forum acusticum 2026
8-12 September · Graz, Austria

A Probabilistic Mixture Model for Evaluating Sound Localisation Performance in the Median Plane

Jakub Sztandera[1], Lorenzo Picinali[1], Roberto Barumerli[1]

[1]*Dyson School of Design Engineering, Imperial College London, UK*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Accurately evaluating sound localisation performance is essential for developing and validating methods to personalise head-related transfer functions (HRTFs). Localisation in the median plane is typically summarised with complementary metrics: polar error for local precision, polar gain for spectral decoding ability, and quadrant error rate for front-back confusions. These metrics rely on hard thresholds applied to subsets of trials, which limits statistical robustness and compresses individual differences (for instance, polar error excludes responses beyond 90° of the target, capping polar error at 52° for listeners responding at chance). Here, we propose a probabilistic model of the complete polar response distribution, based on a four-component von Mises mixture, that requires no data partitioning and negligible computation. The model free parameters are jointly identifiable, as confirmed by a synthetic recovery study, and can be estimated from as few as 45 trials per listener. Fitted to 33 listeners, each parameter correlates with a classical metric: confusion rate onto quadrant error rate, concentration onto polar error, and a spatial prior parameter onto polar gain. Model comparison further supports including the spatial prior for the large majority of listeners. The framework lends itself naturally to Bayesian extensions, offering a principled basis for more data-efficient individualised HRTF evaluation.

Keywords: *sound localisation, head-related transfer function, localisation metric, mixture model, vertical dimension*

Corresponding author: r.barumerli@imperial.ac.uk

1. Introduction

Recently, the field of spatial audio research saw the need to redefine virtual acoustic methods driven by the emergence of multifarious databases of individualised Head-Related Transfer Functions (e.g., [1]). In addition to the standard approach of recording acoustic signals at the ear canal, recent technological advancements have enabled the generation of synthetic HRTFs, e.g., using 3D scanning technology or photogrammetry, which has the potential of minimising the need for complex setups. Classic studies have long established that the benefit of individual over synthetic HRTF measurements lies in the precision of elevated/median-plane localisation, where listener-specific spectral cues dominate [2]. In practice, localisation accuracy for these sources became a de facto benchmark assessment of the HRTF individualisation, with metrics assessing performance against local and global errors. Given their established role in psychoacoustic studies of HRTF, their limitations motivate a closer examination and the development of an alternative metric.

Common metrics used to assess sound localisation performance are local polar RMS error (PE), quadrant error rate (QE) [3], and polar gain (G; [4]). The PE measures angular deviation between the reported and true location within the correct quadrant. Instead, polar error QE captures the proportion of polar responses where error exceeds 90°, thereby capturing front-back confusions. Lastly, responses with deviations ≤40° are regressed against target locations to yield an elevation-tracking metric (G), where higher values indicate better tracking. While PE mainly quantifies localisation precision, G is used as a measure of precision-accuracy trade-off and has been shown to predict the spatial prior where participants tend to respond towards the horizontal plane [5], [6].

Notably, the computation of these metrics relies on dichotomising continuous data. In the case of QE, a continuous variable is reduced to a binary result, whereas PE maintains the continuous property but implicitly relies on a 90° threshold. From a statistical perspective, dichotomising continuous localisation errors into categorical outcomes discards information about error magnitude, reduces statistical power, and increases sensitivity to arbitrary thresholds [7]. In result, QE is a coarse proportional measure that ignores the magnitude and graded structure of underlying errors. Furthermore, the effect of the threshold is particularly significant for the PE, where the ±90 ° limits make it inherently bound to converge around 52° under random sampling. Effectively, this compresses differences at poor performance levels, especially near chance. Together, these limitations highlight the need for an alternative metric that preserves the continuous structure of localisation performance and remains sensitive across the full range.

