A Computational Framework for Sound Localization in Auditory Scenes

Lakshmi Narasimhan Govindarajan, Ajani Stewart, Sagarika Alavilli, Josh H. McDermott
{lakshmin, ajani, salavill, jhm}@mit.edu

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

Speech and Hearing Biosciences and Technology, Harvard, Cambridge MA 02318, USA

Abstract

Humans routinely localize sounds in the world, but little is known about localization abilities in the presence of concurrent sources. We developed a model of multi-source localization by training a model to generate a probability distribution over locations given binaural audio input. We conducted an experiment to measure human multi-source localization in scenes composed of multiple natural sounds at different locations in azimuth and elevation. Human localization became less accurate as the number of sources was increased, showing marked impairments even for two sources compared to one. The model replicated this dependence on the number of sources, suggesting that human limitations are likely inevitable consequences of sampling the spatial world with only two sensors.

Keywords: Auditory scenes; multi-source sound localization; circular regression; von Mises–Fisher distribution; directional statistics

Introduction

The location of a sound in the world is not explicit in our sensory input, and instead must be estimated from cues, in part those derived by comparing sound from our two ears. The localization of individual sound sources has been studied for decades, and is known to rely on both binaural (interaural time/level differences) and monoaural (spectral filtering) cues. By comparison, much less is known about how we localize sounds in scenes containing multiple sources. The few instances in which human localization of multiple concurrent sources has been tested suggest that the problem is challenging for humans (Zhong & Yost, 2017).

Recent progress in computational modeling of sound localization has yielded performant deep neural network models that rival humans in their ability to localize *single* sources (Francl & McDermott, 2022). Such models have treated the continuous space of the three-dimensional world as discretized bins, modeling single source localization as a deterministic discriminative process. This approach is not ideal for modeling localization in auditory scenes as it requires transforming a multi-label task into multiple (independent) single-label tasks.

We developed a new class of localization model to circumvent these challenges. Given an auditory scene, we estimate a probability distribution over spatial locations, inferring multiple modes for scenes with multiple sources. We conducted two experiments on human listeners to test the model on single and multi-source localization.

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 space between 40Hz and 20kHz; 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 model architecture was adapted from prior literature (Francl & McDermott, 2022), replacing the readout layer to facilitate a likelihood-based training objective.

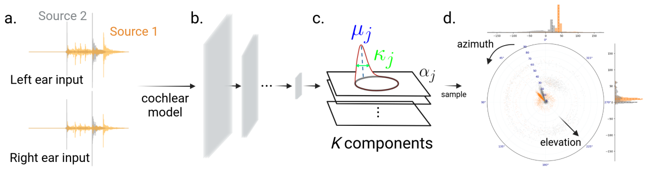

Figure 1: **Localizing sounds in auditory scenes.** (a) Input binaural waveforms are filtered by simulated human ears (Francl & McDermott, 2022). (b-c) From the resulting 'cochleagrams', a neural network model extracts a low dimensional embedding that we interpret as the parameters (circular means $\{\mu_j\}_{j=1..K}$, 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. (d) The model reports perceived source locations by sampling from this density.

Model readouts were factorized to represent the parameters of a bivariate (azimuth/elevation) von Mises mixture density (Figure 1c) 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 train our model to perform heteroskedastic regression by minimizing the negative log-likelihood of the true locations of the sources in a scene.

Dataset generation We used a room acoustic simulator to generate spatialized scenes in different rooms (Shinn-Cunningham, Desloge, & Kopco, 2001) with the listener at randomly selected positions and angles within a room. 1800 rooms were used in training and a different set of 200 rooms were used in validation. Source locations were generated every 5° in azimuth (0° to 355°) and 10° in elevation (0° to 60°), resulting in a total of 504 locations. The source distance was varied from 1.4 meters to the furthest distance within the room.

Training scenes were composed of natural sounds from the GISE-51 dataset (12,465 training sounds and 1,716 validation sounds, grouped into 51 source categories (Yadav & Foster, 2021)). Each scene contained between 1 and 5 sources, each of a different category. For each scene a random room and source locations within the room where chosen. These sources were then spatialized and combined to form a 2 second binaural audio clip. We generated a total of 1,000,000

