Deep neural network model of sound localization replicates "what" and "where" representations in auditory cortex

Chenggang Chen CHENG-GANG.CHEN@JHU.EDU
Zhiyu Yang YANGZHIYUZYYY@GMAIL.COM
Xiaoqin Wang XIAOQIN.WANG@JHU.EDU
Department of Biomedical Engineering, Johns Hopkins University

Editors: List of editors' names

Abstract

Unlike visual cortex, whether the auditory cortex has parallel pathways for sound identification ("what") and localization ("where") is debated. It also lacks a topographic map of auditory space, like the retinotopy in visual cortex. Here, we built a deep neural network to model auditory "what" and "where" representations. We trained our model for localization only, using two-channel audio waveforms from six sound types presented from 394 locations at three sound levels. Surprisingly, the model learned six well-separated clusters by sound type, but not by sound level, in the middle layer. In the model's last layer, sounds were further organized by spectrogram similarity: harmonic types clustered together, single-band types formed a separate group, and broadband noise lay apart from the single-band group. Sound-location representations were random in the first layer but gradually organized into patches, and occasionally into a map, in the last layer. However, formation of a spatial map did not improve localization performance. Together, our model suggests that the auditory cortex does not need to dissociate "what" and "where" or create a space map.

Keywords: Auditory Cortex; Sound Localization; Dual Pathway; What and Where

1. Introduction

In the primate visual cortex, information is processed in a hierarchical manner using two parallel pathways (Mishkin et al., 1983): the ventral, or "where" pathway, and the dorsal, or "what" pathway (Figure 1a). These two pathways are specialized for visual identification/categorization and localization/movement, respectively. The existence of parallel pathways in the auditory cortex (AC) is highly debated, when it was first proposed around the 2000s (Rauschecker and Tian, 2000) (Figure 1a). Anatomical studies found that caudal and rostral streams of auditory afferents target dorsal and ventral domains in the macaque monkey prefrontal cortex (Romanski et al., 1999). Neurophysiology study in anesthetized macaque also found that single neurons in the caudal AC were highly tuned for sound locations, whereas neurons in the rostral AC are more selective for conspecific vocalization (Tian et al., 2001). Functional magnetic resonance imaging (fMRI) in humans also revealed a similar dichotomy in AC. Behavior studies in cats and humans further show causality of modulation of these two streams in what and where discrimination tasks (Lomber and Malhotra, 2008; Ahveninen et al., 2013). On the other hand, the theory of auditory parallel stream has been argued since it was first proposed (Belin and Zatorre, 2000; Hall, 2003). Multiple pieces of evidence are against the parallel stream hypothesis. First, the distribution of 'what' is everywhere since neurons in both caudal and rostral areas of awake macaque

auditory cortex are equally selective for vocalization (Recanzone, 2008; Bizley and Walker, 2009). Second, although neurons in the caudal AC are more selective for sound locations, highly spatially tuned neurons were also identified in the rostral AC (Woods et al., 2006; Remington and Wang, 2019). Third, AC neurons show multiplexed representation of sound features, i.e., where (location) and what (pitch and timbre) information (Bizley et al., 2009; Walker et al., 2011). Last, behavior studies using the advanced optogenetics tool show that inhibiting one area of AC impaired both spatial and non-spatial hearing (Town et al., 2023). Task-optimized deep neural networks (DNN) that were trained to perform sensory

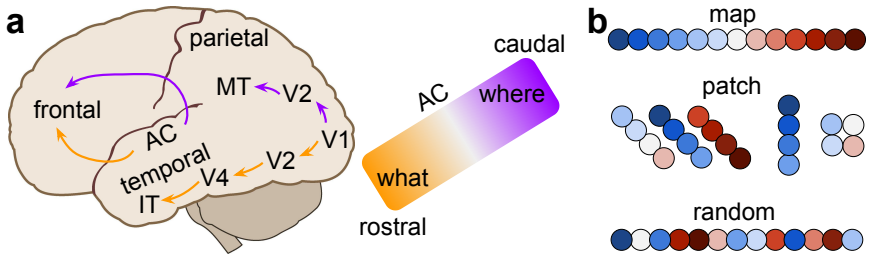

Figure 1: **a**. In visual cortices, the dorsal "where" pathway (purple arrows) starts from the primary visual cortex (V1), and projects to V2 and middle temporal (MT) cortex in the parietal lobe. The ventral "what" pathway (orange arrows) also starts from V1, then projects to V2, V4, and inferior temporal (IT) cortex in the temporal lobe. In the auditory cortex (AC), caudal "where" and rostral "what" pathways project to dorsal and ventral frontal cortex, respectively. **b**. Representations of sound locations (color dots) formed three candidate organizations in the brain.

tasks have shown promise as computational models of the sensory cortex (see Related work). However, to our best knowledge, there is no DNN model on the auditory dual pathway or a space map (Figure 1b). In this study, we trained a DNN to localize sound locations and examined the learnt representation of sound locations, types, and levels.