To address these limitations, we introduce a probabilistic model that captures the full structure of localisation behaviour in the median plane without reducing it to threshold-based outcomes [8]. The proposed four-component von Mises mixture model jointly accounts for front-back reversals, polar deviations and overall response accuracy within a unified circular representation [9]. Responses are modelled as a continuous distribution along the vertical dimension, preserving error magnitude, avoiding arbitrary dichotomisation, and ensuring sensitivity across the full performance range [8]. In contrast to the standard metrics, the model provides interpretable parameters that reflect processes underlying localisation errors in the median plane. In the remainder of this paper, we formalise the model and its components, evaluate its performance against standard metrics, and demonstrate its applicability to empirical localisation data.

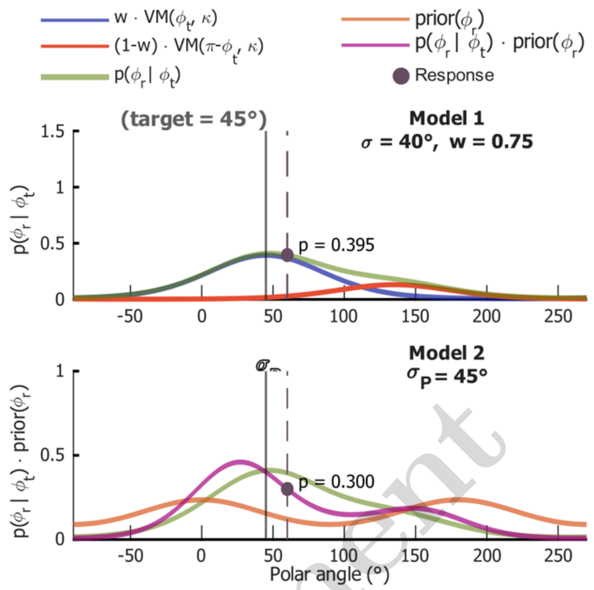

Figure 1. Response distribution of the mixture model for a target at a 45° polar angle. Top: two-parameter model ($\sigma = 40°$, $w = 0.75$); the distribution is a weighted sum of a correct-hemifield component (blue) and a mirror-hemifield component (red), yielding the marginal density (green). Bottom: three-parameter model with a spatial prior of width $\sigma_P = 45°$; the prior (orange) pulls probability mass toward the horizontal plane (0°) and its antipode (180°), reshaping the response distribution (magenta) relative to the flat-prior case. A sample response (filled circle, $p = 0.395$ and $p = 0.300$, respectively) illustrates the likelihood assigned to a near-target observation under each model.

2. Material and Methods

2.1 Empirical dataset and localisation metrics.

We evaluated the models on 33 participants (aged 18–43, median 26) who provided written informed consent under ethics approval (Imperial College London SETREC reference: 7046527, Declaration of Helsinki). The data were collated from three experiments conducted in the same sound-insulated room in static conditions using broadband and flat-spectrum sounds with an identical protocol (stimuli, apparatus, and familiarisation procedure as described in [10], [11]). Each experiment included a common baseline condition: static localisation of broadband sounds rendered binaurally with individual, acoustically measured HRTFs. Selecting only these trials yielded a variable number per participant: 9 participants contributed 297 trials (33 directions × 9 repetitions), 6 contributed 198 trials (6 repetitions), and the remaining 18 contributed 99 trials (3 repetitions), covering the polar range [−45°, 225°]. The localisation metrics were computed for the subset of trials with lateral targets within ±30° of the median plane to account for the dominance of spectral cues in such a region, yielding 45, 78 and 123 trials per participant.

We computed three localisation metrics over all available trials. Polar RMS error (PE) was calculated as the root-mean-square of polar angle differences between target and response, restricted to trials within the correct quadrant (polar error < 90°). Quadrant error rate (QE) was defined as the proportion of responses deviating more than 90° in polar angle from the target. Polar gain (G) was estimated as the slope of a linear regression of polar response on polar target, computed via the iterative sphere-fitted regression procedure of, excluding front-back confused trials. Each metric captures a distinct aspect of localisation behaviour, but none characterises the full response distribution jointly. We therefore fitted a probabilistic mixture model directly to the polar response angles.

