity at overall higher levels compared to the biophysical and phenomenological models. A closer look at this aspect revealed that this change occurs because `relanoiborra2019` and `king2019` only apply compression after the bandpass filtering and, therefore, lower level signals are used as input for their compression (broken-stick) module.

4.1.3 Frequency selectivity: Number of filters

The number of filters in a filter bank is relevant for several model applications because too few filters can lead to a loss of signal information (e.g., [84]) and too many filters may unnecessarily increase the computational costs. The number of filters is a free parameter in `zilany2014`/`bruce2018`, but is fixed for `verhulst2015` and `verhulst2018` to yield an accurate precision of the transmission-line solver [85]. The remaining models use by default one ERB-wide bands (`dau1997`, `king2019`, and `osses2021`), or have an overlap every 0.5 ERB_N (`relanoiborra2019`).

Here, we report the minimum number of filters that are required to obtain a filter bank with overlapping at −3-dB points of the individual filter responses. Using the empirical Q-factors of Figure 6, we assessed the number of filters that would be required to cover a frequency range between 126 Hz (n = 396 in Eq. (3)) and the first filter with its upper cut-off frequency equal or greater than 8000 Hz. The number of filters derived from the 40-dB and 100-dB frequency tuning curves (Fig. 6, panels c and d) are shown

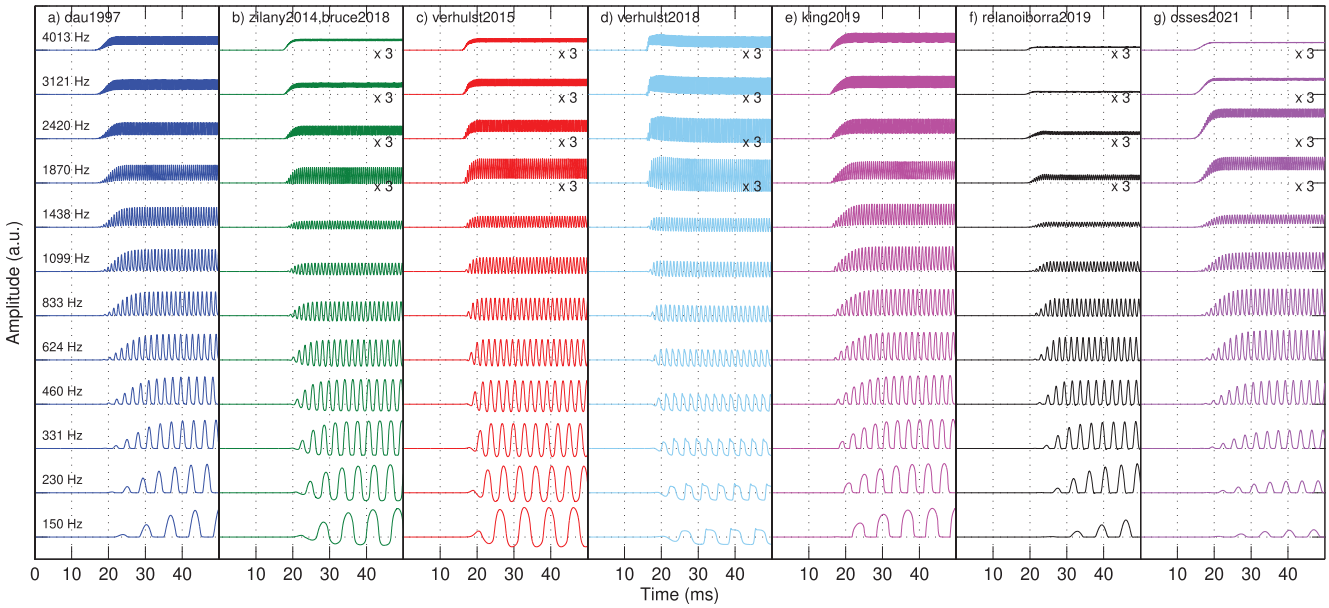

Figure 7. Simulated IHC responses to pure tones of different frequencies evaluated at the corresponding on-frequency bin. The amplitudes were normalised with respect to their maximum value to allow a direct comparison across models. *Literature*: Figure 9 from [82] and Figure 7 from [83].

in Table 2, including the average filter bandwidth in ERB_N for the corresponding model.

For the biophysical models, the filters were much wider at the higher level than for the other models, with average bandwidths being as wide as 3.05 ERB_N for `verhulst2015` and 2.30 ERB_N for `verhulst2018`. This contrasts with the 1.57 ERB_N for `zilany2014` and `bruce2018` and the 1.15 ERB_N or less for the remaining models. These bandwidths are a consequence of the fast-acting (sample-by-sample) compression that is applied just before the transmission-line in the biophysical models and the slower-acting bandwidth control in `zilany2014` (denoted as the "control path" in the chirp filter bank). While cochlear filters are generally wider at high sound levels (e.g., [76, 86]), the appropriate tuning must be evaluated depending on the species' characteristics, the tested CFs, and the type of evaluated excitation signals.

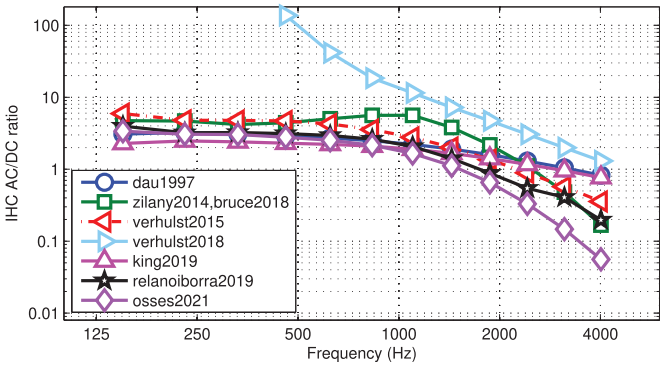

Figure 8. Ratio between simulated AC and DC components ($V_{\mathrm{AC}}/V_{\mathrm{DC}}$, see the text) in response to 80-dB pure tones. *Literature*: Figure 10 from [82] and Figure 8 from [83].

4.2 IHC processing: Phase locking to temporal fine structure

To illustrate the loss in phase locking to temporal fine structure with increasing stimulus frequency, we simulated IHC responses to pure tones with frequencies between 150 Hz ($n = 387$ in Eq. (3)) and 4013 Hz ($n = 112$ in Eq. (3)) spaced at $n = 25$ bins, resulting in 12 test frequencies. The tones were generated at 80 dB SPL, with a duration of 100 ms, and were gated on and off with 5-ms raised-cosine ramps. The simulated waveforms, that are assumed to approximate the IHC potential, are displayed and described in terms of AC (fast-varying) and DC (average bias) components, and the simulated resting potentials (V_{rest}). The AC potential was assessed from the peak-to-peak amplitudes as $V_{\mathrm{AC}} = V_{\mathrm{peak,max}} - V_{\mathrm{peak,min}}$. The DC potential was obtained as $V_{\mathrm{DC}} = (V_{\mathrm{peak,max}} + V_{\mathrm{peak,min}})/2 - V_{\mathrm{rest}}$ [82, 83].

The obtained IHC waveforms are shown in Figure 7. Within each panel, bottom to top waveforms represent on-frequency simulations for the test signals, from low to high frequency carriers, respectively. For some model outputs, the four highest carriers ($1870 \leq f_c \leq 4013$ Hz) were amplified by a factor of 3 to improve waveform visibility. The simulated voltages before the tone onset, i.e., the resting potential V_{rest}, were equal to 0 for all models except for `verhulst2018`, where V_{rest} was −57.7 mV (not schematised in Fig. 7). It seems clear, however, that the decrease of peak-to-peak AC voltage towards high frequencies—a measure of the residual amount of temporal fine structure—is significantly different across models. When increasing the CFs from 1099 to 4013 Hz, three models showed

V_{AC} reductions of less than 76.0% (king2019: 59.7%-decrease from $3.12 \cdot 10^{-3}$ to $1.26 \cdot 10^{-3}$ a.u.; dau1997: 62.6%-decrease from 0.097 to 0.037 a.u.; and verhulst2018: 76.0%-decrease from 39.2 to 9.4 mV), while the other five models showed V_{AC} reductions of at least 92.5%. From the low-frequency IHC waveforms (bottom-most waveforms in each panel), it can be seen that the simulated amplitudes of dau1997, king2019, relanoiborra2019, and osses2021 did not go below their V_{rest} (horizontal grid lines in Fig. 7) as a result of the applied half-wave rectification process. Furthermore, zilany2014/bruce2018 and verhulst2015 have $V_{peak,min}$ amplitudes of −66 mV and −4.7 mV, respectively. Despite the different range in their minimum voltages, there is a strong qualitative resemblance between waveforms (green and red traces in the figure). In fact, both these models use the same type of IHC nonlinearity (compare Eqs. (17)–(18) from [55] with Eqs. (4)–(5) from [34]) and the same LP filter implementation, albeit with a different filter order and cut-off frequency (see Tab. 2).

The obtained AC/DC ratios are shown in Figure 8, where a reduction in phase locking is given by a lower ratio. For all models, the ratio decreased with increasing frequency. All AC/DC curves, except those for verhulst2018, overlap well at low frequencies with ratios between 2.1 and 5.9 (below 1000 Hz), decreasing to ratios between 0.06 (osses2021) and 0.83 (dau1997) at 4013 Hz. Although the AC/DC curve for verhulst2018 showed the highest overall ratios between 137.4 at 460 Hz down to 1.3 at 4013 Hz, due to its nearly zero DC voltages towards low frequencies (see the fairly symmetric waveforms around the horizontal grid in Fig. 7d), we still observed the systematic decrease in ratio with increasing frequency. If we further focus on the AC/DC curves in the frequency range between 600 and 1000 Hz, where the phase-locking is expected to start declining [82], all models showed monotonically decreasing curves starting from about 833 Hz (except for verhulst2018, that always showed a decreasing tendency). The lowest ratios were observed for osses2021, followed by the similarly steep curve of zilany2014. Finally, a similar AC/DC curve was obtained for relanoiborra2019 and verhulst2015.