2. Results

Here, we modeled the representation of sound locations with a DNN trained for the sound localization task. We used an open-source bioacoustics sound localization dataset (Peterson et al., 2024). In a 40 by 60-centimeter arena, sounds were played from 394 locations in the bottom (Figure 2a, colored dots) and captured by four microphones at the corner that were 35 cm from the bottom. The distance between one speaker and multiple microphones creates cues for sound localization (Supplementary Figure 8). The arriving time is also shortest for the same microphone. These interaural time/level difference (ITD/ILD) cues contribute to sound localization in this horizontal plane. The monaural spectral cues that result from pinna and torso reflection mainly contribute to vertical sound locations (Francl and McDermott, 2022). However, they also play roles in horizontal localizations. One example is that humans with one ear deaf can still localize horizontal locations (Van Wanrooij and Van Opstal, 2004). Each location has a median of 144 stimuli, including a fixed number

of six sound types and three sound levels, and eight (median) different samples of the same sound type. Sound types include five different classes of gerbil vocalization and one artificial white noise (Figure 2b). Notice that "dfm", "sc", and "stack" calls have equally spaced frequency bands in their spectrograms. The first or bottom frequency band has a fundamental frequency of f0. Other bands are called harmonics and have a frequency that is an integer multiple of f0. For example, there are three, two, and five harmonics in three sound types. "upfm" and "warble" only have one narrow band of frequency modulations. In contrast, sound energy was distributed uniformly in "white" noise call. Notice that the audio waveforms are very different from each other (Supplementary Figure 8). The medium sound level is calibrated to be approximately the same level as a natural vocalization. The soft and loud sound levels are around 6 dB lower or above the medium level.

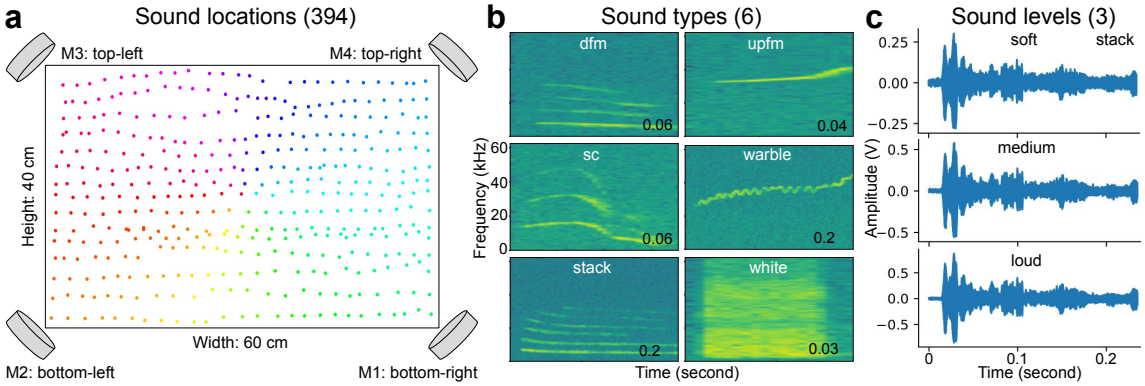

Figure 2: **a**. Four microphones at the corners of the arena will record sounds coming from 394 locations. **b**. Spectrograms of six example sound types. Here, "fm" stands for frequency modulation, "d" for downwards, and "sc" for soft chirp. **c**. Three different sound levels for the call type "stack".

We feed two raw audio waveforms to a DNN that has five one-dimensional convolutional layers (Figure 3a). Our five-layer model is inspired by the five stations in the ascending auditory pathway (Figure 3b). It begins from the cochlear nucleus (CN), superior olivary complex (SOC) that includes lateral superior olive (LSO) for ILD and medial superior olive (MSO) for ITD, central IC (ICC), to either medial geniculate body (MGB) and AC, or to external IC (ICX) and SC. Notice that our goal here is not to match the responses between the DNN and the animal brain. In addition, we did not simulate the head-related transfer function (HRTF) and cochleagram to simulate human or animal (i.e., gerbil) hearing.

We trained this model for a sound localization task (Figure 3c). In the trained DNN, we extracted the 512-dimensional embeddings after each layer. To visualize the high-dimensional embeddings, we used uniform manifold approximation and projection (UMAP), an unsupervised dimensionality reduction method. Because we only trained this DNN model to localize sound locations ("where"), we hypothesized that the model's representations of sound types and levels ("what") would form random patterns (Figure 3d). Surprisingly, this DNN model, which was trained to localize sound locations, formed six well-separated

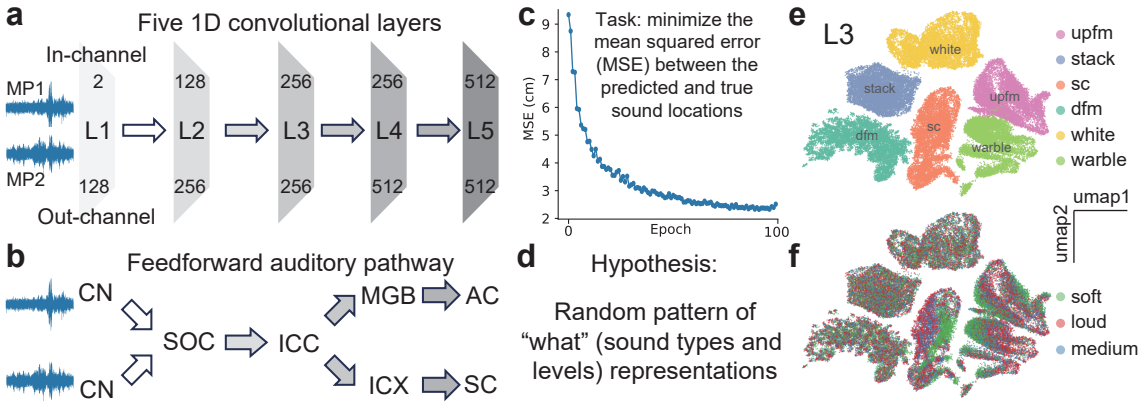

Figure 3: **a**. Network structure. **b**. Major nuclei in the ascending auditory pathway. **c**. Validation loss. **d**. Hypothesis. **e**. 2D visualization of the representation of five sound types in layer 3 using UMAP. **f**. Representations of sound levels.

clusters for six sound types (Figure 3e) in the middle layer. In contrast, the representations of three sound types were not organized globally (Figure 3f).