2.2 Probabilistic model and its evaluation

We propose a probabilistic model of the polar-angle response distribution that replaces the three threshold-dependent metrics reported above with three continuously estimated parameters. The listener's response ϕ_r on a trial with target elevation ϕ_t is modelled as a draw from a four-component von Mises mixture obtained by factoring the response distribution into a perceptual term and a spatial prior term [9]:

$$p(\phi_r \mid \phi_t) \propto \left[w \cdot \mathrm{VM}(\phi_r|\phi_t,\kappa) + (1-w)\cdot \mathrm{VM}(\phi_r|\pi-\phi_t,\kappa)\right] \left[.5\cdot \mathrm{VM}(\phi_r|0,\kappa_\mathrm{P}) + .5\cdot \mathrm{VM}(\phi_r|\pi,\kappa_\mathrm{P})\right] \quad (1)$$

The von Mises distribution is the circular analogue of the Gaussian distribution, parameterised by a mean direction and a concentration κ that controls the sharpness of the peak [9]. The perceptual term captures the two dominant response modes: with probability w the listener responds near the correct hemifield (concentrated around ϕ_t with precision κ), and with probability $1-w$ near the mirror elevation $\pi - \phi_t$ (a front-back confusion, same κ). The prior term encodes a listener-specific bias toward the horizontal plane as a symmetric mixture of two von Mises distributions centred at 0° and 180° with concentration κ_P; when $\kappa_P = 0$, the prior is flat, and the model reduces to a two-component case. We will use this model simplification to assess whether the spatial prior improves the model's ability to describe individual response patterns.

The loglikelihood for a specific participant with N trials is:

$$\mathrm{NLL} = \sum_{n=1}^{N} \log(\mathrm{p}(\phi_{r,n}|\phi_{t,n})) \quad (2)$$

Parameters are estimated by maximising Eq. 1 and 2 via the Bayesian Adaptive Direct Search algorithm (BADS) [12], a derivative-free optimiser suited to stochastic objectives, with w optimised in logit space and κ, κ_P in log space. For reporting, we converted the concentration parameters κ and κ_P to the equivalent standard deviation in degrees (σ and σ_P, respectively), obtained via the inverse of the standard von Mises variance formula. The three parameters carry direct interpretations: σ is response uncertainty, $1-$ w is front-back confusion rate, and σ_P is the width of the horizontal prior (i.e. a smaller value indicates a stronger pull toward the horizontal plane). Fig. 1 illustrates the response density for a target at $\phi_t = 45°$ under a flat prior and a strong prior.

To verify that all three parameters (i.e. κ, κ_P and w) are jointly identifiable before fitting to real data, we generated 28 synthetic listeners via Latin Hypercube Sampling over ($w \in [0.55, 0.95]$, $\sigma \in [10°, 55°]$, $\sigma_P \in [10°, 90°]$), ensuring near-zero pairwise correlations among true parameters (Pearson correlation coefficient $|r| < 0.32$). Each synthetic listener received $N = 45$ trials with target elevations drawn with replacement from the empirical target distribution, matching the number of trials available in the median plane from actual data. Responses were sampled from the generative model (Eq. 1) and from these, the three-parameter model was fitted. Recovery quality was assessed by Pearson correlation and mean signed bias between synthetic and retrieved parameters.