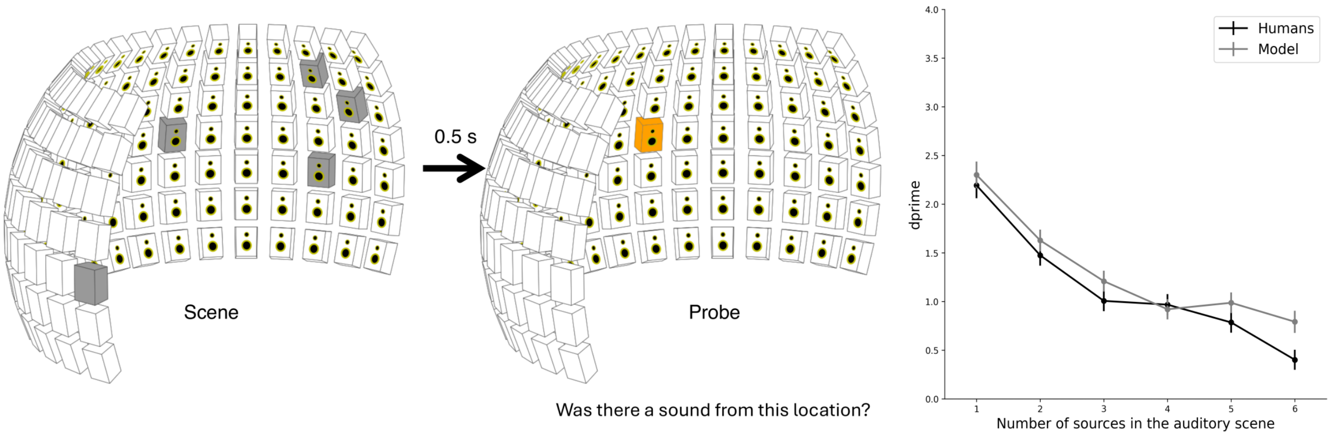

Figure 2: **Comparing human and model performance on a location discrimination task.** (a) Schematic of discrimination experiment trial for humans. Humans were presented with a 1 second scene followed by a white-noise probe. There was a 500 ms silence between the scene and probe. Models were presented a spatialized auditory scene and the spatialized probe, presented separately. Both were tasked with responding whether there was a source present at the probe location. (b) Location discrimination performance (quantified as d') as a function of scene size for humans (black, n=6) and our model (gray). Error bars are s.e.m.

training scenes and 500,000 validation scenes.

Human localization experiments. We assessed human multi-source localization using a discrimination task. Sounds were presented using a speaker array of 133 speakers, arranged in a hemisphere of around the participant. The speakers spanned 180° in azimuth (-90° to 90°) and 60° in elevation (-20° to 40°) with 10° of separation between adjacent speakers in both dimensions. The participant sat in the center of the array.

Stimuli consisted of 160 natural sounds, each 1 second long. On each trial, participants listened to a scene composed of 1 to 6 sounds played concurrently from different speakers, followed by 0.5 seconds of silence, followed by a 1 second white noise "probe" from a single speaker. The probe's location either coincided with one of the scene's sounds or was at least 30° away. Participants judged whether the probe's location overlapped with any of the locations of the sounds in the scene. Feedback was not provided. To test the model on the same experiment, we rendered the same stimuli in a virtual replica of the speaker array room with similar room acoustics.

To obtain model judgments on the multi-source discrimination task we presented the scene and probe from each trial individually to the model. The maximum a posteriori (MAP) estimate from the probe trial was used as the model's detected probe location. We then evaluated the likelihood of the probe location under the density predicted for the scene. We set the model criterion to be the median likelihood value (per scene size) across all trials. Performance was then expressed as d', computed from hits (trials where the probe was at the same location as one of the scene sources, and was correctly identified as such) and false alarms (trials judged as same but where the probe was at a different location from the scene sources).

Results & Discussion

Location discrimination in auditory scenes. Human discrimination was good for single sources, but became less accurate as the number of sources in the scene increased (Figure. 2a). This result indicates that human localization is substantially impaired when multiple sources make sounds concurrently. The primary prior experiment on limits on human multi-source localization was conducted with speech sources that varied only in azimuth (Zhong & Yost, 2017). The present results are consistent with these prior results but show that multi-source localization is limited even when sources are diverse (being drawn from a large set of natural sounds) and when they vary in elevation as well as azimuth.

The model qualitatively matched the dependence of performance on the number of sound sources, and showed a rough quantitative match to human performance (Figure. 2a). This result suggests that the limits of human performance in this domain reflect intrinsic limits on the information available in the acoustic array, in that a system optimized for the problem of multi-source localization exhibits similar limitations.

General conclusion We introduced a model for localizing sounds in auditory scenes. The model can express distributions over location, which here we used to model multi-source localization. It exhibited human-like behavior on a multi-source location discrimination task. The model's ability to represent uncertainty in location should allow it to also represent scenes that are ambiguous, or sound sources that are diffuse in the world.

Acknowledgments

Work supported by National Institutes of Health grant number R01DC017970.

References

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.

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).

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

Zhong, X., & Yost, W. A. (2017). How many images are in an auditory scene? *The Journal of the Acoustical Society of America*, *141*(4), 2882–2892.