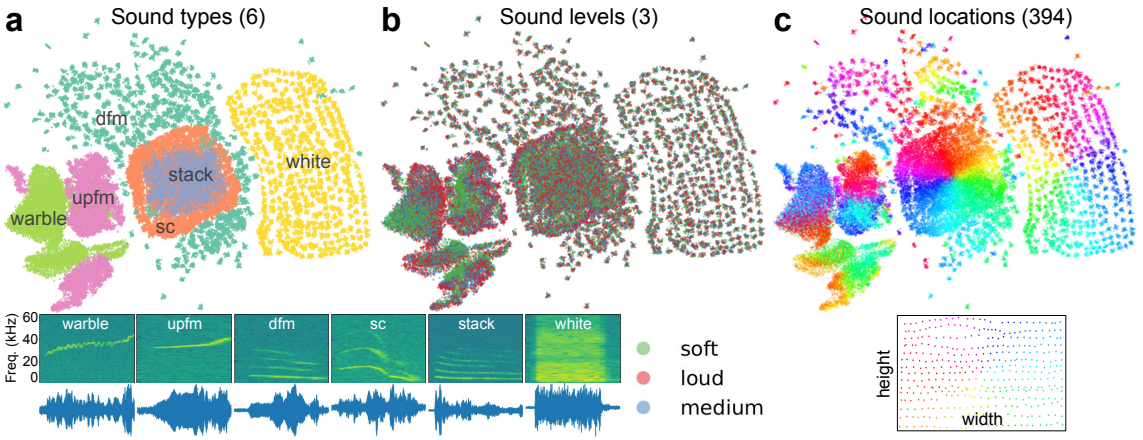

Figure 4: **a-c**. Representations of sound types, sound levels and sound locations.

Next, we compared the representations of sound types, levels, and locations in the final layer (Figure 4). The representations of sound types in the final layer do not form clear clusters as before in the middle layer (Figure 3e), as three sound types ("dfm", "stack", and "sc") were overlapped. Interestingly, those three sound types all contain multiple harmonics. Clusters from two sound types ("warble" and "upfm") with narrow frequency bands are also near each other. The cluster of white noise, which has wide frequency bands, was next to sound types with multiple frequency bands but far away from sound types with only one band. Therefore, in this final layer, the DNN model further clusters sound types

with similar features instead of treating them as independent categories as in the middle layer. Together, our findings suggest that AC and DNN converge on a common strategy for representing "what" and "where" information.

References

Jyrki Ahveninen, Samantha Huang, Aapo Nummenmaa, John W Belliveau, An-Yi Hung, Iiro P Jääskeläinen, Josef P Rauschecker, Stephanie Rossi, Hannu Tiitinen, and Tommi Raij. Evidence for distinct human auditory cortex regions for sound location versus identity processing. *Nature communications*, 4(1):2585, 2013.

Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34:25164–25178, 2021.

Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

Pascal Belin and Robert J Zatorre. 'what','where'and 'how'in auditory cortex. *Nature neuroscience*, 3(10):965–966, 2000.

Jennifer K Bizley and Kerry MM Walker. Distributed sensitivity to conspecific vocalizations and implications for the auditory dual stream hypothesis. *Journal of Neuroscience*, 29 (10):3011–3013, 2009.

Jennifer K Bizley, Kerry MM Walker, Bernard W Silverman, Andrew J King, and Jan WH Schnupp. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *Journal of neuroscience*, 29(7):2064–2075, 2009.

Nicholas M Blauch, Marlene Behrmann, and David C Plaut. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, 119(3):e2112566119, 2022.

Chenggang Chen and Sen Song. Distinct neuron types contribute to hybrid auditory spatial coding. *Journal of Neuroscience*, 44(43), 2024.

Chenggang Chen, Sheng Xu, Yunyan Wang, and Xiaoqin Wang. Location-specific neural facilitation in marmoset auditory cortex. *Nature communications*, 16(1):2773, 2025.

Mayukh Deb, Mainak Deb, and N Murty. Toponets: High performing vision and language models with brain-like topography. *arXiv preprint arXiv:2501.16396*, 2025.

Amirozhan Dehghani, Xinyu Qian, Asa Farahani, and Pouya Bashivan. Credit-based self organizing maps: training deep topographic networks with minimal performance degradation. In *The Thirteenth International Conference on Learning Representations*, 2024.

Fenil R Doshi and Talia Konkle. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023.

Richard Durbin and Graeme Mitchison. A dimension reduction framework for understanding cortical maps. *Nature*, 343(6259):644–647, 1990.

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

Umut Güçlü and Marcel AJ van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017.

Deborah A Hall. Auditory pathways: are 'what'and 'where'appropriate? *Current Biology*, 13(10):R406–R408, 2003.

Lloyd A Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35, 1948.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

ANDREW J King. The wellcome prize lecture. a map of auditory space in the mammalian brain: neural computation and development. *Experimental Physiology: Translation and Integration*, 78(5):559–590, 1993.

