A Bayesian model for human directional localization of broadband and 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*

Human listeners estimate the spatial direction of a sound source from multiple auditory features and prior information on the sound direction. In this work, we describe a model of directional localization of a broadband and stationary sound source presented in an anechoic environment to a static listener. The model is based on Bayesian inference and it infers the sound direction from the spatial percept built by weighting the sensory evidence with prior beliefs. With the imposed restrictions, the sensory evidence is composed by a simplistic extraction mechanisms for two binaural features evaluating interaural disparities in time and intensity with a more detailed extraction mechanisms of monaural spectral-shape features. Our analysis considered two model variants for monaural features: amplitude spectra and their spectral gradient profiles. In addition, the model accounts for two types of uncertainties: the sensory noise in spatial cues and the response noise of pointing towards the direction. The proposed model variants were fitted to individual performance of human listeners localizing noise bursts. We then tested the predictive power of those models on the effects of localizing with other ears and sounds with rippled spectrum. This evaluation, especially with the deviation from a flat source spectrum, showed a clear preference for the model variant with spectral gradient cues, also in comparison to previously proposed localization models. Hence, the proposed model provides a tool for the evaluation of head-related transfer functions and a solid basis for future extensions towards modeling dynamic listening conditions.

©2022 [https://doi.org(DOI number)]

[XYZ] Pages: 1–14

I. INTRODUCTION

Human listeners estimate the direction of a sound source [1, 2] by decoding the spatial information provided by acoustic filtering between the listener body and the sound source [3]. In this estimation process, the ears, a pair of spatially separated sensors, enable the auditory system to extract directional information from the acoustic filtering by the listener's body.

When inferring the direction of a sound source, the auditory system uses various acoustic features [3], with their main contribution depending on the source direction. Figure 1 shows the so-called modified interaural-polar coordinate system, in which the contribution of the features is represented by the lateral and polar angles [5]. For the sound localization along the lateral angle, i.e., sound lateralization, human listeners rely on the interaural time and level differences (ITDs and ILDs) [6, 7]. Under the conditions of broadband and stationary sources, sound lateralization seems to work even when the interaural features are heavily approximated by means of broadband ITDs and ILDs [3, 8].

For the localization along the polar angle, the interaural differences are acoustically ambiguous and can only encode the sound direction along a cone, resulting in the so-called cone of confusion centered on the interaural axis [5]. Thus, in order to estimate the sound direction along the polar angle, including resolving front from back, human listeners rely on monaural spectral features provided in the HRTFs [9]. A variety of definitions for the spectral features has been proposed [10–16], with the magnitude profiles [13, 16] and the extraction of spectral gradients [11, 14] being the most prominent.

Many sound-localization models assume that the auditory system estimates the sound direction by comparing observed acoustic features with their association to a direction that has been learned over time [12, 17]. This process is usually described as the *template selection* and relies on the hypothesis that the auditory system has access to the learned knowledge, i.e., internal templates, in order to decode the latent information from the sensory evidence and to mitigate sensory uncertainty introduced by the neural system [18, 19].

Bayesian inference is a flexible concept describing such cognitive processes [20] and has been proposed to model the task of directional sound localization [21, 22]. For example, a probabilistic template selection was proposed to model the directional sound-localization task as an ideal observer [21]. This model extracts noisy auditory features which are then mapped to the most-probable

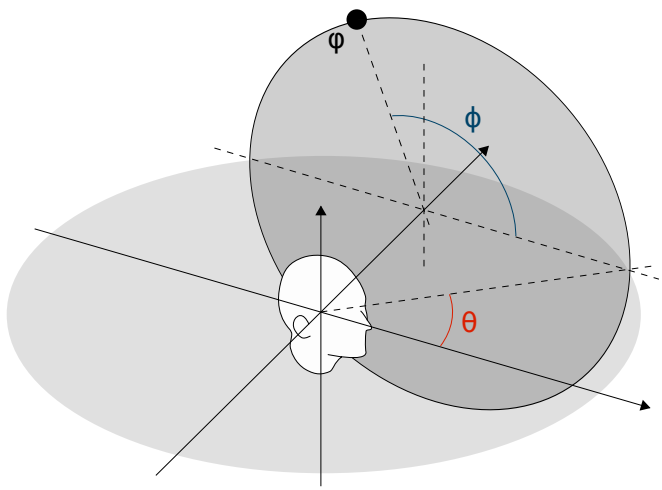

FIG. 1: The so-called modified interaural-polar coordinate system [4?]. The source direction $\boldsymbol{\varphi} = (\theta, \phi)$ is represented by the lateral angle $\theta \in [-90°, 90°]$ (left hand side is positive) describing the position of the sound source along the left/right direction of the listener and the polar angle $\phi \in [-90°, 270°]$ (0° is to the front, 90° to the top) describes the position of the sound source on the sagittal plane along the front/top/back/bottom direction of the listener. The frontal position is identified by the coordinates $(0, 0)°$.