4.3 AN firing patterns

Simulations included AN responses to pure tones and to amplitude-modulated (AM) tones from which rate-level functions expressed as onset and steady-state responses were obtained. With these benchmarks we attempt to characterise model responses at the output of the AN synapse stage or their equivalent, with a particular interest in the phenomenon of adaptation [60, 75]. We comment on how adaptation is affected by the type of output of Stage 5, using either the approximations from the effective models, the average or instantaneous firing rate estimates of the phenomenological models (zilany2014, bruce2018), or the average rates of the biophysical models (verhulst2015, verhulst2018).

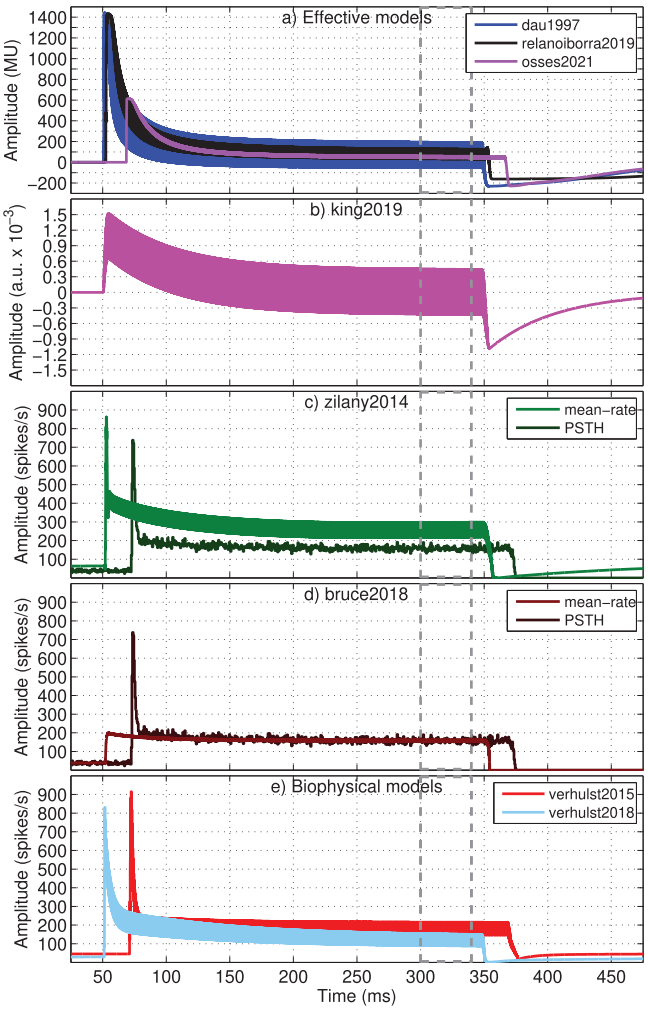

Figure 9. Simulated AN responses to a 4000-Hz pure tone of 70 dB SPL. For ease of visualisation, the responses from osses2021, verhulst2015, and the PSTHs are horizontally shifted by 20 ms. *Literature*: Figure 1 from [87] and Figures 3 and 10 from [36].

4.3.1 Adaptation

To illustrate the effect of auditory adaptation, we obtained AN model responses to a 4000-Hz pure tone of 70 dB SPL, duration of 300 ms, that was gated on and off with a cosine ramp of 2.5 ms. The obtained AN responses are shown in Figure 9. All responses had an amplitude overshoot just after the tone onset which then decreased to a plateau (e.g., between 300 and 340 ms, grey dashed lines). After the tone offset ($t = 350$ ms), the AN responses showed an undershoot with decreased amplitudes that subsequently returned to their resting level. This stereotypical behaviour is related to the AN adaptation process (e.g., [60]).

The waveforms from effective models using the adaptation loops (dau1997, relanoiborra2019, osses2021) are shown in Figure 9a, where their amplitudes had values between −230.5 MU and 1440.2 MU

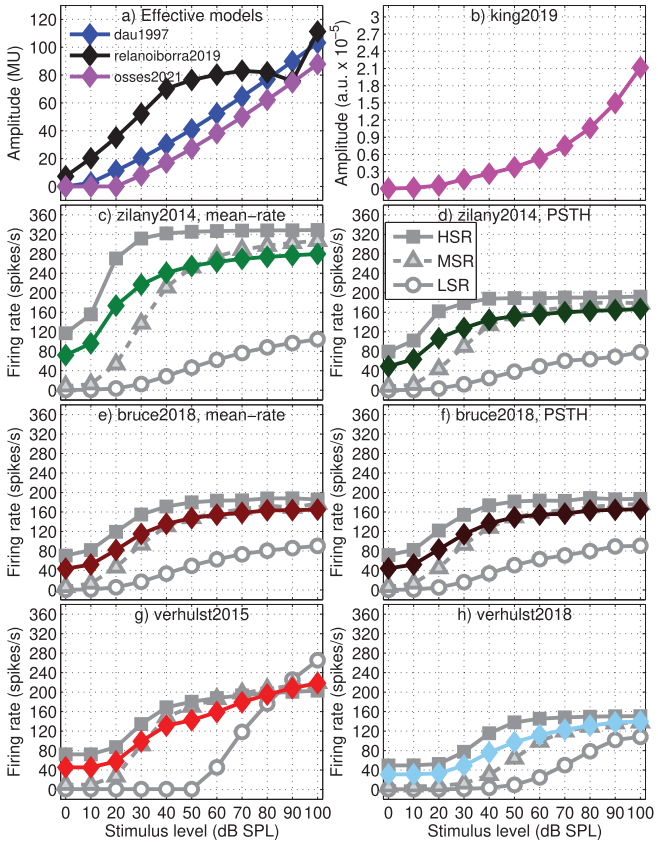

Figure 10. Simulated rate-level functions derived from the steady-state AN responses of 4000-Hz pure tones. For all models, average responses are shown (coloured traces). For the biophysical and phenomenological models, the responses for HSR, MSR, and LSR neurons are also shown (grey traces). *Literature*: Figure 7 from [36], Figure 5A from [35], and Figure 3 from [44].

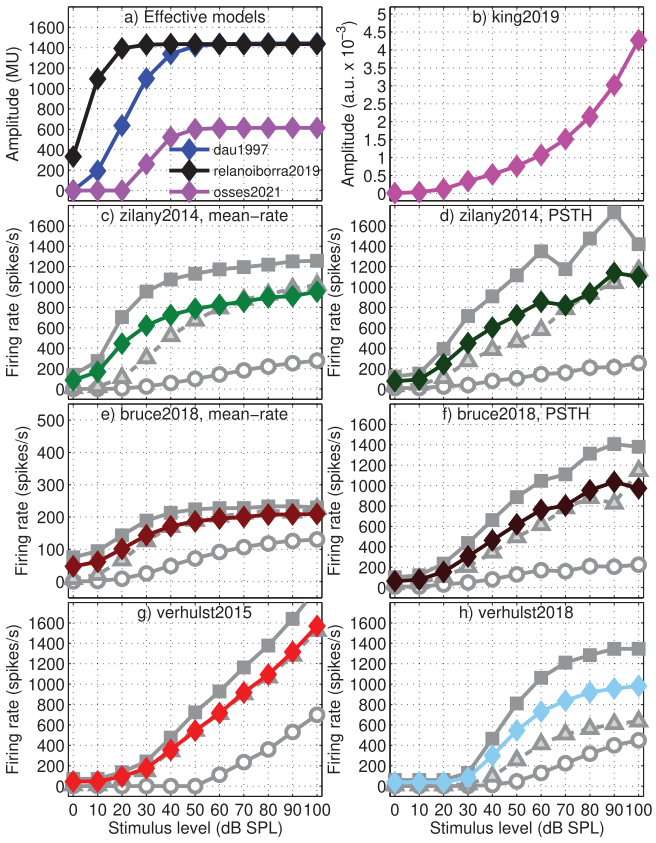

Figure 11. Simulated rate-level functions derived from the onset (maximum) AN responses of 4000-Hz pure tones. The colour codes and legends are as in Figure 10. *Literature*: Figure 3 from [87].

(`dau1997`), with a strong onset overshoot and a resting position at 0 MU. For `king2019` (Fig. 9b), a mild overshoot was observed, whose maximum amplitude ($1.52 \cdot 10^{-3}$ a.u.) was higher in absolute value than that for the undershoot ($-1.08 \cdot 10^{-3}$ a.u.). With an observed steady-state peak-to-peak amplitude of $0.87 \cdot 10^{-3}$ a.u. `king2019` is, at this stage, the model that preserves the most temporal fine structure.

For the phenomenological models (`zilany2014` and `bruce2018`), the simulated waveforms using their two types of AN synapse outputs are shown in Figure 9c–d, based on a PSTH (dark green or brown curves) and mean-rate synapse output (light green or brown curves). The obtained PSTH and mean rate responses in `zilany2014` differ in their steady-state values (lower values for the PSTH estimate), while for `bruce2018` the difference is in their onset responses, with almost no onset adaptation in the simulated mean-rate output. For the biophysical models (Fig. 9e), the AN synapse outputs represent mean firing rates where a stronger effect of adaptation was observed for `verhulst2018` (sky blue), with a plateau after onset that was reached after about 150 ms (at $t \approx 200$ ms) while for `verhulst2015` (red) the plateau is reached shortly after the tone onset.

4.3.2 Rate-level functions

Rate-level functions were simulated for a 4000-Hz pure tone presented at levels between 0 and 100 dB SPL with a duration of 300 ms, gated on and off with 2.5-ms cosine ramps. The obtained results are shown in Figures 10 and 11 for rate-level curves in the steady-state regime and for onset responses, respectively. For all models, average rates are shown (coloured traces) while for the phenomenological and biophysical models (panels c–h), the simulated response for the three types of neurons (HSR, MSR, and LSR) are shown (grey traces).

For the phenomenological and biophysical models, the discharge curves in Figure 10c–h tend to saturate towards higher levels, which is in line with experimental evidence (e.g., [87]). One difference between these curves is that they start to increase at slightly higher levels for the biophysical (from ~20 dB SPL) than for the phenomenological models (from ~0 dB SPL).

For the effective models (Fig. 10a and b), with the exception of `relanoiborra2019`, the simulated rates did not show saturation as a function of level.