After parameter identification, we fitted two model variants (with and without spatial prior) to the 33 listeners using trials on the median plane. To test

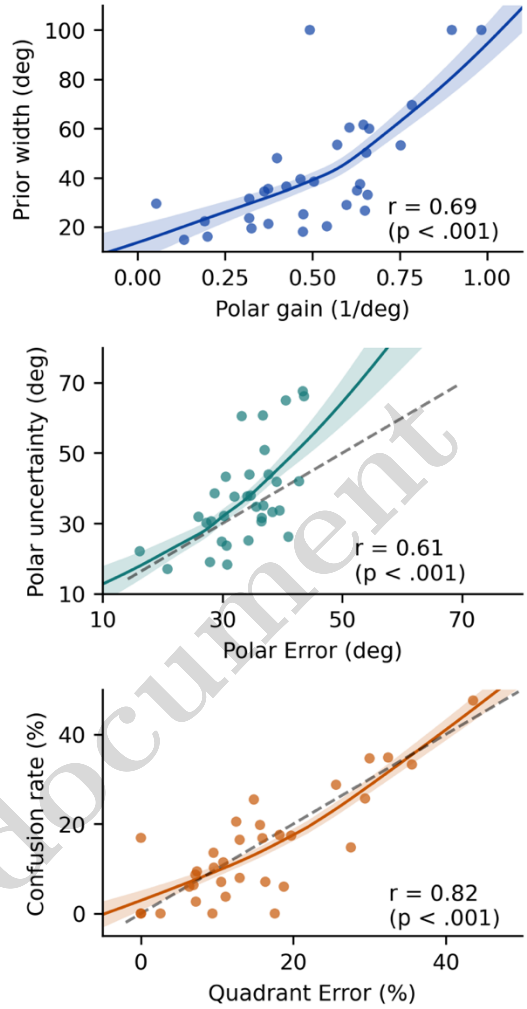

Figure 2. Correspondence between model-derived parameters and localisation metrics for 33 listeners. Each panel shows a scatter of one model parameter against its classical counterpart, with a LOESS smooth (solid line) with the ribbon reporting the standard error over participants. [Top] Prior width σ_P vs. polar gain; [Middle] polar uncertainty σ vs. polar RMS error (PE); [Bottom] confusion rate (1−w) vs. quadrant error rate (QE). Identity line shown dashed.

whether σ_P captures variance in elevation-tracking ability that classical metrics do not, we computed Pearson correlations between each model parameter and the localisation metrics (i.e. polar error, quadrant error and polar gain). Extended and reduced models were compared by computing the difference of their Bayesian Information Criterion [13]:

$$\mathrm{BIC} = k \cdot \ln(N) + 2 \cdot \mathrm{NLL} \quad (3)$$

where NLL is the maximised likelihood obtained from Eq. 2, penalised by the number of free parameters (k) and corrects for differences in trial counts used for fitting with ln(N).

3. Results

3.1 Parameter recovery

All three parameters were recoverable from synthetic data with N = 45 trials, confirming joint identifiability. Pearson correlations on the transformed scale were $r = 0.83$ for w, $r = 0.89$ for κ, and $r = 0.84$ for κ_P (all $p < 0.001$). Estimated parameters were mutually uncorrelated ($r = -0.08$, $p = 0.678$, between σ and σ_P), indicating that these two parameters do not trade off against each other in estimation. Mean bias was small for w (−0.005) and σ (−0.18°) and more negative for σ_P (−4.22°), the latter reflecting boundary accumulation for synthetic listeners with $\sigma_P > 60°$, where estimates approached the upper optimisation bound of 100°. Crucially, ignoring the spatial prior did not bias the other two parameters: the mean difference in σ between the reduced and extended models was only −1.00°, and in w only −0.012, indicating that the two-parameter model remains a reliable descriptor of precision and confusion rate even when a spatial prior is present in the data.

3.2 Parameter estimation from empirical data

The extended model converged for all 33 listeners (see Table 1 for parameter estimates). Estimated confusion rates ranged widely across participants (1 − w: 0–47.5%), consistent with the known inter-individual variability in front-back confusion rates for individual HRTFs. Response precision showed a similarly broad spread (σ: 17.0–67.5°), reflecting differences in spectral cue sensitivity across the group. The spatial prior parameter σ_P ranged from 14.7° to 100°, with three participants reaching the upper bound, indicating an uninformative horizontal prior for those individuals.