source direction. While this model is able to generally describe the spatial localization patterns of human listeners, unfortunately, the predictions seem to diverge from the human performance in terms of localization errors and the rate of front-back confusions [23]. One of the potential reasons can be the lack of the prior information usually included in the process of Bayesian inference [20]. Thus, in an other sound-localization model, prior information on the source direction was introduced as a bias towards the horizontal plane [22]. This prior information increased the localization precision at the eye-level for the sounds along the sagittal planes. Unfortunately, their model was only proposed and evaluated within the frontal hemisphere, thus being not able to deal with the commonly observed confusions between front and back [24]. Further, this model's predictions deviate from the localization performance of actual listeners potentially because of the missing stage modeling response scattering in order to account for the spatial noise introduced by the transformation from the perceived direction to the motor response [14]. To our knowledge, a model based on Bayesian inference predicting the performance of a listener in a task of directional broadband sound localization is missing.

Here, we introduce a model based on Bayesian inference to predict the performance of a listener in a task of directional broadband sound localization. Similar to [21], our model considers a noisy feature extraction combining interaural and monaural features which are used in the selection of the internal templates obtained from listener-specific HRTFs. With the prior information on the source direction emphasizing directions at the eye level [22], the resulting spatial representation is used to perform Bayesian inference as an estimate of the source direction. The model further incorporates response scattering as a final stage [14]. The model's free parameters were fitted to the performance of individual listeners obtained in a sound-localization task [25]. In the evaluation, we considered a model variant based on spectral amplitudes and a model variant based on spectral gradients. The two variants were evaluated by simulating sound-localization experiments investigating the effects of non-individual HRTFs [26] and ripples in the source spectrum [27].

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 generalizability of the model and compares the predictions to the actual performance of listeners. Finally, Sec. IV discusses the model's ability to predict human performance and outlines suggestions for future research.

II. METHODS

A. Model description

The proposed auditory model consists of three main elements, as shown in Fig. 2: 1) the feature extraction stage determines the acoustic spatial information represented as a set of spatial features altered by noise; 2) the Bayesian inference integrates the likelihood resulting from the template selection procedure 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 (DTF) transformed in the time domain (i.e. the HRTF processed to remove the direction independent component [28]). We follow [21] in that we represent the spatial information for lateral and polar dimensions provided by a single sound source via the binaural stimulus as a vector of spatial features:

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

where $\boldsymbol{x}_{itd}$ denotes a scalar ITD feature, $\boldsymbol{x}_{ild}$ 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 meaningful spatial information to higher levels of the auditory system [7, 19].

Prior to the computation of the spatial cues and since the model's aim is direction estimation, The ILD was approximated as the time averaged broadband level difference between the left and right channels [3]. The ITD was estimated by first processing each channel of the binaural

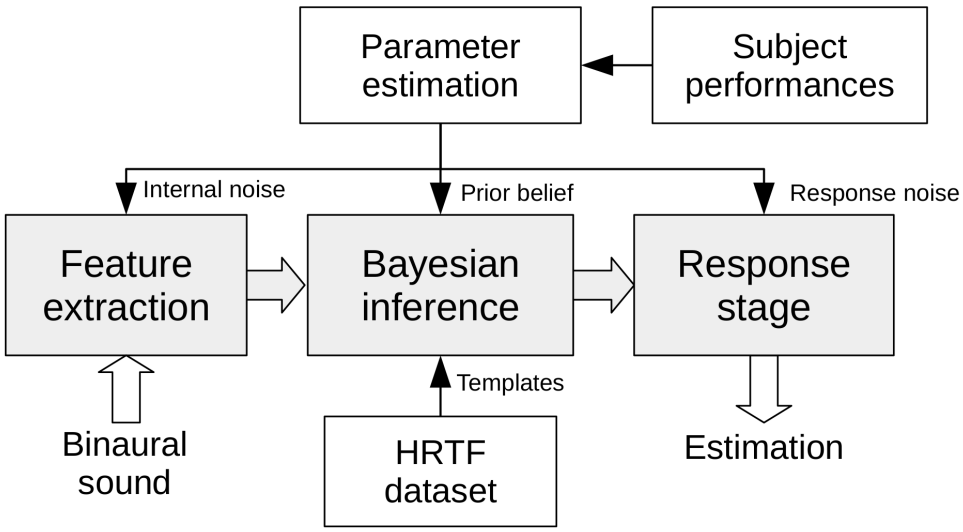

FIG. 2: Model structure. Gray blocks: Model's processing pipeline consisting of 1) the *feature extraction* stage to compute spatial features from the binaural sound; 2) the *Bayesian inference* stage integrating the template selection and prior belief to estimate the most probable direction; and 3) 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 1) internal noise corrupting the feature space; 2) internal templates with subject-specific HRTFs; 3) prior belief as response bias; and 4) response noise affecting the subject's estimate.

signal with a low-pass Butterworth filter (10th 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 [8]. In addition, we applied the transformation proposed by Reijniers *et al.* [21] to compensate the increasing uncertainty levels for increasing ITD amplitudes [29] resulting in a quantity with a more isotropic variance:

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

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 [21]). An example of the interaural features as functions of the lateral angle is shown in Fig. 3.