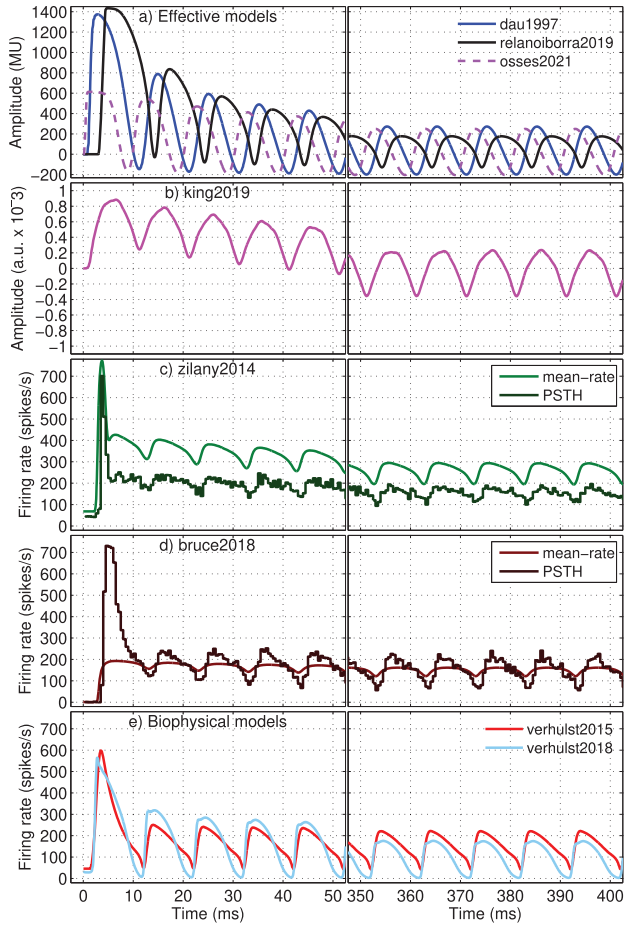

Figure 12. Simulated on-frequency AN responses to a 4000-Hz AM tone of 60 dB SPL (100% modulation, $f_{\mathrm{mod}} = 100$ Hz). Left: Onset responses. Right: Steady-state responses. *Literature*: Figure 12 from [75] and Figure 3C from [35].

In `relanoiborra2019`, the simulated rates were between 70.2 and 83 MU for signal levels beyond 40 dB. This saturation effect results from the combined action of the nonlinear cochlear filter (Stage 3) with the later expansion stage (Stage 5, Fig. 2) that precedes the adaptation loops. Despite the overall lack of saturation in the evaluated effective models when looking at the steady-state outputs, a different situation is observed for the onset responses of Figure 11, where the responses of the models using adaptation loops had a prominent onset saturation (`dau1997`: 1443 MU for levels above 50 dB; `relanoiborra2019`: 1435 MU for levels above 30 dB; `osses2021`: 614 MU for levels above 50 dB). Other interesting aspects to highlight are that: (1) almost no onset effect is observed in the mean-rate output of `bruce2018`; (2) `king2019` does not account for any type of saturation as the signal level increases (Figs. 10b and 11b). It should be noted that although hard saturation (as in Fig. 11a) has not been experimentally observed for onset AN responses, a decrease in the rate of growth of onset-rate curves with level is expected [87], a condition that is not met in `king2019` or `verhulst2015`.

4.3.3 AM model responses

Model responses were obtained for a 500-ms 4000-Hz pure tone that was sinusoidally modulated in amplitude (modulation index of 100%) at a rate $f_{\mathrm{mod}} = 100$ Hz, presented at 60 dB SPL, including up/down ramps of 2.5 ms. The initial (0–50 ms) and later (350–400 ms) portions of the simulated responses are shown in the left and right panels of Figure 12, respectively. In all models, the modulation rate of 100 Hz is visible as amplitude fluctuations with the corresponding periodicity of 10 ms. In addition, adaptation was observed with stronger simulated responses immediately after the tone onset (left panels) than during the steady-state portion of the response (right panels).

For the effective models with adaptation loops (`dau1997`, `relanoiborra2019`, `osses2021`), the maximum amplitudes (Fig. 12a, left) were much lower in `osses2021` than for `dau1997` and `relanoiborra2019`, due to the stronger overshoot limitation. For these models, it was also observed that their phases are not perfectly aligned due to the outer- and middle-ear filters that introduced a delay into `relanoiborra2019` (black traces run "ahead" the blue traces of `dau1997`), while the group-delay compensation in `osses2021` (Sect. 2.2) seemed to overcompensate the alignment of the simulated waveforms (purple traces run "behind" the blue traces). In the right panel, the dynamic range of `relanoiborra2019` (black traces) is lower than for `osses2021` and `dau1997`, which have very similar steady-state amplitudes. The reduced dynamic range in `relanoiborra2019` is mainly due to the nonlinear cochlear compression of the filter bank that interacts further with the expansion stage. In `king2019` (Fig. 12b), a small effect of adaptation was observed with a maximum onset response of $0.88 \cdot 10^{-3}$ a.u. (left panel) that decreases to a local maximum amplitude of $0.24 \cdot 10^{-3}$ a.u. during the steady-state response (right panel).

The AN responses produced by `verhulst2015` and `verhulst2018` (Fig. 12e) showed an overshoot reaching firing rates of 598.5 and 565.2 spikes/s, respectively. After the onset, the overshoot effect quickly disappeared in `verhulst2015`, reaching a maximum local rate of 251 spikes/s during the second modulation cycle and 222 spikes/s between 370 and 400 ms. In contrast, `verhulst2018` adapted more slowly after the onset with a maximum rate of 319 spikes/s in response to the second modulation cycle, while the response continued adapting reaching a maximum rate of 176 spikes/s between times 370 and 400 ms.

For `zilany2014` (Fig. 12c) and `bruce2018` (Fig. 12d), the mean-rate and PSTH outputs are shown as lighter and darker traces, respectively. It can be observed that in `zilany2014`, the AM modulations showed a similar mean-rate and PSTH excursions of about 100 spikes/s (Fig. 12b, right: mean rates between 194 and 295 spikes/s; PSTHs with rates between 94 and 196 spikes/s), but the PSTHs had overall lower rates. In `bruce2018`, a greater AM fluctuation is observed for the PSTH outputs

(darker brown traces) with an excursion of 185 spikes/s (Fig. 12d, right: rates between 56 and 241 spikes/s) compared with the 40 spikes/s (rates between 121 and 161 spikes/s) of its mean-rate output. Additionally, `bruce2018` showed a limited effect of adaptation in its mean-rate outputs, along with a shallower AM response in comparison to the obtained PSTH. We will not focus on the mean-rate output of this model, because (1) their authors validated the model primarily using PSTHs, recommending the use of the AN synapse output for further processing [36], (2) the model using PSTH outputs can be used as input for subcortical processing stages from the UR EAR toolbox [66], and (3) all the studies that we have so far identified using `bruce2018` consistently used PSTH outputs [88, 89].

It should be noted that `zilany2014`, from the same model family, has been extensively validated using both mean-rate and PSTHs outputs. In fact, for studies where psychoacoustic aspects have been investigated (e.g., [65]) there is a tendency to use the mean-rate model outputs.

4.3.4 Synchrony capture

Model responses were obtained for a complex tone of 50 dB SPL formed by three sinusoids of equal peak amplitude and frequencies of 414 Hz (9.6 ERB_N), 650 Hz (12.6 ERB_N), and 1000 Hz (15.6 ERB_N). This type of complex tone with more carriers and greater range of frequencies is commonly used in studies of profile analysis (e.g., [65]) and it is useful to explain an AN property named "synchrony capture" [57, 63] that is believed to play a relevant role in the neural coding of supra-threshold speech sounds [33, 63]. When synchrony capture occurs, the neural activity in on-frequency channels is driven primarily by one frequency component in the harmonic complex, such that there are minimal fluctuations due to the fundamental-frequency envelope, while off-frequency channels exhibit fluctuating AN patterns at the fundamental frequency. To illustrate whether the evaluated models account for synchrony capture, the model outputs in response to the described complex tone were obtained for frequencies between 415 Hz ($n = 320$ in Eq. (3)) and 1007 Hz ($n = 245$ in Eq. (3)) for CFs spaced at approximately 1 ERB_N ($\Delta n = 12$ or 13), resulting in three on-CF and four off-CF channels. The obtained simulations are shown in Figure 13 for a 30-ms window (between 220 and 250 ms). For each waveform, a schematic metric of envelope fluctuation was obtained and shown as thick grey lines. Those envelope fluctuations were constructed by connecting consecutive local maxima that had amplitudes above the mean responses (onset excluded) of each simulated channel. Subsequently, the standard deviation of the obtained envelope estimate was (arbitrarily) divided by one thirtieth of the amplitude scales shown in the insets of each panel (e.g., divided by 800/30 MU for `dau1997`, `relanoiborra2019`, and `osses2021`). The obtained estimates were drawn as maroon circles and connected with dashed lines along the right vertical axes in Figure 13

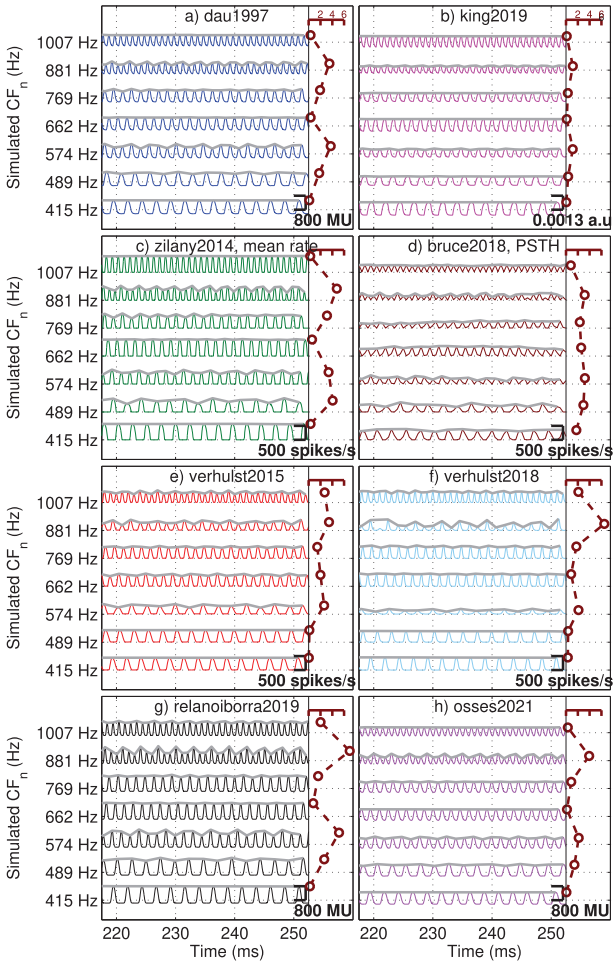

Figure 13. Simulated AN responses to a complex tone with three frequency components at 414, 650, and 1000 Hz. The model simulations were obtained at on- and off-frequency CFs spaced at 1 ERB_N. The thick grey lines represent the envelope of the AN responses. The maroon circle markers represent a metric that is proportional to the standard deviation of the corresponding envelope (see the text for details). *Literature*: Figures 7–8 from [90] and Figure 1 from [91].

