A Bayesian model for human directional localization of broadband static sound sources

Roberto Barumerli,[1] Piotr Majdak,[1] Michele Geronazzo,[2,3] Federico Avanzini,[4] David Meijer,[1] and Robert Baumgartner[1]

[1] *Acoustics Research Institute, Austrian Academy of Sciences, 1040 Vienna, Austria*
[2] *Dept. of Management and Engineering, University of Padova, Italy*
[3] *Dyson School of Design Engineering, Imperial College London, London, United Kingdom*
[4] *Dept. of Computer Science, University of Milano, Italy*

Humans estimate sound-source directions by combining prior beliefs with sensory evidence. Prior beliefs represent statistical knowledge about the environment while sensory evidence is acquired from auditory features such as interaural disparities and monaural spectral shapes. Models of directional sound localization often impose constraints on the contribution of these features to either the horizontal or vertical dimension. Instead, we propose a Bayesian model that more flexibly incorporates each feature according to its spatial precision and integrates prior beliefs in the inference process. We applied the model to directional localization of a single, broadband, stationary sound source presented to a static human listener in an anechoic environment. We simplified interaural features to be broadband and compared two model variants, each considering a different type of monaural spectral features: magnitude profiles and gradient profiles. Both model variants were fitted to the baseline performance of five listeners and evaluated on the effects of localizing with non-individual head-related transfer functions (HRTFs) and sounds with rippled spectrum. The model variant with spectral gradient profiles outperformed other localization models. This model variant appears particularly useful for the evaluation of HRTFs and may serve as a basis for future extensions towards modeling dynamic listening conditions.

 [https://doi.org(DOI number)]

[XYZ] Pages: 1–14

I. INTRODUCTION

When localizing a sound source, human listeners have to deal with numerous sources of uncertainty [1]. Uncertainties originate from ambiguities in the acoustic signal encoding the source position [2] as well as the limited precision of the auditory system in decoding the received acoustic information [3, 4]. Bayesian inference describes a statistically optimal solution to deal with such uncertainties in the process of perceptual decision making [5] and has been applied to model sound localization in various ways [6–9].

Common approaches of sound localization models rely on the evaluation of several spatial auditory features. Head-related transfer functions (HRTFs) describe all the spatially dependent acoustic filtering produced by the listener's ears, head, and body [10] and have been used to derive spatial auditory features. The way to quantify or extract those features is a matter of debate. In particular, a large variety of monaural spectral-shape features have been studied [11–17], with spectral magnitude profiles [14, 17] and spectral gradient profiles [12, 15] being the most established ones. Despite such details, there is consensus that the interaural time and level differences (ITDs and ILDs) [1] as well as some form of monaural spectral shapes are important features for the directional localization of broadband sound sources [18].

In order to decode the spatial direction from the auditory features, models rely on the assumption that listeners have learned to associate acoustic features with spatial directions [13, 19]. In fact, the interaural features are particularly informative about the horizontal dimension [1] though rather ambiguous with respect to the vertical dimension, where evidence from the monaural spectral features is more important [12]. This anisotropic relevance of different features is the reason for why specific auditory features are often studied along a single dimension of the so-called modified interaural-polar coordinate system [20], with the lateral angle along the horizontal left/right dimension and the polar angle along the vertical and front/back dimension. However, this geometric separation is a simplification. Monaural spectral features, for instance, can also contribute to the inference process in the direction estimation along the lateral dimension [21–23]. Hence, directional sound localization models should rather exploit the joint information encoded by all auditory features.

Such joint information has already been considered in a model of directional sound localization based on Bayesian inference [6]. This model computes a spatial likelihood function from a precision-weighted integration of a set of noisy acoustic features. Then, the perceived source direction is assumed to be at the maximum of that likelihood function. While this model was built to assess

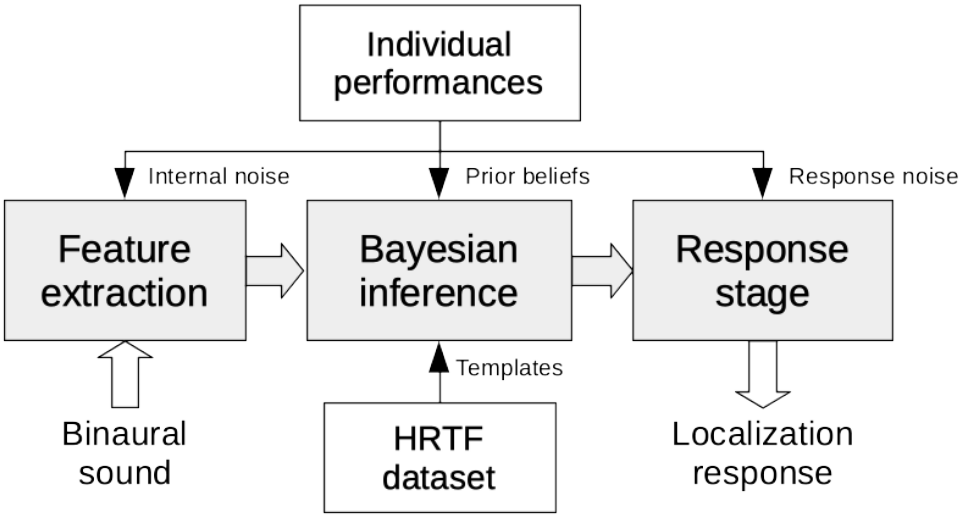

FIG. 1: Model structure. Gray blocks: Model's processing pipeline consisting of 1) the *feature extraction* stage to compute spatial features from the binaural sound; 2) the *Bayesian inference* stage integrating the sensory evidence obtained by comparison with feature templates and with prior beliefs in order to estimate the most probable direction; and 3) the *response stage* transforming the internal estimate to the final localization response. White blocks: Elements required to fit the model to an individual subject consisting of listener performances in estimating sound direction and individual HRTF dataset.

which spatial information can be accessible to the auditory system, its predictions overestimate the actual human performance yielding unrealistically low front-back confusion rates and localization errors [24]. Still, in order to model human performance, this model can serve as a solid basis for improvements such as the consideration of monaural spectral features, the integration of response noise involved in typical localization tasks, and the incorporation of prior beliefs.

Prior beliefs are important in the process of Bayesian inference because they reflect the listener's statistical knowledge about the environment, helping to compensate for uncertainties in the sensory evidence [25]. For example, listeners seem to effectively increase precision in a frontal localization task by assuming source directions to be more likely located at the eye-level rather than at extreme vertical positions [8]. However, such an increase in precision may come at the cost of decreasing accuracy. As it seems, the optimal accuracy-precision trade-off in directional localization depends on the statistical distribution of sound sources [26]. While listeners seem to adjust their prior beliefs to changes in the sound-source distribution [26, 27], they may also establish long-term priors reflecting the distribution of sound sources in their everyday environment.

Here, we introduce a Bayesian inference model to predict the performance of a listener in estimating the direction of static broadband sounds. Similar to [6], our model implements a noisy feature extraction and probabilistically combines interaural and monaural spatial features. These features are then compared with templates of spatial features, obtained from listener-specific HRTFs, to generate the sensory evidence in the form of a likelihood function. Subsequently, the sensory evidence is combined with prior beliefs emphasizing directions at the eye level [8]. The estimated source direction is selected from the resulting (posterior) spatial representation according to a Bayesian decision function. In a final stage, the model incorporates response scattering [15] to account for the uncertainty introduced by pointing responses in localization experiments.

For evaluation, we considered a model variant based on spectral amplitudes and a model variant based on spectral gradients [15]. Each model's free parameters were fitted to the sound-localization performance of individual listeners [28]. We then tested the simulated responses of both model variants against human responses from sound-localization experiments investigating the effects of non-individual HRTFs [29] and ripples in the source spectrum [30].

The paper is organized as follows: Sec. II describes the auditory model (Sec. II A) and explains the parameter estimation (Sec. II B). Then, Sec.III evaluates the model's performance by comparing its estimations to the actual performance of human listeners. Finally, Sec. IV discusses the model's relevance as well as its limitations, and outlines its potential for future extensions.

II. METHODS

A. Model description

The proposed auditory model consists of three main stages, as shown in Fig. 1: 1) The feature extraction

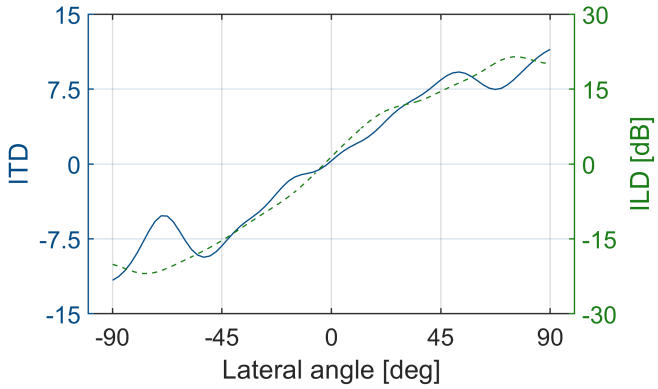

FIG. 2: Interaural features as functions of lateral angle in the horizontal frontal plane ($\phi = 0°$). Left axis (blue solid line): Transformed ITD x_{itd} (dimensionless), see Eq. 2. Right axis (green dashed line): ILD (in dB) obtained from the magnitude profiles. Example for subject NH12 [28].