Monaural spectral features, $\boldsymbol{x}_{mon}^{\{L,R\}}$, 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 [30, 31], resulting in $N_B = 27$ bands within the interval $[0.7, 18]$ kHz [32, 33]. Followed by half-wave rectification and square-root compression to model hair-cell transduction [e.g., 34, 35], it results in the unit-less excitation:

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

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 $h_{\zeta}^{\boldsymbol{\varphi}}[n]$ 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 first spectral feature vector $\boldsymbol{x}_{\zeta,DTF}$ as the collection of root mean square amplitudes across time in decibels for each of the spectral bands for each ear:

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

An alternative spectral feature can be computed by positive gradient extraction (PGE) 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 [14]. Therefore, we define a second possible monaural feature vector $\boldsymbol{x}_{\zeta,PGE}$ that includes PGE as an additional pro-

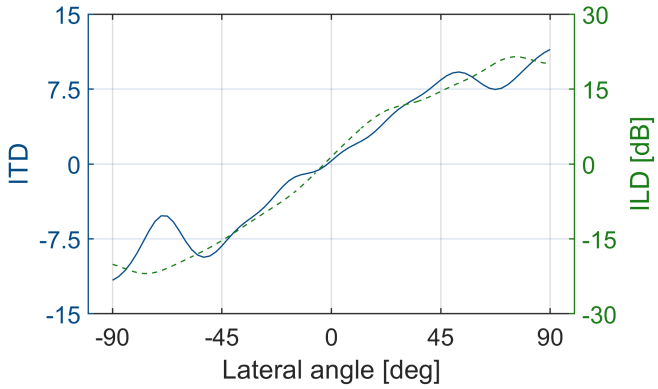

FIG. 3: Interaural features as a function of lateral angle in the horizontal frontal plane ($\phi = 0°$). The features are the transformed interaural time difference (ITD) (left axis, blue and solid line) and interaural level difference (ILD) in dB (right axis, green and dashed line) obtained from the directional transfer functions (DTFs). Example taken from subject NH12 [25].

cessing step:

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

A visualization of these monaural features is shown in Fig. 4.

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:

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

Limited precision in the feature extraction process leads to corruption of the features and can be modelled as additive internal noise [21]. 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 [7]. Hence, the covariance matrix $\boldsymbol{\Sigma}$ is defined as:

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

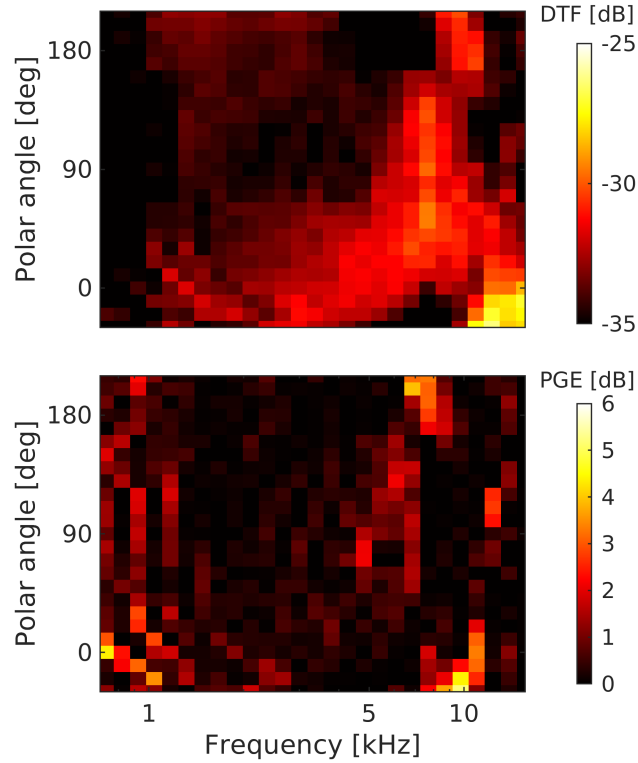

FIG. 4: Monaural spectral features as a function of polar angle in the median plane ($\theta = 0°$). Top: Features obtained from the directional transfer functions (DTFs). Bottom: Features obtained by additional positive gradient extraction (PGE). Example taken from left ear of subject NH12 [25]

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.

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 [36], 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 as a template selection procedure. Similarly to [21], 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.

