Modeling and Evaluating Human Sound Localization in the Natural Environment

by

Andrew Francl

B.S., Boston College (2016)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author .....
Department of Brain and Cognitive Sciences
August 23, 2022

Certified by .....
Josh McDermott
Associate Professor
Thesis Supervisor

Accepted by .....
Mark Harnett
Graduate Officer, Department of Brain and Cognitive Sciences

Modeling and Evaluating Human Sound Localization in the Natural Environment

by

Andrew Francl

Submitted to the Department of Brain and Cognitive Sciences
on August 23, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

Humans locate sounds in their environment to avoid danger and identify objects of interest. In a ten-minute bike ride, a person might take note of a car approaching from behind, a tree where a bird is singing, and pedestrians walking from around a blind corner.

Research on human sound localization has greatly advanced our understanding of binaural hearing but leaves us some ways from a complete understanding. In particular, it has been difficult to assess human sound localization in ways that align with humans experience on an everyday basis. This thesis aims to more closely align research methods and modeling approaches with the natural sound localization tasks that humans perform in the real world.

In the first study, we show that a model trained to localize sounds in naturalistic conditions exhibits many features of human spatial hearing. But when trained in unnatural environments without reverberation, noise, or natural sounds, the model's performance characteristics deviate from those of humans. The results show how biological hearing is adapted to the challenges of real-world environments and illustrate how artificial neural networks can reveal the real-world constraints that shape perception.

In the second study, we ran a behavioral experiment to evaluate human sound localization in a naturalistic setting with natural sounds and identified specific sounds that are difficult for humans to localize. We assessed whether the model of sound localization from the first study could predict the accuracy with which individual sounds are localized. We found that the model predicted human localization accuracy well above chance. However, the model biases were distinct from those evident in humans, suggesting room for future improvement.

In the third study, we constructed a model that uses a biologically inspired learning approach to localizing sounds, relying on self-motion cues from head movements to learn representations of sound locations. We show that this strategy can learn a representation that enables accurate decoding of sound location without having access to the ground truth location for sounds during training.

In the fourth study, we used a model of human speech perception as a perceptual metric to improve speech denoising. We found that while this perceptual metric improved denoising over standard approaches, a simple model of the cochlea performed similarly, suggesting much of the benefit of this approach may be in using a frequency-based overcomplete representation of the signal.

Thesis Supervisor: Josh McDermott
Title: Associate Professor

Acknowledgments

I thank my advisor, Josh McDermott. His unwavering dedication to his students and willingness to invest so much in their scientific growth is truly remarkable. Over the past six years, I have been struck by his ability to always make time for me and consistently provide thoughtful advice to aid my development. His investment often felt like a vote of confidence, particularly during difficult periods where that unwavering support gave me the self-confidence to push forward.

I thank my committee, Nancy Kanwisher, Jim DiCarlo, Josh Tenenbaum, and Steve Colburn, for their insights, advice, and guidance over the past six years. Their encouragement and insights greatly improved my research program.

I thank the McDermott Lab, whose members are kind, intelligent, and supportive. The countless conversations with lab members opened my eyes to new ideas and ways of thinking about fundamental scientific questions. The generosity of members with their time helped me advance scientifically and made me feel incredibly welcomed and supported. I could not imagine a better environment in which to do a Ph.D.

I thank my undergraduate advisor, Laura Anne Lowery. She introduced me to the wonders of the brain and the joy of scientific discovery. Through her kindness and optimism, she cultivated a positive and welcoming environment, which I benefitted from greatly as an undergraduate.

I thank my family for their support over the years. They have always nurtured my curiosity and encouraged me to strive for my goals. Their guidance, support, and love have made me who I am today.

Lastly, I thank my partner Leyla. She has both encouraged me and helped me maintain a broader perspective over the last six years. I am grateful for such a wonderful partner and look forward to what our future holds.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Background

Humans locate sounds in their environment to avoid danger and identify objects of interest. In a ten-minute bike ride, a person might take note of a car approaching from behind, a tree where a bird is singing, and pedestrians walking from around a blind corner. Human sound localization is both remarkable for its utility in our daily lives and how quickly and automatically we perform it. The mechanisms underlying this ability are not straightforward because the sensory periphery does not provide explicit location information. This contrasts with vision, where the retina provides fine-grained spatial information. This spatial readout is possible due to retinotopy: light travels in rays which allows the retina to refract light from different directions to stereotyped regions of the retina [155]. Spatial information is not as readily available in audition, where each sound source produces diffuse waves, as opposed to rays, that propagate through the environment. Extracting auditory spatial information is also difficult because waves from each source sum together when they come in contact, which results in one final waveform reaching each ear of a human listener. In addition, this final waveform is a linear combination of all waves from all sources in a scene as well as reflections of waves off of other surfaces [16]. The brain must infer source location and identity from the pair of waveforms entering the two ears despite there being an infinite number of combinations of sources and positions that could

lead to the received waveforms. This problem is ill-posed and results in a difficult computational challenge, yet is one the brain solves seemingly effortlessly.

Understanding how human listeners solve this problem has long been a subject of scientific investigation, dating back over 100 years ago to foundational work by Lord Rayleigh in 1907 [182]. This early work explored how well individuals could localize tones produced by tuning forks in a quiet environment. It concluded that human sound localization relies on level differences between the left and right ear for high-frequency sounds and differences in arrival time between the ears for low-frequency sounds. Over the next 100 years, scientists documented the details of and limits on the sensitivity to these cues [182, 26, 85, 94]. In addition, they discovered new types of cues for identifying a sound's vertical position [11, 16, 220], azimuthal position [101], and distance [12]. This body of work has proven critical in understanding the basic organization of human sound localization strategies.

In addition to characterizing human sound localization, another field of study emerged with the goal of building mathematical and computational models to understand and replicate the mechanisms underlying localization. Examples in this line of work include the Interaural Time Difference (ITD) delay line model[115], ITD interaction model[44], weighted-image model[202], a contralateral inhibition model[141], and a model of auditory distance perception [23]. Although significant contributions, these models were designed and tuned to explain behavioral or neural responses for a single task and did not take waveform input, instead operating on precomputed features [65, 23, 141], and thus could not be tested on natural sounds in natural conditions.

These previous scientific approaches greatly advanced our understanding of binaural hearing but leave us some ways from a complete understanding. One gap in understanding derives from the field's approach to modeling. Most binaural hearing models are hand-designed to replicate a specific behavior or interest. Researchers often design the models by relying on intuition to identify a chain of signal processing steps that might lead to the observed human behavior. By fitting a model to one task at a time, the model's details may be overfit to their specific behavioral task. And

by using hand-designed and precomputed features for the model input, the model is restricted to operating on a small subset of all possible sounds. We localize sounds with a single auditory system but have thus far lacked a single model that can account for many aspects of human sound localization.

A second gap in understanding is that the field has tended to assess human sound localization behavior using stimuli that deviate significantly from a person's everyday experience. Specifically, most stimuli used to test binaural hearing are unnatural, such as variations on noise bursts or sinusoids. There are several reasons that researchers used these types of stimuli. One previous issue was technological limitations on replicating natural sounds. However, the most common is that traditional psychophysics [**?**, 83] emphasizes holding all aspects of the stimulus constant except one variable of interest, which is systematically varied while human responses are recorded. This method is designed to measure human sensitivity to a specific variable of a stimulus. However, the resulting stimuli lack the structure present in ecologically valid sounds. This raises the possibility that the resulting scientific characterization of sound localization may deviate from what would be observed in more realistic situations.

This thesis aims to more closely align research methods and modeling approaches with the natural sound localization tasks that humans perform daily. We intend this approach to advance binaural hearing research along its two primary axes. To extend and unify models, we explore approaches and applications that are constrained and inspired by the natural world. To better understand behavior, we evaluate human sound localization in a naturalistic setting with natural sounds and measure the accuracy of human listeners.

1.2 Organization of Thesis

This thesis consists of four studies: three computational studies and one behavioral study.

The first study [63] builds a neural network-based model of human sound localization. A core goal of binaural hearing models is to accurately predict and explain

human behavior in sound localization tasks. We hypothesized that localizing sounds in naturalistic conditions is a significant constraint on the solution space for humans. To test this idea, we simulated a naturalistic environment using a virtual auditory world and optimized a model to localize natural sounds in this environment. We found that the resulting model was also able to accurately localize real-world binaural recordings, indicating that the virtual acoustic world simulator captures enough aspects of the real world to allow the optimized model to generalize beyond the artificial training data. The model also replicated human behavior in a range of psychophysical experiments, including sensitivity to monaural spectral cues and interaural time and level differences, integration across frequency, biases for sound onsets, and limits on localization of concurrent sources. The similarity between human and model behavior suggests that many aspects of human sound localization behaviors may be a consequence of optimizing performance in a natural environment. Lastly, deviating from natural training conditions during training caused the model to deviate from human behavioral data. In some cases, these deviations from human behavior were specific to a single psychophysics experiment, such as the precedence effect, which only emerged when the model learned to localize in a reverberant environment. The approach provides a tool that can be used to discover links between specific behavioral traits and challenges posed by specific properties of the natural environment.

The second study measured human localization of natural sounds presented in a realistic environment. In addition to quantifying the spatial accuracy of localization for natural sounds, we identified specific sounds that are difficult to localize. Lastly, we evaluated how well the model in chapter 2 could predict which sounds would be difficult to localize by measuring model localization errors for the same set of natural sounds that were used in the human experiment. Model errors were correlated with human errors, with correlation coefficients around 0.6-0.7, but the model also made errors substantially larger than human listeners in some cases.

The third chapter explored a biologically inspired approach to learning to localize sounds that relies on head self-motion cues. Specifically, we constructed a neural-network model that receives the simulated binaural audio for an auditory scene at

many different head positions. The model uses contrastive learning to find a representation in which binaural audio from nearby head positions and the same auditory scene is represented similarly but where audio from distant head positions or different auditory scenes is dissimilar. Specifically, the model compares pairs of binaural audio excerpts and calculates the cosine similarity between the representations for each pair. If the excerpts are from the same auditory scene and similar head positions, the model uses gradient descent to increase the cosine similarity between the representations in that pair. In all other cases, the model uses gradient descent to minimize cosine similarity between pairs of binaural audio excerpts. We show that this strategy can learn a representation that enables accurate linear decoding of sound location without having access to the ground truth location for sounds during training.

The fourth study [187] explores using a neural-network model of human speech perception as a perceptual metric to improve speech denoising. Specifically, we measured the distance between features from a pre-trained model of speech or environmental sound classification to quantify how much a stimulus deviated from a target signal. We used the distance as an error signal to train a second neural network to remove background noise from excerpts of noisy speech. We found that while this perceptual metric improved denoising over standard waveform-based approaches, it performed no better than a simple model of the cochlea. This suggests that much of the benefit derived from this perceptual metric can be attributed to simply using a frequency-based overcomplete signal representation of the signal.

Together, these studies suggest that natural sounds, environments, and behaviors provide important constraints on human sound localization and suggest a promising path forward for incorporating ecological constraints to advance the study of sound localization.

Chapter 2

Deep neural network models of sound localization reveal how perception is adapted to real-world environments

Abstract

Mammals localize sounds using information from their two ears. Localization in real-world conditions is challenging, as echoes provide erroneous information and noises mask parts of target sounds. To better understand real-world localization, we equipped a deep neural network with human ears and trained it to localize sounds in a virtual environment. The resulting model localized accurately in realistic conditions with noise and reverberation. In simulated experiments, the model exhibited many features of human spatial hearing: sensitivity to monaural spectral cues and interaural time and level differences, integration across frequency, biases for sound onsets and limits on localization of concurrent sources. But when trained in unnatural environments without reverberation, noise or natural sounds, these performance characteristics deviated from those of humans. The results show how biological hearing is adapted to the challenges of real-world environments and illustrate how artificial neural networks can reveal the real-world constraints that shape perception.

2.1 Introduction

Why do we see or hear the way we do? Perception is believed to be adapted to the world, shaped over evolution and development to help us survive in our eco-

logical niche. Yet adaptedness is often difficult to test. Many phenomena are not obviously a consequence of adaptation to the environment, and perceptual traits are often proposed to reflect implementation constraints rather than the consequences of performing a task well. Well-known phenomena attributed to implementation constraints include aftereffects[45, 117], masking[52, 150], poor visual motion and form perception for equiluminant colour stimuli[146] and limits on the information that can be extracted from high-frequency sound[7, 114, 113].

Evolution and development can be viewed as an optimization process that produces a system that functions well in its environment. The consequences of such optimization for perceptual systems have traditionally been revealed by ideal observer models—systems that perform a task optimally under environmental constraints[73, 72] and whose behavioural characteristics can be compared to actual behaviour. Ideal observers are typically derived analytically, but as a result are often limited to simple psychophysical tasks[199, 99, 224, 77, 29, 28]. Despite recent advances, such models remain intractable for many real-world behaviours. Rigorously evaluating adaptedness has thus remained out of reach for many domains. Here we extend ideas from ideal observer theory to investigate the environmental constraints under which human behaviour emerges, using contemporary machine learning to optimize models for behaviourally relevant tasks in simulated environments. Human behaviours that emerge from machine learning under a set of naturalistic environmental constraints, but not under alternative constraints, are plausibly a consequence of optimization for those natural constraints (that is, adapted to the natural environment) (Fig. 2-1a).

Sound localization is one domain of perception where the relationship of behaviour to environmental constraints has not been straightforward to evaluate. The basic outlines of spatial hearing have been understood for decades[182, 11, 32, 85]. Time and level differences in the sound that enters the two ears provide cues to a sound's location, and location-specific filtering by the ears, head and torso provide monaural cues that help resolve ambiguities in binaural cues (Fig. 2-1b). However, in real-world conditions, background noise masks or corrupts cues from sources to be localized and reflections provide erroneous cues to direction[16]. Classical models based on

these cues thus cannot replicate real-world localization behaviour[18, 65, 40]. Instead, modelling efforts have focused on accounting for observed neuronal tuning in early stages of the auditory system rather than behaviour[115, 44, 17, 93, 245, 203, 56], or have modelled behaviour in simplified experimental conditions using particular cues[40, 203, 190, 179, 202, 213, 62]. Engineering systems must solve localization in real-world conditions, but typically adopt approaches that diverge from biology, using more than two microphones and/or not leveraging cues from ear/head filtering[156, 230, 233, 184, 34, 148, 1, 116]. As a result, we lack quantitative models of how biological organisms localize sounds in realistic conditions. In the absence of such models, the science of sound localization has largely relied on intuitions about optimality. Those intuitions were invaluable in stimulating research, but on their own are insufficient for quantitative predictions.

Here we exploit the power of contemporary artificial neural networks to develop a model optimized to localize sounds in realistic conditions. Unlike much other contemporary work using neural networks to investigate perceptual systems[128, 88, 236, 41, 58, 126], our primary interest is not in potential correspondence between internal representations of the network and the brain. Instead, we aim to use the neural network as a way to find an optimized solution to a difficult real-world task that is not easily specified analytically, for the purpose of comparing its behavioural characteristics to those of humans. Our approach is thus analogous to the classic ideal observer approach, but harnesses modern machine learning in place of an ideal observer for a problem where one is not analytically tractable.