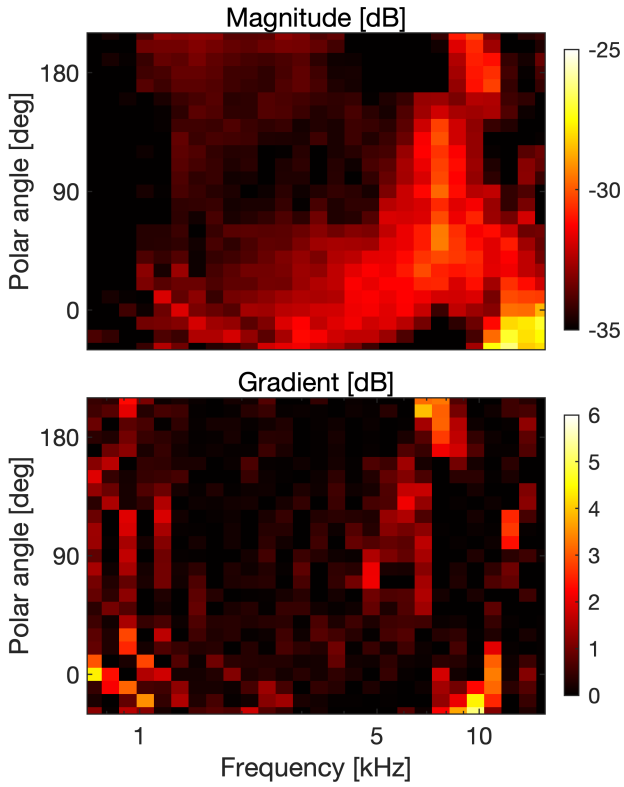

FIG. 3: Monaural spectral features as a function of polar angle in the median plane ($\theta = 0°$). Top: Features obtained from the magnitude profiles. Bottom: Features obtained from the gradient profiles. Example for the left ear of subject NH12 [28].

stage determines the encoded acoustic spatial information represented as a set of spatial features altered by noise; 2) The Bayesian inference integrates the sensory evidence resulting from the decoding procedure based on feature templates with the prior belief and forms a perceptual decision, and 3) The response stage transforms the perceptual decision in a directional response by corrupting the estimation with uncertainty in the pointing action.

1. Feature extraction

The spatial auditory features are extracted from the sensory input which is provided by the directional transfer function transformed in the time domain (i.e. the HRTF processed to remove the direction independent component [31]). We follow [6] in that we decode the spatial information provided by a single sound source via the binaural stimulus from a vector of spatial features:

$$\bar{\boldsymbol{t}} = [x_{itd}, x_{ild}, \boldsymbol{x}_{L,mon}, \boldsymbol{x}_{R,mon}]\,, \tag{1}$$

where x_{itd} denotes a scalar ITD feature, x_{ild} a scalar ILD feature, and a vector that concatenates monaural spectral features for left ear, $\boldsymbol{x}_{L,mon}$, and right ear, $\boldsymbol{x}_{R,mon}$. Each feature is assumed to be extracted by different neural pathways responsible to deliver encoded spatial information to higher levels of the auditory system [1, 4].

Assuming broadband and spatially stationary sources, interaural features can be heavily approximated by means of wideband estimators [18, 32]. The ILD was approximated as the time-averaged broadband level difference between the left and right channels [18]. The ITD was estimated by first processing each channel of the binaural signal with a low-pass Butterworth filter (10^{th} order and cutoff 3000 Hz) and an envelope extraction step based on the Hilbert transform. Then, the ITD value is computed with the interaural cross-correlation method which is a good estimator of perceived lateralization in static scenarios with noise bursts [32]. In addition, we applied the transformation proposed by Reijniers *et al.* [6] to compensate the increasing uncertainty levels for increasing ITDs [33] resulting in a dimensionless quantity with a more isotropic variance:

$$x_{itd} = \frac{\mathrm{sgn}(itd)}{b_{itd}} \log\left(1 + \frac{b_{itd}}{a_{itd}} \cdot |itd|\right), \tag{2}$$

with itd denoting ITDs in μs and the parameters $a_{itd} = 32.5\mu s$ and $b_{itd} = 0.095$ and 'sgn' indicating the sign function (for details on the derivation based on signal detection theory, see Supplementary Information from [6]). An example of the interaural features as functions of the lateral angle is shown in Fig. 2.

Monaural spectral features, $\boldsymbol{x}_{\{L,R\},mon}$, were derived from approximate neural excitation patterns. To approximate the spectral resolution of the human cochlea, we processed the binaural signal by the gammatone filterbank with non-overlapping equivalent rectangular bandwidths [34, 35], resulting in $N_B = 27$ bands within the interval $[0.7, 18]$ kHz [36, 37]. Followed by half-wave rectification and square-root compression to model hair-cell transduction [e.g., 38, 39], it results in the unit-less exci-

tation:

$$\begin{aligned} \overline{\mathrm{c}}^{\boldsymbol{\varphi}}_{\zeta,b}[n] &= (h^{\boldsymbol{\varphi}}_{\zeta} * g_b)[n], \\ \mathrm{c}^{\boldsymbol{\varphi}}_{\zeta,b}[n] &= \begin{cases} \sqrt{\overline{\mathrm{c}}^{\boldsymbol{\varphi}}_{\zeta,b}[n]} & \text{if } \overline{\mathrm{c}}^{\boldsymbol{\varphi}}_{\zeta,b}[n] \geq 0 \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{3}$$

where subscripts $\zeta \in \{L, R\}$ indicate the left and right ears, $n = 1, \ldots, N$ is the time index, $b = 1, \ldots, N_B$ is the band index, $g_b[n]$ is the corresponding gammatone filter and $h^{\boldsymbol{\varphi}}_{\zeta}[n]$ is the binaural signal in a normalized scale with sound direction $\boldsymbol{\varphi}$ (i.e. a pair of head-related impulse responses or their convolution with a source signal).

We thus define the spectral feature for the magnitude profiles (MPs) with the vector $\boldsymbol{x}_{\zeta,MP}$. This vector is the collection of root mean square amplitudes across time in decibels for each of the spectral bands for each ear:

$$\begin{aligned} \mathrm{mp}^{\boldsymbol{\varphi}}_{\zeta,b} &= 10 \log_{10} \left(\tfrac{1}{N} \textstyle\sum_{n=1}^{N} \mathrm{c}^{\boldsymbol{\varphi}}_{\zeta,b}[n]^2 \right), \\ \boldsymbol{x}_{\zeta,mp} &= \left[\mathrm{mp}^{\boldsymbol{\varphi}}_{\zeta,1}, \ldots, \mathrm{mp}^{\boldsymbol{\varphi}}_{\zeta,N_B} \right], \end{aligned} \tag{4}$$

where the function $\mathrm{c}^{\boldsymbol{\varphi}}_{\zeta,b}[n]$ is defined in Eq. 3.

An alternative spectral feature can be computed by positive gradient extraction over the frequency dimension. It has previously been shown that integrating such spectral features in an auditory model provides good agreement with human localization performance [15]. Therefore, we define a second possible spectral feature based on gradient profiles (GPs) with the vector $\boldsymbol{x}_{\zeta,GP}$. It includes the gradient extraction as an additional processing step:

$$\begin{aligned} \overline{\mathrm{gp}}^{\boldsymbol{\varphi}}_{\zeta,b} &= \mathrm{mp}^{\boldsymbol{\varphi}}_{\zeta,b+1} - \mathrm{mp}^{\boldsymbol{\varphi}}_{\zeta,b}, \\ \mathrm{gp}^{\boldsymbol{\varphi}}_{\zeta,b} &= \begin{cases} \overline{\mathrm{gp}}^{\boldsymbol{\varphi}}_{\zeta,b} & \text{if } \overline{\mathrm{gp}}^{\boldsymbol{\varphi}}_{\zeta,b} \geq 0 \\ 0 & \text{otherwise,} \end{cases} \\ \boldsymbol{x}_{\zeta,GP} &= \left[\mathrm{gp}^{\boldsymbol{\varphi}}_{\zeta,1}, \cdots, \mathrm{gp}^{\boldsymbol{\varphi}}_{\zeta,N_B-1} \right]. \end{aligned} \tag{5}$$

A visualization of these monaural features is shown in Fig. 3.