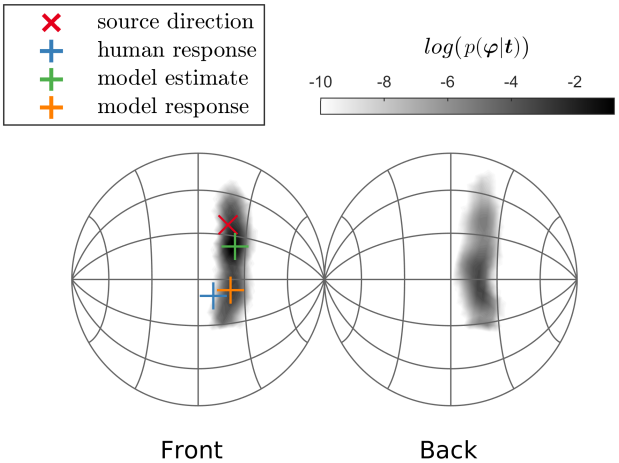

FIG. 5: Example of model's posterior probability distribution $p(\boldsymbol{\varphi}|\boldsymbol{t})$ for a broadband sound source. Red: Actual direction of the sound source. Blue: Response of the actual subject. Green: Direction inferred by the model right after the Bayesian-inference stage. Orange: Direction inferred by the model considering the response noise. Note that to increase the readability, the posterior distribution is shown on a logarithmic scale.

To accommodate non-uniform HRTF acquisition grids, we performed spatial interpolation based on spherical harmonics with order $N_{SH} = 15$, followed by Tikhonov regularization [37].

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 [22]. 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. 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 [20]. As shown in Fig. 5, 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 [14], 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. 5, 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 entails the following free parameters: σ_{ild}, σ_{mon} (amount of noise per feature; σ_{itd} was fixed to 0.569 as in [21]), $\sigma_{P,\epsilon}$ (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 [38], 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 [25, 26, 39], for a summary see [40]. 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) [26]. 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

TABLE I: Parameters estimated to fit the models to actual subjects' direction estimation performance [25].

Variant	Subject	$\sigma_{P,\epsilon}$ [deg]	σ_{ild} [dB]	σ_{mon} [deg]	σ_m [deg]
$\bar{\boldsymbol{t}}_{\mathrm{DTF}}$	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
$\bar{\boldsymbol{t}}_{\mathrm{PGE}}$	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

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 [41].

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 [22]. The interaural feature noise was set to $\sigma_{ild} = 1$ dB, reflecting the range of ILD thresholds for pure tones [42]. The starting value for the monaural feature noise was set to $\sigma_{mon} = 3.5$ dB similarly to in [21]. The response noise standard deviation was set to $\sigma_m = 17°$ as the sensorimotor scatter found in [14]. 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 $[5°, 20°]$ 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 [25] 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.

TABLE II: Predicted performance metrics averaged across all subjects and directions (±1 standard deviation across subjects). Actual data from [25].

Metric	Actual	Predicted	
		$\bar{\boldsymbol{t}}_{\mathrm{DTF}}$	$\bar{\boldsymbol{t}}_{\mathrm{PGE}}$
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

A. Parameter fits

The parameter estimation procedure was done with both model variants, based on either $\bar{\boldsymbol{t}}_{\mathrm{DTF}}$ or $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$, and for five individuals tested in a previous study [25]. 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 DTFs. 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. Notably, 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. 6 compares predicted localization performance to the actual performance of subjects estimating the direction of a noise burst for different spherical segments [25]. 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 $\bar{\boldsymbol{t}}_{\mathrm{DTF}}$ and $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$.

Contribution of model stages

Fig. 7 illustrates the effects of different model stages on target-specific predictions. The example shows direction estimations from subject NH16 localizing broadband noise bursts [25] and the corresponding predictions of the model based on $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$ 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 the *prior* in the polar dimension, we computed the polar gain as a measure of accuracy [12] for both simulated and the actual re-