To obtain sufficient labelled data with which to train the model, and to enable the manipulation of training conditions, we used a virtual acoustic world[197]. The virtual world simulated sounds at different locations with realistic patterns of surface reflections and background noise that could be eliminated to yield unnatural training environments. To give the model access to the same cues available to biological organisms, we trained it on a high-fidelity cochlear representation of sound, leveraging recent technical advances[35] to train the large models that are required for such high-dimensional input. Unlike previous generations of neural network models[40,

156, 184, 148, 116], which were reliant on hand-specified sound features, we learn all subsequent stages of a sound localization system to obtain good performance in real-world conditions.

When tested on stimuli from classic laboratory experiments, the resulting model replicated a large and diverse array of human behavioural characteristics. We then trained models in unnatural conditions to simulate evolution and development in alternative worlds. These alternative models deviated notably from human-like hearing. The results indicate that the characteristics of human hearing are indeed adapted to the constraints of real-world localization, and that the rich panoply of sound localization phenomena can be explained as consequences of this adaptation. The approach we use is broadly applicable to other sensory modalities, providing a way to test the adaptedness of aspects of human perception to the environment and to understand the conditions in which human-like perception arises.

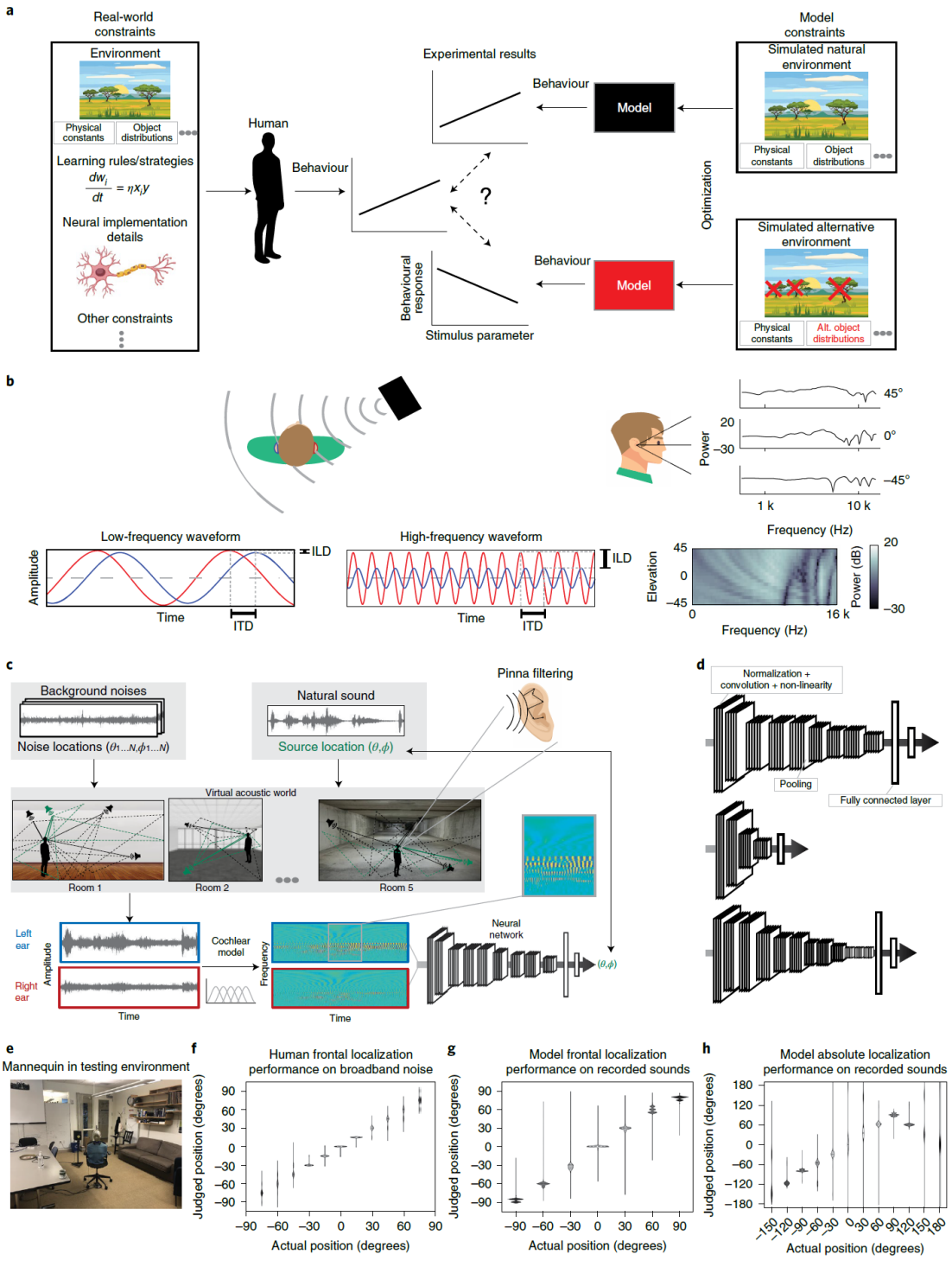
2.2 Results

2.2.1 Model Construction

We began by building a system that could localize sounds using the information available to human listeners. The system thus had outer ears (pinnae), and a simulated head and torso, along with a simulated cochlea. The outer ears and head/torso were simulated using head-related impulse responses (HRIRs) recorded from a standard physical model of the human[69]. The cochlea was simulated with a bank of band-pass filters modelled on the frequency selectivity of the human ear[78, 159], whose output was rectified and low-pass filtered to simulate the presumed upper limit of phase locking in the auditory nerve[171]. The inclusion of a fixed cochlear front-end (in lieu of trainable filters) reflected the assumption that the cochlea evolved to serve many different auditory tasks rather than being primarily driven by sound localization. As such, the cochlea seemed a plausible biological constraint on localization.

The output of the two cochleae formed the input to a standard convolutional

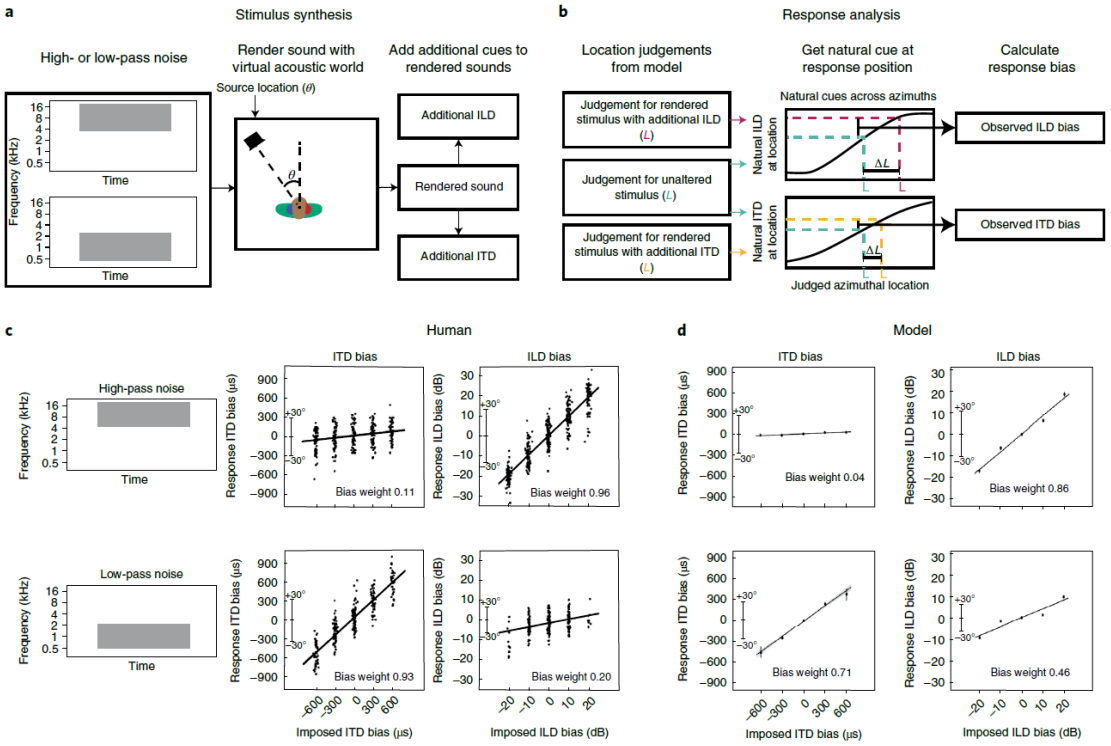
neural network (CNN) (Fig. 2-1c). This network instantiated a cascade of simple operations—filtering, pooling and normalization—culminating in a softmax output layer with 504 units corresponding to different spatial locations (spaced 5° in azimuth and 10° in elevation). The parameters of the model were tuned to maximize localization performance on the training data. The optimization procedure had two phases: an architecture search in which we searched over architectural parameters to find a network architecture that performed well (Fig. 2-1d), and a training phase in which the filter weights of the selected architectures were trained to asymptotic performance levels using gradient descent.



(Caption on next page.)

Figure 2-1:
a, Illustration of the method. A variety of constraints (left) shape human behaviour. Models optimized under particular environmental constraints (right) illustrate the effect of these constraints on behaviour. Environment simulators can instantiate naturalistic environments as well as alternative environments in which particular properties of the world are altered, to examine the constraints that shape human behaviour. b, Cues to sound location available to humans: interaural time and level differences (ITDs and ILDs) (left and centre) and spectral cues to elevation (right). Time and level differences are shown for low and high-frequency sinusoids (left and centre, respectively). The level difference is small for the low frequency, and the time difference is ambiguous for the high frequency. c, Training procedure. natural sounds (green) were rendered at a location in a room, with noises (natural sound textures, black) placed at other locations. Rendering included direction-specific filtering by the head/torso/pinnae, using head- related transfer functions from the KEMAR mannequin. neural networks were trained to classify the location of the natural sound source (azimuth and elevation) into one of a set of location bins (spaced 5° in azimuth and 10° in elevation). d, Example neural network architectures from the architecture search. Architectures consisted of sequences of 'blocks' (a normalization layer, followed by a convolution layer, followed by a non-linearity layer) and pooling layers, culminating in fully connected layers followed by a classifier that provided the network's output. Architectures varied in the total number of layers, the kernel dimensions for each convolutional layer, the number of blocks that preceded each pooling layer and the number of fully connected layers preceding the classifier. Labels indicate an example block, pooling layer and fully connected layer. The model's behaviour was taken as the average of the results for the ten best architectures (assessed by performance on a held-out validation set of training examples). e, Recording setup for real-world test set. The mannequin was seated on a chair and rotated relative to the speaker to achieve different azimuthal positions. Sound was recorded from microphones in the mannequin ears. f, Free-field localization of human listeners, replotted from a previous publication[240]. Participants heard a sound played from one of 11 speakers in the front horizontal plane and pointed to the location. Graph plots kernel density estimate of participant responses for each actual location. g, Localization judgements of the trained model for the real-world test set. Graph plots kernel density estimates of response distribution. For ease of comparison with f, in which all locations were in front of the listener, positions were front–back folded. h, Localization judgements of the model without front–back folding. Model errors are predominantly at front–back reflections of the correct location.

Figure 2-2:
a, Schematic of stimulus generation. noise bursts filtered into high or low-frequency bands were rendered at a particular azimuthal position, after which an additional ITD or ILD was added to the stereo audio signal. b, Schematic of response analysis. Responses were analysed to determine the amount by which the perceived location (L) was altered (Δ L) by the added ITD/ILD bias, expressed as the amount by which the ITD/ILD would have changed if the actual sound's location changed by Δ L. c, Effect of added ITD and ILD bias on human localization. The y axis plots amount by which the perceived location was altered, expressed in ITD/ILD as described above. Each dot plots a localization judgement from one trial. Data reproduced from a previous publication[151]. d, Effect of additional ITD and ILD on model localization. Same conventions as b. Error bars plot s.e.m., bootstrapped across the ten networks.



The architecture search consisted of training each one of a large set of possible architectures for 15,000 training steps with 16 1-s stimulus examples per step (240,000 total examples; see Extended Data Fig. 2-9 for distribution of localization performance across architectures and Extended Data Fig. 2-10 for the distributions from which architectures were chosen). We then chose the ten networks that performed best on a validation set of data not used during training (Extended Data Fig. 2-11). The parameters of these ten networks were then reinitialized and each trained for 100,000 training steps (1.6 million examples). Given evidence that internal representations can vary across different networks trained on the same task[161], we present results aggregated across the top ten best-performing architectures, treated akin to different participants in an experiment[228]. Most results graphs present the average results for these ten networks, which we collectively refer to as 'the model'.

The training data were based on a set of roughly 500,000 stereo audio signals with associated three-dimensional (3D) locations relative to the head (on average 988 examples for each of the 504 location bins, Methods). These signals were generated from 385 natural sound source recordings (Extended Data Fig. 2-12) rendered at a spatial location in a simulated room. The room simulator used a modified source-image method[197, 3] to simulate the reflections off the walls of the room. Each reflection was then filtered by the (binaural) HRIR for the direction of the reflection[69]. Five different rooms were used, varying in their dimensions and in the material of the walls (Extended Data Fig. 2-13). To mimic the common presence of noise in real-world environments, each training signal also contained spatialized noise. Background noise was synthesized from the statistics of a natural sound texture[160], and was rendered at between three and eight randomly chosen locations using the same room simulator to produce noise that was diffuse but non-uniform, intended to replicate common real-world sources of noise. At each training step, the rendered natural sound sources were randomly paired with rendered background noises. The neural networks were trained to map the binaural audio to the location of the sound source (specified by its azimuth and elevation relative to the model's 'head').

2.2.2 Model evaluation in real-world conditions

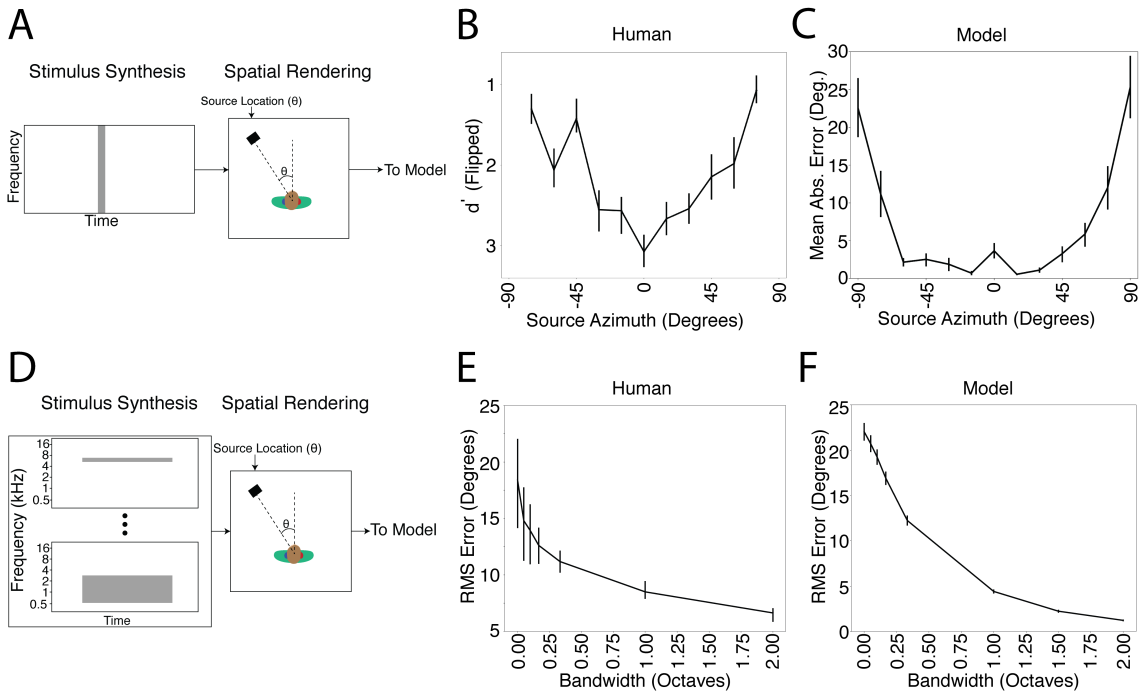
The trained networks were first evaluated on a held-out set of 70 sound sources rendered using the same pipeline used to generate the training data (yielding a total of around 47,000 stereo audio signals). The best-performing networks produced accurate localization for this validation set (the mean error was 5.3° in elevation and 4.4° in azimuth, front–back folded: that is, reflected about the coronal plane to discount front–back confusions).