Table 1. Summary of the two model fits. Values are mean ± standard error over 33 listeners. The two-parameter model serves as reference for goodness of fit (ΔBIC).

Parameter	Generative model: without spatial prior	Generative model: with spatial prior
Prior width σ_P (deg)	-	41.6 ± 4.1
Polar uncertainty σ (deg)	38.5 ± 1.9	37.4 ± 2.4
Confusion rate (1−w) (%)	14.6 ± 2.0	14.3 ± 2.1
ΔBIC	0	-34.1 ± 6.3

3.3 Relationship between model parameters and classical metrics

We first examined the correspondence between model-derived parameters and the localisation metrics QE and PE (see Fig. 2). The mixture weight w showed a strong linear relationship with QE (Pearson $r = +0.83$, $p < 0.001$), confirming that the confusion rate 1 − w captures the same phenomenon as the quadrant error count but within a generative framework that uses all trials rather than partitioning responses by a fixed threshold. The uncertainty parameter σ was correlated with PE ($r = +0.61$, $p < 0.001$), consistent with σ representing the spread of responses around the target in the correct hemifield. However, PE is bounded above by approximately 52°, where the RMS error expected under a uniform response distribution on a half-circle, so for listeners with high confusion rates who are near chance on polar responses, PE saturates and can no longer distinguish between different degrees of imprecision. The model does not share this limitation: σ is estimated from the full response distribution, with the confusion component absorbed by 1 − w rather than discarded. These correspondences establish that the model parameters are behaviourally grounded: w and σ recover the same signal as their classical counterparts while avoiding their threshold-induced distortions and using all trials for parameter estimation [8]. Another key difference is that the model simultaneously estimates a third parameter, σ_P, that has no classical equivalent (i.e., it is not a linear combination of QE and PE).

Results also show that the spatial prior width σ_P, estimated directly from individual response distributions via maximum likelihood, recovers this relationship at the level of individual listeners: σ_P correlates positively with polar gain across listeners (Pearson $r = +0.685$, $p < 0.001$; regression slope 0.0061 per degree of σ_P). Listeners with a stronger horizontal prior (i.e., smaller σ_P, meaning their responses are more strongly attracted toward the horizontal plane) track target elevation less faithfully. This relationship is illustrated in Fig. 2.

3.4 Model comparison

As a final evaluation, we compared the fitting quality of two model variants: one with and one without a spatial prior. Over the 33 listeners, the extended model with spatial prior was preferred for 26/33 participants (ΔBIC < −2), with a mean ΔBIC of −34.1. The four participants for whom the reduced model was preferred (ΔBIC > +2) all had large σ_P estimates (≥ 69°), with three reaching the upper optimisation bound of 100°, indicating a flat or near-flat prior for those individuals; consistently, the BIC penalty for the extra parameter reverses model preference precisely when that parameter carries no information. The strongest BIC improvements were concentrated among listeners with $\sigma_P \leq 22°$ (e.g. P0181: ΔBIC = −140.65, P0204: −92.01, P0015: −86.67), confirming that the model accounting for the spatial prior is most valuable when the horizontal prior is strong.

4. Discussion

Our comparison using empirical behavioural data demonstrated that the proposed probabilistic model notably outperformed the standard localisation metrics. The three fitted model parameters, i.e., prior width, polar uncertainty, and confusion rate, correlated with their conventional counterparts (G, PE, and QE, respectively), confirming that they capture interpretable phenomena underlying localisation behaviour. Crucially, whereas the classical approach that relies on ad-hoc summaries of differently partitioned data, these parameters are obtained from a jointly fitted mixture model, are mutually orthogonal, and are derived from the full response distribution [8]. This eliminates the need for arbitrary data exclusion and ensures that all trials contribute simultaneously to the estimation of distinct behavioural components, including uncertainty, front–back confusions, and spatial bias. Independence is a particularly important distinction between polar gain and polar width proposed here, as the latter reflects dispersion directly rather than a slope-based summary. Furthermore, the model's sensitivity across the full performance range allows it to provide informative characterisation of behaviour beyond the saturation point of PE. Taken together with the possibility of goodness-of-fit assessment and formal hypothesis testing through likelihood-based metrics, the proposed framework enables richer inference in comparison to classical metrics.