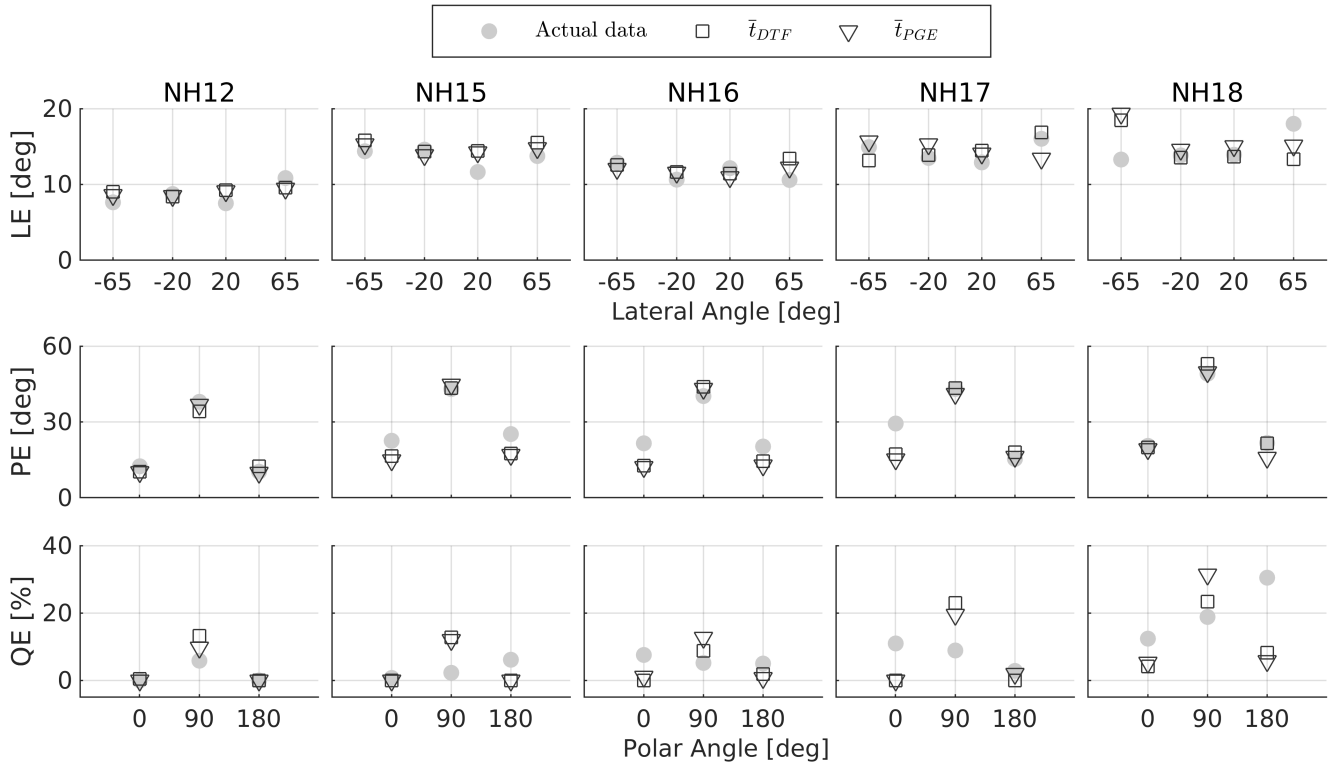

FIG. 6: Sound-localization performance as functions of sound-source direction. Open symbols: Predictions obtained by models based on either $\bar{\boldsymbol{t}}_{\mathrm{DTF}}$ or $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$. Filled grey symbol: Actual data from [25]. Top row: Lateral error (LE), calculated 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 (PE) and quadrant error (QE) rates, respectively, calculated for all median-plane ($\pm 30°$) targets with polar angles of $0° \pm 30°$, $90° \pm 60°$, and $180° \pm 40°$.

sponses. 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 unitless localization gain g_ϕ, which can be seen as a measure of accuracy [12, 22]. 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, the predicted polar gain was 1.00 (Fig. 7a). The polar gain obtained by the model including the prior was 0.62 (Fig. 7b 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. [21], 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 prior distribution $p(\boldsymbol{\varphi})$, does not include response noise (Eq. 13) and does not fit individualized parameters. The second model, described by Baumgartner et al. [14], estimates sound positions only in the polar dimension. Neverthe-