(dimensionless scale with labels between 0 and 6, as indicated in panels a and b), where higher values indicate greater envelope fluctuation variability. The resulting envelope scale allows for a direct comparison between models. In Figure 13 it can be observed that for all models, the on-frequency channels had nearly flat envelope fluctuations. The variability estimate averaged across on-frequency bins (at 415, 662, and 1007 Hz) ranged between 0.11 (`king2019`) and 1.71 (`bruce2018`). The variability estimate across off-frequency bins (at 489, 574, 769, and 881 Hz) ranged between 0.74 (`king2019`) and 4.09 (`relanoiborra2019`, with a maximum deviation of 6.95 at CF = 881 Hz in Fig. 13g). For all models the off-CF variability was greater than the on-CF variability, with `king2019` being the least sensitive model to code envelope fluctuations.

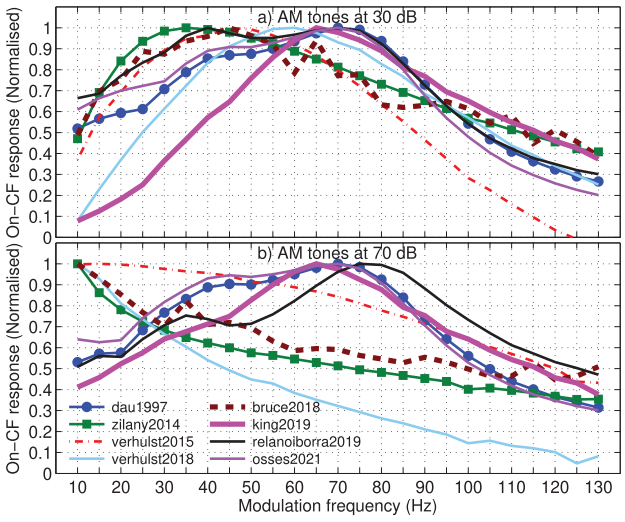

Figure 14. Modulation transfer functions (MTFs) of a modulation filter with a BMF ≈ 80 Hz, assessed using 1000-Hz AM tones presented at 30 (panel a) or 70 dB SPL (panel b) that were sinusoidally modulated with f_{mod} frequencies between 10 and 130 Hz. The MTFs are normalised to the maximum model response across the tested f_{mod} frequencies. *Literature*: Figures 4–6 from [92] and Figures 1 and 4 from [93].

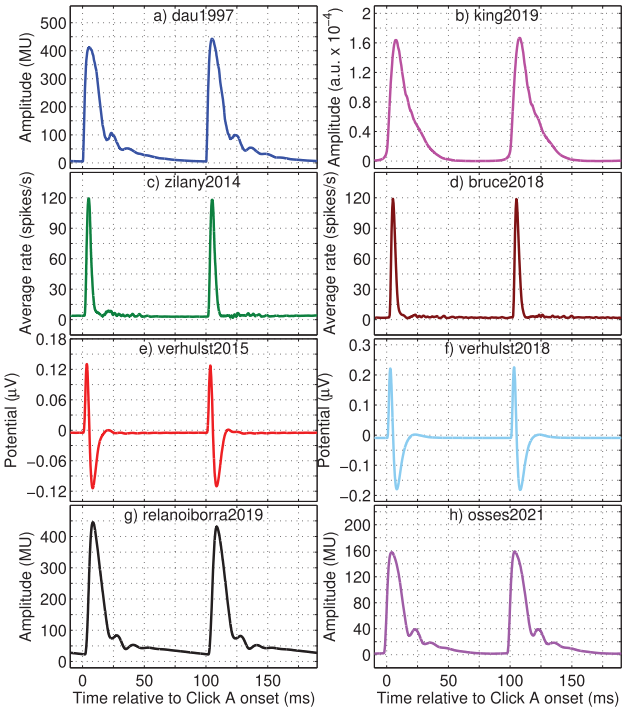

Figure 15. Simulated IC responses using one modulation filter (BMF ≈ 80 Hz) to a click train of alternating polarity with a total duration of 1 s, repetition rate of 10 Hz and click duration of 100 μs. Only the responses to the two last clicks are shown, whose peak-to-peak amplitudes are indicated in Table 2. *Literature*: Figure 1 from [94] and Figures 8–9 from [95].

4.4 Subcortical neural processing

We show two sets of figures to schematise the subcortical processing of the evaluated models.

4.4.1 Modulation transfer function

The first set of figures represents a modulation transfer function (MTF) in response to 100% AM tones modulated at f_{mod} rates between 10 and 130 Hz (steps of 5 Hz). The tones were centred at 1000 Hz, had a duration of 300 ms, included 5-ms up/down ramps, and were presented at 30 and 70 dB SPL. For this analysis, 100 ms in the last portion of the simulated responses were used (between times 190 and 290 ms). The MTFs were derived from the maximum of the simulated responses. The responses were normalised to the corresponding maximum estimate over the set of tested f_{mod} values, so that the MTF of each model had a maximum value of 1. The resulting MTFs are shown in Figure 14.

The results in Figure 14 show that the models produce bandpass-shaped MTFs with estimated BMFs between 35 Hz (`zilany2014`) and 70 Hz (`dau1997`, `relanoiborra2019`, and `osses2021`) that are below the theoretical BMFs (see Tab. 2). It is interesting to observe that the sharpest MTFs were obtained not only for `dau1997` and `osses2021` (both designed with Q = 2), but also for `king2019` (which has a Q = 1), while a wider tuning was observed for the remaining models, including `relanoiborra2019` (which has a Q = 2).

For the biophysical and phenomenological models, the MTFs obtained for the 70-dB AM tones (Fig. 14b) were different than those obtained for 30 dB (Fig. 14a). For these models, the MTFs were no longer bell-shaped and seemed to act as lowpass filters, which is inline with physiological evidence indicating that regions of "enhancement" in MTFs of low level-signals can become regions of "suppression" for higher presentation levels (see, e.g., Fig. 4 from [92]).

The effective models were more insensitive to the change in presentation levels. The only exception to this is `relanoiborra2019`, where a narrower MTF was obtained in Figure 14b (compared with panel a). The models `dau1997`, `osses2021`, and `king2019` have MTFs that are qualitatively similar across presentation levels.

4.4.2 Response to clicks of alternating polarity

The second set of figures focuses on simulating the response to a typical click train as used in the assessment of auditory brainstem responses (ABRs) [95]. We used a click train with a repetition rate of 10 Hz and a duration of 1 s (i.e., containing 10 clicks). The clicks had an alternating polarity (amplitude A or $-A$) and were presented at 70 dB peak-equivalent SPL (dB peSPL) [96], i.e., using $A = 0.1789$ Pa. Each individual click had a duration of 100 μs. For this processing, the simulated outputs of Stage 6 of each model (see Fig. 2) were averaged across CFs to obtain a broadband representation, i.e., all simulated representations were added together and then divided by the

number of CFs [34, 35]. This type of output can be used to derive a peak-to-peak or peak-to-trough amplitude correlate of the wave-V ABR component [95].

For this processing, we used the default number of CFs for the biophysical and effective models, while for `zilany2014` and `bruce2018`, 50 CFs were obtained between $\mathrm{CF}_n = 133.7$ Hz ($n = 393$, Eq. (3)) and $\mathrm{CF}_n = 12010$ Hz ($n = 1$), spaced at $n = 8$ bins to roughly meet the number of filters from Table 2. The obtained click responses are shown in Figure 15 and are illustrated for the last two clicks (of amplitudes A and $-A$) of the test click train.

The biophysical models provided click responses that had positive and negative amplitudes (Figs. 15e–f), which was not the case for the phenomenological models that also use the SFIE model. This is because `verhulst2015` and `verhulst2018` assume that a population response can be obtained from the sum of single neuron activity (as, e.g., in [56]), with no half-wave rectification in the SFIE model (a non-explicit choice of the authors [34, 35]) that, after scaling [35, 41], results in a simplified neural representation that correlates with changes in electrical dipoles visible in scalp-recorded potentials [35].

The effective models, that use the modulation-filterbank concept, showed only positive amplitudes for all filters with BMFs ≥ 10 Hz [31] due to their envelope extraction, a phase-insensitive ("venelope") processing [37, 70]. For modulation frequencies below 10 Hz, the perceptual models preserve the phase information, something that is not illustrated in Figure 15 (nor in Fig. 14).

Finally, the simulated peak-to-peak amplitudes in response to the last positive and negative clicks of the pulse train (ninth and tenth click, shown in Fig. 15) are shown in the entries "Click A" and "Click $-A$" of Table 2. From those amplitudes, it can be observed that there are models that have higher peak-to-peak amplitudes in response to positive clicks (`zilany2014`, `verhulst2015`, `bruce2018`, `relanoiborra2019`) and others where higher amplitudes are observed in response to the clicks of negative polarity (`daul997`, `verhulst2018`, `king2019`, `osses2021`). Although we do not discuss the significance of this polarity sensitivity, this aspect has been a matter of discussion, in particular for electrical hearing, where it has been found that evoked potentials in response to positive and negative polarity clicks represent one of the differences between humans (e.g., [97]) and other mammals (e.g., [98]), whose responses are more sensitive to stimulation with clicks of negative and positive polarities, respectively.