To demonstrate the impact of monaural spectral feature type, we will analyze the results of both variants with the corresponding feature spaces defined as follows:

$$\begin{aligned} \bar{\boldsymbol{t}}_{\mathrm{MP}} &= [x_{itd}, x_{ild}, \boldsymbol{x}_{L,MP}, \boldsymbol{x}_{R,MP}], \\ \bar{\boldsymbol{t}}_{\mathrm{GP}} &= [x_{itd}, x_{ild}, \boldsymbol{x}_{L,GP}, \boldsymbol{x}_{R,GP}]. \end{aligned} \tag{6}$$

Limited precision in the feature extraction process leads to corruption of the features and can be modelled as additive internal noise [6]. Hence, we define the noisy internal representation of the target features as:

$$\begin{aligned} \boldsymbol{t} &= \bar{\boldsymbol{t}} + \boldsymbol{\delta}, \\ \boldsymbol{\delta} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{7}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the multivariate Gaussian noise. Furthermore, we assume each spatial feature to be processed independently and thus to be also corrupted by independent noise [1]. Hence, the covariance matrix $\boldsymbol{\Sigma}$ is defined as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{itd} & 0 & 0 \\ 0 & \sigma^2_{ild} & 0 \\ 0 & 0 & \sigma^2_{mon}\boldsymbol{I} \end{bmatrix}, \tag{8}$$

with σ^2_{itd} and σ^2_{ild} being the variances associated with the ITDs and ILDs and $\sigma^2_{mon}\boldsymbol{I}$ being the covariance matrix for the monaural features where $\boldsymbol{I}$ is the identity matrix and the scalar σ_{mon} represents a constant and identical uncertainty for all frequency bands.

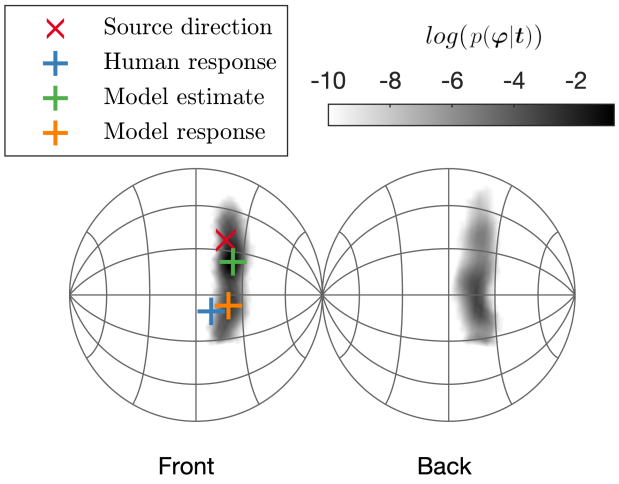

FIG. 4: Example of the model estimating the direction of a broadband sound source. Red: Actual direction of the sound source. The grayscale represents the posterior probability distribution $p(\boldsymbol{\varphi}|\boldsymbol{t})$, shown, in order to increase the readability, on a logarithmic scale. Green: Direction inferred by the Bayesian-inference stage (without the response stage). Orange: Direction inferred by the model (with the response stage). Blue: Actual response of the subject.

2. Bayesian inference

The observer infers the sound direction $\boldsymbol{\varphi}$ from the spatial features in $\boldsymbol{t}$ while taking into account potential prior beliefs about the sound direction. Within the Bayesian inference framework [5], this requires to weight the likelihood $p(\boldsymbol{t}|\boldsymbol{\varphi})$ with the prior $p(\boldsymbol{\varphi})$ to obtain the posterior distribution by means of Bayes' law as:

$$p(\boldsymbol{\varphi}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\varphi}) p(\boldsymbol{\varphi}). \tag{9}$$

The likelihood function is implemented by comparing $\boldsymbol{t}$ with the feature templates. Similarly to [6], the template $\boldsymbol{T}(\boldsymbol{\varphi})$ contains noiseless features of Eq. 1 for every sound direction $\boldsymbol{\varphi}$. While the sound direction is defined on a continuous support, our implementation sampled

it over a uniform spherical grid with a spacing of 4.5° between points ($N_\varphi = 1500$ over the full sphere). Template features were computed from the listener-specific HRTFs. To accommodate non-uniform HRTF acquisition grids, we performed spatial interpolation based on spherical harmonics with order $N_{SH} = 15$, followed by Tikhonov regularization [40].

Since the templates are constructed without noise there exists a one-to-one mapping between direction and template features. This allows us to write the likelihood function for each point of the direction grid as:

$$p(\boldsymbol{t}|\boldsymbol{\varphi}) = p(\boldsymbol{t}|\boldsymbol{T}(\boldsymbol{\varphi})) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{T}(\boldsymbol{\varphi}), \boldsymbol{\Sigma}), \tag{10}$$

where $\boldsymbol{\Sigma}$ represents the learned precision of the auditory system (i.e. the sensory uncertainty $\boldsymbol{\delta}$ reported in Eq. 7). Finally, we interpret the a-priori probability $p(\boldsymbol{\varphi})$ to reflect long-term expectations of listeners where prior probabilities are modelled as uniformly distributed along the horizontal dimension but centered towards the horizon as [8]. In particular, we extend the results from Ege *et al.* for sources positioned in the front and as well as back positions with:

$$p(\boldsymbol{\varphi}) \propto \exp\left(-\frac{\epsilon^2}{2\sigma_{P,\epsilon}^2}\right), \tag{11}$$

with ϵ denoting the elevation angle of $\boldsymbol{\varphi}$ and $\sigma_{P,\epsilon}^2$ the variance of the prior distribution [8]. For simplicity, the prior definition is based for the spherical coordinate system. Importantly, the origin of that prior is currently unknown and its implications are discussed in Sec. IV.

According to Eq. 9, a posterior spatial probability distribution is computed for every sound by optimally combining sensory evidence with prior knowledge [25]. As shown in Fig. 4, the most probable direction of the source $\boldsymbol{\varphi}$ is then selected as the maximum a-posteriori (MAP) estimate:

$$\hat{\boldsymbol{\varphi}} = \arg\max_{\boldsymbol{\varphi}} p(\boldsymbol{t}|\boldsymbol{T}(\boldsymbol{\varphi}))p(\boldsymbol{\varphi}). \tag{12}$$

3. Response stage

After a sound direction estimate has been inferred, experiments usually require the listener to provide a motor response (e.g. manual pointing). To account for the uncertainty introduced by such responses, we incorporate post-decision noise in the model's response stage. Following the approach from previous work [15], we distort the location estimate by additive, direction-independent (i.e. isotropic noise) Gaussian noise:

$$\hat{\boldsymbol{\varphi}}_r = \hat{\boldsymbol{\varphi}} + \boldsymbol{m}, \tag{13}$$

where $\boldsymbol{m} \sim \mathrm{vMF}(0, \kappa_m)$ is a von-Mises-Fisher distribution with zero mean and concentration parameter κ_m. The concentration parameter κ_m can be interpreted as a standard deviation $\sigma_m = \kappa_m^{-2} \cdot 180\pi^{-1}$ [deg]. The contribution of the response noise is also visible in Fig. 4, where the final estimate is scattered independently of the spatial information provided by the a-posteriori distribution. With Eq. 13, the model outputs the response of the estimated sound source direction.

B. Parameter estimation

The model includes the following free parameters: σ_{ild}, σ_{mon} (amount of noise per feature; σ_{itd} was fixed to 0.569 as in [6]), $\sigma_{P,\epsilon}$ (directional prior), and σ_m (amount of response noise). Because of the structure of the model, these parameters jointly contribute to the prediction of performance in both lateral and polar dimensions. To roughly account for listener-specific differences in localization performance [2], the parameters were fitted to match individual listener performance.

As for the objective fitting function, we selected a set of performance metrics widely used in the analysis of behavioral localization experiments [28, 29, 41], for a summary see [42]. A commonly used set of metrics contains the quadrant error rate (QE, i.e., frequency of polar errors larger than 90°), local polar errors (PE, i.e., root mean square error in the polar dimension that are smaller than 90°, limited to lateral angles in the range of ±30°), and lateral errors (LE, i.e., root mean square error in the lateral dimension) [29]. We accounted for the inherent stochasticity of the model estimations by averaging the simulated performance metrics over 300 repetitions of the $N_\varphi = 1550$ directions in the HRTF dataset (i.e. Monte Carlo approximation with 465000 model simulations). Model parameters were jointly adjusted in an iterative procedure (see below) until the relative residual between the actual performance metric E_a and the predicted performance metric E_p was minimized below a metric-specific threshold τ_E, i.e.,

$$|E_a - E_p|\frac{1}{E_a} < \tau_E. \tag{14}$$

We set the thresholds to $\tau_{LE} = 0.1$, $\tau_{PE} = 0.15$, and $\tau_{QE} = 0.2$ because those values were feasible for all subjects. In addition, the QE was transformed with the rationalized arcsine function to handle small and large values adequately [43].

We ran the estimation procedure separately for each feature space in Eq. 6 and each listener. First, initial values of the parameters were derived from previous literature: the variance of the prior distribution was set to $\sigma_{P,\epsilon} = 11.5°$ as in [8]. The interaural feature noise was set to $\sigma_{ild} = 1$ dB, reflecting the range of ILD thresholds for pure tones [44]. The starting value for the monaural feature noise was set to $\sigma_{mon} = 3.5$ dB similarly to in [6]. The response noise standard deviation was set to $\sigma_m = 17°$ as the sensorimotor scatter found in [15]. Second, in an iterative procedure, σ_m was optimized to minimize the residual error relative to the PE metric and, similarly, σ_{mon} was adjusted to match the QE metric. Then, σ_{ild} was decreased to reach the LE metric. These steps were reiterated until the residual errors between actual and simulated metrics was less than the respective threshold. This procedure limited the σ_m to the interval

TABLE I: Parameters estimated to fit the models to actual subjects' performance [28] for both model variants where either magnitude profiles (MPs) or gradient profiles (GPs) were the monaural spectral features.