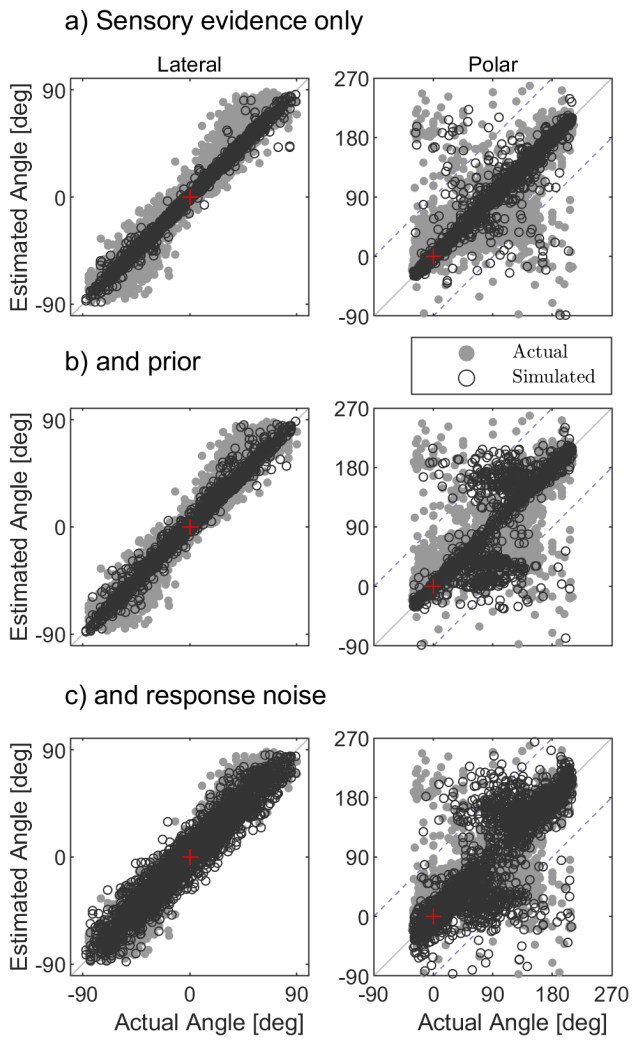

FIG. 7: Effects of likelihood, prior, and response noise on predicted response patterns as a result of modeling the localization of the direction of broadband noise bursts. **a)** Likelihood obtained by sensory evidence (i.e, no spatial prior and response noise). **b)** Bayesian inference including the spatial prior (no response noise). **c)** Full model, also including the response noise. Gray: actual data of NH16 from [25]. Black: estimation obtained by the model considering spectral gradients ($\bar{\boldsymbol{t}}_{\mathrm{PGE}}$). Each estimation is a new instantiation of the Monte Carlo sampling. Red cross: frontal position. Identity and front-back confusion lines are reported in grey and solid style and blue and dashed style, respectively.

less, 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 selection 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., 43, 44]. 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 [26]. Originally, eleven listeners localized Gaussian white noise bursts with a duration 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. 8 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 [26] and our model predictions. This mismatch reflects the fact that the digital observers represent a different pool of subjects (taken from [25]) 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 [23, 45].

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 feature spaces $\bar{\boldsymbol{t}}_{\mathrm{DTF}}$ and $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$ resulted in increased PEs and QEs, but the amount of increase was larger and more similar to the actual data for $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$, especially with respect to QE. The predictions from `baumgartner2014` were very similar to those based on $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$, 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 [23].

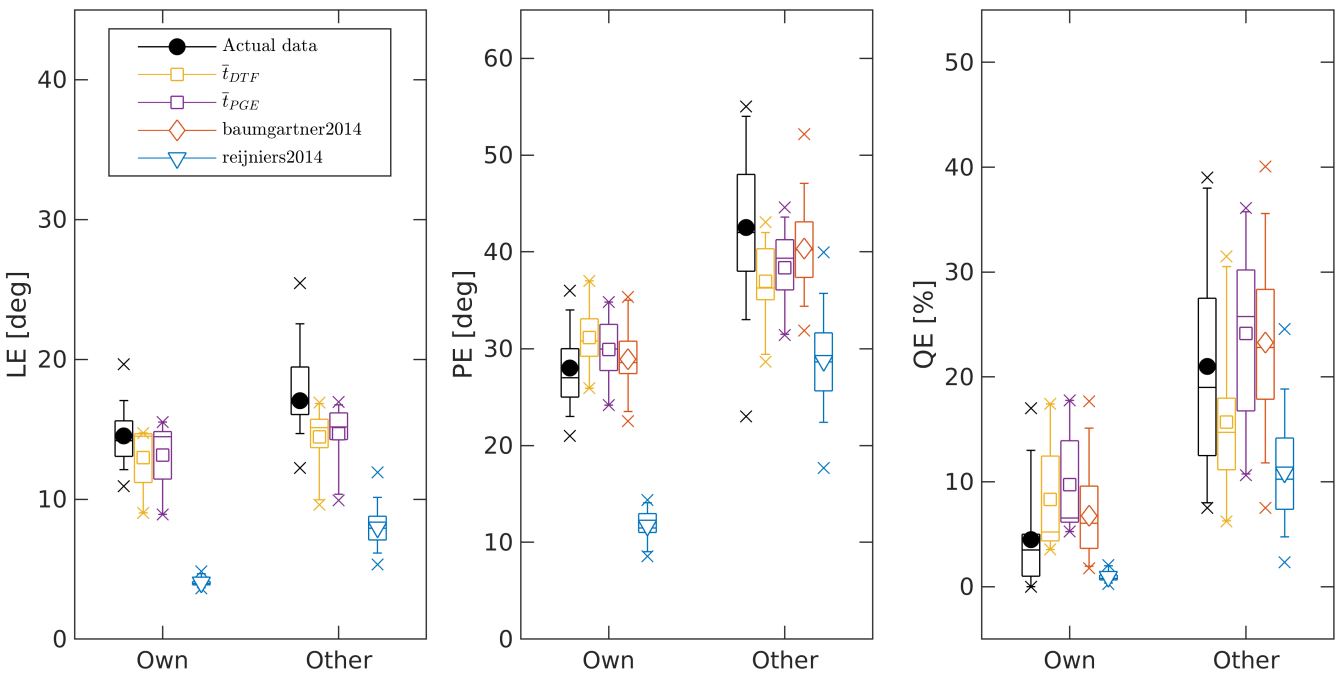

FIG. 8: Localization performance with individual (*Own*) and non-individual (*Other*) HRTFs. LE: Lateral error. PE: Polar error. QE: Quadrant error rate. Actual data from [26] (`data_middlebrooks1999`). Model predictions were based on two feature spaces, $\bar{\boldsymbol{t}}_{\mathrm{DTF}}$ and $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$. Additionally, predictions by the models `reijniers2014` [21] and `baumgartner2014` [14] are shown. Note that `baumgartner2014` does not predict the LE.