4.5 Computational costs

The computational cost required to run each model was measured using the same click train as described in the previous section. Therefore, we assessed the time required to process an input signal of 1-s duration between Stages 1 and 6 of each model (Fig. 2). This metric aims at providing a relative notion of the processing times across models. Note that some model implementations can use parallel processing, which was disabled in this evaluation. The assessment was performed on a personal computer equipped with an Intel Core i5-10210UR, 1.6-GHz processor with 16 GB of RAM memory.

The results of the computational costs used by each model are given in the entry "Performance" of Table 2. The time required by the models to process one frequency channel ranged between ~0.02 s (`osses2021`, `daul997`) and 2.5 s (`bruce2018`). For individual frequency channels, the biophysical models (`verhulst2015` and `verhulst2018`) showed moderate calculation times between 0.3 and 0.8 s. However, these models always require (internally) the simulation of the whole discretised cochlea with 1000 cochlear sections, independent of the number of user-requested cochlear channels (default number of 401 for the Verhulst models). This means for the current simulations, that the reported processing times of 122.9 and 319.5 s for `verhulst2015` and `verhulst2018`, respectively, cannot be further reduced, even if the user requests the simulation of fewer CFs. In contrast, in any model based on a parallel filter bank, including `zilany2014` and `bruce2018`, each cochlear section is independent of each other, and a user-defined number of frequency channels can be simulated, which vastly reduces the computation time for different model configurations.

Due to the long processing time of the evaluated biophysical models, their implementations include an option of parallel processing (also available in the original implementation of `bruce2018` [36]), where multiple input signals can be processed simultaneously. The number of signals that can be processed in parallel will depend on the number of threads of the host computer. As a further solution to the long processing time, Stages 2-5 of `verhulst2018` (transmission-line, IHC, and AN modules) and `bruce2018` (generating mean PSTHs) have been approximated using deep neural networks in [99, 100] and [101], respectively.

5 Models in perspective

The stimuli and comparison measures used in our evaluation (Sect. 4) were chosen to reflect relevant temporal and spectral properties of the models in a normal-hearing condition. Our evaluation provides an objective view, accompanied by a graphical representation of how the model responses reflect specific aspects of the hearing process in their model structure.

The content of this study might be considered as a guideline for model selection, but the motivation was not to select a "winner" among the different evaluated models. We compared models which were verified in different experimental conditions or in connection to different hearing applications, and we only presented raw model outputs (Figs. 7, 9, 12, 13, 15) or outputs transformed to characterise specific hearing properties (Figs. 4–6, 8, 10, 11, 14). These outputs reflect model responses to a very specific dataset that may not be suitable to appropriately verify all models. Any model, though, to be informative, needs to be verifiable and falsifiable, such that it is possible to understand both its essential characteristics and features

that explain the predictive power in a given range of conditions, as well as its limitations that make transparent where the model fails. In the following sections we provide a brief overview of the context in which each of the selected models has been used and include general recommendations for further applications.

5.1 Applications of the evaluated auditory models

`Dau1997` is a monaural model that has been used to simulate a number of psychoacoustic tasks including tone-in-noise and AM detection experiments using a forced-choice paradigm (e.g., [31, 44]). To enable the model for the comparison between two or more sounds, the output of Stage 6 (Fig. 2) is used as input to a decision back-end based on a signal-detection-theory (SDT) framework, the template-matching approach. This framework, extended to adopt two templates, has been recently validated to account for the perceptual similarity between two sounds using `osses2021` [39].

The models `zilany2014` and `bruce2018` can account for elevated hearing thresholds due to OHC ("Cochlear gain loss" in Stage 3) or IHC impairment ("IHC loss" in Stage 4) [55]. The AN stage (Stage 5) includes two types of outputs: An actual spike generator and an analytical mean-rate synapse output. The spike generator has been primarily used to simulate physiological data, including the phenomenon of short- and long-term adaptation [75]. The mean-rate synapse output using `zilany2014` has been used to simulate specific psychoacoustic tasks [65, 102], including speech intelligibility predictions [103].

The models `verhulst2015` and `verhulst2018` were initially designed to simulate otoacoustic emissions [26] and can account for elevated hearing thresholds due to OHC impairment ("Cochlear gain loss" in Stage 3). Furthermore, they allow to study effects of the gradual disconnection of AN fibres, known as synaptopathy, on auditory brainstem responses [35, 104]. When coupled with a decision back-end, they have been used to simulate psychoacoustic performance in simultaneous tone-in-noise and high-rate AM tasks (f_{mod} ~100–120 Hz) [105, 106].

The model `relanoiborra2019` can predict speech intelligibility [38] when coupled with a decision back-end stage [38, 107]. Relying on the prediction power of earlier model implementations [77, 108], `relanoiborra2019` should be able to (1) account for elevated thresholds based on OHC and IHC impairment [108], and (2) to predict a number of psychoacoustic tasks including simultaneous and forward masking and amplitude modulation [77]. Our results showed that `relanoiborra2019` accounts well for hearing properties such as nonlinearities in the cochlear processing and auditory adaptation, including a saturation behaviour similar to that of the AN physiological models.

The model `king2019` was designed to simulate perceptual tasks of amplitude- and frequency-modulation detection, primarily at low modulation rates ($f_{\mathrm{mod}} \leq 20$ Hz). The model's decision back-end includes deterministic limitations (suboptimal template matching strategies) or stochastic limitations such as internal additive noise, multiplicative noise [109], and memory noise [72, 110]. The model can be adapted to simulate hearing impairment by modifying its compression parameters (knee point and compression rate), and by increasing the bandwidth of the underlying cochlear filters. Despite the simplicity of this model—in fact one of its strengths—we have shown in this paper that the model can account for several of the comparison metrics, with the exception of the broadening of cochlear filters at higher presentation levels (Fig. 6), the adaptation saturation (Figs. 10–11), and the coding of fluctuation profiles (with minimal difference in amplitude fluctuations in Fig. 13).

5.2 Other applications of auditory models

Apart from the listed applications, auditory models have also been used in several other applications such as sound quality assessment (e.g., [10, 111–113]), prediction of speech intelligibility (e.g., [103, 114]), and automatic speech recognition (e.g., [115]).

In the context of this special issue on binaural hearing, it is worth mentioning a number of binaural applications that rely on the evaluated monaural auditory models: The low-pass modulation filter (similar to `dau1997`) [31, 44] served as the basis for a model of binaural masking that uses a decision stage based on the equalisation-cancellation theory [116]. This model was later extended to predict perceptual attributes of room acoustics [117–119]. The model `zilany2014` has been used to predict (1) the sensitivity to interaural time and level differences by estimating the disparity between left and right AN responses using a decision back-end based on shuffled cross-correlograms [120], and (2) the median-plane sound localisation for various profiles of sensorineural hearing loss (OHC impairment) [121]. Finally, `bruce2018` has been used to simulate the lateralisation of high-frequency stimuli in a coincidence-counting model [88].

5.3 Simplified auditory representations

When an auditory model is used to broaden our understanding of auditory processes [1, 2], it is required that the model be as complete as possible. More details in the model often come at the price of a more computationally-expensive implementation. Such a level of detail is represented in the selected biophysical and phenomenological models, that attempt to shed light on the mechanisms behind the cellular and neural elements included in auditory processing. On the other hand, effective models have a more epistemic status providing an intelligible but simplified representation of the process. These models can guide the design of new experiments or facilitate the development of listener-targeted products. Such model simplification, however, potentially reduces the number of effects a model can account for, leading to an actual narrowing of its application field. An example of a successful model simplification is presented in [27], where MTFs were simulated

using only a stage of envelope extraction followed by a modulation filter bank, omitting the stages of cochlear filtering and auditory adaptation. This model, however, is not thought to predict the performance in listening conditions where the omitted model stages do play a role as it is the case (in this example) for forward-masking tasks.

Peripheral auditory models are often combined with a decision back-end module converting simulated responses into (1) a behavioural response that reflects detectability or discriminability of a sound (e.g., [2]), or into (2) perceptual metrics to estimate, e.g., loudness [28], perceived reverberation [117, 118], and sound-source localisation [122, 123]. For successful simulations, the decision stage should appropriately weight the information contained in the model representations. An analysis of weighted time-frequency representations (time, audio frequency, and/or modulation frequency) can reveal what portions of the simulated responses are more relevant (e.g., [39, 124]).

It is important to note that the simplification of auditory models based on statistical methods or machine learning processes requires a careful interpretation. While these approaches might be well suited to achieve goals such as real-time processing (e.g., [99]) in applications of speech perception (e.g., [101]) or in the prediction of evoked potentials [89], they limit the modular comprehension of each auditory stage, especially if multiple model stages are approximated (as in [89, 101]).

In a recent study [89], firing rates of cortical A1 neurons in ferrets were approximated using several time-frequency representations ranging from simple short-time Fourier transforms to more detailed models of AN synapses (including `bruce2018`) to which a linear-nonlinear (LNL) encoder was used. Based on their separately-fitted encoders, the authors concluded that cortical processing in ferrets perform a "very simple signal transformation," without discussing how different the linear and nonlinear components in each of their encoders were. Despite the success of the authors in approximating neural responses in ferrets, we believe that it is difficult to know whether the "simple transformation" is indeed related to the underlying mechanisms of hearing (the cortical processing in ferrets) or rather is related to the complexity of operations in the fitted encoders.

