Confidence in Sound Localization Reflects Calibrated Uncertainty Estimation

Lakshmi Narasimhan Govindarajan[1,2,3], **Sagarika Alavilli**[2,4], **Josh H. McDermott**[1,2,3,4]
{lakshmin, salavill, jhm}@mit.edu

[1]Department of Brain and Cognitive Sciences
[2]McGovern Institute for Brain Research
[3]K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center
MIT, 43 Vassar Street, Cambridge, MA 02139, USA

[4]Speech and Hearing Biosciences and Technology, Harvard, Cambridge MA 02318, USA

Abstract

Humans localize sounds using a combination of binaural and monaural cues. However, the location of a sound remains ambiguous under many conditions. Because sound localization is often used to guide behavior, representing the uncertainty of a sound's location is likely to be critical to decisions about where and when to act. However, little is known about whether humans represent the uncertainty associated with a sound's location and whether any such representations are calibrated to the accuracy of localization. To study these issues, we developed a new class of stimulus-computable models to enable the representation of uncertainty. We optimized the model for sound localization in natural conditions and then compared its uncertainty estimates to those of humans.

Keywords: Sound localization; perceptual uncertainty; circular regression; von Mises distribution; directional statistics

Introduction

The location of a sound in the world is not directly specified in the sensory input, but rather must be *inferred* from acoustic cues in the signals arriving at the two ears. These include interaural time and level differences, as well as spectral shaping by the outer ear (Blauert, 1997). However, such cues are often ambiguous due to (1) the geometric symmetry of the head and ears and (2) corruptions from noise, reverberation, and concurrent sources. Together, these factors introduce perceptual uncertainty that affects spatial judgments and, consequently, behavior (Obleser, 2025; Van den Berg, Zylberberg, Kiani, Shadlen, & Wolpert, 2016).

Recent work has yielded deep neural network models that rival humans in their ability to localize single sources (Francl & McDermott, 2022; Saddler & McDermott, 2024). However, these models are typically optimized for discriminative performance and do not explicitly represent uncertainty in their spatial estimates, making them ill-suited to account for trial-by-trial human confidence. Models of uncertainty from other domains have almost exclusively been limited to few-alternative decision-making and so are not naturally adapted to real-world perceptual problems (Li & Ma, 2020; Keshvari, Van den Berg, & Ma, 2012; Kepecs & Mainen, 2012).

We developed a novel class of neural network models that represent uncertainty by predicting full probability distributions over an estimated variable. When applied to sound localization, the model estimates a distribution over location from binaural audio. When a sound's location is ambiguous, the model produces broader, or potentially even multimodal, distributions. To evaluate how these predictions relate to human confidence, we conducted two experiments measuring human listeners' localization confidence in different conditions.

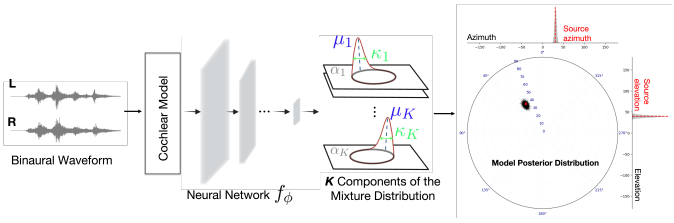

Figure 1: **Computational framework for uncertainty-aware sound localization** | Binaural waveforms are filtered by simulated human ears (Francl & McDermott, 2022). The resulting cochleagrams are transformed into a low-dimensional embedding by a neural network model. We interpret this embedding to represent the parameters (circular means $\{\mu_j\}_{j=1..K}$, concentrations $\{\kappa_j\}_{j=1..K}$, and component weights $\{\alpha_j\}_{j=1..K}$) of a K-component von Mises mixture that denotes a probability distribution over sound location. The model reports perceived source locations by sampling from this density.

Methods

Model and training details

Architecture and training objective. Binaural audio waveforms were processed by a gammatone filter bank ($N_f = 40$ frequency channel bins with filters uniformly spaced between 40Hz and 20kHz on an ERB_N scale, with bandwidths approximating those of a healthy human ear). Filter bank outputs were half-wave rectified and low-pass filtered with a 4kHz cutoff frequency to simulate the upper cutoff of phase locking in the mammalian ear. The base model architecture was adapted from prior literature (Francl & McDermott, 2022), replacing the readout layer to facilitate the likelihood-based training objective.