To assess whether the model would generalize to real-world stimuli outside the training distribution, we made binaural recordings in an actual conference room using a mannequin with in-ear microphones (Fig. 2-1e). Humans localize relatively well in such free-field conditions (Fig. 2-1f). The trained networks also localized real-world recordings relatively well (Fig. 2-1g), on par with human free-field localization, with errors mostly limited to the front–back confusions that are common to humans when they cannot move their heads (Fig. 2-1h)[242, 220].

For comparison, we also assessed the performance of a standard set of two-microphone localization algorithms from the engineering literature[223, 193, 55, 54, 238, 217]. In addition, we trained the same set of neural networks to localize sounds from a two-microphone array that lacked the filtering provided to biological organisms by the ears, head and torso but that included the simulated cochlea (Extended Data Fig. 2-14a). Our networks that had been trained with biological pinnae, head and torso filtering outperformed the set of standard two-microphone algorithms from the engineering community, as well as the neural networks trained with stereo microphone input without a head and ears (Extended Data Fig. 2-14b,c). This latter result confirms that the head and ears provide valuable cues for localization. Overall, performance on the real-world test set demonstrates that training a neural network in a virtual world produces a model that can accurately localize sounds in realistic conditions.

2.2.3 Model behavioural characteristics

To assess whether the trained model replicated the characteristics of human sound localization, we simulated a large set of behavioural experiments from the literature, intended to span many of the best-known and largest effects in spatial hearing. We replicated the conditions of the original experiments as closely as possible (for example, when humans were tested in anechoic conditions, we rendered experimental stimuli in an anechoic environment). We emphasize that the networks were not fit to human data in any way. Despite this, the networks reproduced the characteristics of human spatial hearing across this broad set of experiments.



(Caption on next page.)

Figure 2-3:
a, Schematic of stimuli from experiment measuring localization accuracy at different azimuthal positions. b, Localization accuracy of human listeners for broadband noise at different azimuthal positions. Data were scanned from a previous publication[229], which measured discriminability of noise bursts separated by 15° (quantified as d'). Error bars plot s.e.m. c, Localization accuracy of our model for broadband noise at different azimuthal positions. Graph plots mean absolute localization error (Mean abs. error) of the same noise bursts used in the human experiment in b. Error bars plot the s.e.m. across the ten networks. d, Schematic of stimuli from experiment measuring effect of bandwidth on localization accuracy. noise bursts varying in bandwidth were presented at particular azimuthal locations; participants indicated the azimuthal position with a keypress. e, Effect of bandwidth on human localization of noise bursts. Accuracy was quantified as r.m.s. error. Error bars plot the s.d. Data are replotted from a previous publication[241]. f, Effect of bandwidth on model localization of noise bursts. networks were constrained to report only the azimuth of the stimulus. Error bars plot s.e.m. across the ten networks.

2.2.4 Sensitivity to interaural time and level differences

We began by assessing whether the networks learned to use the binaural cues known to be important for biological sound localization. We probed the effect of interaural time differences (ITDs) and interaural level differences (ILDs) on localization behaviour using an experiment in which additional time and level differences are added to high- and low-frequency sounds rendered in virtual acoustic space[151] (Fig. 2-2a). This experimental method has the advantage of using realistically externalized sounds and an absolute localization judgement (rather than the left/right lateralization judgements of simpler stimuli that are common to many other experiments[249, 89, 102, 26]).

In the original experiment[151], the change to perceived location imparted by the additional ITD or ILD was expressed as the amount by which the ITD or ILD would change in natural conditions if the actual location were changed by the perceived amount (Fig. 2-2b). This yields a curve whose slope indicates the efficacy of the manipulated cue (ITD or ILD). We reproduced the stimuli from the original study, rendered them in our virtual acoustic world, added ITDs and ILDs as in the original study and analysed the model's localization judgements in the same way.

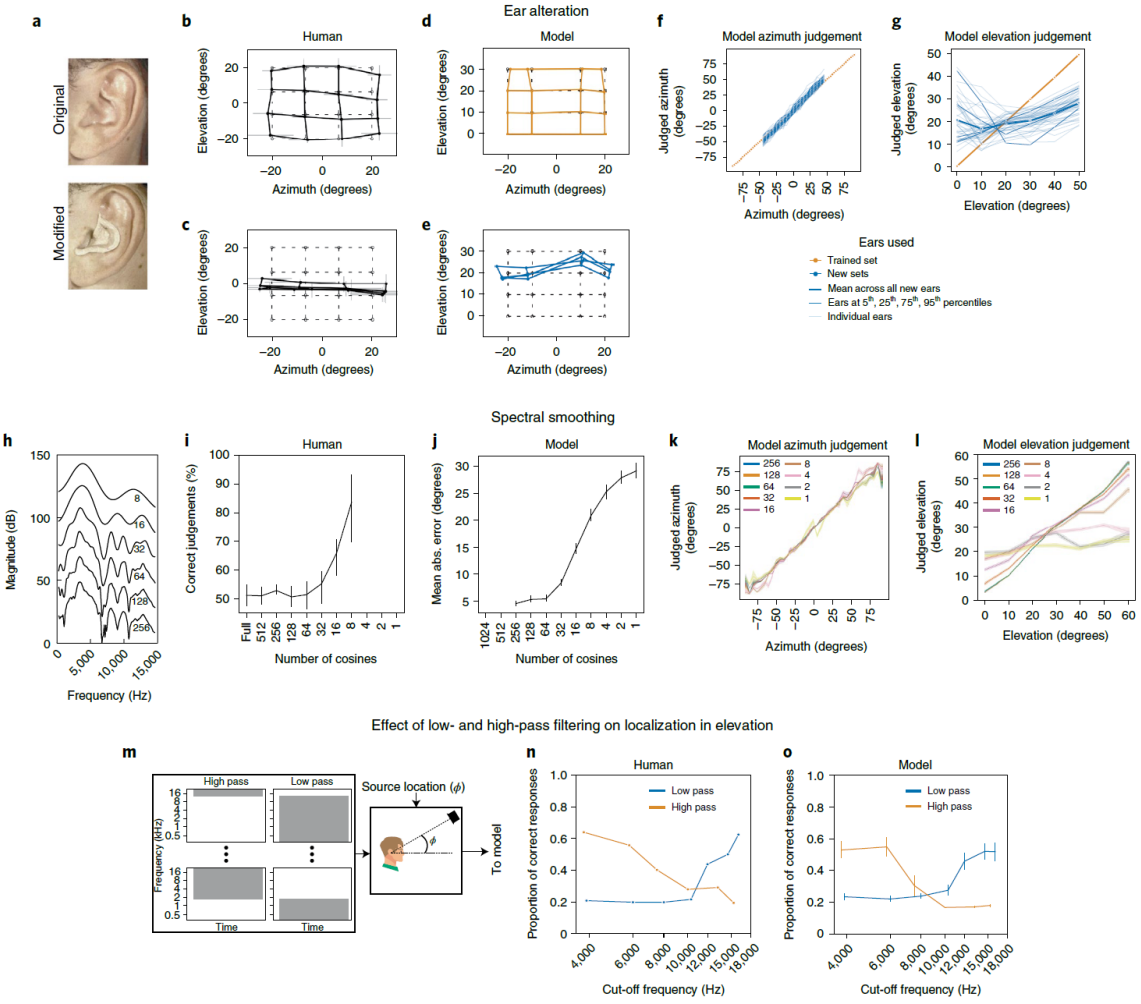
For human listeners, ITD and ILD have opposite efficacies at high and low frequencies (Fig. 2-2c), as predicted by classical 'duplex' theory[182]. An ITD bias imposed on low-frequency sounds shifts the perceived location of the sound substantially (bottom left), whereas an ITD imposed on high-frequency sound does not (top left). The

opposite effect occurs for ILDs (right panels), although there is a weak effect of ILDs on low-frequency sound. This latter effect is inconsistent with the classical duplex story but consistent with more modern measurements indicating small but reliable ILDs at low frequencies[31] that are used by the human auditory system[90, 239, 94].

As shown in Fig. 2-2d, the model qualitatively replicated the effects seen in humans. Added ITDs and ILDs had the largest effect at low and high frequencies, respectively, but ILDs had a modest effect at low frequencies as well. This produced an interaction between the type of cue (ITD/ILD) and frequency range (difference of differences between slopes significantly greater than 0; $P < 0.001$, evaluated by bootstrapping across the ten networks). However, the effect of ILD at low frequencies was also significant (slope significantly greater than 0; $P < 0.001$, via bootstrap). Thus, a model optimized for accurate localization both exhibits the dissociation classically associated with duplex theory, but also its refinements in the modern era.

2.2.5 Azimuthal localization of broadband sounds

We next measured localization accuracy of broadband noise rendered at different azimuthal locations (Fig. 2-3a). In humans, localization is most accurate near the midline (Fig. 2-3b), and becomes progressively less accurate as sound sources move to the left or right of the listener[188, 165, 229]. One explanation is that the first derivatives of ITD and ILD with respect to azimuthal location decrease as the source moves away from the midline[16], providing less information about location[17]. The model qualitatively reproduced this result (Fig. 2-3c).



(Caption on next page.)

Figure 2-4:
a, Photographs of ear alteration in humans (reproduced from a previous publication[105]). b, Sound localization by human listeners with unmodified ears. Graph plots mean and s.e.m. of perceived locations for four participants, superimposed on grid of true locations (dashed lines). Data scanned from the original publication[105]. c, Effect of ear alteration on human localization. Same conventions as b. d, Sound localization in azimuth and elevation by the model, using the ears (HRIRs) from training, with broadband noise sound sources. Graph plots mean locations estimated by the ten networks. Tested locations differed from those in the human experiment to conform to the location bins used for network training. e, Effect of ear alteration on model sound localization. Ear alteration was simulated by substituting an alternative set of HRIRs when rendering sounds for the experiment. Graph plots average results across all 45 sets of alternative ears (averaged across the ten networks). f, Effect of individual sets of alternative ears on localization in azimuth. Graph shows results for a larger set of locations than in d and e to illustrate the generality of the effect. g, Effect of individual sets of alternative ears on localization in elevation. Bolded lines show ears at 5th, 25th, 75th and 95th percentiles when the 45 sets of ears were ranked by accuracy. h, Smoothing of HRTFs, produced by varying the number of coefficients in a discrete cosine transform. Reproduced from the original publication, ref. [137]. i, Effect of spectral smoothing on human perception. Participants heard two sounds, one played from a speaker in front of them and one played through open-backed earphones, and judged which was which. The earphone-presented sound was rendered using HRTFs smoothed by various degrees. In practice, participants performed the task by noting changes in apparent sound location. Data scanned from the original publication[137]. Error bars plot s.e.m. Conditions with 4, 2 and 1 cosine coefficients were omitted from the experiment, but are included on the x axis to facilitate comparison with the model results in j. j, Effect of spectral smoothing on model sound localization accuracy (measured in both azimuth and elevation, as the mean absolute localization error). Conditions with 512 and 1,024 cosine components were not realizable given the length of the impulse responses we used. k, Effect of spectral smoothing on model accuracy in azimuth. l, Effect of spectral smoothing on model accuracy in elevation. m, Stimuli from experiment in n and o. noise bursts varying in low- or high-pass cut-off were presented at particular elevations. n, Effect of low- and high-pass cut-off on accuracy in humans. Data scanned from the original publication[98]; error bars were not provided in the original publication. o, Effect of low- and high-pass cut-off on model accuracy. networks were constrained to report only elevation. Here and in j, k and l, error bars plot s.e.m. across the ten networks.

2.2.6 Integration across frequency

Because biological hearing begins with a decomposition of sound into frequency channels, binaural cues are thought to be initially extracted within these channels[85, 115]. However, organisms are believed to integrate information across frequency to achieve more accurate localization than could be mediated by any single frequency channel. One signature of this integration is improvement in localization accuracy as the bandwidth of a broadband noise source is increased (Fig. 2-3d,e)[30, 241]. We replicated one such experiment on the networks and they exhibited a similar effect, with accuracy increasing with noise bandwidth (Fig. 2-3f).

2.2.7 Use of ear-specific cues to elevation

In addition to the binaural cues that provide information about azimuth, organisms are known to make use of the direction-specific filtering imposed on sound by the ears, head and torso[11, 227]. Each individual's ears have resonances that 'colour' a sound differently depending on where it comes from in space. Individuals are believed to learn the specific cues provided by their ears. In particular, if forced to listen with altered ears, either via moulds inserted into the ears[105] or via recordings made in a different person's ears[225], localization in elevation degrades even though azimuthal localization is largely unaffected (Fig. 2-4a–c).

To test whether the trained networks similarly learned to use ear-specific elevation cues, we measured localization accuracy in two conditions: one where sounds were rendered using the HRIR set used for training the networks, and another where the impulse responses were different (having been recorded in a different person's ears). Because we have unlimited ability to run experiments on the networks, in the latter condition we evaluated localization with 45 different sets of impulse responses, each recorded from a different human. As expected, localization of sounds rendered with the ears used for training was good in both azimuth and elevation (Fig. 2-4d). But when tested with different ears, localization in elevation generally collapsed (Fig. 2-4e), much like what happens to human listeners when moulds are inserted in their

ears (Fig. 2-4c), even though azimuthal localization was nearly indistinguishable from that with the trained ears. Results for individual sets of alternative ears revealed that elevation performance transferred better across some ears than others (Fig. 2-4f,g), consistent with anecdotal evidence that sounds rendered with head-related transfer functions (HRTFs) other than one's own can sometimes be convincingly localized in three dimensions.

2.2.8 Limited spectral resolution of elevation cues

Elevation perception is believed to rely on the peaks and troughs introduced to a sound's spectrum by the ears/head/torso[11, 16, 227] (Fig. 2-1b, right). In humans, however, perception is dependent on relatively coarse spectral features — the transfer function can be smoothed substantially before human listeners notice abnormalities[137] (Fig. 2-4h,i), for reasons that are unclear. In the original demonstration of this phenomenon, human listeners discriminated sounds with and without smoothing, a judgement that was in practice made by noticing changes in the apparent location of the sound. To test whether the trained networks exhibited a similar effect, we presented sounds to the networks with similarly smoothed transfer functions and measured the extent to which the localization accuracy was affected. The effect of spectral smoothing on the networks' accuracy was similar to the measured sensitivity of human listeners (Fig. 2-4j). The effect of the smoothing was most prominent for localization in elevation, as expected, but there was also some effect on localization in azimuth for the more extreme degrees of smoothing (Fig. 2-4k,l), consistent with evidence that spectral cues affect azimuthal space encoding[111].