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 [27]. 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 [27]. 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. 9 shows the results of testing the fitted models with rippled spectra. In the baseline condition, our model 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 based on $\bar{\boldsymbol{t}}_{\mathrm{PGE}}$, similar to the predictions from `baumgartner2014`. In contrast, both `reijnier2014` and our model based on $\bar{\boldsymbol{t}}_{\mathrm{DTF}}$ were not able to reflect the effects of ripple density and depth as present in the actual data. Hence, the PGE appears crucial for predicting sagittal-plane localization of sources with a non-flat spectrum.

IV. DISCUSSION

The proposed functional model aims at reproducing listeners' performances in sound-source direction estimation including both the lateral and polar angles. The model formulation relies on Bayesian inference [20] as it integrates the likelihood obtained by template selection [12] with a spatial prior [22]. Our approach considers multiple uncertainties for the various sensory cues as in [21] as well as the noise introduced by pointing responses [14]. These model's components enabled to

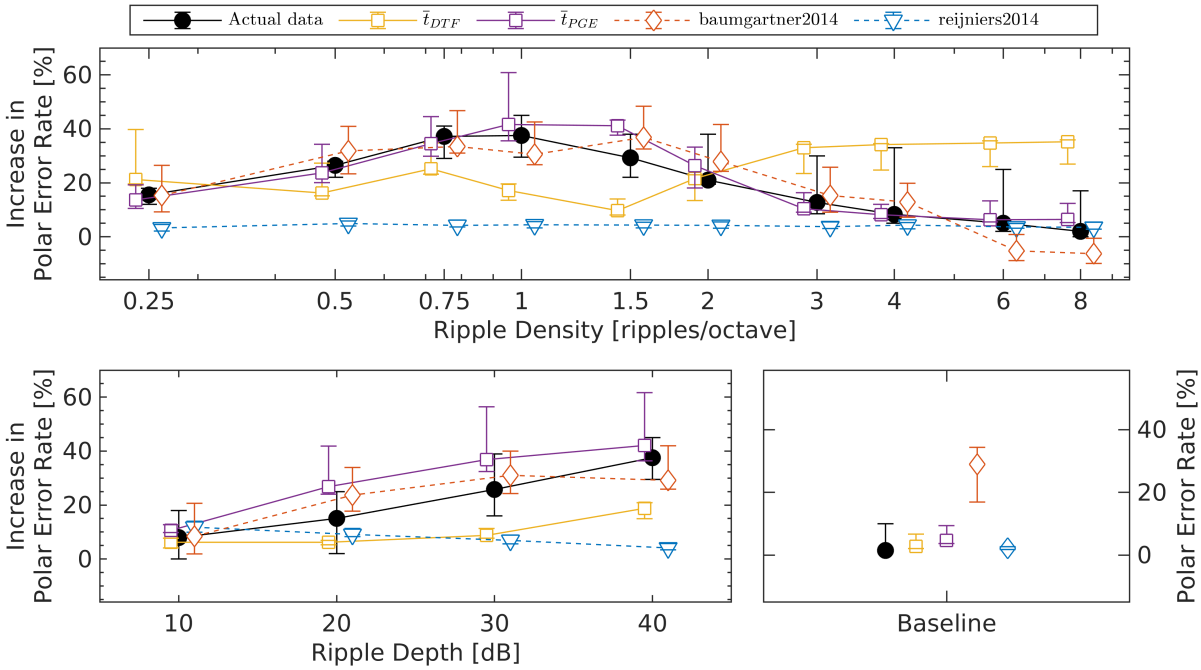

FIG. 9: Effect of spectral ripples in source spectrum on sound localization performance in medial sagittal plane. Top and bottom left panels show differences to the reference condition in the right-most bottom panel which reports the polar error rates obtained without spectral ripples. All panels show, in addition to predictions from our models, predictions from `reijniers2014` [21] and `baumgartner2014` [14] as well as actual data from [27] (`data_macpherson2003`). 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.

successfully match overall performance metrics (LE, PE, and QE) of five subjects (see Tab. II), also within spatially restricted areas (Fig. 6). The consideration of a spatial prior was necessary to explain listeners' response biases towards the horizontal plane. Compared to previous models [14, 21], our model better predicted the group-level effects of non-individualized HRTFs and rippled source spectra, yet only if positive spectral gradient profiles ($\bar{\boldsymbol{t}}_{\mathrm{PGE}}$), 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 immersed in an acoustic free-field. Additionally, the model evaluation only considered broadband and stationary sounds. The extraction of ITDs and ILDs relied on simple approximations that are sufficient for this set of conditions, as demonstrated both in our present work as well as in previous work [3, 8] by the high accuracy of predicted lateral angles, where those cues are arguable most important. The model, as presented here, does not account for feature-specific non-linearities required to predict phenomena like the precedence effect [2, 46] and should not be applied to non-stationary sounds like speech.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 integrate our model to quantify how rendering a spatial stimulus by means of generic HRTFs can result in a deviation from individual performances [47].