Model readouts were factorized to represent the parameters of a bivariate (azimuth/elevation) von Mises mixture density (Figure 1) specified as

$$p(\Theta|\{\alpha_\mathbf{j},\mu_\mathbf{j},\kappa_\mathbf{j}\}_{j=1..K}) = \sum_{j=1}^{K} \alpha_j \frac{e^{\kappa_\mathbf{j} cos(\Theta-\mu_\mathbf{j})}}{2\pi I_0(\kappa_\mathbf{j})}, \qquad (1)$$

where Θ is the true location, $\{\alpha_\mathbf{j},\mu_\mathbf{j},\kappa_\mathbf{j}\}_{j=1..K}$ are neural network outputs and $I_0(.)$ is the Bessel function of order 0. We trained the model to perform heteroskedastic regression by minimizing the negative log-likelihood of the true source locations for a scene.

Dataset generation. We used a room simulator to generate spatialized scenes in different rooms (Shinn-Cunningham, Desloge, & Kopco, 2001) with the listener at random positions and angles within a room. 1800 rooms were used in training and a different set of 200 rooms were used for validation. Source locations were generated every 5° in azimuth (0° to 355°) and 10° in elevation (0° to 60°). Source distance was varied from 1.4m to the furthest distance within the room.

Each training example consisted of a 1-second binaural audio clip featuring a natural foreground sound from the GISE-

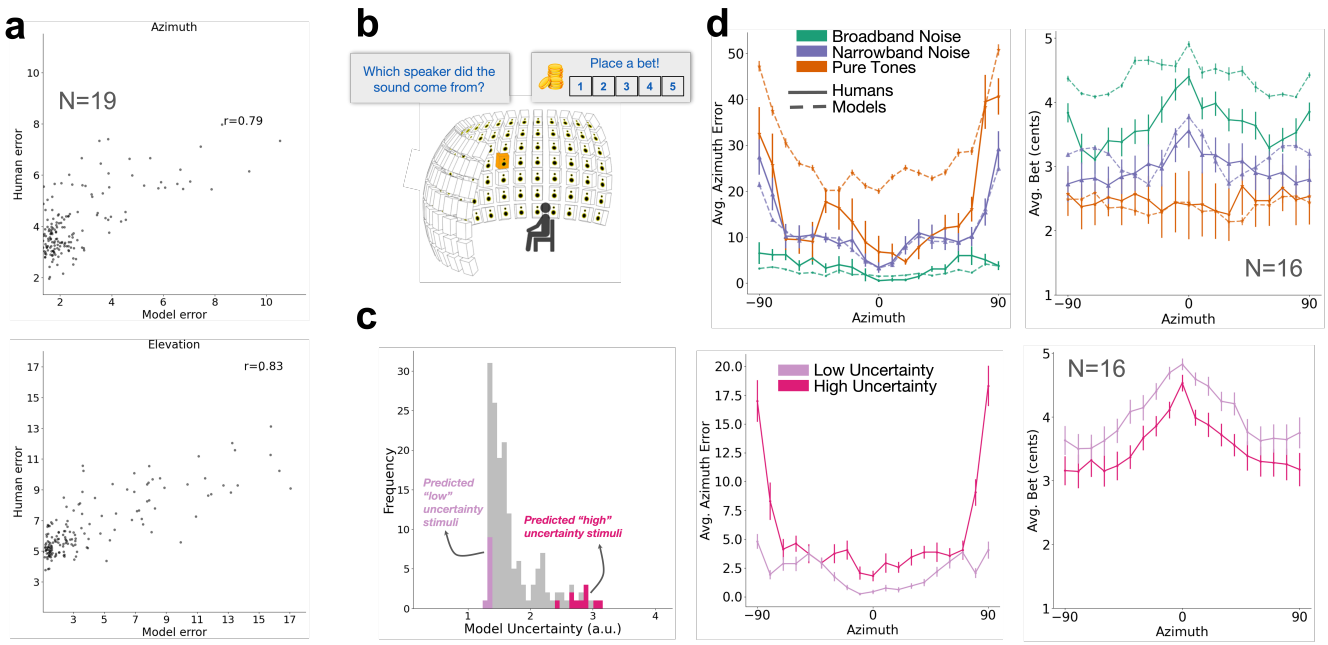

Figure 2: **Human and model psychophysics reveal calibrated uncertainty estimation** | **a.** Item-wise human-model alignment in azimuth (top) and elevation (bottom) errors. **b.** Localization betting paradigm for Experiments 2-3. **c.** Histogram of model uncertainties for 160 natural sounds. Highlighted stimuli were subsequently presented to human listeners in Experiment 3. **d.** Absolute azimuthal error (left) and average bet (right) vs. azimuth, for Experiment 2 (top) and Experiment 3 (bottom).