2.2.9 Dependence on high-frequency spectral cues to elevation

The cues used by humans for localization in elevation are primarily in the upper part of the spectrum[139, 14]. To assess whether the trained networks exhibited a similar dependence, we replicated an experiment measuring the effect of high- and low-pass filtering on the localization of noise bursts[98] (Fig. 2-4m). Model performance varied

with the frequency content of the noise in much the same way as human performance (Fig. 2-4n,o).

2.2.10 The precedence effect

Another hallmark of biological sound localization is that judgements are biased towards information provided by sound onsets[18, 200]. The classic example of this bias is known as the precedence effect[221, 143, 24]. If two clicks are played from speakers at different locations with a short delay (Fig. 2-5a), listeners perceive a single sound whose location is determined by the click that comes first. The effect is often suggested to be an adaptation to the common presence of reflections off environmental surfaces (Fig. 2-1c) — reflections arrive from an erroneous direction but traverse longer paths and arrive later, such that basing location estimates on the earliest arriving sound might avoid errors[18]. To test whether our model would exhibit a similar effect, we simulated the classic precedence experiment, rendering two clicks at different locations. When clicks were presented simultaneously, the model reported the sound to be centred between the two click locations, but when a small inter-click delay was introduced, the reported location switched to that of the leading click (Fig. 2-5b). This effect broke down as the delay was increased, as in humans, although with the difference that the model could not report hearing two sounds and so instead reported a single location intermediate between those of the two clicks.

To compare the model results to human data, we simulated an experiment in which participants reported the location of both the leading and lagging click as the interclick delay was varied[144]. At short but non-zero delays, humans accurately localize the leading but not the lagging click (Fig. 2-5c, because a single sound is heard at the location of the leading click). At longer delays, the lagging click is more accurately localized and listeners start to mislocalize the leading click, presumably because they confuse which click is first[144]. The model qualitatively replicated both effects, in particular the large asymmetry in localization accuracy for the leading and lagging sound at short delays (Fig. 2-5d).

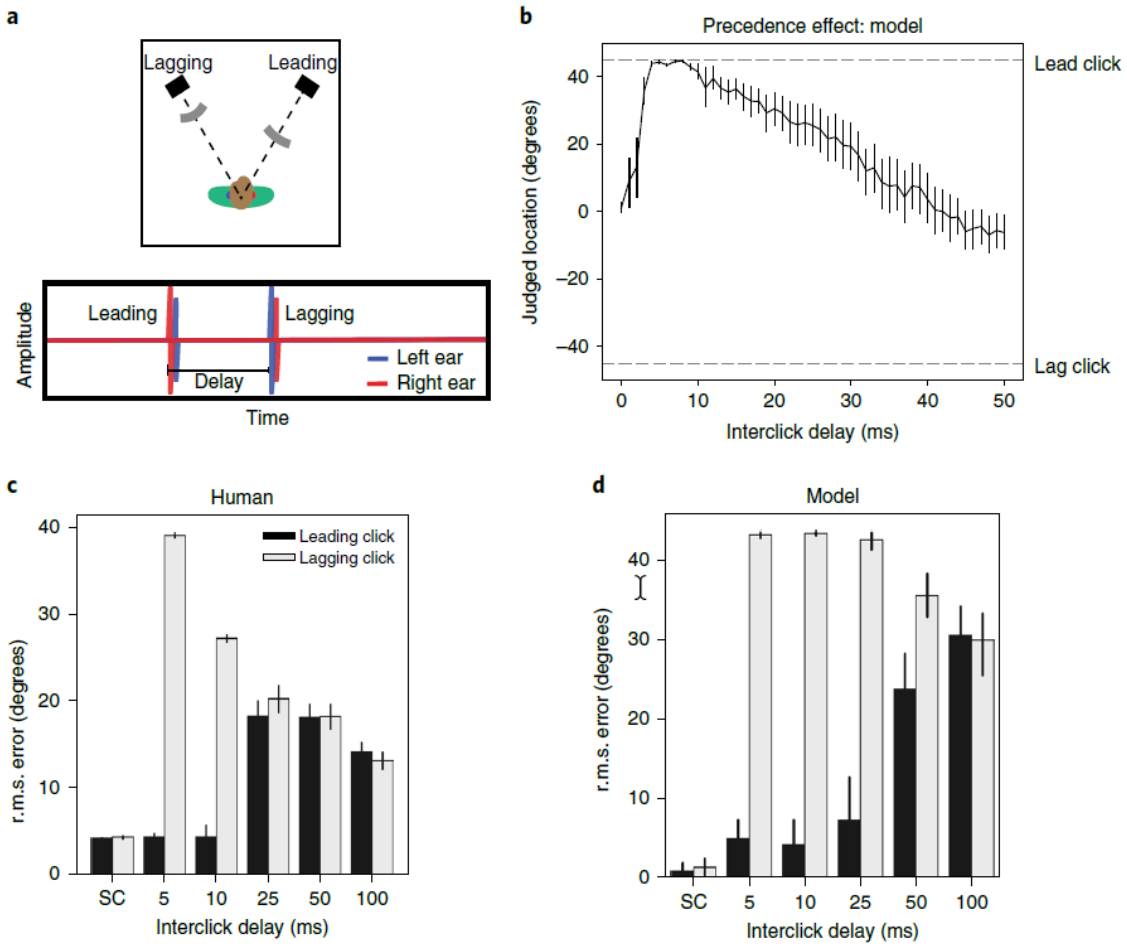
2.2.11 Multi-source localization

Humans are able to localize multiple concurrent sources, but only to a point[189, 122, 244]. The reasons for the limits on multi-source localization are unclear[122]. These limitations could reflect human-specific cognitive constraints. For instance, reporting a localized source might require attending to it, which could be limited by central factors not specific to localization. Alternatively, localization could be fundamentally limited by corruption of spatial cues by concurrent sources or other ambiguities intrinsic to the localization problem.

To assess whether the model would exhibit limitations like those observed in humans, we replicated an experiment[244] in which humans judged both the number and location of a set of speech signals played from a subset of an array of speakers (Fig. 2-6a). To enable the model to report multiple sources we fine-tuned the final fully connected layer to indicate the probability of a source at each of the location bins, and set a probability criterion above which we considered the model to report a sound at the corresponding location (Methods). The weights in all earlier layers were 'frozen' during this fine-tuning, such that all other stages of the model were identical to those used in all other experiments. We then tested the model on the experimental stimuli.

Humans accurately report the number of sources up to three, after which they undershoot, only reporting about four sources in total regardless of the actual number (Fig. 2-6b). The model reproduced this effect, also being limited to approximately four sources (Fig. 2-6c). Human localization accuracy also systematically drops with the number of sources (Fig. 2-6d): the model again quantitatively reproduced this effect (Fig. 2-6e). The model–human similarity suggests that these limits on sound localization are intrinsic to the constraints of the localization problem, rather than reflecting human-specific central factors.

Figure 2-5:
a, Diagram of stimulus. Two clicks are played from two different locations relative to the listener. The time interval between the clicks is manipulated and the listener is asked to localize the sound(s) that they hear. When the delay is short but non-zero, listeners perceive a single click at the location of the first click. At longer delays, listeners hear two distinct sounds. b, Localization judgements of the model for two clicks at + 45 and - 45°. The model exhibits a bias for the leading click when the delay is short but non-zero. At longer delays, the model judgements (which are constrained to report the location of a single sound, unlike those of humans) converge to the average of the two click locations. Error bars plots s.e.m. across the ten networks. c, Error in localization of the leading and lagging clicks by humans as a function of interclick delay. SC denotes a single click at the leading or lagging location. Bars plot r.m.s. localization error. Error bars plot s.d. Data scanned from the original publication[144]. d, Error in localization of the leading and lagging clicks by the model as a function of interclick delay. Bars plot r.m.s. localization error. Error bars plots s.e.m. across the ten networks.



2.2.12 Effect of optimization for unnatural environments

Despite having no previous exposure to the stimuli used in the experiments and despite not being fit to match human data in any way, the model qualitatively replicated a wide range of classic behavioural effects found in humans. These results raise the possibility that the characteristics of biological sound localization may be understood as a consequence of optimization for real-world localization. However, given these results alone, the role of the natural environment in determining these characteristics is left unclear.

To assess the extent to which the properties of biological hearing are adapted to the constraints of localization in natural environments, we took advantage of the ability to optimize models in virtual worlds altered in various ways, intended to simulate the optimization that would occur over evolution and/or development in alternative environments (Fig. 2-1a). We altered the training environment in one of three ways (Fig. 2-7a): (1) by eliminating reflections (simulating surfaces that absorb all sound that reaches them, unlike real-world surfaces), (2) by eliminating background noise and (3) by replacing natural sound sources with artificial sounds (narrowband noise bursts). In each case, we trained the networks to asymptotic performance, then froze their weights and ran them on the full suite of psychophysical experiments described above. The psychophysical experiments were identical for all training conditions; the only difference was the strategy learned by the model during training, as might be reflected in the experimental results. We then quantified the dissimilarity between the model psychophysical results and those of humans as the mean squared error between the model and human results, averaged across experiments (normalized to have uniform axis limits, Methods).

Figure 7b shows the average dissimilarity between the human and model results on the suite of psychophysical experiments, computed separately for each model training condition. The dissimilarity was lowest for the model trained in natural conditions, and significantly higher for each of the alternative conditions ($P < 0.001$ in each case, obtained by comparing the dissimilarity of the alternative conditions to a null

distribution obtained via bootstrap across the ten networks trained in the naturalistic condition; results were fairly consistent across networks, Extended Data Fig. 2-15). The effect size of the difference in dissimilarity between the naturalistic training condition results and each of the other training conditions was large in each case (d = 2.13, anechoic; d = 2.75, noiseless; d = 3.06, unnatural sounds). This result provides additional evidence that the properties of spatial hearing are consequences of adaptation to the natural environment — human-like spatial hearing emerged from task optimization only for naturalistic training conditions.

To get an insight into how the environment influences perception, we examined the human–model dissimilarity for each experiment individually (Fig. 2-7c). Because the absolute dissimilarity is not meaningful (in that it is limited by the reliability of the human results, which are not perfect; Extended Data Fig. 2-16), we assessed the differences in human–model dissimilarity between the natural training condition and each unnatural training condition. These differences were most pronounced for a subset of experiments in each case.

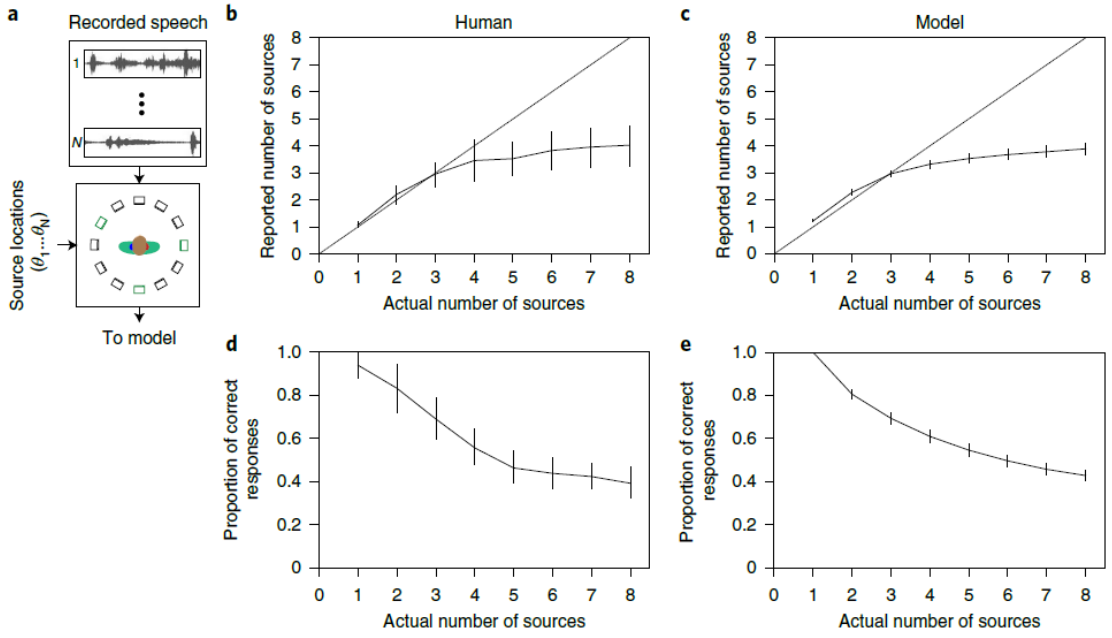
The anechoic training condition produced most abnormal results for the precedence effect, but also produced substantially different results for ITD cue strength. The effect size for the difference in human–model dissimilarity between anechoic and natural training conditions was significantly greater in both these experiments (precedence effect d = 4.16; ITD cue strength d = 3.41) than in the other experiments ($P < 0.001$, by comparing the effect sizes of one experiment to the distribution of the effect size for another experiment obtained via bootstrap across networks). The noiseless training condition produced most abnormal results for the effect of bandwidth (d = 4.71; significantly greater than that for other experiments, $P < 0.001$, via bootstrap across networks). We confirmed that this result was not somehow specific to the absence of internal neural noise in our cochlear model, by training an additional model in which noise was added to each frequency channel (Methods). We found that the results of training in noiseless environments remained very similar. The training condition with unnatural sounds produced most abnormal results for the experiment measuring elevation perception (d = 4.4 for the ear alteration experiment;

d = 4.28 for the high-frequency elevation cue experiment; $P < 0.001$ in both cases, via bootstrap across networks), presumably because without the pressure to localize broadband sounds, the model did not acquire sensitivity to spectral cues to elevation. These results indicate that different worlds would lead to different perceptual systems with distinct localization strategies.

The most interpretable example of environment-driven localization strategies is the precedence effect. This effect is often proposed to render localization robust to reflections, but others have argued that its primary function might instead be to eliminate interaural phase ambiguities, independent of reflections[248]. This effect is shown in Fig. 2-7d for models trained in each of the four virtual environments. Anechoic training completely eliminated the effect, even though it was largely unaffected by the other two unnatural training conditions. This result substantiates the hypothesis that the precedence effect is an adaptation to reflections in real-world listening conditions. See Extended Data Figs. 9 and 10 for full psychophysical results for models trained in alternative conditions.

In addition to diverging from the perceptual strategies found in human listeners, the models trained in unnatural conditions performed more poorly at real-world localization. When we ran models trained in alternative conditions on our real-world test set of recordings from mannequin ears in a conference room, localization accuracy was substantially worse in all cases (Fig. 2-7e, $P < 0.001$ in all cases). This finding is consistent with the common knowledge in engineering that training systems in noisy and otherwise realistic conditions aids performance[156, 148, 116, 91]. Coupled with the abnormal psychophysical results of these alternatively trained models, this result indicates that the classic perceptual characteristics of spatial hearing reflect strategies that are important for real-world localization, in that systems that deviate from these characteristics localize poorly.