Variant	Subject	$\sigma_{P,\epsilon}$ [deg]	σ_{ild} [dB]	σ_{mon} [deg]	σ_m [deg]
MP	NH12	11.50	0.50	3.40	8.50
	NH15	10.00	0.50	3.20	14.27
	NH16	11.50	1.00	3.60	11.00
	NH17	11.50	0.50	4.10	14.30
	NH18	11.50	1.00	6.50	14.00
GP	NH12	11.50	0.50	1.10	8.50
	NH15	11.00	0.50	1.25	14.30
	NH16	11.50	1.00	1.25	11.50
	NH17	11.50	1.00	1.60	14.00
	NH18	11.50	1.00	2.10	15.00

TABLE II: Predicted performance metrics averaged across all subjects and directions (±1 standard deviation across subjects) for both model variants. Actual data from [28].

Metric	Actual	Predicted	
		MP	GP
LE [deg]	12.25 ± 2.43	12.97 ± 2.50	13.18 ± 2.66
PE [deg]	32.73 ± 3.44	31.20 ± 4.04	29.78 ± 4.01
QE [%]	7.83 ± 7.11	8.32 ± 5.75	9.80 ± 5.23

$[5°, 20°]$ and used a step-size of 0.1°, σ_{mon} was defined in the interval $[0.5, 10]$ dB with a step-size of 0.05 dB; σ_{ild} was defined in the interval of $[0.5, 2]$ dB with a step-size of 0.5 dB. If the procedure did not converge, we decreased $\sigma_{P,\epsilon}$ by 0.5° and reattempted the parameter optimization procedure.

III. RESULTS

We first report the quality of model fits to the calibration data itself [28] in Sec. III A. Then, Sec. III B quantitatively evaluates the simulated performances of our two model variants and of two previously proposed models against data from two additional sound localization experiments.

A. Parameter fits

The parameter estimation procedure was done with both model variants, based on either $\bar{\boldsymbol{t}}_{\mathrm{MP}}$ or $\bar{\boldsymbol{t}}_{\mathrm{GP}}$, and for five individuals tested in a previous study [28]. In that experiment, naive listeners were asked to localize broadband noise bursts of 500 ms duration presented from various directions on the sphere via binaural rendering through headphones based on listener-specific directional transfer functions. The subjects were wearing a head-mounted display and were asked to orient the pointer in their right hand to the perceived sound-source direction. The fitting procedure converged for both models and all subjects. Notably, subject NH15 required to reduce the step size of σ_m to 0.1° to meet the convergence criteria. Tab. I reports the estimated parameters σ_m, $\sigma_{P,\epsilon}$, σ_{mon} and σ_{ild} for every listener. The amount of response noise was similar for both model types. Tab. II contrasts the predicted performance metrics with the actual ones, averaged across listeners.

More in detail, Fig. 5 compares predicted localization performance to the actual performance of subjects estimating the direction of a noise burst for different spherical segments [28]. The predicted LEs and PEs, both as functions of the actual lateral and polar angles, respectively, were in good agreement with those from the actual experiment. Instead, the simulated QE metric failed to mimic the front back asymmetries present in four subjects. Finally, only small differences were observed between the two feature spaces $\bar{\boldsymbol{t}}_{\mathrm{MP}}$ and $\bar{\boldsymbol{t}}_{\mathrm{GP}}$.

Contribution of model stages

Fig. 6 illustrates the effects of different model stages on target-specific predictions. The example shows direction estimations from subject NH16 localizing broadband noise bursts [28] and the corresponding predictions of the model based on $\bar{\boldsymbol{t}}_{\mathrm{GP}}$ with different configurations of priors and response noise: without both (a), with priors only (b), and with both (c). While adding response noise scatters the estimated directions equally across spatial dimensions (compare c to b), including the spatial prior only affects the polar dimension (compare b to a). As observed in the actual responses, the prior causes more of the simulated estimations to be biased towards the horizon (0° and 180°).

In order to quantify the effect of introducing the spatial prior in the polar dimension, we computed the polar gain as a measure of accuracy [13] for both simulated and the actual responses. This metric relies on two regressions performed on the baseline condition, separating between targets in the front and back. The linear fits for the baseline condition are defined as:

$$\phi_e = g_\phi \cdot \phi_a + b_\phi \tag{15}$$

with ϕ_e being the estimated polar angles and ϕ_a being the actual polar angles. The parameters are the localization bias b_ϕ in degrees, which is typically very small, and the dimensionless localization gain g_ϕ, which can be seen as a measure of accuracy [8, 13]. The regression fits only incorporate ϕ_e that deviate from the regression line by less than 40°. Since that definition of outliers depends on the regression parameters, this procedure is initialized with $b_\phi = 0°$ and $g_\phi = 1$ and re-iterated until convergence. In our analysis, only the frontal positions were considered. The polar gain of the actual responses, averaged over subjects, was 0.50, indicating that our subjects showed a localization error increasing with the angular distance to the horizontal plane. For the models without the prior,

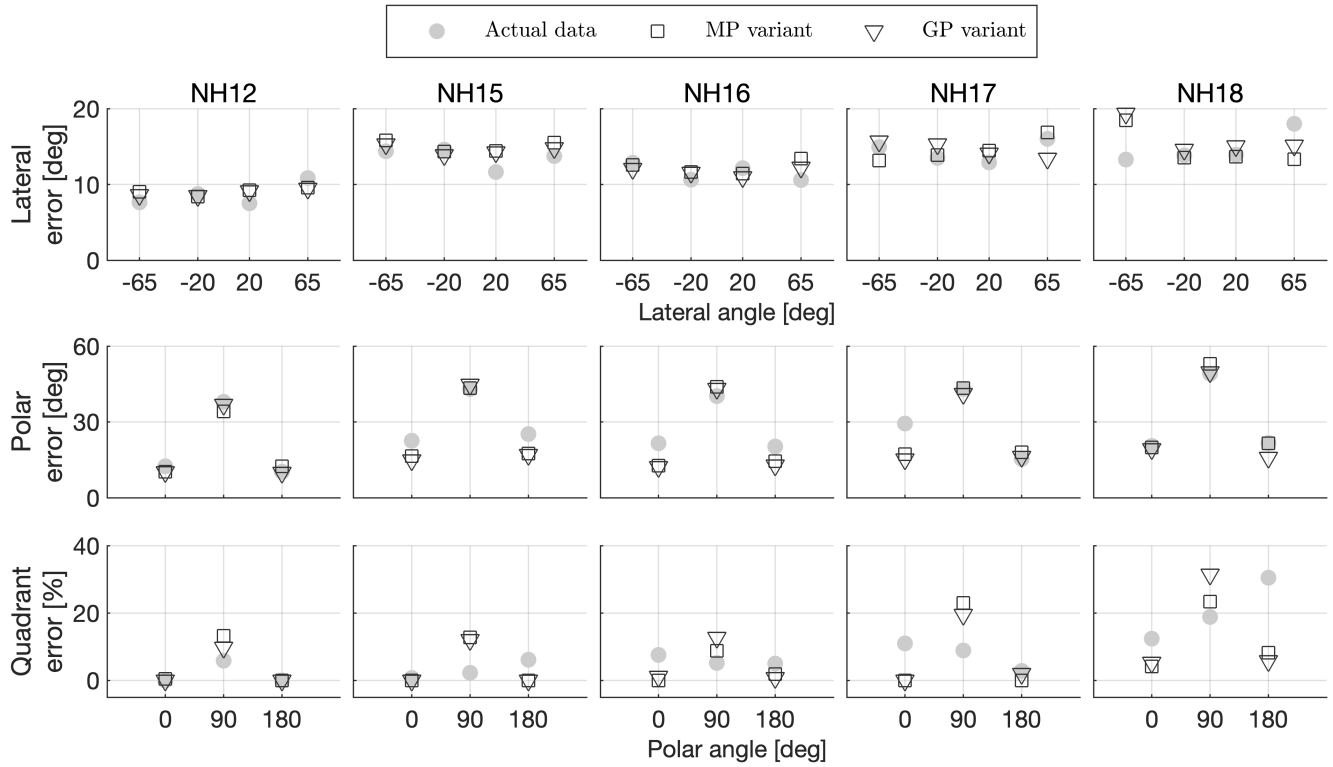

FIG. 5: Sound-localization performance as functions of sound-source direction. Open symbols: Predictions obtained by the two model's variants based on either spectral magnitude profiles (MPs) or gradient profiles (GPs). Filled grey symbol: Actual data from [28]. Top row: Lateral error, calculated for all targets with lateral angles of $-65° \pm 25°$, $-20° \pm 20°$, $20° \pm 20°$, and $65° \pm 25°$. Center and bottom rows: Polar error and quadrant error rates, respectively, calculated for all median-plane ($\pm 30°$) targets with polar angles of $0° \pm 30°$, $90° \pm 60°$, and $180° \pm 40°$.

the predicted polar gain was 1.00 (Fig. 6a). The polar gain obtained by the model including the prior was 0.62 (Fig. 6b and c) showing a better correspondence to the actual polar gain. Hence, the introduction of the prior belief improved the agreement with the actual localization responses by biasing them towards the horizon.