51 dataset (Yadav & Foster, 2021), spatialized to a randomly chosen room and location. Between 10 and 20 background sounds from AudioSet, spatialized to randomly sampled locations within the same room, were added at an SNR $\in [-15, +25]$dB. We generated a total of ~ 1.8M training scenes and 500K validation scenes.

Human experiments

General overview. Participants heard sounds played from a speaker array. The participant sat in the center of the array and entered their responses through an iPad interface. To test the models, we rendered the same stimuli from each experiment in a virtual replica of the speaker array room.

Experiment 1. Stimuli consisted of 160 unique natural sounds, each 1 second long. Participants were asked to judge the location of the source (Francl, 2022).

Experiments 2 and 3. Participants localized the source and placed a bet (from 1-5 cents) on their answer. Trial-level feedback was not provided. The model's "bet" was a monotonic function of the posterior entropy for that trial. **Experiment 2:** Stimuli were broadband noise, pure tones ($600 - 4000$ Hz), and narrowband noise (same center frequencies as the tones; half octave bandwidth). **Experiment 3:** 10 high uncertainty and 10 low uncertainty natural sound stimuli were selected from the original 160 sounds based on model predictions (Fig. 2c).

Results & Discussion

Experiment 1: Model validation. Model and human error patterns were correlated across test stimuli (Fig. 2a). After correcting for attenuation, the item-wise correlation between model and human errors was $r = 0.79, p < .001$ for azimuth, and $r = 0.83, p < .001$ for elevation.

Experiment 2: Uncertainty in humans is calibrated. Human confidence judgments varied systematically with stimulus conditions, with lower bets on stimuli that elicited higher localization errors (Fig. 2d). Confidence was lower for (1) peripheral sound sources and (2) pure tones compared to noise. These patterns suggest that human confidence estimates are well calibrated. The model showed similar condition-dependent variations in uncertainty despite being optimized purely for localization performance.

Experiment 3: Model predicts human confidence. Human accuracy and bets differed between the high- and low-confidence stimulus groups of the model (Fig. 2c,d).

General conclusion Human confidence estimates for sound localization are similar to those of a model whose uncertainty representations are optimized for accurate localization, indicating that human confidence is normatively appropriate in this domain. The modeling framework provides a way to investigate confidence in realistic perceptual problems.

Acknowledgments

This work was supported by National Institutes of Health (Grant number R01DC021464). LNG was supported by the ICoN Postdoctoral Fellowship. SA was supported by NIH National Institute on Deafness and Other Communication Disorders (Grant T32 DC000038). We thank Adanna Thomas for assisting in data collection.

References

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT press.

Francl, A. (2022). *Modeling and evaluating human sound localization in the natural environment*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Francl, A., & McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature human behaviour*, *6*(1), 111–133.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337.

Keshvari, S., Van den Berg, R., & Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS One*, *7*(6), e40216.

Li, H.-H., & Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the bayesian confidence hypothesis. *Nature communications*, *11*(1), 2004.

Obleser, J. (2025). Metacognition in the listening brain. *Trends in Neurosciences*.

Saddler, M. R., & McDermott, J. H. (2024). Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, *15*(1), 1–29.

Shinn-Cunningham, B. G., Desloge, J. G., & Kopco, N. (2001). Empirical and modeled acoustic transfer functions in a simple room: Effects of distance and direction. In *Proceedings of the 2001 ieee workshop on the applications of signal processing to audio and acoustics (cat. no. 01th8575)* (pp. 183–186).

Van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology*, *26*(23), 3157–3168.

Yadav, S., & Foster, M. E. (2021). Gise-51: A scalable isolated sound events dataset. *arXiv preprint arXiv:2103.12306*.