Figure 2-6:
a, Diagram of experiment. On each trial, between one and eight speech signals (each spoken by a different talker) was played from a subset of the speakers in a 12-speaker circular array. The lower panel depicts an example trial in which three speech signals were presented, with the corresponding speakers in green. Participants reported the number of sources and their locations. b, Average number of sources reported by human listeners, plotted as a function of the actual number of sources. Error bars plot standard deviation across participants. Here and in d, graph is reproduced from original paper[244] with permission of the authors. c, Same as b, but for the model. Error bars plot standard deviation across the ten networks. d, Localization accuracy (measured as the proportion of sources correctly localized to the actual speaker from which they were presented), plotted as a function of the number of sources. Error bars plot s.d. across participants. e, Same as d, but for the model. Error bars plot s.d. across the ten networks.



2.2.13 Model predictions of sound localizability

One advantage of a model that can mediate actual localization behaviour is that one can run large numbers of experiments on the model, searching for 'interesting' predictions that might then be tested in human listeners. Here we used the model to estimate the accuracy with which different natural sounds would be localized in realistic conditions. We chose to examine musical instrument sounds as these are both diverse and available as clean recordings in large numbers. We took a large

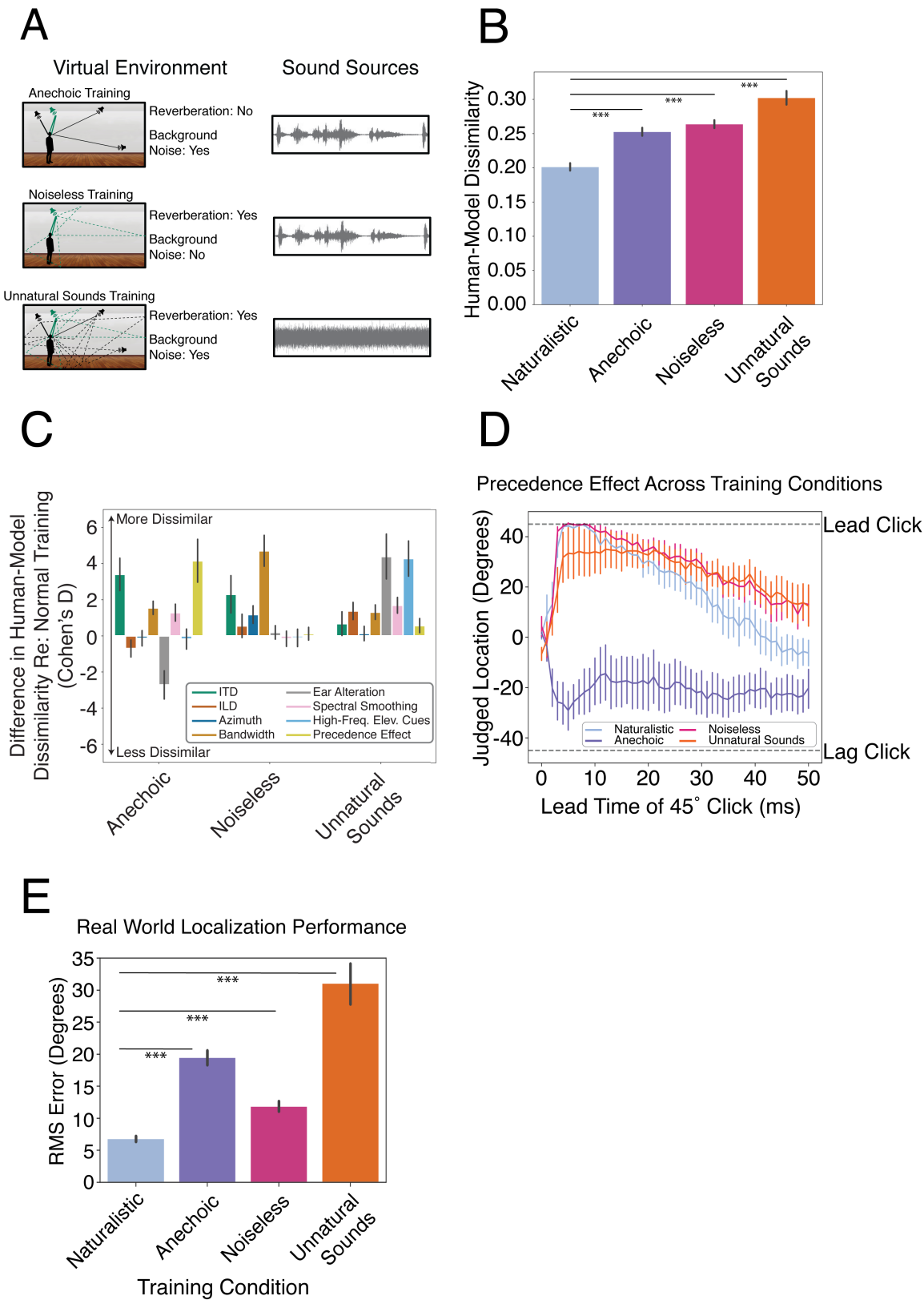
set of instrument sounds[59] and rendered them at a large set of randomly selected locations. We then measured the average localization error for each instrument.

As shown in Fig. 2-8a, there was reliable variation in the accuracy with which instrument sounds were localized by the model. The median error was as low as 1.06° for reed instrument no. 3 and as high as 40.02° for mallet no. 1 (folded to discount front–back confusions: without front–back folding, the overall error was larger, but the ordinal relations among instruments were similar). The human voice was also among the most accurately localized sounds in the set we examined, with a mean error of 2.39° (front–back folded).

Figure 8b displays spectrograms for example notes for the three best- and worst-localized instruments. The best-localized instruments are spectrally dense, and thus presumably take advantage of cross-frequency integration (which improve localization accuracy in both humans and the model, Fig. 2-3e,f). This result is consistent with the common idea that narrowband sounds are less well localized, but the model provides a quantitative metric of localizability that we would not otherwise have.

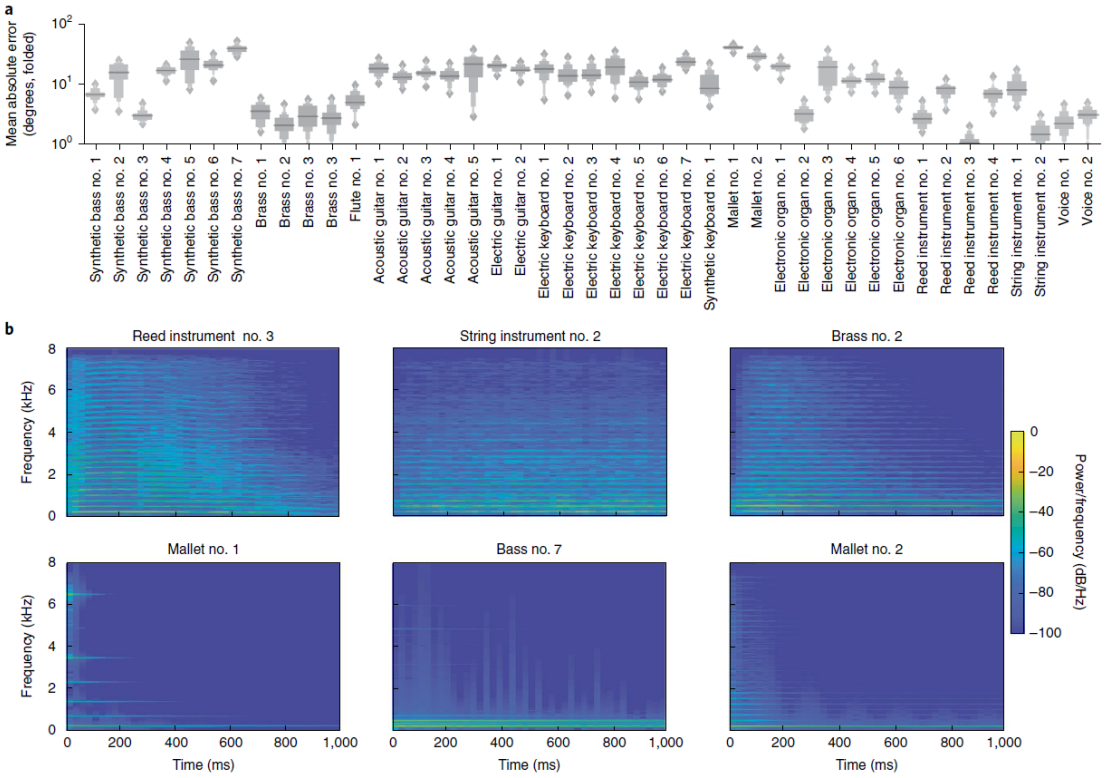
To assess whether the results could be predicted by simple measures of spectral sparsity, we measured the spectral flatness[118] of each instrument sound (the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum). The average spectral flatness of an instrument was significantly correlated with the model's localization accuracy ($rs = 0.77$,$P < 0.001$), but this correlation was well below the split-half reliability of the model's accuracy for an instrument ($rs = 0.99$). This difference suggests that there may be sound features above and beyond spectral sparsity that determine a sound's localizability, and illustrates the value of an optimized system to make perceptual predictions.

We had intentions of running a free-field localization experiment in humans to test these predictions, but had to halt experiments due to COVID-19. We have hopes of running the experiment in the future. However, we note that informal observation by the authors listening in free-field conditions suggest that the sounds that are poorly localized by the model are also difficult for humans to localize.



(Caption on next page.)

Figure 2-7:
a, Schematic depiction of altered training conditions, eliminating echoes or background noise or using unnatural sounds. b, Overall human–model dissimilarity for natural and unnatural training conditions. Error bars plot s.e.m., bootstrapped across networks. Asterisks denote statistically significant differences between conditions ($P < 0.001$, two-tailed), evaluated by comparing the human–model dissimilarity for each unnatural training condition to a bootstrapped null distribution of the dissimilarity for the natural training condition. c, Effect of unnatural training conditions on human–model dissimilarity for individual experiments, expressed as the effect size of the difference in dissimilarity between the natural and each unnatural training condition (Cohen's *d*, computed between human–model dissimilarity for networks in normal and modified training conditions). Positive numbers denote a worse resemblance to human data compared to the model trained in normal conditions. Error bars plot s.e.m., bootstrapped across the ten networks d, The precedence effect in networks trained in alternative environments. e, Real-world localization accuracy of networks for each training condition. Error bars plot s.e.m., bootstrapped across the ten networks. Asterisks denote statistically significant differences between conditions ($P < 0.001$, two-tailed), evaluated by comparing the mean localization error for each unnatural training condition to a bootstrapped null distribution of the localization error for the natural training condition.



(Caption on next page.)

Figure 2-8:
a, Mean model localization error for each of 43 musical instruments. Each of a set of instrument notes was rendered at randomly selection locations. Graph shows letter-value plots[106] of the mean azimuthal localization error across notes, measured after actual and judged positions were front–back folded. Letter-value plots are boxplots with additional quantiles. The widest box depicts the middle two quartiles (1/4) of the data distribution, as in a box plot, the second widest box depicts the next two octiles (1/8), the third widest box depicts the next two hexadeciles (1/16) and so on, up to the upper and lower 1/64 quantiles. Horizontal line plots median value and diamonds denote outliers. b, Spectrograms of an example note (middle C) for the three most and least accurately localized instruments (top and bottom, respectively).

2.3 Discussion

We trained artificial neural networks to localize sounds from binaural audio rendered in a virtual world and heard through simulated ears. When the virtual world mimicked natural auditory environments, with surface reflections, background noise and natural sound sources, the trained networks replicated many attributes of spatial hearing found in biological organisms. These included the frequency-dependent use of ITDs and ILDs, the integration of spatial information across frequency, the use of ear-specific high-frequency spectral cues to elevation and robustness to spectral smoothing of these cues, localization dominance of sound onsets and limitations on the ability to localize multiple concurrent sources. The model successfully localized sounds in an actual real-world environment better than alternative algorithms that lacked ears. The model also made predictions about the accuracy with which different types of real-world sound could be localized. But when the training conditions were altered to deviate from the natural environment by eliminating surface reflections, background noise or natural sound source structure, the behavioural characteristics of the model deviated notably from human-like behaviour. The results indicate that most of the key properties of mammalian spatial hearing can be understood as consequences of optimization for the task of localizing sounds in natural environments. Our approach extends classical ideal observer analysis to new domains, where provably optimal analytic solutions are difficult to attain but where supervised machine learning can nonetheless provide optimized solutions in different conditions.

The general method involves two nested levels of computational experiments: opti-

mization of a model under particular conditions, followed by a suite of psychophysical experiments to characterize the resulting behavioural phenotype. This approach provides an additional tool with which to examine the constraints that yield biological solutions[38, 125], and thus to understand evolution[194]. It also provides a way to link experimental results with function. In some cases, these links had been proposed but not definitively established. For example, the precedence effect was often proposed to be an adaptation to reverberation[16, 221], although other functional explanations were also put forth[248]. Our results indicate it is indeed an adaptation to reverberation (Fig. 2-7d). We similarly provide evidence that sensitivity to spectral cues to elevation emerges only with the demands of localizing broadband sounds[164]. In other cases, the model provided explanations for behavioural characteristics that previously had none. One such example is the relatively coarse spectral resolution of elevation perception (Fig. 2-4h–j), which evidently reflects the absence of reliable information at finer resolutions. Another is the number of sources that can be concurrently localized (Fig. 2-6b,c), and the dependence of localization accuracy on the number of sources (Fig. 2-6d,e). Without an optimized model there would be no way to ascertain whether these effects reflect intrinsic limitations of localization cues in auditory scenes or some other human-specific cognitive limit.

Previous models of sound localization required cues to be hand-coded and provided to the model by the experimenter[18, 66, 40, 62]. In some cases, previous models were able to derive optimal encoding strategies for such cues[93], which could be usefully compared to neural data[201]. In other cases, models were able to make predictions of behaviour in simplified conditions using idealized cues[62]. However, the idealized cues that such models work with are not well-defined for arbitrary real-world stimuli[166], preventing the modelling of general localization behaviour. In addition, ear-specific spectral cues to elevation (Fig. 2-1b, right) are not readily hand-coded, and as a result have remained largely absent from previous models. It has thus not previously been possible to derive optimal behavioural characteristics for real-world behaviour.

Our results highlight the power of contemporary machine learning coupled with virtual training environments to achieve realistic behavioural competence in compu-

tational models. Supervised learning has traditionally been limited by the need for large amounts of labelled data, typically acquired via painstaking human annotation. Virtual environments allow the scientist to generate the data, with the labels coming for free (as the parameters used to generate the data), and have the potential to greatly expand the settings in which supervised learning can be used to develop models of the brain[67]. Virtual environments also allow tests of optimality that would be impossible in biological systems, because they enable environmental conditions to be controlled, and permit optimization on rapid timescales.

Our approach is complementary to the long tradition of mechanistic modelling of sound localization. In contrast with mechanistic modelling, we do not produce specific hypotheses about underlying neural circuitry. However, the model gave rise to rich predictions of real-world behaviour, and normative explanations of a large suite of perceptual phenomena. It should be possible to merge these two approaches, both by training model classes that are more faithful to biology (for example, spiking neural networks, or networks with biologically constrained weights)[86, 214], and by building in additional known biological structures to the neural network (for example, replicating brainstem circuitry)[119, 27].

One limitation of our approach is that optimization of biological systems occurs in two distinct stages of evolution and development, which are not obviously mirrored in our model optimization procedure. The procedure we used had separate stages of architectural selection and weight training, but these do not cleanly map onto evolution and development in biological systems. This limitation is shared by classical ideal observers, but limits the ability to predict effects that might be specific to one stage or the other, for instance involving plasticity[120].