B. Model evaluation

The performance evaluation was done at the group-level. For our model, we used the five calibrated parameter sets with templates $\boldsymbol{T}(\boldsymbol{\varphi})$ based on the individuals' HRTFs as "digital observers". Group-level results of these digital observers were then evaluated for two psychoacoustic experiments with acoustic stimuli as input that differed from the baseline condition with a flat source spectrum and individual HRTFs.

In addition, we compared our results with the ones of two previously published models. The first one, described by Reijniers et al. [6], is probabilistic and able to jointly estimate the lateral and polar dimensions similar to the model described in this work. Reijniers' model deviates from the current model since it relies on a different feature extraction stage, uses a uniform spatial prior distribution, does not include response noise (Eq. 13) and does not fit individualized parameters. The second model, described by Baumgartner et al. [15], estimates sound positions only in the polar dimension. Nevertheless, it shares a similar processing pipeline with the current model in that it considers both a perceptually relevant feature extraction stage, includes response noise, and considers individualized parameters. The main differences with our model resides on the incorporation of a directional prior and of the lateral dimension per se, and on how the distance between target and templates is computed. Particularly, this previous work resorted on the l^2-norm to implement the template comparison procedure which is substantially different from our likelihood function. At the moment, this model is commonly used by the scientific community that is interested in elevation perception based on monaural spectral features for sound direction estimation [e.g., 45, 46]. We will refer to these two models as `reijniers2014` and `baumgartner2014`, respectively.

1. Effects of non-individual HRTFs

In first evaluation, sounds were spatialized using non-individualized HRTFs [29]. Originally, eleven listeners localized Gaussian white noise bursts with a duration

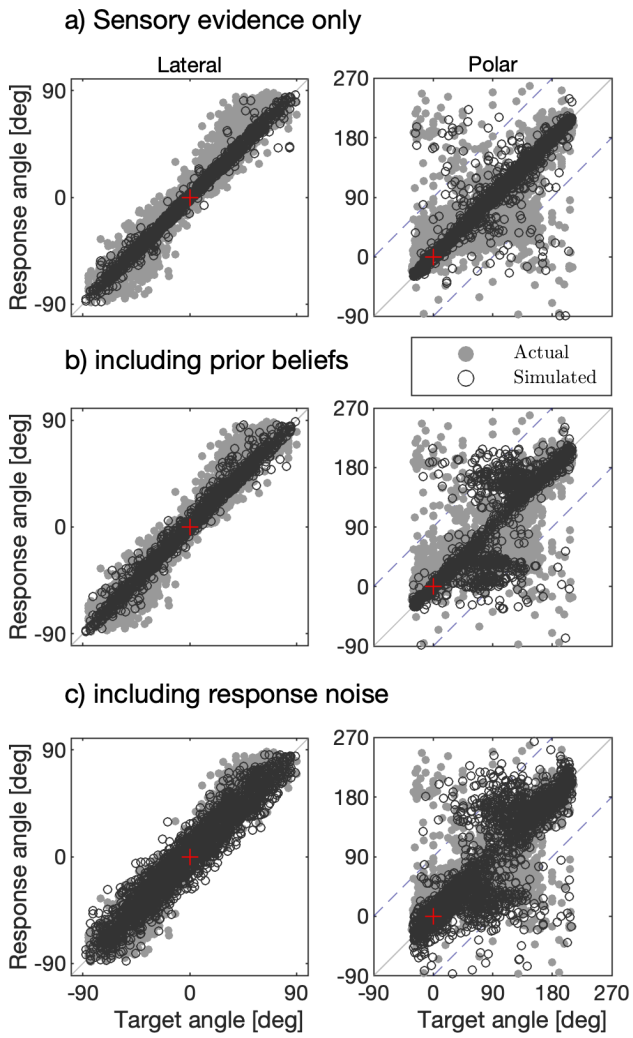

FIG. 6: Effects of likelihood, prior, and response noise on predicted response patterns as a result of modeling the directional localization of broadband noise bursts. **a)** Likelihood obtained by sensory evidence (i.e, no spatial prior and no response noise). **b)** Bayesian inference (with the spatial prior but no response noise). **c)** Full model (with prior and response noise). Gray: actual data of NH16 from [28]. Black: estimation obtained by the model considering spectral gradient profiles (GPs). Red cross: frontal position. Blue dashed lines separate regions of front-back confusions.

of 250 ms and sound directions were randomly sampled from the full sphere. Subjects were asked to estimate the direction of sounds that were spatialized using their own HRTFs in addition to sounds that were spatialized using up to 4 HRTFs from other subjects (21 cases in total). With the aim to reproduce these results, we had our pool of five digital listeners localize sounds from all available directions that were spatialized with their own individual HRTFs (*Own*) as well as sounds that were spatialized with HRTFs from the other 4 individuals (*Other*). We thus considered all inter-listener HRTF combinations for the non-individual condition.

Fig. 7 summarizes the results obtained for localization experiments with own and other HRTFs. In the *Own* condition, there is a small deviation between the actual results from [29] and our model predictions. This mismatch reflects the fact that the digital observers represent a different pool of subjects (taken from [28]) tested on a slightly different experimental protocol and setup. Differences in performance metrics are small between the two feature spaces, as already reported during parameter fitting. Predictions from the `baumgartner2014` model are only possible for the polar dimension. Instead, the model `reijniers2014` predicted too small errors, as also observed in previous simulations employing this model [24, 47].

In the *Other* condition, both of our model variants predicted a smaller degradation for the lateral dimension as compared to the actual data. The lateral errors predicted by `reinjiers2014` increased moderately but remained too small in comparison to the actual data. In the polar dimension, both model variants resulted in increased PEs and QEs, but the amount of increase was larger and more similar to the actual data for the variant equipped with gradients profiles, especially with respect to QE. The predictions from `baumgartner2014` were very similar to the model based on spectral gradients, as expected given the similar method to extract monaural spectral feature. Instead, the simulations form `reijniers2014` demonstrates how this model reported super-human performances as already demonstrated in previous analysis [24].

2. Effects of rippled-spectrum sources

The second evaluation tested the effect of spectral modulation of sound sources on directional localization in the polar dimension [30]. In that study, localization performance was probed by using noises in the frequency band $[1, 16]$ kHz which spectral shape were distorted with a sinusoidal modulation in the log-magnitude domain. The conditions considered different ripple depths, defined as the peak-to-peak difference of the log-spectral magnitude, and ripple densities, defined as the sinusoidal period along the logarithmic frequency scale. The actual experiment tested six trained subjects in a dark, anechoic chamber listening to the stimuli via loudspeakers. The sounds lasted 250 ms and were positioned between lateral angles of $\pm 30°$ and polar angles of either $0 \pm 60°$ for the front or $180 \pm 60°$ for the back. A "baseline" condition included a broadband noise without any spectral modulation (ripple depth of 0dB). To quantify the localization performance, we used the polar error rate (PER) as they defined [30]. For every condition, two baseline regressions are computed as in Sec. III A allowing to quantify the PER as the ratio of actual responses deviating by more than 45° from the predicted values of the baseline regression.

Fig. 8 shows the results of testing the fitted models with rippled spectra. In the baseline condition, our model