The relevance of this framework becomes particularly evident in the context of recent developments in HRTF generation and individualisation. A wide range of approaches is currently being explored, spanning from measured [1] and anthropometry-based [14] to fully synthetic or data-driven HRTFs [15]. Many approaches might exhibit more subtle and heterogeneous degradations on localisation performances since only a specific portion of spectral cues is degraded [15]. By contrast, the proposed framework might provide the flexibility necessary to disentangle these effects, enabling a more diagnostic comparison in which different HRTF generation methods can be characterised in terms of distinct behavioural changes in uncertainty, bias, and confusion rate. Furthermore, by leveraging all trials, our framework offers the possibility of reducing the number of trials required to obtain a reliable estimate. These metrics can also improve the definition of feature space for the selection of publicly available HRTFs. More broadly, this parameterisation allows localisation performances to be interpreted in terms of underlying perceptual and prior-driven mechanisms, offering a principled tool for studying how sensory information is integrated, degraded, or adapted across experimental conditions [16].

Despite these advantages, several limitations of the current framework should be acknowledged. The present formulation assumes a shared concentration parameter for the von Mises components and is restricted to median-plane localisation, which may limit its ability to capture more complex response patterns observed in full-sphere tasks. Extending the model to account for lateral dimensions and additional perceptual phenomena, such as lateral uncertainty or combined spatial accuracy measures, would broaden its applicability. Furthermore, the current analysis is based on a static scenario and does not capture temporal dynamics; incorporating time-varying parameters could provide a principled way to study deviations from individual HRTFs [17], or learning/adaptation effects [18]. From a methodological perspective, future work could explore hierarchical formulations to better account for inter-listener variability and enable population-level inference, as well as Bayesian estimation approaches (e.g., via Markov Chain Monte Carlo) to support more comprehensive uncertainty quantification and hypothesis testing [19]. Finally, validating the framework on larger and more heterogeneous datasets will be essential to assess its robustness and generalisability across experimental conditions [20].

5. Conclusions

This work proposes a four-component von Mises mixture model for evaluating median-plane sound localisation that replaces three threshold-dependent metrics with three jointly estimated, interpretable parameters. Applied to 33 listeners, the model recovered the signal captured by standard metrics while additionally revealing individual differences in horizontal prior width, which predicts polar gain across listeners. The framework is computationally lightweight, requires no data partitioning, and is naturally extensible to hierarchical and Bayesian formulations, offering a principled basis for data-efficient HRTF evaluation.

6. Acknowledgments

This work was supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101201118 (project MIA), and the Horizon 2020 research and innovation funding programme under grant agreement No. 101017743 (project SONICOM).

The code implementing the proposed model and analysis is available at https://github.com/robaru/fa2026-localisation-mixture. The behavioural data are available at https://ecosystem.sonicom.eu/databases/58.

References

[1] I. Engel *et al.*, "The SONICOM HRTF Dataset," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 241–253, May 2023.

[2] J. C. Middlebrooks, "Sound localization," in *Handbook of Clinical Neurology*, vol. 129, Elsevier, 2015, pp. 99–116. doi: 10.1016/B978-0-444-62630-1.00006-8.

[3] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, Aug. 1999, doi: 10.1121/1.427147.