Our model also shares many limitations common to current deep neural network models of the brain[138]. The learning procedure is unlikely to have much in common with biological learning, both in the extent and nature of supervision (which involves millions of explicitly labelled examples) and in the learning algorithm, which is often argued to lack biological plausibility[86]. The model class is also not fully consistent with biology, and so does not yield detailed predictions of neural circuitry.

The analogies with the brain thus seem most promising at the level of behaviour and representations. Our results add to growing evidence that task-optimized models can produce human-like behaviour for signals that are close to the manifold of natural sounds or images[126, 135, 187]. However, artificial neural networks also often exhibit substantial representational differences with humans, particularly for unnatural signals derived in various ways from a network[80, 61, 70, 112, 79], and our model may exhibit similar divergences.

We chose to train models on a fixed representation of the ear. This choice was motivated by the assumption that the evolution of the ear was influenced by many different auditory tasks, such that it may not have been strongly influenced by the particular demands of sound localization, instead primarily serving as a constraint on biological solutions to the sound localization problem[187]. However, the ear itself undoubtedly reflects properties of the natural environment[140]. It could thus be fruitful to 'evolve' ears along with the rest of the auditory system, particularly in a framework with multiple tasks[126]. Our cochlear model also does not replicate the fine details of cochlear physiology[246, 25, 8] due to practical constraints of limited memory resources. These differences could in principle influence the results, although the similarity of the model results to those of humans suggests that the details of peripheral physiology beyond those that we modelled do not figure critically in the behavioural traits we examined.

The virtual world we used to train our models also no doubt differs in many ways from real-world acoustic environments. The rendering assumed point sources in space, which is inaccurate for many natural sound sources. The distribution of source locations was uniform relative to the listener, and both the listener and the sound sources were static, all of which are often not true of real-world conditions. And although the simulated reverberation replicated many aspects of real-world reverberation, it probably did not perfectly replicate the statistical properties of natural environmental impulse responses[212], or their distribution across environments. Our results indicate that the virtual world approximates the actual world in many of the respects that matter for spatial hearing, but the discrepancies with the real world

could make a difference for some behaviours.

We also emphasize that despite presenting our approach as an alternative to ideal observer analysis[71, 72], the resulting model almost surely differs in some respects from a fully ideal observer. The solutions reached by our approach are not provably optimal like classic ideal observers, and the model class and optimization methods could impose biases on the solutions. It is also likely that the architecture search was not extensive enough to find the best architectures for the task. Those caveats aside, the similarity to human behaviour, along with the strong dependence on the training conditions, provides some confidence that the optimization procedure is succeeding to a degree that is scientifically useful.

Our focus in this paper has been to study behaviour, as there is a rich set of auditory localization behaviours for which normative explanations have traditionally been unavailable. However, it remains possible that the model we trained could be usefully compared to neural data. There is a large literature detailing binaural circuitry in the brainstem[53] that could be compared to the internal responses of the model. The model could also be used to probe for functional organization in the auditory cortex, for instance by predicting brain responses using features from different model stages[128, 88, 236, 41, 58, 126], potentially helping to reveal hierarchical stages of localization circuitry.

A model that can predict human behaviour should also have useful applications. Our model showed some transfer of localization for specific sets of ears (Fig. 2-4g), and could be used to make predictions about the extent to which sound rendering in virtual acoustic spaces (which may need to use a generic set of HRTF) should work for a particular listener. It can also predict which of a set of sound sources will be most compellingly localized, or worst localized (Fig. 2-8). Such predictions could be valuable in enabling better virtual reality, or in synthesizing signals that humans cannot pinpoint in space.

One natural extension of our model would be to incorporate moving sound sources and head movements. We modelled sound localization in static conditions because most experimental data have been collected in this setting. But in real-world con-

ditions sound sources often move relative to the listener and listeners move their heads[211, 21], often to better disambiguate front from back[220] and more accurately localize. Our approach could be straightforwardly expanded to moving sound sources in the virtual training environment and a model that can learn to move its head[148], potentially yielding explanations of auditory motion perception[82, 33, 247]. The ability to train models that can localize in realistic conditions also underscores the need for additional measurements of human localization behaviour — front–back confusions, localization of natural sounds in actual rooms, localization with head movements and so on — with which to further evaluate models.

Another natural next step is to instantiate both recognition and localization in the same model, potentially yielding insight into the segregation of these functions in the brain[15], and to the role of spatial cues in the 'cocktail party problem'[46, 50, 22, 95, 129, 157, 195]. More generally, the approach we take here — using deep learning to derive optimized solutions to perceptual or cognitive problems in different operating conditions — is broadly applicable to understanding the forces that shape complex, real-world, human behaviour.

2.4 Methods

2.4.1 Training data generation

Virtual acoustic simulator: image/source method

We used a room simulator[197] to render binaural room impulse responses (BRIRs). This simulator used the image-source method, which approaches an exact solution to the wave equation if the walls are assumed to be rigid[3], as well as an extension to that method that allowed for more accurate calculation of the arrival time of a wave[176]. This enabled the simulator to correctly render the relative timing between the signals received by the two simulated ears, including reflections (enabling both the direct sound and all reflections to be rendered with the correct spatial cues). Our specific implementation was identical to that used in the original paper[197], except

for some custom optimization to take advantage of vectorized operations and parallel computation.

The room simulator operated in three separate stages. First, the simulator calculated the positions of reflections of the source impulse forward in time for 0.5 s. For each of these positions, the simulator placed an image symmetrically reflected about the wall of last contact. Second, the simulator accounted for the absorption spectra of the reflecting walls for each image location and filtered a broadband impulse sequentially using the absorption spectrum of the simulated wall material. Third, the simulator found the direction of arrival for each image and convolved the filtered impulse with the HRIR in the recorded set whose position was closest to the computed direction. This resulted in a left and right channel signal pair for each path from the source to the listener. Last, each of these signal pairs was summed together, factoring in both the delay from the time of arrival and the level attenuation given the total distance travelled by each reflection. The original authors of the simulator previously assessed this method's validity and found that simulated BRIRs were good physical approximations to recorded BRIRs provided that sources were rendered more than 1 m from the listener[197].

We used this room simulator to render BRIRs at each of a set of listener locations in five different rooms varying in size and material (listed in Extended Data Fig. 2-15) for each of the source location bins in the output layer of the networks: all pairings of seven elevations (between 0° and 60°, spaced 10°) and 72 azimuths (spaced 5° in a circle around the listener), at a distance of 1.4 m. This yielded 504 source positions per listener location and room. Listener locations were chosen subject to three constraints. First, the listener location had to be at least 1.4 m from the nearest wall (because sounds were rendered 1.4 m from the listener). Second, the listener locations were located on a grid whose axes ran parallel to the walls of the room, with locations spaced 1 m apart in each dimension. Third, the grid was centred in the room. These constraints yielded four listener locations for the smallest room and 81 listener locations for the largest room. This resulted in 71,064 pairs of BRIRs, each corresponding to a possible source–listener–room spatial configuration. Each

BRIR took approximately 4 min to generate when parallelized across 16 cores. We parallelized[209] the generation of the full set of BRIRs across approximately 1,000 cores on the MIT OpenMind Cluster, which allowed us to generate the full set of BRIRs in approximately 4 days.

Virtual acoustic simulator: HRIRs

The simulator relied on empirically derived HRIRs to incorporate the effect of pinna filtering, head shadowing and time delays without solving wave equations for the ears, head and/or torso. Specifically, the simulator used a set of HRIRs recorded with KEMAR: a mannequin designed to replicate the acoustic effects of head and torso filtering on auditory signals. These recordings consisted of 710 positions ranging from 40° to +90° elevation at 1.4 m (ref. [69]). A subset of these positions corresponded to the location bins into which the network classified source locations.

Virtual acoustic simulator: two-microphone array

For comparison with the networks trained with simulated ears, we also trained the same neural network architectures to localize sounds using audio recorded from a two-microphone array (Extended Data Fig. 2-14). To train these networks, we simulated audio received from a two-microphone array by replacing each pair of HRIRs in the room simulator with a pair of fractional delay filters (that is, that delayed the signal by a fraction of a sample). These filters consisted of 127 taps and were constructed via a sinc function windowed with a Blackman window, offset in time by the desired delay. Each pair of delay filters also incorporated signal attenuation from a distance according to the inverse square law, with the goal of replicating the acoustics of a two-microphone array. After substituting these filters for the HRIRs used in our main training procedure, we simulated a set of BRIRs as described above.

Natural sound sources

We collected a set of 455 natural sounds, each cut to two seconds in length. Of these sounds, 300 were drawn from a set used in previous work in the laboratory[169]. An-

other 155 sounds were drawn from the BBC Sounds Effects Database, selected by the first author to be easily identifiable. The sounds included human and animal vocalizations, human actions (chopping, chewing, clapping and so on), machine sounds (cars, trains, vacuums and so on) and nature sounds (thunder, insects, running water, etc.). The full list of sounds is given in Extended Data Fig. 2-12. All sounds were sampled at 44.1 kHz. Of this set, 385 sounds were used for training and another 70 were withheld for model validation and testing. To augment the dataset, each of these was bandpass-filtered with a two-octave-wide second-order Butterworth filter with centre frequencies spaced in one-octave steps starting from 100 Hz. This yielded 2,492 (2,110 training, 382 testing) sound sources in total.

Background noise sources

Background noise sources were synthesized using a previously described texture generation method that produced texture excerpts rated as highly realistic[159]. The specific implementation of the synthesis algorithm was that used in ref. [160], with a sampling rate of 44.1 kHz. We used 50 different source textures obtained from in-laboratory collections[158]. Textures were selected that synthesized successfully, both subjectively (sounding perceptually similar to the original texture) and objectively (the ratio between mean squared statistic values for the original texture and the mean squared error between the statistics of the synthesized and original texture was greater than 40 dB). We then rendered 1,000 5-s exemplars for each texture (subsequently cut to 2 s in length) for a total of 50,000 unique waveforms (1,000 exemplars $\times$ 50 textures). Background noises were created by spatially rendering between three and eight exemplars of the same texture at randomly chosen locations using the virtual acoustic simulator described above. We made this choice on grounds of ecological validity, on the basis of the intuition that noise sources are typically not completely spatially uniform[189] despite being more diffuse than sounds made by single organisms or objects. By adding noises rendered at different locations we obtained background noise that was not as precisely localized as the target sound sources, which seemed a reasonable approximation of common real-world conditions.

Generating training exemplars

To reduce the storage footprint of the training data, we separately rendered the sound sources to be localized, and the background noise, and then randomly combined them to generate training exemplars. For each source, room and listener location, we randomly rendered each of the 504 positions with a probability:

$$P = \frac{0.025 \ \times no.\ of\ listener\ locations\ in\ smallest\ room}{no.\ of\ listener\ locations\ in\ room\ being\ rendered}$$

We used base probability of 2.5% to limit the overall size of the training set and normalized by the number of listener locations in the room being used to render the current stimulus so that each room was represented equally in the dataset. This yielded 545,566 spatialized natural sound source stimuli in total (497,935 for training, 47,631 for testing). This resulted in 988 examples per training location, on average.

For each training example, the audio from one spatialized natural sound source and one spatialized background texture scene was combined (with a signal-to-noise ratio (SNR) sampled uniformly from 5 to 30 dB SNR) to create a single auditory scene that was used as a training example for the neural network. The resulting waveform was then normalized to have an root-mean-square (r.m.s.) amplitude of 0.1. Each training example was passed through the cochlear model before being fed to the neural network.

Stimulus preprocessing: cochlear model

Training examples were preprocessed with a cochlear model to simulate the human auditory periphery. The output of the cochlear model is a time-frequency representation intended to represent the instantaneous mean firing rates in the auditory nerve. The cochlear model was chosen to approximate the time and frequency information in the human cochlea subject to practical constraints on the memory footprint of the model and the dataset. Cochleagrams were generated using a filter bank similar to that used in previous work from our laboratory[159]. However, the cochleagrams we used provided fine timing information to the neural network by passing recti-

fied subbands of the signal instead of the envelopes of the subbands. This came at the cost of substantially increasing the dimensionality of the input relative to an envelope-based cochleagram. The dimensionality was nonetheless considerably lower than what would have resulted from a spiking model of the auditory nerve, which would have been prohibitive given our hardware.

The waveforms for the left and right channels were first upsampled to 48 kHz, then separately passed through a bank of 39 bandpass filters. These filters were regularly spaced on an equivalent rectangular bandwidth scale[78] with bandwidths matched to those expected in a healthy human ear. Filter centre frequencies ranged from 45 to 16,975 Hz. Filters were zero-phase, with transfer functions in the frequency domain shaped like the positive portion of a cosine function. These filters perfectly tiled the frequency axis such that the summed squared response of all filters was flat and allowed for reconstruction of the signal in the covered frequency range. Filtering was performed by multiplication in the frequency domain, yielding a set of subbands. The subbands were then transformed with a power function (0.3 exponent) to simulate the outer hair cells' non-linear compression. The results were then half-wave rectified to simulate auditory nerve firing rates and were low-pass filtered with a cut-off frequency of 4 kHz to simulate the upper limit of phase locking in the auditory nerve[171], using a Kaiser-windowed sinc function with 4,097 taps. The results of the low-pass filtering were then downsampled to 8 kHz to reduce the dimensionality of the neural network input (without information loss because the Nyquist limit matched the low-pass filter cut-off frequency). Because the low-pass filtering and downsampling were applied to rectified filter outputs, the representation retained information at all audible frequencies, just with limits on fidelity that were approximately matched to those believed to be present in the ear. We note also that the input was not divided into 'frames' as are common in audio engineering applications, as these do not have an obvious analogue in biological auditory systems. All operations were performed in Python but made heavy use of the NumPy and SciPy library optimization to decrease processing time. Code to generate cochleagrams in this way is available on the McDermott laboratory webpage (http://mcdermottlab.mit.edu).

To minimize artificial onset cues at the beginning and end of the cochleagram that would not be available to a human listener in everyday listening conditions, we removed the first and last 0.35 s of the computed cochleagram and then randomly excerpted a 1-s segment from the remaining 1.3 s. The neural network thus received 1 s of input from the cochlear model, as a 39 $\times$ 8,000 $\times$ 2 tensor (39 frequency channels $\times$ 8,000 samples at 8 kHz $\times$ 2 ears).

For reasons of storage and implementation efficiency, the cochlear model stage was in practice implemented as follows, taking advantage of the linearity of the filter bank. First, the audio from each spatialized natural sound source and each spatialized background texture scene was run through the cochlear filter bank. Second, we excerpted a 1-s segment from the resulting subbands as described in the previous paragraph. Third, the two sets of subbands were stored in separate data structures. Fourth, during training, the subbands for a spatialized natural sound source and a spatialized background scene were loaded, scaled to achieve the desired SNR (sampled uniformly from 5 to 30 dB), summed and scaled to correspond to a waveform with r.m.s. amplitude of 0.1. The resulting subbands were then half-wave rectified, raised to the power of 0.3 to simulate cochlear compression, and downsampled to 8 kHz to simulate the upper limit of auditory nerve phase locking. This 'cochleagram' was the input to the neural networks.

2.4.2 Environment modification for unnatural training conditions

In each unnatural training condition, one aspect of the training environment was modified.

Anechoic environment

All echoes and reflections in this environment were removed. This was accomplished by setting the room material parameters for the walls, floor and ceiling to completely absorb all frequencies. This can be conceptualized as simulating a perfect anechoic

chamber.