Eric I Knudsen and Masakazu Konishi. A neural map of auditory space in the owl. *Science*, 200(4343):795–797, 1978.

Yuanning Li, Gopala K Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H Carney, Junfeng Lu, Jinsong Wu, and Edward F Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, 2023.

Stephen G Lomber and Shveta Malhotra. Double dissociation of'what'and'where'processing in auditory cortex. *Nature neuroscience*, 11(5):609–616, 2008.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451, 2024.

John C Middlebrooks. A search for a cortical map of auditory space. *Journal of Neuroscience*, 41(27):5772–5778, 2021.

John C Middlebrooks and John D Pettigrew. Functional classes of neurons in primary auditory cortex of the cat distinguished by sensitivity to sound location. *The Journal of neuroscience*, 1(1):107, 1981.

Patrick Mineault, Shahab Bakhtiari, Blake Richards, and Christopher Pack. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in neural information processing systems*, 34:28757–28771, 2021.

Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983.

Ralph Peterson, Aramis Tanelus, Christopher Ick, Bartul Mimica, Niegil Francis Muttath Joseph, Violet Ivan, Aman Choudhri, Annegret Falkner, Mala Murthy, David Schneider, et al. Vocal call locator benchmark (vcl) for localizing rodent vocalizations from multi-channel audio. *Advances in Neural Information Processing Systems*, 37:106370–106382, 2024.

Christopher I Petkov, Christoph Kayser, Thomas Steudel, Kevin Whittingstall, Mark Augath, and Nikos K Logothetis. A voice region in the monkey brain. *Nature neuroscience*, 11(3):367–374, 2008.

Josef P Rauschecker and Sophie K Scott. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6):718–724, 2009.

Josef P Rauschecker and Biao Tian. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences*, 97(22): 11800–11806, 2000.

Gregg H Recanzone. Representation of con-specific vocalizations in the core and belt areas of the auditory cortex in the alert macaque monkey. *Journal of Neuroscience*, 28(49): 13184–13193, 2008.

Gregg H Recanzone and Yale E Cohen. Serial and parallel processing in the primate auditory cortex revisited. *Behavioural brain research*, 206(1):1–7, 2010.

Evan D Remington and Xiaoqin Wang. Neural representations of the full spatial field in auditory cortex of awake marmoset (callithrix jacchus). *Cerebral Cortex*, 29(3):1199–1216, 2019.

Lizabeth M Romanski, Biao Tian, Jonathan Fritz, Mortimer Mishkin, Patricia S Goldman-Rakic, and Josef P Rauschecker. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature neuroscience*, 2(12):1131–1136, 1999.

Mark R Saddler and Josh H McDermott. Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, 15(1):10590, 2024.

Frédéric E Theunissen, Kamal Sen, and Allison J Doupe. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of neuroscience*, 20(6):2315–2331, 2000.

Biao Tian, David Reser, Amy Durham, Alexander Kustov, and Josef P Rauschecker. Functional specialization in rhesus monkey auditory cortex. *Science*, 292(5515):290–293, 2001.

Stephen M Town, Katarina C Poole, Katherine C Wood, and Jennifer K Bizley. Reversible inactivation of ferret auditory cortex impairs spatial and nonspatial hearing. *Journal of Neuroscience*, 43(5):749–763, 2023.

Marc M Van Wanrooij and A John Van Opstal. Contribution of head shadow and pinna cues to chronic monaural sound localization. *Journal of Neuroscience*, 24(17):4163–4171, 2004.

Kerry MM Walker, Jennifer K Bizley, Andrew J King, and Jan WH Schnupp. Multiplexed and robust representations of sound features in auditory cortex. *Journal of Neuroscience*, 31(41):14565–14576, 2011.

Timothy M Woods, Steve E Lopez, James H Long, Joanne E Rahman, and Gregg H Recanzone. Effects of stimulus azimuth and intensity on the single-neuron activity in the auditory cortex of the alert macaque monkey. *Journal of neurophysiology*, 96(6):3323–3337, 2006.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Yiyuan Zhang, Ke Zhou, Pinglei Bao, and Jia Liu. A biologically inspired computational model of human ventral temporal cortex. *Neural Networks*, 178:106437, 2024.

Appendix A. Additional Introduction

What makes the above arguments of auditory parallel streams more complex is the other mystery about the auditory space map in AC (Figure 1b). According to the parallel pathway, spatially tuned neurons should form an orderly space map in the dorsal AC. Neurons selective for sound types should form several clusters in the rostral AC, like the ones in the visual ventral pathway (Bao et al., 2020). Although there is some evidence of voice patch existence in rostral AC (Petkov et al., 2008), no evidence exists for a space map in any area of AC (Middlebrooks, 2021). A neural map of auditory space was first found in the barn owl (Knudsen and Konishi, 1978), then in the mammalian inferior/superior colliculus (IC/SC) (Chen and Song, 2024; King, 1993). Middlebrooks and colleagues and other groups started to search for such a map in AC since 1981 in cats (Middlebrooks and Pettigrew, 1981). After 40 years of research, based on his and others' work, there is still no evidence for such a map in the AC for all the species that have been examined (Middlebrooks, 2021). Unlike the visual and somatosensory systems with orderly mapped receptors in the retina and skin, the cochlea has a map of sound frequency instead of location. Since the sound locations were computed inside the central nervous system, the representation of those computed but not relayed locations could be in any structure. After two decades of research on the auditory parallel pathway (Rauschecker and Scott, 2009; Recanzone and Cohen, 2010), and four decades of research on the auditory space map, our understanding of two questions is still very limited, partially due to that we lack a computational model for them. There are many models about either only sound spatial and nonspatial attributes. For example, the most famous delay line model (Jeffress, 1948) for computing interaural time difference (ITD), and the spectral-temporal receptive field (STRF) model (Theunissen et al., 2000) for explaining natural sound. Although mechanistic models are useful for explaining how