5.4 Considerations for further modelling work

The following is a list of aspects that we recommend to keep in mind for further auditory modelling work, based on the general observations of this study:

- If the evaluated sounds are assumed to be reproduced via loudspeakers, or supra-aural or circumaural headphones, we recommend to use an outer-ear module as in `relanoiborra2019`, `osses2021`, or to apply an HRTF (as in [57]). Although we did not evaluate this effect, such an omission implicitly assumes that the outer ear (compare the grey and black lines for `relanoiborra2019` and `osses2021` in Fig. 3) does not influence the coding of incoming signals in the ascending auditory pathway.
- The results in Figures 6 and 14 show that there are nonlinear interactions between model stages as a function of level and for different types of signals. This suggests that different sets of stimuli are required to characterise the behaviour of complex processes such as that of nonlinear filter banks. In other words, models may not always act as a linear time invariant (LTI) system.
- In Section 4.1.3, we suggested a minimum number of filters for each filter bank to roughly meet a −3-dB filter crossing (Tab. 2, "40 dB: Number of bands"). The required number of bands may vary from application to application and depend on the type of sounds that are to be simulated. This choice can be particularly critical in models where the number of bands are a free parameter (here `zilany2014` and `bruce2018`). For models that are used as front-ends to machine-learning applications, Lyon [22] suggested a "not-too-sparse set of channels" with about a 50% overlap between filters, i.e., twice the number of channels that we recommend in Table 2. It is important to keep in mind, however, that our estimation was based on model responses to white noises, which are sustained signals in time and broadband in frequency. At higher presentation levels, where nonlinear filter banks act as compressors, similar estimations using sine tones (sustained narrowband signals, as in Fig. 5) or clicks (transient broadband sounds) may result in a different number of required bands.
- Different simulation results can be expected when evaluating mean-rate and PSTH outputs of models including AN synapse stages as shown in Figures 9–12. The particular choice of the type of output depends on the target application of the model. The spike generator is primarily used to simulate physiological data (e.g., [36, 75], while the mean-rate synapse output is typically used to simulate specific psychoacoustic tasks (e.g., [65, 102]).
- The choice of a set of stimuli to test and validate a specific model is crucial. As we stated in Section 1, the simulation of "unseen" (arbitrary) sounds may produce model outputs that have not been previously validated (or at least not reported) by the model developers. Actually, an unexpected model behaviour may not be strictly related to an unseen sound, but rather to an unseen sound property. For example, the models with adaptation loops have historically had an oversensitivity to transient sounds (e.g., [39, 125, 126]), leading to model versions with limited onset responses to counteract this effect [31, 39] or have used stimuli with smoother onsets in their evaluation.

When using large datasets, where the stimuli are split into training and validation data (e.g., [16, 101, 112]), the stimuli should contain representative samples of the relevant sound properties that the model user wishes to test. A practice like this can help to support (or not) the applicability of a specific model to sounds that may have not been even validated before, the "unseen sounds", reducing (or generating awareness of) the potential limitations of the test model.

6 Conclusions

In this study we compared eight monaural models of human auditory processing that simulate responses—with different levels of accuracy—up to the level of the inferior colliculus in the midbrain. We described and quantified the similarities and differences among model implementations and derived a minimum number of filters required for those stages to ensure the preservation of auditory information based on our estimates of frequency selectivity.

We showed that despite the differences in model design that result in more physiologically- (biophysical and phenomenological models) or perceptually-plausible approximations (effective models), all the models can account for a number of basic hearing properties. Examples of these properties are the phase-locking reduction in inner-hair-cell processing and the phenomenon of auditory adaptation. Still, an in-depth understanding of each of the model stages is required when selecting a model for a specific application. We encourage future users to be explicitly aware of the specific datasets of sounds and experimental paradigms upon which their models have been evaluated, as well as of other underlying model limitations. To this end, a comparison across model implementations provides a guideline for their selection and an excellent way to challenge the capabilities of different models.

Data availability statement

The implementations of the evaluated models (see Tab. 1) and the model comparison (function `exp_osses2022` to reproduce Figs. 4–15) are publicly available as part of the AMT toolbox (https://www.amtoolbox.org) [8] as of version 1.1.0 [9].

Acknowledgments

We are grateful to several colleagues who participated in technical discussions during the writing process: Enrique Lopez-Poveda, Fotios Drakopoulos, Alessandro Altoè, Armin Kohlrausch, Thomas Biberger, and Richard Lyon. We are particularly grateful to Clara Hollomey, who provided immense support for the integration of all models into AMT 1.1. The authors AOV and LV received support from ANR (project: 17-EURE-0017), LHC received support from NIH (R01-DC010813), SV received support from the ERC project RobSpear (Grant No. 678120), and PM received support from the H2020 project SONICOM (EC Grant No. 101017743).

References

1. R. Meddis, E. Lopez-Poveda, R. Fay, A. Popper (Eds.): Computational Models of the Auditory System. Springer Handbook of Auditory Research. Springer, 2010.
2. T. Dau, Auditory Processing Models. In: Havelock D., Kuwano S., Vorländer M., Eds. Handbook of Signal Processing in Acoustics, Springer, 2008: 175–196. https://doi.org/10.1007/978-0-387-30441-0_12.
3. M. Dietz, S. Ewert, V. Hohmann: Auditory model based direction estimation of concurrent speakers from binaural signals, Speech Communication 53 (2011) 592–605. https://doi.org/10.1016/j.specom.2010.05.006.
4. G. Bustamante, P. Danès, T. Forgue, A. Podlubne, J. Manhès: An information based feedback control for audio-motor binaural localization. Autonomous Robots 42 (2018) 477–490. https://doi.org/10.1007/s10514-017-9639-8.
5. R. Peng: Reproducible research in computational science. Science 334, 6060 (2011) 1226–1227. https://doi.org/10.1126/science.1213847.
6. R. Patterson, M. Allerhand, C. Giguère: Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. Journal of the Acoustical Society of America 98 (1995) 1890–1894. https://doi.org/10.1121/1.414456.
7. B. Fontaine, D. Goodman, V. Benichoux, R. Brette: Brian hears: Online auditory processing using vectorization over channels. Frontiers in Neuroinformatics 5 (2011). https://doi.org/10.3389/fninf.2011.00009.
8. P. Majdak, C. Hollomey, R. Baumgartner: AMT 1.0: the toolbox for reproducible research in auditory modeling. Submitted to Acta Acustica (2021).
9. The AMT team: The Auditory Modeling Toolbox full package (version 1.1.0) [code] (2021), https://sourceforge.net/projects/amtoolbox/files/AMT1.x/amtoolbox-full-1.1.0.zip/download.
10. T. Biberger, S. 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.
11. T. Biberger: GPSM_2016: Generalized Power Spectrum Model (GPSM), 2021. Available at https://gitlab.uni-oldenburg.de/kuxo2262/GPSM_2016, last accessed February 27, 2022.
12. A. Saremi, R. Beutelmann, M. Dietz, G. Ashida, J. Kretzberg, S. Verhulst: A comparative study of seven human cochlear filter models. Journal of the Acoustical Society of America 140 (2016) 1618–1634. https://doi.org/10.1121/1.4960486.
13. E. Lopez-Poveda: Spectral processing by the peripheral auditory system: Facts and models. International Review of Neurobiology 70 (2005) 7–48. https://doi.org/10.1016/S0074-7742(05)70001-5.
14. T. Anderson: A comparison of auditory models for speaker independent phoneme recognition. International Conference on Acoustics, Speech, and Signal Processing 2 (1993) 231–234. https://doi.org/10.1109/ICASSP.1993.319277.
15. J. Breebaart, S. van de Par, A. Kohlrausch: On the difference between cross-correlation and EC-based binaural models, in Forum Acusticum, Sevilla, Spain, 2002, pp. 1–6.
16. N. Harlander, R. Huber, S. Ewert: Sound quality assessment using auditory models. Journal of the Audio Engineering Society 62 (2014) 324–336. https://doi.org/10.17743/jaes.2014.0020.
17. K. Steinmetzger, J. Zaar, H. Relaño-Iborra, S. Rosen, T. Dau: Predicting the effects of periodicity on the intelligibility of masked speech: an evaluation of different modelling approaches. Journal of the Acoustical Society of America 146 (2019) 2562–2576. https://doi.org/10.1121/1.5129050.
18. M. Rudnicki, O. Schoppe, M. Isik, F. Völk, W. Hemmert: Modeling auditory coding: from sound to spikes. Cell and Tissue Research 361 (2015) 159–175. https://doi.org/10.1007/s00441-015-2202-z.