The model incorporates several non-acoustic parameters because such components have been found crucial to explain human performances [38, 48]. Extending the `reijniers2014` model [21] by incorporating a spatial prior and a response scatter appears vital to explain listeners' estimation patterns. Without these components, fitting the model to the polar performance metrics was unfeasible [23]. The response noise allows to control the precision at a local level (LE and PE) while leaving global errors (QE) largely affected by the noise variance added to the monaural features. The spatial prior shapes the response patterns by biasing the simulated responses towards the horizon [39]. As shown in Fig. 6, the prior contribution is visible in the PE metric where the precision is higher for directions near the eye-level while the accuracy is lower in the positions near the zenith. Additional evidence is given by the polar gain measure as reported in Sec. III A where the integration of prior beliefs better matches performances in the vertical dimension. Our formulation of the spatial prior was extrapolated from previous work [22] in the sense that its spatial distribution was assumed to be front-back symmetric. Discrepancies observed between actual and predicted global errors (Fig. 6) indicate that this assumption is inadequate. The reasons behind the existence of long-term priors in spa-

tial hearing can only be speculated at this point and may be manifold: it could reflect the spatial distribution of sound sources during everyday exposure [49], might have higher relevance from an evolutionary point of view [19], or could be related to the center of gaze as observed in barn owls [50].

While the model considers a simple static scenario, it sets the foundations for future work to predict sound localization behavior in realistic environments. To this end, the model not only needs to consider the dynamic interaction between the listener and the acoustic field. There is a rich set of modulators that influence spatial hearing and the model's prior belief is an entry point to account for those, c.f.: evidence accumulation to track source statistics [51], auditory attention to segregate sources [52], and visual influences on auditory spatial perception [53]. Evaluating the environment's dynamics as a chain of consecutive events with Sequential Monte Carlo approaches [54] as in [55] may be a promising approach. Furthermore, selective temporal integration appears important to deal with the spatial information of many natural sources and their reflections competing in realistic scenarios. This limitation could be partially addressed by integrating recent findings related to interaural cue extraction [56]. Together, these extensions will potentially enable the model to account for subject movements [57] and simultaneous tracking of source movements [58] while extracting spatial information from echoic scenarios [59].

V. CONCLUSIONS

We proposed a computational auditory model for the perceptual inference of sound-source direction based on a binaural input. Interaural and monaural spatial features are jointly used to predict the perceived direction, while not enforcing a separation of their contributions to lateral and polar angles. 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 feature extraction, the model variant evaluating gradients from the magnitude profile performed best.

The proposed model seems useful to assess the perceptual validity of HRTFs. Its validity is limited to static conditions, but it seems to provide a good basis for future extensions considering 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 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) and the Austrian Science Fund (FWF) within the project Dynamates (grant number: ZK66).

DATA AVAILABILITY STATEMENT

The implementation of the model presented in this manuscript is available in the Auditory Modeling Toolbox (AMT 1.1) as `barumerli2022` [60]. Data from [25] 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 [14], the model `reijniers2014` from [21] 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] J. W. Lord Rayleigh, Strutt, "On our perception of sound direction," Philosophical Magazine **13**, 214–232 (1907).

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

[3] 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.

[4] P. Majdak, T. Walder, and B. Laback, "Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions," The Journal of the Acoustical Society of America **134**(3), 2148–2159 (2013) http://link.aip.org/link/?JAS/134/2148/1 doi: 10.1121/1.4816543.

[5] M. Morimoto and H. Aokata, "Localization cues of sound sources in the upper hemisphere," Journal of the Acoustical Society of Japan (E) **5**(3), 165–173 (1984) doi: 10.1250/ast.5.165.

[6] J. Vliegen and A. J. van Opstal, "The influence of duration and level on human sound localization," The Journal of the Acoustical Society of America **115 (4)**, 1705–1703 (2004).

[7] 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).

[8] 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).

[9] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," The Journal of the Acoustical Society of America **88**(1), 159–168 (1990) https://asa.scitation.org/doi/abs/10.1121/1.399963 doi: 10.1121/1.399963.

[10] 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).

[11] 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).

[12] 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.

[13] 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).

[14] 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.

[15] 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).

[16] 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).

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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).

[24] P. X. Zhang and W. M. Hartmann, "On the ability of human listeners to distinguish between front and back," Hearing research **260**(1-2), 30–46 (2010).

[25] 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.

[26] 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.

[27] 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.

[28] 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).

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] A. Doucet, N. de Freitas, and N. Gordon, eds., *An Introduction to Sequential Monte Carlo Methods*, 3–14 (Springer New York, New York, NY), doi.org/10.1007/978-1-4757-3437-9_1, doi: 10.1007/978-1-4757-3437-9_1.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.