Noiseless environment

In this environment, the background noise was removed by setting the SNR of the scene to 85 dB. No other changes were made.

Unnatural sound sources

In this environment, we replaced the natural sound sources with unnatural sounds consisting of repeating bandlimited noise bursts. For each 2-s sound source, we first generated a 200 ms 0.5 octave-wide noise burst with a 2 ms half-Hanning window at the onset and offset. We then repeated that noise burst separated by 200 ms of silence for the duration of the signal. The noise bursts in a given source signal always had the same centre frequency. The centre frequencies (the geometric mean of the upper and lower cut-offs) across the set of sounds were uniformly distributed on a log scale between 60 and 16.8 kHz.

2.4.3 Neural network models

The 39 $\times$ 8,000 $\times$ 2 cochleagram representation (representing 1 s of binaural audio) was passed to a CNN, which instantiated a feedforward, hierarchically organized set of linear and non-linear operations. The components of the CNNs were standard; they were chosen because they have been shown to be effective in a wide range of sensory classification tasks. In our CNNs, there were four different kinds of layer, each performing a distinct operation: (1) convolution with a set of filters, (2) a point-wise non-linearity, (3) batch normalization and (4) pooling. The first three types of layer always occurred in a fixed order (batch normalization, convolution and a point-wise non-linearity). We refer to a sequence of these three layers in this order as a 'block'. Each block was followed by either another block or a pooling layer. Each network ended with either one or two fully connected layers feeding into the final classification layer. Below, we define the operations of each type of layer.

Convolutional layer

A convolutional layer consists of a bank of K linear filters, each convolved with the input to produce K separate filter responses. Convolution performs the same operation at each point in the input, which in our case was the cochleagram. Convolution in time is natural for models of sensory systems as the input is a temporal sequence whose statistics are translation invariant. Convolution in frequency is less obviously natural, as translation invariance does not hold in frequency. However, approximate translation invariance holds locally in the frequency domain for many types of sound signal, and convolution in frequency is often present, implicitly or explicitly, in auditory models[51, 39]. Moreover, imposing convolution greatly reduces the number of parameters to be learned, and we have found that neural network models train more readily when convolution in frequency is used, suggesting that it is a useful form of model regularization.

The input to a convolutional layer is a three-dimensional array with shape n_{in}, m_{in}, d_{in} where n_{in} and m_{in} are the spectral and temporal dimensions of the input, respectively, and d_{in} is the number of filters. In the case of the first convolutional layer, $n_{in} = 36$ and $m_{in} = 8{,}000$, corresponding to the spectral and temporal dimensions of the cochleagram, and $d_{in} = 2$, corresponding to the left and right audio channels.

A convolution layer is defined by five parameters:

1. n_k, the height of the convolutional kernels (that is, their extent in the frequency dimension)

2. m_k, the width of the convolutional kernels (that is, their extent in the time dimension)

3. K, the number of different kernels

4. W, the kernel weights for each of the K kernels; this is an array of dimensions $(n_{in}, m_{in}, d_{in}, K)$

5. $\mathbf{B}$, the bias vector, of length K

For any input array X of shape n_{in}, m_{in}, d_{in}, the output of this convolutional layer is an array Y of shape $(n_{in},\ m_{in} - m_k + 1,\ K)$ (due to the boundary handling choices described below):

$$Y[i,j,k] = \mathbf{B}[k] + \sum_{n=-n_k/2, m=-m_k/2, d=0}^{n_k/2, m_k/2, d_{in}} W[n,m,d,k] \odot X[i+n, j+n, d]$$

where i ranges from $(1, \ldots, n_{in})$, j ranges $(1, \ldots, m_{in})$ and $\odot$ represents point-wise array multiplication.

Boundary handling via valid padding in time

There are several common choices for boundary handling during convolution operations. For the output of a convolution to be the same dimensionality as the input, the input signal is typically padded with zeros. This approach —often termed 'same' convolution — has the downside of creating an artificial onset in the data that would not be present in continuous audio in the natural world, and that might influence the behaviour of the model. To avoid this possibility, we used 'valid' convolution in the time dimension. This type of convolution only applies the filter at positions where every element of the kernel overlaps with the actual input. This eliminates artificial onsets at the start/end of the signal but means that the output of the convolution will be slightly smaller than its input, as the filters cannot be centred over the first and last positions in the input without having part of the filter not overlap with the input data.

We used 'same' convolution in the frequency dimension because the frequency dimension has lower and upper limits in the cochlea, such that boundary effects are less obviously inconsistent with biology. In addition, the frequency dimension was much smaller than the time dimension, such that it seemed advantageous to preserve channels at each convolution stage.

Point-wise non-linearity

If a neural network consists of only convolution layers, it can be mathematically reduced to a single matrix operation. A non-linearity is needed for the neural network to learn more complex functions. We used rectified linear units (a common choice in current deep neural networks) that operate point wise on every element in the input map according to a piecewise linear function:

$$f(x) = \begin{cases} x \; x \; > 0 \\ 0 \; else \end{cases}$$

Normalization layer

The normalization layer applied batch normalization[110] in a point-wise manner to the input map. Specifically, for a batch B of training examples, consisting of examples $\{X_1, \ldots, X_M\}$, with shape n_{in}, m_{in}, d_{in}, each example is normalized by the mean and variance of the batch:

$$\mu_B[n, m, d] = \frac{1}{M} \Sigma_{i=0}^{M} X_{in}[n, m, d]$$

$$\sigma_B^2[n, m, d] = \frac{1}{M} \Sigma_{i=0}^{M} (X_{in}[n, m, d] - \mu_B[n, m, d])^2$$

$$\hat{X}_i[n, m, d] = \frac{X_i[n,m,d] - \mu_B[n,m,d]}{\sqrt[2]{\sigma_B^2[n,m,d] + \epsilon}}$$

where $\hat{X}_i$ is the normalized three-dimensional matrix of the same shape as the input matrix and $\epsilon = 0.001$ to prevent division by zero.

Throughout training, the batch normalization layer maintains a cumulative mean and variance across all training examples, μ_{Total} and σ_{Total}^2. At test time $\hat{X}_i$ is calculated using μ_{Total} and σ_{Total}^2 in place of μ_B and σ_B^2.

Pooling layer

A pooling layer allows downstream layers to aggregate information across longer periods of time and wider bands of frequency. It downsamples its input by aggregating values across nearby time and frequency bins. We used max pooling, which is defined via four parameters:

1. P_h, the height of the pooling kernel
2. P_w, the width of the pooling kernel
3. s_h, the stride in the vertical dimension
4. s_w, the stride in the horizontal dimension

A pooling layer takes array X of shape (n_{in}, m_{in}, d_{in}) and returns array Y with shape $(n_{in}/sw, m_{in}/s_h, d_{in})$ according to:

$$Y(i, j, k) = max(N_{p_w p_h}\ (X, is_w, js_h, k))$$

where $N_{p_w p_h}$ *(X, i, j, k)* is a windowing function that takes a (P_w, P_h) excerpt of X of centred at *(i,j)* from filter k. The maximum is over all elements in the resulting excerpt.

Fully connected layer

A fully connected layer, also often called a dense layer, does not use the weight sharing found in convolutional layers, in which the same filter is applied to all positions within the input. Instead, each (input unit, output unit) pair has its own learned weight parameter and each output unit has its own bias parameter. Given input X with shape (n_{in}, m_{in}, d_{in}), a fully connected layer produces output Y with shape n_{out}. It does so in two steps:

1. Flattens the input dimensions, creating an input X_{flat} of shape $(n_{in} \times m_{in} \times d_{in})$

2. Multiplies X_{flat} by weight and bias matrices of shape $(n_{out}, n_{in} \times m_{in} \times d_{in})$ and n_{out}, respectively. This is implemented as:

$$Y(n_i) = \mathbf{B}(n_i) + \sum_{l=1}^{n_{in} \times m_{in} \times d_{in}} \mathrm{W}\ (\mathrm{n}_{in}, l)\ X_{flat}\ (l)\ ;\ n_i\ \epsilon\ \{1\ \ldots\ \ \mathrm{n}_{out}\}$$

where $\mathbf{B}(n_{out})$ is the bias vector, $W(n_{out,l})$ is the weight matrix and l ranges from 1 to $(n_{in} \times m_{in} \times d_{in})$ and indexes all positions in the flattened input matrix.

Softmax classifier

The final layer of every network was a classification layer, which consists of a fully connected layer where n_{out} is the number of class labels (in our case 504). The output of that fully connected layer was then passed through a normalized exponential (softmax) function. Together this was implemented as:

$$\mathbf{y}(i) = \frac{exp\ (\sum_{j=0}^{n_T}\ w_{ij}x_j)}{\sum_{k=0}^{n_{out}}\ exp\ (\sum_{j=0}^{n_T}\ w_{kj}x_j)}$$

The vector $\mathbf{y}$ sums to one and all entries are greater than zero. This is often interpreted as a vector of label probabilities conditioned on the input.

Dropout during training

For each new batch of training data, dropout was applied to all fully connected layers of a network. Dropout consisted of randomly choosing 50% of the weights in the layer and temporarily setting them to zero, thus effectively not allowing the network access to the information at those positions. The other 50% of the weights were scaled up such that the expected value of the sum over all inputs was unchanged. This was implemented as:

$$dropout\ (W_{i,j}) = \begin{cases} W_{i,j}\frac{1}{(1-0.5)} & j \notin \text{weights to drop} \\ 0 & j \in \text{weights to drop} \end{cases}$$

Dropout is common in neural network training and can be viewed as a form of model averaging where exponentially many models using different subsets of the input vector are being trained simultaneously[104]. During evaluation, dropout was turned off (and no weight scaling was performed) so that all weights were used.

2.4.4 Neural network optimization

Architecture search: overview

When neural networks are applied to a new problem it is common to use architectures that have previously produced good results on similar problems. However, most standard CNN architectures that operate on two-dimensional inputs have been designed for visual tasks and make assumptions based on the visual world. For example, most architectures assume that the units in the x and y dimension are equivalent, such that square filter kernels are a reasonable choice. However, in our problem the two input dimensions are not comparable (frequency versus time). Additionally, our input dimensionality was several orders of magnitude larger than standard visual stimuli (70,000 versus 1.1 million), even though some relevant features occur on the scale of a few samples. For example, an ITD of 400 μs (a typical value) corresponds to only a six-sample offset between channels. Given that our problem was distinct from many previous applications of standard neural network architectures, we performed an architecture search to find architectures that were well-suited to our task. First, we defined a space of architectures described by a small number of hyperparameters. Next, we defined discrete probability distributions for each hyperparameter. Last, we independently sampled from these hyperparameter distributions to generate architectures. We then trained each architecture for a brief period and selected the architectures that performed best on our task for further training.

Architecture search: distribution over hyperparameters

To search over architectures, we defined a space of possible architectures that were encoded via a set of hyperparameters. The space had the following constraints:

- There could be between three and eight pooling layers for any given network.
- A pooling layer was preceded by between one and three blocks. Each block consisted of batch normalization, followed by convolution, followed by a rectified linear layer
- The number of channels (filters) in the network was always 32 in the first convolutional layer and could either double or remain the same in each successive convolutional layer.
- The penultimate stage of each network consisted of one or two fully connected layers containing 512 units each. Each of these was followed by a dropout layer.
- The final stage of each network was always a Softmax Classifier with 504 output units, corresponding to the 504 locations the network could report.

We picked the pooling and convolutional kernel parameters at each layer by uniformly sampling from the lists of values in Extended Data Fig. 2-10. We chose these distributions to skew toward smaller values at deeper layers, approximately in line with the downsampling that resulted from pooling operations. Multiple copies of the same number increased the probability of that value being chosen for the kernel size. Note that differences between the time and frequency dimensions of the cochlear input motivate the use of filters that are not square.

Filter weight training

Throughout training, the parameters in each convolutional kernel and all weights from fully connected layers were iteratively adjusted to improve task accuracy via mini-batch stochastic gradient descent (SGD)[19]. Training was performed with 1.6 million sounds (100,000 training steps each with a batch of 16 training examples) generated by looping over the 500,000 foreground sounds and combining each with a randomly

selected background sound. Networks were assessed via a held-out set of 50,000 test stimuli created by looping over the 48,000 sound sources in the validation set in the same manner. We used a Softmax Cross-Entropy loss function. The trainable weights in the convolutional layers and fully connected layers were updated using the gradient of the loss function, computed using backpropagation.

Gradient checkpointing

The dimensionality of our input was sufficiently large (due to the high sampling rates needed to preserve the fine timing information in the simulated auditory periphery) as to preclude training neural networks using standard methodology. For example, consider training a network consisting of four pooling layers (2×1 kernel), each preceded by one block. If there are 32 convolutional filters in the first layer, and double the number of filters in each successive layer, this network would require approximately 80 GB of memory at peak usage, which exceeded the maximum memory of graphical processing units (GPUs) that were standard at the time of model training (available GPUs varied between 12 and 32 GB). We addressed this problem using a previously proposed solution called gradient checkpointing[35].

In the standard backpropagation algorithm, we must retain the output from each layer of a network in memory because it is needed to calculate gradients for each updatable parameter. The gradient checkpointing algorithm we used trades speed for lower memory usage by not retaining each layer's output during the forward pass, instead recomputing it a second time during the backward pass when gradients are computed. In the most extreme version, this would result in laboriously recomputing each layer starting with the original network input. Instead, the algorithm creates sparse, evenly spaced checkpoints throughout the network that save the output of selected layers. This strategy allows recomputation during backpropagation to start from one of these checkpoints, saving compute time. In practice, it also provides users with a parameter that allows them to select a speed/memory trade-off that will maximize speed subject to a network fitting onto the available GPU. We created checkpoints at every pooling layer and found it kept our memory use below the 16-GB

limit of the hardware we used for all networks in the architecture search.

Network architecture selection and training

We performed our architecture search on the Department of Energy's Summit Supercomputer at Oak Ridge National Laboratory. First, we randomly drew 1,500 architectures from our hyperparameter distribution. Next, we trained each architecture (that is, optimized the weights of the convolutional and fully connected layers) using mini-batch SGD for 15,000 steps, each with a batch size of 16, for a total of 240,000 unique training examples, randomly drawn from the training set described above. We then evaluated the performance of each architecture on left-out data. The length of this training period was determined by the job limits on Summit; however, it was long enough to see substantial reductions in the loss function for many networks. We considered the procedure adequate for architecture selection given that performance early in training is a good predictor of training performance late in training[57]. In total, this architecture search took 2.05 GPU years and 45.2 CPU years.

We selected the ten best-performing architectures. They varied significantly, ranging from four to six pooling layers. We then retrained these ten architectures until a point where performance on the withheld validation set began to decrease, evaluating every 25,000 iterations. This occurred at 100,000 iterations for the naturalistic, anechoic and noiseless training conditions and at 150,000 iterations for the unnatural sounds training condition. Model architectures and the trained weights for each model are available online in the associated codebase:

www.github.com/afrancl/BinauralLocalizationCNN.

2.4.5 Real-world evaluation

We tested the model in real-world conditions to verify generalization from the virtual training environment. We created a series of spatial recordings in an actual conference room (part of our laboratory space, with dimensions distinct from the rooms in our virtual training environment) and then presented those to the trained networks.