individual neurons are tuned to spatial and nonspatial sound features, they cannot model a spatial organization like parallel streams or a space map. The self-organizing map (SOM) has been used successfully to model orientation and direction maps in the visual cortex (Durbin and Mitchison, 1990). Because their inputs are only handcrafted simple features like line directions, SOM could not take inputs from two sensory attributes or compute the ITD between two audio channels.

Appendix B. Additional Results

To quantify their clustering differences between layers, we used the normalized mutual information (NMI). NMI is an information-theoretic metric used to evaluate the similarity between two clusters. The NMI ranges from 0 to 1, with 0 representing no mutual information. Since the representations of sound levels are random globally, their NMI in layer 3 (Figure 3f) and 5 (Fig. 4b) were both near 0. The NMI for sound type representations decreased from 0.89 in layer 3 (Figure 3e) to 0.62 in layer 5 (Figure 4a). The reason is due to the reorganizations of representations for the sound locations (Figure 4c) because the task is to localize sound locations based on the features extracted from the last layer. Notice our model is a task-driven, hypothesis-free: we ask it to localize sound locations, we do not impose on it to create a patch or map. Surprisingly, our model chose to represent sound locations with maps ("white", "sc", and "stack"), patches ("dfm"), and random patterns ("warble" and "upfm") (Figure 4c). Figure 5 shows the representations of all three sound attributes (row) across all five layers (column). The sound type representations were not well-separated in layer 1 (NMI: 0.54) as two sound types overlapped and all clusters were close to each other. The NMI increases from layer 2 (0.61) to peak at layer 3 and then decreases (0.55). The sound level representations are purely organized following the sound types or locations. Two clusters of "soft" sounds were preserved consistently for sound types of "upfm" and "warble". The sound location representations show the clearest changes from layer 1 to layer 5. Within layers 1 to 3, the organizations follow the sound type clusters. The "white" noise type of sound begins to show a twisted map since layer 2. After layer 3, the representations become scrambled again to better classify sound locations. The map for "white" noise is very clear with high spatial resolution. In contrast, the maps for "sc" and "stack" are very tight with low resolution. We observed very similar patterns in the other two microphones (Supplementary Figure 9): six sound types form six well-separated clusters only in layer 3 but not in layer 1 or layer 5. In addition, the "soft" sound levels also form clusters in two narrow-band sound types. However, the representation of "white" noise did not form a map in the last layer. Among all the six microphone pairs, only microphone pairs M12 and M24 form clusters for the sound type of white noise (Figure 6). Although these two pairs form a spatial map, their sound localization performances were not the best. For example, the pairs M23 and M34 do not contain a spatial map, but their MSEs (2.45 and 2.46 cm) are even smaller than the pair M12 which contains a map (2.51 cm). To quantify how well the representation was organized (i.e., organization strength), we used the explained variance R2, which is the linear regression between 2D UMAP embeddings against 2D sound locations (height and width) (Supplementary Figure 10). The explained variance of two example sessions was very different (0.99 vs 0.22), with one having a map (top) and the other one not (bottom). Surprisingly, their sound localization performance

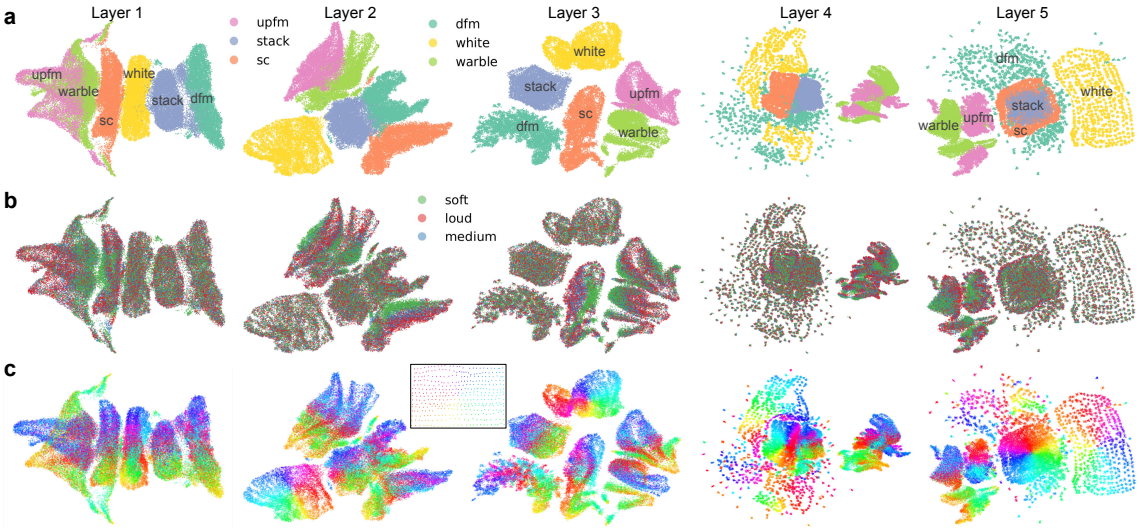

Figure 5: Representations of "what" and "where" attributes of sounds in all layers. a-c. Representations of six sound types, three sound levels, and 394 sound locations. Data of sound types and levels in layer 3 were the same as Figure 2e, f. Data in layer 5 was the same as Fig. 4a-c. The microphone pair is M24. Supplementary Fig. 2 shows another microphone pair M13. The NMIs for sound types are 0.5454, 0.6140, 0.8898, 0.5488, and 0.6233. The NMIs for sound levels are 0, 0.0001, 0.0025, 0, and 0.