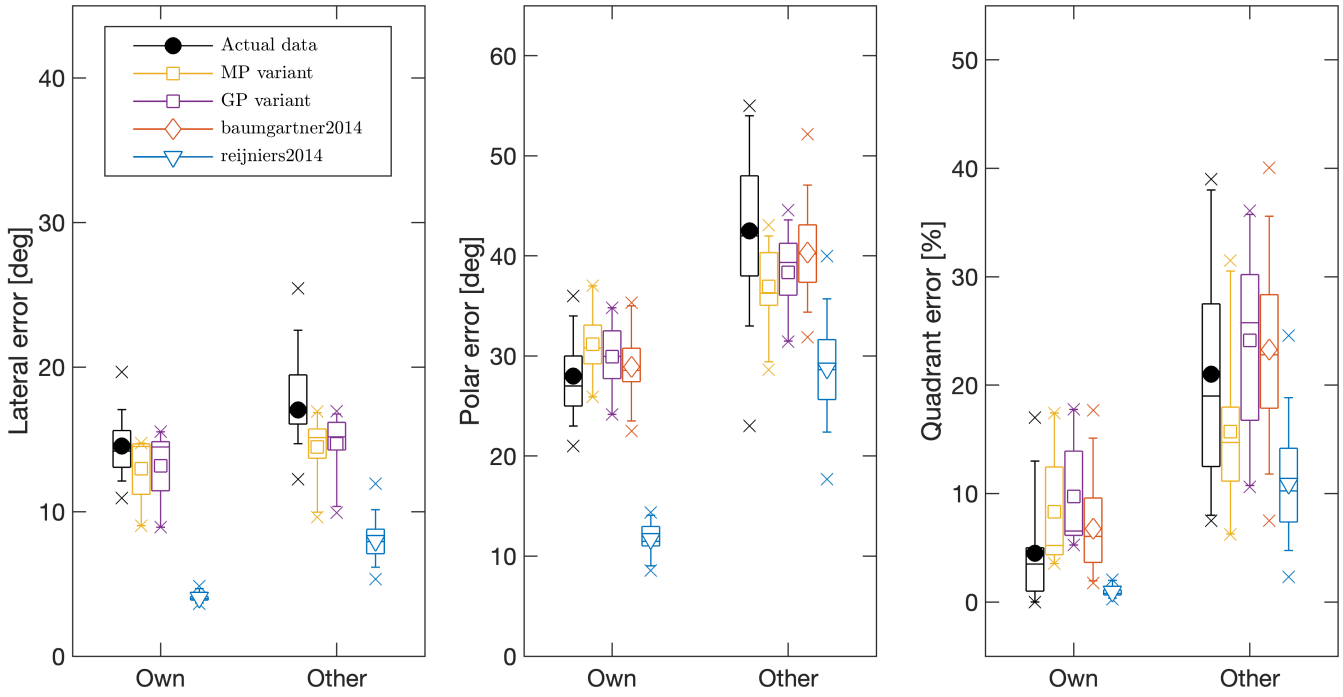

FIG. 7: Localization performance with individual (*Own*) and non-individual (*Other*) HRTFs. Actual data from [29] (`data_middlebrooks1999`). Model predictions for two model variants: spectral magnitude profiles (MPs) and spectral gradient profiles (GPs). As references, predictions by the models `reijniers2014` [6] and `baumgartner2014` [15] are shown. Note that `baumgartner2014` does not predict the lateral error.

exhibited similar performances to those obtained in the actual experiment, whereas `baumgartner2014` underestimates the baseline performance for this particular error metric. In the ripple conditions, actual listeners demonstrated poorest performance for ripple densities around one ripple per octave and a systematic increase in error rate with increasing ripple depth. These effects were well predicted by the model variant based on gradient profiles, similar to the predictions from `baumgartner2014`. In contrast, both `reijnier2014` and our model based on magnitude profiles were not able to reflect the effects of ripple density and depth as present in the actual data. Hence, the positive gradient extraction appears crucial processing step for predicting sagittal-plane localization of sources with a non-flat spectrum.

IV. DISCUSSION

The proposed functional model aims at reproducing listeners' performances when inferring the sound-source direction.The model formulation relies on Bayesian inference [25] as it integrates the sensory evidence for spatial directions obtained by combining binaural and monaural features [13] with a spatial prior [8]. Our approach considers uncertainties about the various sensory features, as in [6], in addition to the noise introduced by pointing responses [15]. These model components enabled us to successfully match overall performance metrics (LE, PE, and QE) for five subjects (see Tab. II) and within spatially restricted areas (Fig. 5). Importantly, with the inclusion of a spatial prior the model was able to adequately explain listeners' response biases towards the horizontal plane. Compared to previous models [6, 15], our model better predicted the group-level effects of non-individualized HRTFs and rippled source spectra, yet only if positive spectral gradient profiles ($\bar{\boldsymbol{t}}_{\mathrm{GP}}$), rather than magnitude profiles, served as monaural spectral features.

The validity of the model is limited to scenarios where both the source and subject are spatially static and situated in an acoustic free-field. Additionally, the current model evaluation only considered broadband and stationary sounds. For this reason, we have restricted the extraction of ITDs and ILDs to simple approximations that are sufficient for this set of conditions, as demonstrated both in our present work as well as in previous literature [18, 32] by the high accuracy of estimated lateral angles, where those features are arguable most important. The model, as presented here, does not currently account for feature-specific non-linearities required to predict phenomena like the precedence effect [48, 49]

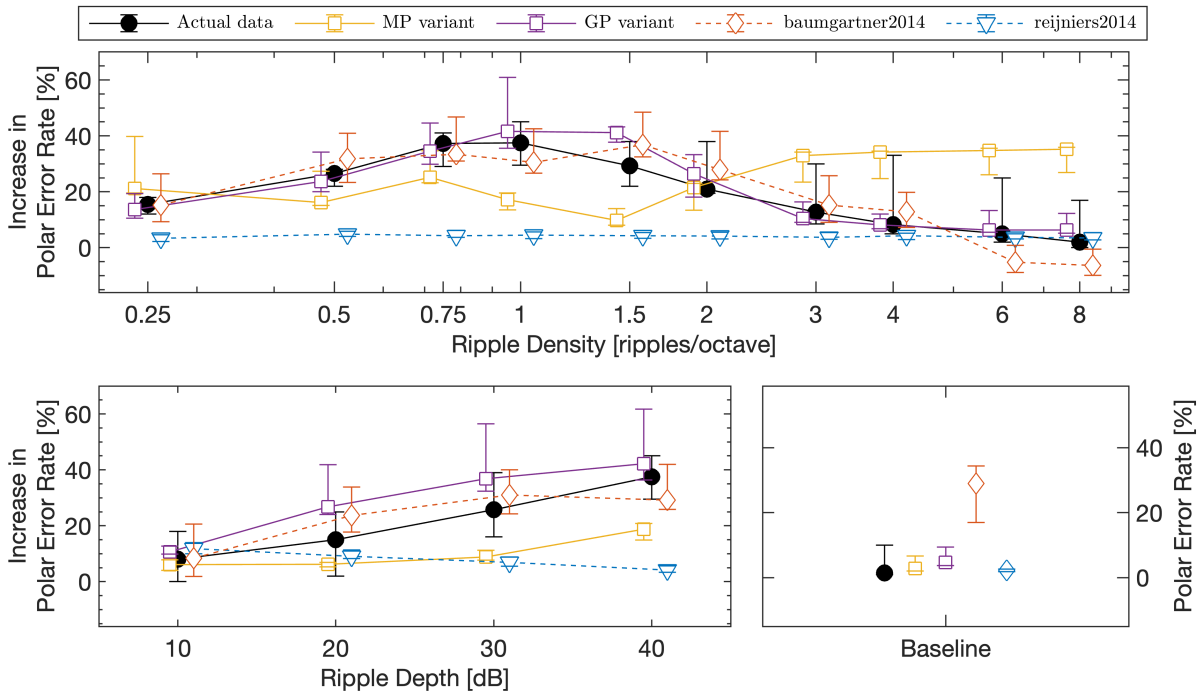

FIG. 8: Effect of spectral ripples in the source spectrum on sound localization performance in the median plane. Right-most bottom panel: localization error rates obtained without spectral ripples serving as reference. Top and bottom left panels: Differences to the reference condition shown in the right-most bottom panel. In addition to predictions from the two model variants (MP and GP), predictions from `reijniers2014` [6] and `baumgartner2014` [15] as well as actual data from [30] (`data_macpherson2003`) as shown. Note that ripple densities were varied at a fixed ripple depth of 40 dB and ripple depths were varied at a fixed ripple density of one ripple per octave.

and cannot readily be applied to non-stationary sounds, such as speech, without extensions to the feature extraction procedure. Nevertheless, from an application point of view, the proposed model can be a useful tool to assess the perceptual validity of a non-individual HRTF dataset. As an example, one can use the model's predictions to quantify differences in expected direction estimation performance for (input) sounds that are spatialized with generic vs. individual HRTFs [50].

In this work, the most probable source position is selected from the posterior distribution via the MAP decision rule. We preferred this widely-used estimator over the equally-common mean estimator to adequately deal with multiple modes of the posterior distribution generated by poor discrimination between front-back positions along sagittal planes. On the other hand, one could argue that the MAP estimator disproportionately biases direction estimates towards the mode where prior probability is large, at least under conditions of high sensory uncertainty. One of many possible alternative decision functions that may better describe stochastic human localization responses is posterior sampling [8]. With this decision function the model would probabilistically sample its best perceptual estimate from the full posterior distribution. Although often considered suboptimal, this strategy would allow an observer faster exploration and flexible adaptation to novel environmental statistics [51]. Importantly, a different decision rule might affect the here-estimated magnitudes of the sensory and motor noise. Therefore, comparative evaluations of different estimators would benefit from a more robust fitting procedure, which falls outside of the scope of the current study.

The model incorporates several non-acoustic components because they are deemed crucial to explain human performances [2, 52]. Extending the `reijniers2014` model [6] by incorporating a spatial prior and response scatter appears vital to explain listeners' estimation patterns. Without these components, fitting the model to the polar performance metrics was unfeasible [24]. First, response noise allowed us to control response precision at a local level (LE and PE) while leaving global errors (QE) largely unaffected. Instead, the occurrence rate of such global errors depends predominantly on the noise variance that is added to the monaural features. Second, the spatial prior shapes the response patterns by introducing a bias towards the horizon [41]. As shown in Fig. 6, the prior contribution is visible in the polar component of the simulated responses which cluster around the eye-level direction. Additional evidence is given by the polar gain measure as reported in Sec. III A where the inte-

gration of prior beliefs leads to better matching performances in the vertical dimension. Our formulation of the spatial prior was extrapolated from previous work [8] in the sense that its spatial distribution was assumed to be front-back symmetric. Discrepancies observed between actual and predicted global errors (Fig. 5) indicate that this assumption is likely incorrect and points towards an asymmetric prior instead. Nonetheless, at this point we can only speculate about the reasons behind the existence of such a long-term prior in spatial hearing. It potentially reflects the spatial distribution of sound sources during everyday exposure [53], or it may stem from an evolutionary emphasis on high relevance auditory signals [4], or could be related to the center of gaze as observed in barn owls [54] although the processing underlying the spatial inference mechanism might be different in mammals [3].