[4] E. A. Macpherson and J. C. Middlebrooks, "Localization of brief sounds: effects of level and background noise," *J Acoust Soc Am*, vol. 108, no. 4, pp. 1834–49, Oct. 2000.

[5] R. Ege, A. J. V. Opstal, and M. M. Van Wanrooij, "Accuracy-Precision Trade-off in Human Sound Localisation," *Sci. Rep.*, vol. 8, no. 1, p. 16399, Dec. 2018, doi: 10.1038/s41598-018-34512-6.
[6] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A Bayesian model for human directional localization of broadband static sound sources," *Acta Acust.*, vol. 7, p. 12, 2023, doi: 10.1051/aacus/2023006.
[7] J. DeCoster, M. Gallucci, and A.-M. R. Iselin, "Best Practices for Using Median Splits, Artificial Categorization, and their Continuous Alternatives," *J. Exp. Psychopathol.*, vol. 2, no. 2, pp. 197–209, May 2011, doi: 10.5127/jep.008310.
[8] R. Barumerli and P. Majdak, "FrAMBI: A Software Framework for Auditory Modeling Based on Bayesian Inference," *Neuroinformatics*, vol. 23, no. 2, p. 20, Feb. 2025, doi: 10.1007/s12021-024-09702-5.
[9] K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009.
[10] R. Daugintis, R. Barumerli, M. Geronazzo, and L. Picinali, "Initial evaluation of an auditory-model-aided selection procedure for non-individual hrtfs," in *Proceedings of the 10th Forum Acusticum*, Turin, Italy, 2023, pp. 1–8.
[11] R. Daugintis, M. Geronazzo, K. C. Poole, and L. Picinali, "Perceptual evaluation of an auditory model–based similarity metric for head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 159, no. 3, pp. 2822–2843, Mar. 2026, doi: 10.1121/10.0043130.
[12] L. Acerbi and W. J. Ma, "Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.
[13] D. L. Weakliem, "A Critique of the Bayesian Information Criterion for Model Selection," *Sociol. Methods Res.*, vol. 27, no. 3, pp. 359–397, Feb. 1999, doi: 10.1177/0049124199027003002.
[14] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Franm, and D. Manocha, "P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound," in *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Sep. 2014, pp. 53–61. doi: 10.1109/ISMAR.2014.6948409.
[15] L. Pirard, L. Picinali, and K. C. Poole, "Photogrammetry-Reconstructed 3D Head Meshes for Accessible Individual Head-Related Transfer Functions," *Acta Acust.*, Mar. 2026, doi: 10.48550/arXiv.2603.24104.
[16] W. J. Ma, K. P. Kording, and D. Goldreich, *Bayesian Models of Perception and Action: An Introduction*. MIT press, 2023.
[17] P. Lladó, R. Barumerli, R. Baumgartner, and P. Majdak, "Predicting the effect of headphones on the time to localize a target in an auditory-guided visual search task," *Front. Virtual Real.*, vol. 5, Jun. 2024, doi: 10.3389/frvir.2024.1359987.
[18] D. Meijer, R. Barumerli, and R. Baumgartner, "How relevant is the prior? Bayesian causal inference for dynamic perception in volatile environments," *eLife*, vol. 14, Feb. 2025, doi: 10.7554/eLife.105385.1.
[19] K. Ignatiadis, D. Baier, R. Barumerli, I. Sziller, B. Tóth, and R. Baumgartner, "Cortical signatures of auditory looming bias show cue-specific adaptation between newborns and young adults," *Nat. Commun. Psychol.*, vol. 2, no. 1, pp. 1–15, Jun. 2024, doi: 10.1038/s44271-024-00105-5.
[20] T. Greif, R. Barumerli, K. Ignatiadis, B. Tóth, and R. Baumgartner, "The role of spatial perception in auditory looming bias: neurobehavioral evidence from impossible ears," *Front. Neurosci.*, vol. 19, Aug. 2025, doi: 10.3389/fnins.2025.1645936.