19. M. Dietz, J.-H. Lestang, P. Majdak, R. Stern, T. Marquardt, S. Ewert, W. Hartmann, D. Goodman: A framework for testing and comparing binaural models. Hearing Research 360 (2018) 92–106. https://doi.org/10.1016/j.heares.2017.11.010.
20. P. Søndergaard, P. Majdak: The Auditory Modeling Toolbox. In: Blauert J (Ed.), The Technology of Binaural Listening, Chap. 2, Berlin Heidelberg. 2013, pp. 33–56.
21. K. Kanders, T. Lorimer, F. Gomez, R. Stoop: Frequency sensitivity in mammalian hearing from a fundamental nonlinear physics model of the inner ear. Scientific Reports 7 (2017) 9931. https://doi.org/10.1038/s41598-017-09854-2.
22. R. Lyon: Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function. Journal of the Acoustical Society of America 130 (2011) 3893–3904. https://doi.org/10.1121/1.3658470
23. V. Hohmann: Frequency analysis and synthesis using a Gammatone filterbank. Acust. Acta Acust. 88 (2002) 433–442.
24. E. Lopez-Poveda, R. Meddis: A human nonlinear cochlear filterbank. Journal of the Acoustical Society of America 110 (2001) 3107–3118. https://doi.org/10.1121/1.1416197.
25. Q. Tan, L. Carney: A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide. Journal of the Acoustical Society of America 114 (2003) 2007–2020.
26. S. Verhulst, T. Dau, C. Shera: Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission. Journal of the Acoustical Society of America 132 (2012) 3842–3848. https://doi.org/10.1121/1.4763989..
27. S. 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
28. B. Moore, B. Glasberg, T. Baer: A model for the prediction of thresholds, loudness and partial loudness. Journal of the Audio Engineering Society 45 (1997) 224–240.
29. A. Osses Vecchi, R. García León, A. Kohlrausch: Modelling the sensation of fluctuation strength. Proceedings of Meetings on Acoustics 28, 050005 (2016) 1–8. https://doi.org/10.1121/2.0000410.
30. C. Taal, R. Hendriks, R. Heusdens, J. Jensen: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing 19, (2011) 2125–2136. https://doi.org/10.1109/TASL.2011.2114881.
31. 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.
32. M. Zilany, I. Bruce, L. Carney: Updated parameters and expanded simulation options for a model of the auditory periphery. Journal of the Acoustical Society of America 135 (2014) 283–286. https://doi.org/10.1121/1.420344.
33. L. Carney, T. Li, J. McDonough: Speech coding in the brain: Representation of vowel formants by midbrain neurons tuned to sound fluctuations. eNeuro 2 (2015) 1–12. https://doi.org/10.1523/ENEURO.0004-15.2015.
34. S. Verhulst, H. Bharadwaj, G. Mehraei, C. Shera, B. Shinn-Cunningham: Functional modeling of the human auditory brainstem response to broadband stimulation. Journal of the Acoustical Society of America 138 (2015) 1637–1659. https://doi.org/10.1121/1.4928305.
35. S. Verhulst, A. Altoè, V. Vasilkov: Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. Hearing Research 360 (2018) 55–75. https://doi.org/10.1016/j.heares.2017.12.018.
36. I. Bruce, Y. Erfani, M. Zilany: A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. Hearing Research 360 (2018) 40–54. https://doi.org/10.1016/j.heares.2017.12.016.
37. A. King, L. Varnet, C. Lorenzi: Accounting for masking of frequency modulation by amplitude modulation with the modulation filter-bank concept. Journal of the Acoustical Society of America 145 (2019) 2277–2293. https://doi.org/10.1121/1.5094344.
38. H. Relaño-Iborra, J. Zaar, T. Dau: A speech-based computational auditory signal processing and perception model. Journal of the Acoustical Society of America 146 (2019) 3306–3317. https://doi.org/10.1121/1.5129114.
39. A. Osses Vecchi, A. Kohlrausch: Perceptual similarity between piano notes: Simulations with a template-based perception model. Journal of the Acoustical Society of America 149 (2021) 3534–3552. https://doi.org/10.1121/10.0004818
40. A. Gelfert: Strategies and trade-offs in model-building, in How to Do Science with Models: A Philosophical Primer, Springer International Publishing. 2016, 43–70. https://doi.org/10.1007/978-3-319-27954-1_3.
41. A. Osses Vecchi, S. Verhulst: Release note on version 1.2 of the Verhulst et al. 2018 model of the human auditory system: Calibration and reference simulations, 2019, arXiv:1912.10026.
42. S. Verhulst, A. Altoè, V. Vasilkov, A. Osses: Verhulst et al. 2018 Auditory Model v1.2, 2020. https://github.com/HearingTechnology/Verhulstetal2018Model/releases/tag/v1.2. https://doi.org/10.5281/zenodo.3717800
43. P. Nelson, L. Carney: A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. Journal of the Acoustical Society of America 116 (2004) 2173–2186.
44. T. Dau, D. Püschel, A. Kohlrausch: A quantitative model of the effective signal processing in the auditory system. I. Model structure. Journal of the Acoustical Society of America 99 (1996) 3615–3622. https://doi.org/10.1121/1.414959.
45. J. Rosowski: The effects of external- and middle-ear filtering on auditory threshold and noise-induced hearing loss. Journal of the Acoustical Society of America 90 (1991) 124–135. https://doi.org/10.1121/1.401306.
46. H. Møller, M. Sørensen, D. Hammershøi, C. Jensen: Head-related transfer functions of human subjects. Journal of the Audio Engineering Society 43 (1995) 300–321.
47. S. Puria: Measurements of human middle ear forward and reverse acoustics: Implications for otoacoustic emissions. Journal of the Acoustical Society of America 113 (2003) 2773–2789. https://doi.org/10.1121/1.1564018.
48. R. Ibrahim, I. Bruce, Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues. In: Lopez-Poveda E., Palmer A., Meddis R., Eds. The Neurophysiological Bases of Auditory Perception, Springer, New York, NY, 2010, pp. 429–438. https://doi.org/10.1007/978-1-4419-5686-6
49. R. Ibrahim: The role of temporal fine structure cues in speech perception, Ph.D. thesis. McMaster University, 2012. http://hdl.handle.net/11375/11980.
50. S. Puria, W. Peake, J. Rosowski: Sound-pressure measurements in the cochlear vestibule of human-cadaver ears. Journal of the Acoustical Society of America 101 (1997) 2754–2770.
51. J. Pascal, A. Bourgeade, M. Lagier, C. Legros: Linear and nonlinear model of the human middle ear. Journal of the Acoustical Society of America 104 (1998) 1509–1516. https://doi.org/10.1121/1.424363.

52. R. Goode, M. Killion, K. Nakamura, S. Nishihara: New knowledge about the function of the human middle ear: Development of an improved analog model. American Journal of Otolaryngology 15 (1994) 145–154.
53. C. Shera, J. Guinan, A. Oxenham: Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proceedings of the National Academy of Sciences 99 (2002) 3318–3323. https://doi.org/10.1073/pnas.032675099.
54. B. Glasberg, B. Moore: Derivation of auditory filter shapes from notched-noise data. Hearing Research 47 (1990) 103–138. https://doi.org/10.1016/0378-5955(90)90170-T.
55. M. Zilany, I. Bruce: Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery, Journal of the Acoustical Society of America 120 (2006) 1446–1466. https://doi.org/10.1121/1.2225512.
56. F. Rønne, T. Dau, J. Harte, C. Elberling: Modeling auditory evoked brainstem responses to transient stimuli. Journal of the Acoustical Society of America 131 (2012) 3903–3913. https://doi.org/10.1121/1.3699171.
57. I. Bruce, M. Sachs, E. Young: An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. Journal of the Acoustical Society of America 113 (2003) 369–388. https://doi.org/10.1121/1.1519544.
58. L. Westerman, R. Smith: A diffusion model of the transient response of the cochlear inner hair cell synapse. Journal of the Acoustical Society of America 83 (1988) 2266–2276. https://doi.org/10.1121/1.396357.
59. A. Altoè, V. Pulkki, S. Verhulst: Model-based estimation of the frequency tuning of the inner-hair-cell stereocilia from neural tuning curves. Journal of the Acoustical Society of America 141 (2017) 4438–4451. https://doi.org/10.1121/1.4985193.
60. B. Moore: An Introduction to the Psychology of Hearing. 6th ed., Koninklijke Brill NV, 2013.
61. A. Peterson, P. Heil: A simple model of the inner-hair-cell ribbon synapse accounts for mammalian auditory-nerve-fiber spontaneous spike times. Hearing Research 363 (2018) 1–27. https://doi.org/10.1016/j.heares.2017.09.005.
62. P. Majdak, R. Baumgartner, C. Jenny: Formation of three-dimensional auditory space, in The Technology of Binaural Understanding. Springer International Publishing, 2020, pp. 115–149. https://doi.org/10.1007/978-3-030-00386-9_5.
63. L. Carney: Supra-threshold hearing and fluctuation profiles: Implications for sensorineural and hidden hearing loss. Journal of the Association for Research in Otolaryngology 19 (2018) 331–352. https://doi.org/10.1007/s10162-018-0669-5.
64. G. Ashida, D. Tollin, J. Kretzberg: Physiological models of the lateral superior olive. PLoS Computational Biology 13 (2017) 1–50. https://doi.org/10.1371/journal.pcbi.1005903.
65. B. Maxwell, V. Richards, L. Carney: Neural fluctuation cues for simultaneous notched-noise masking and profile-analysis tasks: Insights from model midbrain responses. Journal of the Acoustical Society of America 147 (2020) 3523–3537. https://doi.org/10.1121/10.0001226.
66. L. Carney: University of Rochester: Envisioning Auditory Responses (UR EAR 2020b), 2020. https://osf.io/6bsnt/.
67. W. Gerstner, W. Kistler, R. Naud, L. Paninski: Variability of spike trains and neural codes, in Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition, Cambridge University Press, 2014. Chap. 7. https://doi.org/10.1017/CBO9781107447615.
68. 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.
69. J. 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.
70. S. Ewert, J. Verhey, T. Dau: Spectro-temporal processing in the envelope-frequency domain. Journal of the Acoustical Society of America 112 (2002) 2921–2931. https://doi.org/10.1121/1.1515735.
71. D. Greenwood: A cochlear frequency position function for several species–29 years later, Journal of the Acoustical Society of America 87 (1990) 2592–2605. https://doi.org/10.1121/1.399052.
72. N. Wallaert, L. Varnet, B. Moore, C. Lorenzi: Sensorineural hearing loss impairs sensitivity but spares temporal integration for detection of frequency modulation. Journal of the Acoustical Society of America 144 (2018) 720–733. https://doi.org/10.1121/1.5049364.
73. M.C. Liberman: Auditory-nerve response from cats raised in a low-noise chamber. Journal of the Acoustical Society of America 63 (1978) 442–455. https://doi.org/10.1121/1.381736.
74. M.C. Liberman, L. Dodds, S. Pierce: Afferent and efferent innervation of the cat cochlea: Quantitative analysis with light and electron microscopy, he. Journal of Comparative Neurology 301 (1990) 443–460. https://doi.org/10.1002/cne.903010309.
75. M. Zilany, I. Bruce, P. Nelson, L. Carney: A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. Journal of the Acoustical Society of America 126 (2009) 2390–2412. https://doi.org/10.1121/1.3238250.
76. T. Ren: Longitudinal pattern of basilar membrane vibration in the sensitive cochlea. Proceedings of the National Academy of Sciences 99 (2002) 17101–17106. https://doi.org/10.1073/pnas.262663699.
77. M. Jepsen, S. 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.
78. A. Recio, W. Rhode: Basilar membrane responses to broadband stimuli. Journal of the Acoustical Society of America 108 (2000) 2281–2298. https://doi.org/10.1121/1.1318898.
79. L. Robles, M. Ruggero: Mechanics of the mammalian cochlea. Physiological Reviews 81 (2001) 1305–1352. https://doi.org/10.1152/physrev.2001.81.3.1305.
80. D. McFadden, M. Yama: Upward shifts in the masking pattern with increasing masker intensity. Journal of the Acoustical Society of America 74 (1983) 1185–1189. https://doi.org/10.1121/1.390042.
81. B. Moore, B. Glasberg: Behavioural measurement of level-dependent shifts in the vibration pattern on the basilar membrane at 1 and 2 kHz. Hearing Research 175 (2003) 66–74. https://doi.org/10.1016/S0378-5955(02)00711-6.
82. A. Palmer, I. Russell: Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. Hearing Research 24 (1986) 1–15. https://doi.org/10.1016/0378-5955(86)90002-X.
83. E. Lopez-Poveda, A. Eustaquio-Martín: A biophysical model of the inner hair cell: The contribution of potassium currents to peripheral auditory compression. Journal of the Association for Research in Otolaryngology 7 (2006) 218–235. https://doi.org/10.1007/s10162-006-0037-8.