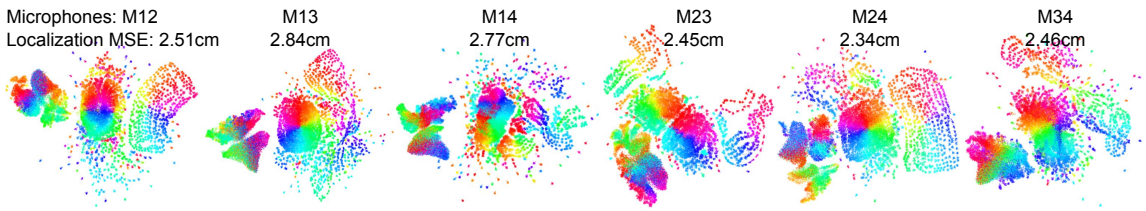

Figure 6: Representations of sound locations in the final layer from all six microphone pairs. Data of M24 was shown previously in Figure 4c and Figure 5c (layer 5). Data of M13 was shown previously in supplementary Figure 9c (layer 5).

(MSE: 0.31 vs 0.37 cm) was similar. Figure 7a shows the representations of four sound types that were trained separately. The sound types "dfm" and "stack" form clear space maps (R2: 0.95 and 0.99), but their task performance was very different (MSE: 0.27 vs 2.09 cm). Although sounds from "upfm" did not form a map, its task performance was better than "stack" sounds, which form a map. Furthermore, the representations of the same sound type could be very different depending on the microphone pair (Figure 7b).

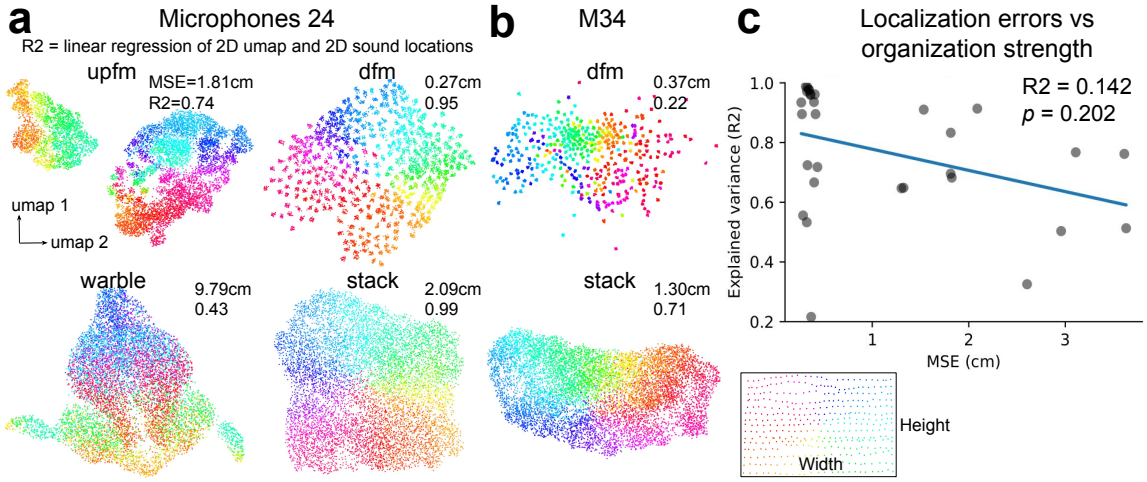

Figure 7: Representations of sound locations in the final layer from a single sound type. a. Same microphone pair as Fig. 4c, but sound stimuli from each sound type were trained separately. b. Tw sound types from a different pair of microphones. c. Each dot represents one session (30 total: five sound types × six microphone pairs). Data from the "warble" call type are excluded because their MSE values exceed 10 cm.

Across all six pairs of microphones and five types of stimuli (Figure 7c), there was an insignificant ($p = 0.202$) and weak (R2 of X-Y axis $= 0.142$) correlation between organization strength (Y-axis) against task performance (X-axis). Therefore, a map of sound locations is not always the best choice in representing sound locations. In summary, a sound-localization task-driven DNN represents sound locations as maps, patches, or random patterns similar to the superior colliculus and cortices in the brain.