While the model currently only considers the static scenario, it sets the foundations for future work on predicting sound localization behavior in realistic environments. Evaluating the environment's dynamics as a chain of consecutive events as in [7] may be a promising approach. Sequential updating of one's beliefs, from prior to posterior to prior again, comes natural under the Bayesian inference scheme [25]. This makes the here-proposed model well suited as a basis for such investigations. Thus, a rich set of modulators might influence spatial hearing and the model's prior belief is an entry point to account for many of those, c.f.: evidence accumulation to track source statistics [26, 55], visual influences on auditory spatial perception [56], and auditory attention to segregate sources [57]. Selective temporal integration appears important to deal with the spatial information of many natural sources and their reflections competing in realistic scenarios. This aspect could be partially addressed by integrating recent findings related to interaural feature extraction [58]. To this end, the model will need to consider the dynamic interaction between the listener and the acoustic field. Consequently, these extensions will potentially enable the model to account for subject movements [9] and simultaneous tracking of source movements [59] while extracting spatial information from echoic scenarios [60].

V. CONCLUSIONS

We proposed a computational auditory model for the perceptual inference of sound-source direction based on Bayesian inference. From a binaural input, interaural and monaural spatial features jointly provide the sensory evidence to estimate the sound direction. The model parameters are interpretable and related to sensory noise, prior uncertainty, and response noise. Having fitted the model parameters to match subject-specific performance in a baseline condition, the model accurately predicted the localization performance observed for test conditions with non-individualized HRTFs and spectrally-modulated source spectra. Regarding spectral monaural feature extraction, the model variant evaluating gradient profiles performed best.

The proposed model seems useful to assess the perceptual validity of HRTFs. The model's domain is currently limited to static conditions, but it seems to provide a good basis for future extensions to spatially dynamic situations, spectro-temporally dynamic signals like speech and music, and reverberant environments.

ACKNOWLEDGEMENTS

We thank Jonas Reijniers for providing the initial implementation of his model. We are also grateful to Günther Koliander and Clara Hollomey for helping with the mathematical formulation of some model aspects. This work was supported by the European Union (EU) within the project SONICOM (grant number: 101017743, RIA action of Horizon 2020, to P.M.) and the Austrian Science Fund (FWF) within the project Dynamates (grant number: ZK66, to R.Bau.).

DATA AVAILABILITY STATEMENT

The implementation of the model presented in this manuscript is available in the Auditory Modeling Toolbox (AMT 1.1) as `barumerli2022` [61]. Data from [28] are also available in the AMT 1.1 as `data_majdak2010`. Individual HRTF datasets of these subjects are publicly available within the ARI database at http://sofacoustics.org/data/database/ari/. Moreover, the implementations of the model `baumgartner2014` presented in [15], the model `reijniers2014` from [6] and the data for the model evaluation procedure are available as `data_middlebrooks1999` and `data_macpherson2003` in the AMT 1.1.

Additional toolboxes were selected for the model implementation. To provide quasi-uniform sphere sampling the model relied on https://github.com/AntonSemechko/S2-Sampling-Toolbox. For the implementation of the von Mises-Fisher distribution, we used https://github.com/TerdikGyorgy/3D-Simulation-Visualization/.

[1] K. van der Heijden, J. P. Rauschecker, B. de Gelder, and E. Formisano, "Cortical mechanisms of spatial hearing," Nature Reviews Neuroscience **20**(10), 609–623 (2019).

[2] P. Majdak, R. Baumgartner, and B. Laback, "Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization," Frontiers in Psychology **5** (2014) www.frontiersin.org/

articles/10.3389/fpsyg.2014.00319/full doi: 10.3389/fpsyg.2014.00319.

[3] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of Sound Localization in Mammals," Physiological Reviews **90**(3), 983–1012 (2010) http://www.physiology.org/doi/10.1152/physrev.00026.2009 doi: 10.1152/physrev.00026.2009.

[4] M. Pecka, C. Leibold, and B. Grothe, "Biological aspects of perceptual space formation," in *The Technology of Binaural Understanding* (Springer, 2020), pp. 151–171.

[5] W. J. Ma, "Organizing probabilistic models of perception," Trends in Cognitive Sciences **16**(10), 511–518 (2012) linkinghub.elsevier.com/retrieve/pii/S136466131200201X doi: 10.1016/j.tics.2012.08.010.

[6] J. Reijniers, D. Vanderelst, C. Jin, S. Carlile, and H. Peremans, "An ideal-observer model of human sound localization," Biological Cybernetics **108**(2), 169–181 (2014) doi.org/10.1007/s00422-014-0588-4 doi: 10.1007/s00422-014-0588-4.

[7] H. Kayser, V. Hohmann, S. D. Ewert, B. Kollmeier, and J. Anemüller, "Robust auditory localization using probabilistic inference and coherence-based weighting of interaural cues," The Journal of the Acoustical Society of America **138**(5), 2635–2648 (2015) asa.scitation.org/doi/10.1121/1.4932588 doi: 10.1121/1.4932588.

[8] R. Ege, A. J. van Opstal, and M. M. van Wanrooij, "Accuracy-precision trade-off in human sound localisation," Scientific Reports **8**(1), 16399 (2018) www.nature.com/articles/s41598-018-34512-6 doi: 10.1038/s41598-018-34512-6.

[9] McLachlan, Glen, Majdak, Piotr, Reijniers, Jonas, and Peremans, Herbert, "Towards modelling active sound localisation based on bayesian inference in a static environment," Acta Acust. **5**, 45 (2021) https://doi.org/10.1051/aacus/2021039 doi: 10.1051/aacus/2021039.

[10] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," Journal of the Audio Engineering Society **43**(5), 300–321 (1995).

[11] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," The Journal of the Acoustical Society of America **92 (5)**, 2607–2624 (1992).

[12] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," The Journal of the Acoustical Society of America **94**(3), 1323–1331 (1993).

[13] P. M. Hofman and A. J. van Opstal, "Spectro-temporal factors in two-dimensional human sound localization," The Journal of the Acoustical Society of America **103**(5), 2634–2648 (1998) asa.scitation.org/doi/10.1121/1.422784 doi: 10.1121/1.422784.

[14] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," The Journal of the Acoustical Society of America **112 (4)**, 1583–1596 (2002).

[15] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," The Journal of the Acoustical Society of America **136**(2), 791–802 (2014) doi: 10.1121/1.4887447.

[16] R. Baumgartner, P. Majdak, and B. Laback, "Modeling the effects of sensorineural hearing loss on sound localization in the median plane," Trends in Hearing **20**, 2331216516662003 (2016).

[17] A. J. van Opstal, J. Vliegen, and T. van Esch, "Reconstructing spectral cues for sound localization from responses to rippled noise stimuli," PloS one **12**(3), e0174185 (2017).

[18] J. C. Middlebrooks, "Sound localization," in *Handbook of Clinical Neurology*, Vol. 129 (Elsevier, 2015), pp. 99–116, linkinghub.elsevier.com/retrieve/pii/B9780444626301000068, doi: 10.1016/B978-0-444-62630-1.00006-8.

[19] M. M. van Wanrooij and A. J. van Opstal, "Relearning Sound Localization with a New Ear," Journal of Neuroscience **25**(22), 5413–5424 (2005) doi: 10.1523/JNEUROSCI.0850-05.2005.

[20] K. Pollack, W. Kreuzer, and P. Majdak, "Perspective Chapter: Modern Acquisition of Personalised Head-Related Transfer Functions – An Overview," in *Advances in Fundamental and Applied Research on Spatial Audio*, edited by B. F. G. Katz and P. Majdak (IntechOpen, Rijeka, 2022), https://doi.org/10.5772/intechopen.102908, doi: 10.5772/intechopen.102908, section: 2.

[21] D. P. Kumpik, O. Kacelnik, and A. J. King, "Adaptive Reweighting of Auditory Localization Cues in Response to Chronic Unilateral Earplugging in Humans," Journal of Neuroscience **30**(14), 4883–4894 (2010) https://www.jneurosci.org/content/30/14/4883 doi: 10.1523/JNEUROSCI.5488-09.2010 publisher: Society for Neuroscience _eprint: https://www.jneurosci.org/content/30/14/4883.full.pdf.

[22] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," The Journal of the Acoustical Society of America **101**(2), 1050–1063 (1997) http://asa.scitation.org/doi/10.1121/1.418029 doi: 10.1121/1.418029.

[23] J. O. Stevenson-Hoare, T. C. A. Freeman, and J. F. Culling, "The pinna enhances angular discrimination in the frontal hemifield," The Journal of the Acoustical Society of America **152**(4), 2140–2149 (2022) https://doi.org/10.1121/10.0014599 doi: 10.1121/10.0014599 _eprint: https://doi.org/10.1121/10.0014599.