84. J. Antoni: Orthogonal-like fractional-octave-band filters. Journal of the Acoustical Society of America 127 (2010) 884–895. https://doi.org/10.1121/1.3273888.
85. A. Altoè, V. Pulkki, S. Verhulst: Transmission line cochlear models: Improved accuracy and efficiency. Journal of the Acoustical Society of America 136 (2014) EL302–EL308. https://doi.org/10.1121/1.4896416.
86. M. Ruggero, N. Rich, A. Recio, S.S. Narayan, L. Robles: Basilar-membrane responses to tones at the base of the chinchilla cochlea. Journal of the Acoustical Society of America 101 (1997) 2151–2163. https://doi.org/10.1121/1.418265.
87. R. Smith, M. Brachman: Operating range and maximum response of single auditory nerve fibers. Brain Research 184 (1980) 499–505. https://doi.org/10.1016/0006-8993(80)90817-3.
88. 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.
89. M. Rahman, B. Willmore, A. King, N. Harper: Simple transformations capture auditory input to cortex. Proceedings of the National Academy of Sciences 117 (2020) 28442–28451. https://doi.org/10.1073/pnas.1922033117.
90. L. Deng, C.D. Geisler: Responses of auditory-nerve fibers to nasal consonant-vowel syllables. Journal of the Acoustical Society of America 82 (1987) 1977–1988. https://doi.org/10.1121/1.395642.
91. L. Carney, D. Kim, S. Kuwada: Speech coding in the midbrain: effects of sensorineural hearing loss. In: P. van Dijk, D. Baskent, E. Gaudrain, de Kleine E., Wagner A., Lanting C., Eds. Physiology, psychoacoustics and cognition in normal and impaired hearing, Springer International Publishing, 2016, pp. 427–435. https://doi.org/10.1007/978-3-319-25474-6.
92. B.S. Krishna, M. Semple: Auditory temporal processing: Responses to sinusoidally amplitude-modulated tones in the inferior colliculus. Journal of Neurophysiology 84 (2000) 255–273. https://doi.org/10.1152/jn.2000.84.1.255.
93. D. Purcell, M.S. John: Evaluating the modulation transfer function of auditory steady state responses in the 65 Hz to 120 Hz range. Ear and Hearing 31 (2010) 667–678. https://doi.org/10.1097/AUD.0b013e3181e0863b.
94. D. Schwartz, M. Morris, J. Spydell, C. Ten Brink, M. Grim, J. Schwartz: Influence of click polarity on the brain-stem auditory evoked response (BAER) revisited. Ear and Hearing 77 (1990) 445–457. https://doi.org/10.1016/0168-5597(90)90005-X.
95. T. Picton: Auditory brainstem responses: peaks along the way, in Human Auditory Evoked Potentials, Chap. 8, Plural Publishing. 2011, 213–245.
96. E. Laukli, R. Burkard: Calibration/standardization of short-duration stimuli. Seminars in Hearing 36 (2015) 3–10. https://doi.org/10.1055/s-0034-1396923.
97. J. Undurraga, A. van Wieringen, R. Carlyon, O. Macherey, J. Wouters: Polarity effects on neural responses of the electrically stimulated auditory nerve at different cochlear sites. Hearing Research 269 (2010) 146–161. https://doi.org/10.1016/j.heares.2010.06.017.
98. D. Ramekers, H. Versnel, S. Strahl, E. Smeets, S. Klis, W. Grolman: Auditory-nerve responses to varied inter-phase gap and phase duration of the electric pulse stimulus as predictors for neuronal degeneration. Journal of the Association for Research in Otolaryngology 15 (2014) 187–202. https://doi.org/10.1007/s10162-013-0440-x.
99. F. Drakopoulos, D. Baby, S. Verhulst: A convolutional neural-network framework for modelling auditory sensory cells and synapses. Communications Biology 4 (2021) 827. https://doi.org/10.1038/s42003-021-02341-5.
100. D. Baby, A. Van Den Broucke, S. Verhulst: A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. Nature Machine Intelligence 3 (2020) 134–143. https://doi.org/10.1038/s42256-020-00286-8.
101. A. Nagathil, F. Göbel, A. Nelus, I. Bruce: Computationally efficient DNN-based approximation of an auditory model for applications in speech processing, in Proc. of ICASSP. 2021, 301–305. https://doi.org/10.1109/ICASSP39728.2021.9413993.
102. F. Bianchi, L. Carney, T. Dau, S. Santurette: Effects of musical training and hearing loss on fundamental frequency discrimination and temporal fine structure processing: Psychophysics and modeling. Journal of the Association for Research in Otolaryngology 20 (2019) 263–277. https://doi.org/10.1007/s10162-018-00710-2.
103. A. Moncada-Torres, A. van Wieringen, I. Bruce, J. Wouters: Predicting phoneme and word recognition in noise using a computational model of the auditory periphery. Journal of the Acoustical Society of America 141 (2017) 300–312. https://doi.org/10.1121/1.4973569.
104. S. Verhulst, A. Jagadeesh, M. Mauermann, F. Ernst: Individual differences in auditory brainstem response wave characteristics: Relations to different aspects of peripheral hearing loss. Trends in Hearing 20 (2016) 1–20. https://doi.org/10.1177/2331216516672186.
105. S. Verhulst, F. Ernst, M. Garrett, V. Vasilkov: Suprathreshold psychoacoustics and envelope-following response relations: normal-hearing, synaptopathy and cochlear gain loss. Acta Acustica united with Acustica 104 (2018) 800–803.
106. A. Osses Vecchi, F. Ernst, S. Verhulst: Hearing-impaired sound perception: What can we learn from a biophysical model of the human auditory periphery? In: Ochmann M., Vorländer M., Fels J., Eds. International Congress on Acoustics. 2019, 678–685. https://doi.org/10.18154/rwth-conv-239764.
107. 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.
108. M. Jepsen, T. Dau: Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss. Journal of the Acoustical Society of America 129 (2011) 262–281. https://doi.org/10.1121/1.3518768.
109. S. 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.
110. N. Wallaert, B. Moore, S. 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.
111. R. Huber, B. Kollmeier: PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception. IEEE Transactions on Audio, Speech, and Language Processing 14, 6 (2006) 1902–1911.
112. T. Biberger, J.H. Flessner, R. Huber, S. 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.
113. T. Biberger, GPSMq, 2019. Available at https://gitlab.uni-oldenburg.de/kuxo2262/GPSMq, last accessed February 27, 2022.

114. I. Bruce: Physiologically based predictors of speech intelligibility. Acoustics Today 131 (2017) 28–35.
115. M.R. Schädler, A. Warzybok, S. Hochmuth, B. Kollmeier: Matrix sentence intelligibility prediction using an automatic speech recognition system. International Journal of Audiology 54 (2015) 100–107. https://doi.org/10.3109/14992027.2015.1061708.
116. J. Breebaart, S. van de Par, A. Kohlrausch: Binaural processing model based on contralateral inhibition. I. Model structure. Journal of the Acoustical Society of America 110 (2001) 1074–1088. https://doi.org/10.1121/1.1383297.
117. J. van Dorp, D. de Vries, A. Lindau: Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. Journal of the Acoustical Society of America 133 (2013) 1572–1585. https://doi.org/10.1121/1.4789357.
118. A. Osses Vecchi, A. Kohlrausch, W. Lachenmayr, E. Mommertz: Predicting the perceived reverberation in different room acoustic environments using a binaural model. Journal of the Acoustical Society of America 141 (2017) EL381–EL387. https://doi.org/10.1121/1.4979853.
119. A. Osses Vecchi: Binaural auditory model RAA. Available at https://github.com/aosses-tue/binaural-auditory-model-RAA, last accessed February 27, 2022 (2017). https://doi.org/10.5281/zenodo.3596007.
120. A. Prokopiou, A. Moncada-Torres, J. Wouters, T. Francart: Functional modelling of interaural time difference discrimination in acoustical and electrical hearing. Journal of Neural Engineering 14 (2017) 046021. https://doi.org/10.1088/1741-2552/aa7075.
121. R. Baumgartner, P. Majdak, B. Laback: Modeling the effects of sensorineural hearing loss on sound localization in the median plane. Trends in Hearing 20 (2016) 1–11. https://doi.org/10.1177/2331216516662003.
122. R. Baumgartner, P. Majdak, B. Laback: Modeling sound-source localization in sagittal planes for human listeners. Journal of the Acoustical Society of America 136 (2014) 791–802. https://doi.org/10.1121/1.4887447.
123. G. McLachlan, P. Majdak, J. Reijniers, H. Peremans: Towards modelling active sound localisation based on Bayesian inference in a static environment. Acta Acustica 5 (2021) 45. https://doi.org/10.1051/aacus/2021039.
124. E. Joosten, S. Shamma, C. Lorenzi, P. Neri: Dynamic reweighting of auditory modulation filters. PLOS Computational Biology 12 (2016) e1005019. https://doi.org/10.1371/journal.pcbi.1005019.
125. T. Dau, D. Püschel, A. Kohlrausch: A quantitative model of the effective signal processing in the auditory system. II. Simulations and measurements. Journal of the Acoustical Society of America 99 (1996) 3623–3631.
126. 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, 2 (2001) 1105–1117. https://doi.org/10.1121/1.1383299.

Cite this article as: Osses Vecchi A. Varnet L. Carney LH. Dau T. Bruce IC, et al. 2022. A comparative study of eight human auditory models of monaural processing. Acta Acustica, **6**, 17.