Appendix C. Related work and our contributions

There are several lines of DNN modeling works that are related to this study. 1) Visual and auditory 'what' pathway. Using learned features from supervised learning tasks (for example, image recognition or sound classification), encoding models predict, with high accuracy, neural responses in the visual and auditory cortex (Yamins et al., 2014; Li et al., 2023). In particular, Kell et al. (2018) used supervised convolutional neural networks (CNNs) to build encoding models for auditory responses in fMRI recordings and showed an aligned hierarchy between the CNNs and the auditory cortex. 2) Visual 'where' pathway. Like sound

frequency, visual locations are faithfully relayed from the retina, thus the visual 'where' pathway's function was mainly on motion processing. Therefore, some studies are focusing on modeling visual motion processing (Güçlü and van Gerven, 2017; Mineault et al., 2021). In particular, Bakhtiari et al. (2021) used self-supervised learning in a single model with a single loss function to capture the properties of both the visual ventral and the dorsal pathways. 3) Visual 'what' pathway topographical organization. Because standard DNNs have no within-area spatial structure beyond retinotopy, their architecture needs to be modified to model spatial topography. There are two major strategies. One is to add the spatial loss to the task loss (Blauch et al., 2022; Margalit et al., 2024; Deb et al., 2025). The other one is to further train the learnt embeddings using SOM (Doshi and Konkle, 2023; Dehghani et al., 2024; Zhang et al., 2024). 4) Sound localization behaviors. Two studies by McDermott and colleagues built DNN models of sound localizations and found they exhibited several characteristics of human psychological behaviors (Francl and McDermott, 2022; Saddler and McDermott, 2024).

We made four contributions in this study. 1) We built the first computational model for both "what" and "where" representations in the auditory system. Our audios contain both "what" (sound types and levels) and "where" (sound locations) attributes of sounds. Our model is a hypothesis-free and task-optimized DNN for sound localization only. Similar to the auditory system in the brain, our model's inputs are only two-channel audio waveforms instead of spectrograms. 2) From the first to middle layers, our model learnt to classify sound types but also sound levels into distinct clusters. In the final layer, representations of sound types with similar spectrograms are further placed together and away from types with different spectrograms. Because our training data were raw waveforms rather than spectrograms, the model implicitly learned spectrotemporal features while being trained to extract spatial features. 3) Representation of sound locations is random at the first layer, then gradually becomes organized in the deep layer. Maps were only found in the final layer in two out of six microphone pairs, and only for specific sound types. In contrast, patchy organization was found in all conditions and sound types. Representations of locations as patches and maps are consistent with experimental results observed in the AC and subcortical IC/SC, respectively. 4) We also trained the DNN for a single sound and found representation of a specific sound type in specific microphone pairs will generate a map. However, there is no significant correlation between localization performance and the organization strength of representation. This explained why AC does not form a map because it has limited benefit. A patch or random pattern allows the AC to dynamically represent sound locations (Chen et al., 2025).

Appendix D. Supplementary Figures

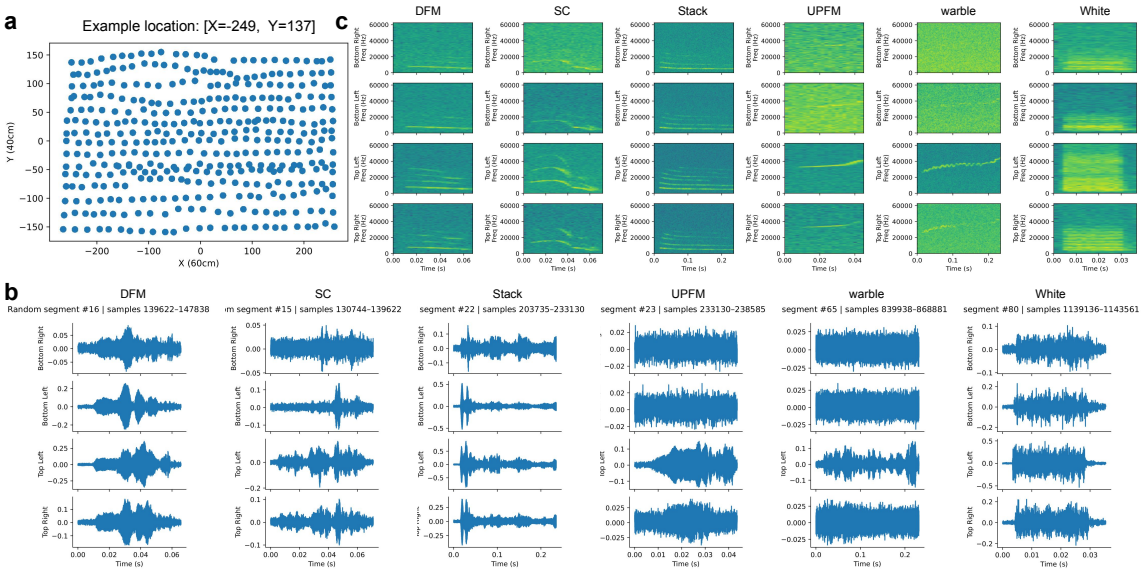

Figure 8: Audio waveforms and spectrograms of six sound types from an example location. a. There are 394 different sound locations, and the example location comes from the top left corner. Each location (big blue dot) has many (i.e., 144) smaller dots inside since different sound types, levels, and a median of eight different samples are presented from there. b. Audio waveforms from six sound types (columns) that were recorded by four microphones at four corners. Notice that the amplitudes are different for each plot. c. Corresponding spectrograms for waveforms shown in b.