[24] R. Barumerli, P. Majdak, R. Baumgartner, M. Geronazzo, and F. Avanzini, "Evaluation of a human sound localization model based on bayesian inference," in *Forum Acusticum* (2020).

[25] W. J. Ma, "Bayesian Decision Models: A Primer," Neuron **104**(1), 164–175 (2019) linkinghub.elsevier.com/retrieve/pii/S0896627319308402 doi: 10.1016/

j.neuron.2019.09.037.

[26] R. Ege, A. J. Van Opstal, and M. M. Van Wanrooij, "Perceived Target Range Shapes Human Sound-Localization Behavior," eneuro **6**(2), ENEURO.0111–18.2019 (2019) https://www.eneuro.org/lookup/doi/10.1523/ENEURO.0111-18.2019 doi: 10.1523/ENEURO.0111-18.2019.

[27] K. Krishnamurthy, M. R. Nassar, S. Sarode, and J. I. Gold, "Arousal-related adjustments of perceptual biases optimize perception in dynamic environments," Nature human behaviour **1**(6), 1–11 (2017).

[28] P. Majdak, M. J. Goupell, and B. Laback, "3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training," Attention, Perception, & Psychophysics **72**(2), 454–469 (2010) link.springer.com/10.3758/APP.72.2.454 doi: 10.3758/APP.72.2.454.

[29] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," The Journal of the Acoustical Society of America **106**(3), 1493–1510 (1999) asa.scitation.org/doi/abs/10.1121/1.427147 doi: 10.1121/1.427147.

[30] E. A. Macpherson and J. C. Middlebrooks, "Vertical-plane sound localization probed with ripple-spectrum noise," The Journal of the Acoustical Society of America **114**(1), 430–445 (2003) asa.scitation.org/doi/10.1121/1.1582174 doi: 10.1121/1.1582174.

[31] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," The Journal of the Acoustical Society of America **91**(3), 1637–1647 (1992).

[32] A. Andreopoulou and B. F. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," The Journal of the Acoustical Society of America **142**(2), 588–598 (2017).

[33] J. E. Mossop and J. F. Culling, "Lateralization of large interaural delays," The Journal of the Acoustical Society of America **104**(3), 1574–1579 (1998) asa.scitation.org/doi/10.1121/1.424369 doi: 10.1121/1.424369.

[34] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," Hearing research **47**(1-2), 103–138 (1990).

[35] A. Saremi, R. Beutelmann, M. Dietz, G. Ashida, J. Kretzberg, and S. Verhulst, "A comparative study of seven human cochlear filter models," The Journal of the Acoustical Society of America **140**(3), 1618–1634 (2016) asa.scitation.org/doi/10.1121/1.4960486 doi: 10.1121/1.4960486.

[36] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," The Journal of the Acoustical Society of America **109**(3), 1110–1122 (2001) scitation.aip.org/content/asa/journal/jasa/109/3/10.1121/1.1349185 doi: 10.1121/1.1349185.

[37] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," The Journal of the Acoustical Society of America **56**(6), 1829–1834 (1974) asa.scitation.org/doi/10.1121/1.1903520 doi: 10.1121/1.1903520.

[38] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," Speech Communication **53**(5), 592–605 (2011) https://www.sciencedirect.com/science/article/pii/S016763931000097X doi: https://doi.org/10.1016/j.specom.2010.05.006.

[39] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**(4), 18 (2003).

[40] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF fitting using spherical harmonics," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, New Paltz, NY, USA (2009), pp. 257–260, ieeexplore.ieee.org/document/5346521/, doi: 10.1109/ASPAA.2009.5346521.

[41] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in sound localization by human listeners," Hearing Research **114**(1), 179–196 (1997) linkinghub.elsevier.com/retrieve/pii/S0378595597001615 doi: 10.1016/S0378-5955(97)00161-5.

[42] P. Majdak, M. J. Goupell, and B. Laback, "Two-dimensional sound localization in cochlear implantees," Ear and hearing **32**(2), 198–208 (2011).

[43] G. A. Studebaker, "A "rationalized" arcsine transform," Journal of Speech, Language, and Hearing Research **28**(3), 455–462 (1985) pubs.asha.org/doi/abs/10.1044/jshr.2803.455 doi: 10.1044/jshr.2803.455.

[44] W. A. Yost and R. H. Dye, "Discrimination of interaural differences of level as a function of frequency," The Journal of the Acoustical Society of America **83**(5), 1846–1851 (1988) asa.scitation.org/doi/10.1121/1.396520 doi: 10.1121/1.396520.

[45] R. Barumerli, M. Geronazzo, and F. Avanzini, "Localization in Elevation with Non-Individual Head-Related Transfer Functions: Comparing Predictions of Two Auditory Models," in *2018 26th European Signal Processing Conference (EUSIPCO)* (2018), pp. 2539–2543, doi: 10.23919/EUSIPCO.2018.8553320, iSSN: 2076-1465.

[46] D. Marelli, R. Baumgartner, and P. Majdak, "Efficient approximation of head-related transfer functions in subbands for accurate sound localization," IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(7), 1130–1143 (2015) doi.org/10.1109/TASLP.2015.2425219 doi: 10.1109/TASLP.2015.2425219.

[47] R. Barumerli, P. Majdak, J. Reijniers, R. Baumgartner, M. Geronazzo, and F. Avanzini, "Predicting Directional Sound-Localization of Human Listeners in both

Horizontal and Vertical Dimensions," in *Audio Engineering Society Convention 148* (2020), www.aes.org/e-lib/browse.cfm?elib=20777.

[48] J. Blauert, *Spatial hearing. The Psychophysics of Human Sound Localization*, revised ed. (The MIT Press, Cambridge, MA, 1997).

[49] R. Ege, A. J. v. Opstal, P. Bremen, and M. M. v. Wanrooij, "Testing the Precedence Effect in the Median Plane Reveals Backward Spatial Masking of Sound," Scientific Reports **8**(1), 8670 (2018) https://www.nature.com/articles/s41598-018-26834-2 doi: 10.1038/s41598-018-26834-2.

[50] M. Geronazzo, S. Spagnol, and F. Avanzini, "Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric," IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**(7), 1243–1256 (2018) doi: 10.1109/TASLP.2018.2821846.

[51] W. Gaissmaier and L. J. Schooler, "The smart potential behind probability matching," Cognition **109**(3), 416–422 (2008) https://linkinghub.elsevier.com/retrieve/pii/S0010027708002151 doi: 10.1016/j.cognition.2008.09.007.

[52] G. Andéol, E. A. Macpherson, and A. T. Sabin, "Sound localization in noise and sensitivity to spectral shape," Hearing Research **304**, 20–27 (2013) http://www.sciencedirect.com/science/article/pii/S0378595513001445 doi: 10.1016/j.heares.2013.06.001.

[53] C. V. Parise, K. Knorre, and M. O. Ernst, "Natural auditory scene statistics shapes human spatial hearing," Proceedings of the National Academy of Sciences **111**(16), 6104–6108 (2014) http://www.pnas.org/cgi/doi/10.1073/pnas.1322705111 doi: 10.1073/pnas.1322705111.

[54] B. J. Fischer and J. L. Peña, "Owl's behavior and neural representation predicted by Bayesian inference," Nature Neuroscience **14**(8), 1061–1066 (2011) www.nature.com/articles/nn.2872 doi: 10.1038/nn.2872.

[55] B. Skerritt-Davis and M. Elhilali, "Detecting change in stochastic sound sequences," PLoS computational biology **14**(5), e1006162 (2018).

[56] B. Odegaard, U. R. Beierholm, J. Carpenter, and L. Shams, "Prior expectation of objects in space is dependent on the direction of gaze," Cognition **182**, 220–226 (2019) linkinghub.elsevier.com/retrieve/pii/S0010027718302695 doi: 10.1016/j.cognition.2018.10.011.

[57] E. M. Kaya and M. Elhilali, "Modelling auditory attention," Philosophical Transactions of the Royal Society B: Biological Sciences **372**(1714), 20160101 (2017) royalsocietypublishing.org/doi/10.1098/rstb.2016.0101 doi: 10.1098/rstb.2016.0101.

[58] M. Dietz, T. Marquardt, N. H. Salminen, and D. McAlpine, "Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds," Proceedings of the National Academy of Sciences **110**(37), 15151–15156 (2013) www.pnas.org/cgi/doi/10.1073/pnas.1309712110 doi: 10.1073/pnas.1309712110.

[59] D. A. Hambrook, M. Ilievski, M. Mosadeghzad, and M. Tata, "A Bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue," PLOS ONE **12**(10), e0186104 (2017) journals.plos.org/plosone/article?id=10.1371/journal.pone.0186104 doi: 10.1371/journal.pone.0186104.

[60] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," IEEE Transactions on Speech and Audio Processing **11**(6), 826–836 (2003) ieeexplore.ieee.org/document/1255469/ doi: 10.1109/TSA.2003.818112.

[61] Majdak, Piotr, Hollomey, Clara, and Baumgartner, Robert, "Amt 1.x: A toolbox for reproducible research in auditory modeling," Acta Acust. **6**, 19 (2022) https://doi.org/10.1051/aacus/2022011 doi: 10.1051/aacus/2022011.