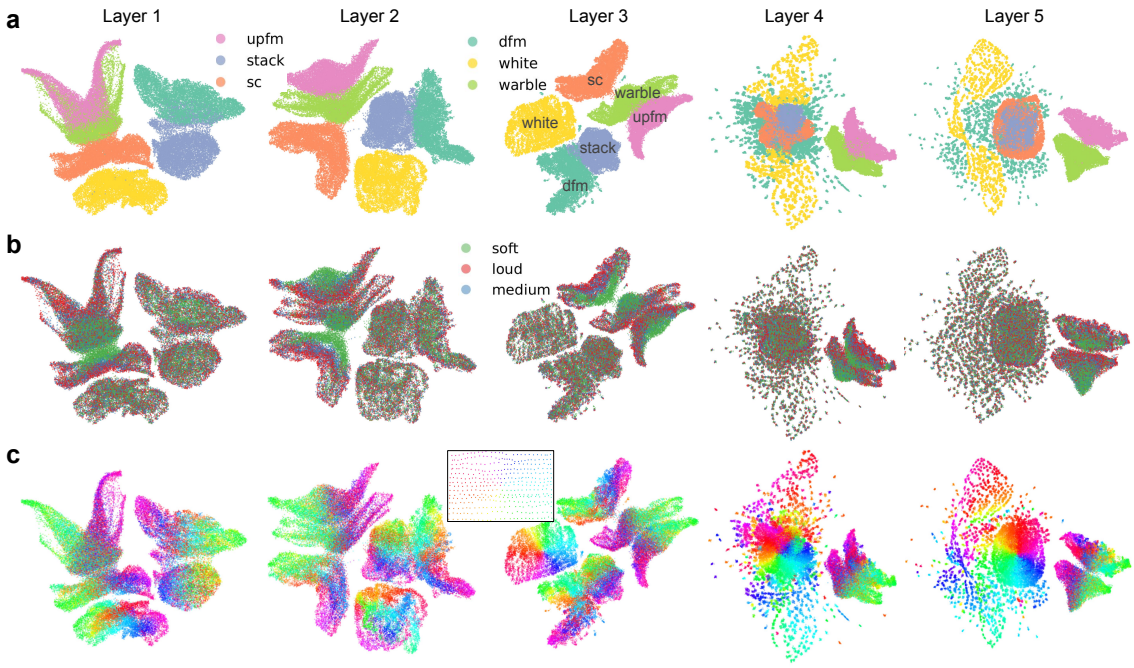

Figure 9: Representations of "what" and "where" attributes of sounds in all layers. Similar to Fig. 5, but for the microphone pair of M1 and M3. The NMIs for sound types are 0.6892, 0.8313, 0.8095, 0.4953, and 0.6582. The NMIs for sound levels are 0.0001, 0.0003, 0.0002, 0, and 0.

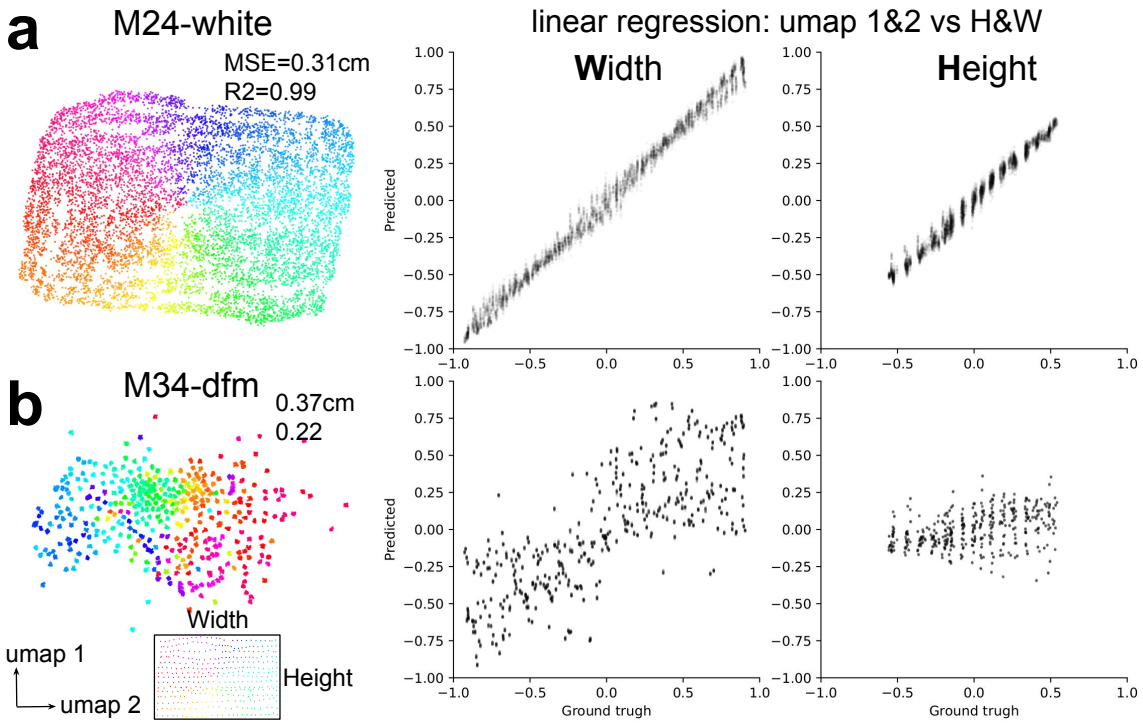

Figure 10: Linear regression of sound location representations and locations of the speaker. a. A white noise sound type from the microphone pair of M2 and M4. Left, 2D UMAP visualization of sound location representation. Middle, linear regression of 2D UMAP against the width of speaker location. Right, regression of UMAP against the height of speaker location. b. Similar to a but for a different microphone pair and sound type.