We also made recordings of the same source sounds and environment with a two-microphone array to test the importance of naturally induced binaural cues (from the ears, head and/or torso).

Sound sources

We used 100 sound sources in total: 50 sound sources were from our validation set of withheld environmental sounds, and the remaining 50 sound sources were taken from the GRID dataset of spoken sentences[9]. For the examples from the GRID dataset, we used five sentences from each of ten speakers (five male and five female). The model performed similarly for stimuli from the GRID dataset as for our validation set stimuli. All source signals were normalized to the same peak amplitude before the recordings were made.

Recording setup

We made the set of real-world evaluation recordings using a KEMAR head and torso simulator mannequin built by Knowles Electronics to replicate the shape and absorbency of a human head, upper body and pinna. The KEMAR mannequin contains a microphone in each ear, recording audio similar to that which a human would hear in natural conditions. The audio from these microphones was then passed through Etymotic Research preamplifiers designed for the KEMAR mannequin before being passed to a Zoom 8 USB to Audio Converter. Finally, it was passed to Audacity where the left and right channels were simultaneously recorded at 48 kHz.

We made recordings of all 100 sounds at every azimuth (relative to the KEMAR mannequin) from 0° to 360° in 30° increments. This led to 1,200 recordings in total. All source sounds were played 1.5 m from the vertical axis of the mannequin using a KRK ROKIT 7 speaker positioned at approximately 0° elevation. The audio was played using Audacity and converted to an analogue signal using a Zoom 8 USB to Audio Converter.

Recordings were made in our main laboratory space in building 46 on the MIT campus, in a room that was roughly 7 × 6 × 3 m. The room was filled with fur-

niture and shelves, and had multiple windows and doors (Fig. 2-1e). This setup was substantially different from any of the simulated rooms in the virtual training environment, in which all rooms were convex, empty and had smooth walls. During the recordings, there was low-level background noise from the ventilation system, the refrigerator and laboratory members talking in surrounding offices. For all recordings, the mannequin was seated in an office chair, with the head approximately 1 m from the ground.

Two-microphone array baseline

We made a second set of recordings using the same sound sources, room and recording equipment as above, but with the KEMAR mannequin replaced with a two-microphone array consisting of two Beyerdynamic MM-1 Omnidirectional Microphones separated by 15 cm (the same distance separating the two microphones in the mannequin ears). The microphone array was also elevated approximately 1 m from the floor using a microphone stand (Extended Data Fig. 2-14a).

Baseline algorithms

We evaluated our trained neural networks against a variety of baseline algorithms. These comprised: steered-response power phase transform (SRP)[55], multiple signal classification (MUSIC)[193], the coherent signal-subspace method (CSSM)[223], weighted average of signal subspaces (WAVES)[54], test of orthogonality of projected subspaces (TOPS)[238] and the WavLoc neural network[217]. With the exception of the WavLoc model, in each case we used the previously validated and published algorithm implementations in Pyroomacoustics[192]. For the WavLoc model, we used a reference GitHub implementation and confirmed that we could reproduce the results of the original paper[217] before testing with our KEMAR mannequin recordings. We also created a baseline model trained using a simulation of the two-microphone array described in the previous section within the virtual training environment (the same ten neural network architectures used for our primary model were trained to localize sounds using audio recorded from simulated a two-microphone array).

The results shown in Extended Data Fig. 2-14b,c for the baselines (aside from our two-microphone array baseline neural network model) all plot localization of the KEMAR mannequin recordings. We found empirically that the baseline methods performed better for the KEMAR recordings than for the two-microphone array recordings, presumably because the mannequin head increases the effective distance between the microphones. The baseline algorithms require previous knowledge of the intermicrophone distance. To make the baselines as strong as possible relative to our method, we searched over all distances shorter than 50 cm and found that an assumed distance of 26 cm yielded the best performance. We then evaluated the baselines at that assumed distance. This optimal assumed distance is greater than the actual intermicrophone distance of 15 cm, consistent with the idea that the mannequin head increases the effective distance between microphones.

Comparison with human listeners

To provide an example of free-field human sound localization, Fig. 2-1f plots the results of an experiment by Yost and colleagues[240]. In that experiment, humans were presented with noise bursts (low-pass filtered white noise with a cut-off of 6 kHz, 200 ms in duration, with 20 ms cosine onset and offset ramps) played from one of 11 speakers in an anechoic chamber. The speakers were spaced every 15°, with the array centred on the midline. Speakers were visible to participants. Participants indicated the speaker from which the sound was played by entering a number corresponding to the speaker. Results are shown for 45 participants (34 female), ages 21–49. Because the human experiment was restricted to speakers in front of the participants, for ease of comparison Fig. 2-1g plots model results after front–back folding of actual and judged positions (Fig. 2-1h shows model results without front–back folding). Figure 1f–h display kernel density estimates of the response distributions, generated using the seaborn statistical data visualization library.

2.4.6 Psychophysical evaluation of model

Overview

We simulated a suite of classic psychoacoustic experiments on the ten trained neural networks, using the same stimuli for each network. We then calculated the mean response across networks for each experimental condition and calculated error bars by bootstrapping across the ten networks. This approach can be interpreted as marginalizing out uncertainty over architectures in a situation in which there is no single obviously optimal architecture (and where the space of architectures is so large that it is probably not possible to find the optimum even if it exists). Moreover, recent work suggests that internal representations across different networks trained on the same task can vary considerably[162], so this approach aided in mitigating the individual idiosyncrasies of any given network. Te approach could also be viewed as treating every network as an individual experimental participant, calculating means and error bars as one would in a standard human psychophysical experiment.

In each experiment, stimuli were run through our cochlear model and passed to each of the networks, whose localization responses were recorded for each stimulus. Stimuli were generated as 2 s sound signals, normalized to have an r.m.s. amplitude of 0.1. The output of the cochlear model was then cropped to 1 s (by excerpting the middle 1 s), which provided the input to the networks.

Front–back folding

For experiments in which human participants judged locations within the frontal hemifield, we front–back folded the model responses to enable a fair comparison. This consisted of treating each model response in the rear hemifield as though it was a response in the corresponding front hemifield. For example, the 10° and 170° azimuthal positions were considered equivalent.

2.4.7 Sensitivity to ITDs and ILDs: stimuli

We reproduced the experimental stimuli from ref. [151], in which ITDs and ILDs were added to 3D spatially rendered sounds. In the original experiment, participants stood in a dark anechoic room and were played spatially rendered stimuli with modified ITDs or ILDs via a set of headphones. After each stimulus presentation, participants oriented their head towards the perceived location of the stimulus and pressed a button. The experiment included 13 participants (five male) ranging in age from 18–35 years old.

Stimulus generation for the model experiment was identical to that in the original experiments apart from using our acoustic simulator to render the sounds. First, we generated high- and low-pass noise bursts with passbands of 4–16 and 0.5–2 kHz, respectively (44.1 kHz sampling rate). Each noise burst was 100 ms long with a 1-ms squared-cosine ramp at the beginning and end of the stimulus. We randomly jittered the starting time of the noise burst by padding the signal to 2,000 ms in total length, constrained such that the entire noise burst was contained in the middle second of the 2-s audio signal (the noise onset was uniformly distributed subject to this constraint). These signals were then rendered at 0° elevation, with azimuth varied from 0 to 355° (in 5° steps) for a total of 72 locations. All signals were rendered using our virtual acoustic simulator in an anechoic environment without any background noise.

Next, we created versions of each signal with an added ITD or ILD bias. ITD biases were ±300 and ±600 µs and ILD biases were ±10 and ±20 dB (Fig. 2-2a). As in the original publication[151], we prevented presentation of stimuli outside the physiological range by restricting the 400 µs/10 dB biases to signals rendered less than 40° away from the midline and restricting the 600 µs/20 dB biases to signals rendered less than 20° away from the midline. In total, there were four stimulus sets (2 passbands × 2 types of bias) of 266 stimuli (72 locations with no bias, 52 locations at ±medium bias, 45 locations ± large bias). We replicated the above process 20 times with different exemplars of bandpass noise, increasing each stimulus set size to 5,320 (20 exemplars of 266 stimuli).

Sensitivity to ITDs and ILDs: analysis

We measured the perceptual bias induced by the added ITD or ILD bias in the same manner as the published analysis of human listeners[151].

We first calculated the naturally occurring ITD and ILD for each sound source position (varying in azimuth, at 0° elevation) from the HRTFs used to train our networks. For ITDs, we ran the HRTFs for a source position through our cochlear model and found the ITD by cross-correlating the cochlear channels whose centre frequency was closest to 600, 700 and 800 Hz and taking the median ITD from the three channels. For ILDs, we computed power spectral density estimates via Welch's method (29 samples per window, 50% overlap, Hamming windowed) for each of the two HRTFs for a source position and integrated across frequencies in the stimulus passband. We expressed the ILD as the ratio between the energy in the left and right channel in decibels, with positive values corresponding to more power in the right ear. This set of natural ILDs and ITDs allowed us to map the judged position onto a corresponding ITD/ILD.

For each stimulus with added ITD, we used the response mapping described above to calculate the ITD of the judged source position. Next, we calculated the ITD for the judged position of the unaltered stimulus using the same response mapping. The perceptual effect of the added ITD was calculated as the difference between these two ITD values, quantifying (in microseconds) how much the added stimulus bias changed the response of the model. The results graphs plot the added stimulus bias on the x axis and the resulting response bias on the y axis. The slope of the best-fitting regression line (the 'bias weight' shown in the subplots of Fig. 2-2c,d) provides a unitless measure of the extent to which the added bias affects the judged position. We repeated an analogous process for ILD bias using the natural ILD response mapping, yielding the bias in decibels. The graphs in Fig. 2-2d plot the mean response across the ten networks with standard error of the mean (s.e.m.) computed via bootstrap over networks.

Azimuthal localization of broadband sounds: stimuli

We reproduced the stimulus generation from ref. [229]. In the original experiment, participants were played six broadband white noise bursts, with three noise bursts (15 ms in duration, 5-ms cosine ramps, repeated at 10 Hz) played from a reference speaker followed by three noise bursts played from one of two target speakers, located 15° to the left or right of the reference speaker. The reference speaker position ranged from -97.5° to +97.5° azimuth in 15° intervals. Participants reported whether the last three noise bursts were played to the left or the right of the reference speaker, and performance was expressed as *d'*. 18 speakers were arranged at 15° intervals from -127.5° to +127.5° azimuth simultaneously played white noise during all trials, producing spatially diffuse background noise that served to bring performance below ceiling. The SNR of the stimulus was set individually for each participant. To determine the SNR, stimuli were played from the speakers at +90° or -90° azimuth and participants judged if each stimulus was to their right or left. The experiment included 16 participants between the ages of 18 and 35 years old.

We measured network localization performance using the same stimuli as in the original paper, but for simplicity rendered the stimulus at a single location and measured performance with an absolute, instead of relative, localization task. The stimuli presented to the networks consisted of three pulses of broadband white noise. Each noise pulse was 15 ms in duration and repeated at 10Hz. A 5-ms cosine ramp was applied to the beginning and end of each pulse. We generated 100 exemplars of this stimulus using different samples of white noise (44.1 kHz sampling rate). The stimuli were zero-padded to 2,000 ms in length, with the temporal offset of the three-burst sequence randomly sampled from a uniform distribution such that all three noise bursts were fully contained in the middle second of audio. We then rendered all 100 stimuli at 0° elevation and azimuthal positions ranging from 0° to 355° in 5° steps. All stimuli were rendered in an anechoic environment without any background noise using our virtual acoustic simulator. This led to 7,200 stimuli in total (100 exemplars at each of 72 locations). The stimuli were presented in spatially diffuse background noise,

generated by presenting white noise from 19 positions at 15° intervals from -135° to +135°. The SNR was set for each network individually by measuring its left/right accuracy on stimuli rendered at +90 or -90 degrees at a range of SNRs spaced in 1 dB increments, and then selecting the highest SNR at which the network performed below 95% accuracy. The SNRs selected in this way ranged from -8 dB to -14 dB depending on the network.

Azimuthal localization of broadband sounds: analysis

Because human participants in the analogous experiment judged relative position in the frontal hemifield, before calculating the model's accuracy we eliminated front–back confusions by mirroring model responses of each stimulus across the coronal plane. We then calculated the difference in degrees between the rendered azimuthal position and the position judged by the model. We calculated the mean absolute error for each rendered azimuth for each network. The graph in Fig. 2-3c plots the mean error across networks. Error bars are s.e.m., bootstrapped over networks.

Integration across frequency: stimuli

We reproduced stimuli from ref. [241]. In the original experiment, human participants were played a single noise burst, varying in bandwidth and centre frequency, from one of eight speakers spaced 15° in azimuth. Participants judged which speaker the noise burst was played from. The experimenters then calculated the localization error in degrees for each bandwidth and centre frequency condition. The experiment included 33 participants (26 female) between the ages of 18 and 36 years old.

The stimuli varied in bandwidth (pure tones, and noise bursts with bandwidths of 1/20, 1/10, 1/6, 1/3, 1 and 2 octaves wide; all with 44.1 kHz sampling rate). All sounds were 200 ms long with a 20-ms squared-cosine ramp at the beginning and end of the sound. All pure tones had random phase. All other sounds were bandpass-filtered white noise with the geometric mean of the passband cut-offs set to 250, 2,000 or 4,000 Hz (as in the original paper[241]).

For the model experiment, the stimuli were zero-padded to 2,000 ms in length, with

the temporal offset of the noise burst randomly sampled from a uniform distribution such that the noise burst was fully contained in the middle second of audio. We generated 30 exemplars of each bandwidth–frequency pair using different exemplars of white noise (or of random phase for the pure tone stimuli). Next, we rendered all stimuli at 0° elevation and azimuthal positions ranging from 0° to 355° in 5° steps. All stimuli were rendered in an anechoic environment without any background noise using the virtual acoustic simulator. This led to 45,360 stimuli in total (30 exemplars $\times$ 72 positions $\times$ 3 centre frequencies $\times$ 7 bandwidths).

Integration across frequency: analysis

Because human participants in the original experiment judged position in the frontal hemifield, before calculating the model's accuracy, we again eliminated front–back confusions by mirroring model responses of each stimulus across the coronal plane. We then calculated the difference in degrees between the rendered azimuthal position and the azimuthal position judged by the model. For each network, we calculated the r.m.s. error for each bandwidth. The graph in Fig. 2-3f plots the mean of this quantity across networks. Error bars are s.e.m., bootstrapped over networks.

Use of ear-specific cues to elevation: stimuli

We simulated a change of ears for our networks, analogous to the ear mould manipulation in ref. [105]). In the original experiment in ref. [105], participants sat in a dark anechoic room and were played broadband white noise bursts from a speaker on a robotic arm that moved ±30° in azimuth and elevation. Participants reported the location of each noise burst by saccading to the perceived location. After collecting a baseline set of measurements, participants were fitted with plastic ear moulds (Fig. 2-4a), which modified the location-dependent filtering of their pinnae. Participants then performed the same localization task a second time. The experimenters plotted the mean judged location for each actual location before and after fitting participants with the plastic ear moulds (Fig. 2-4b,c). The experiment included four participants between the ages of 22 and 44 years old.

For the model experiment, instead of ear moulds we substituted HRTFs from the CIPIC dataset[2]. The CIPIC dataset contains 45 sets of HRTFs, each of which is sampled at azimuths from 80 to +80 in 25 steps of varying size, and elevations from 0 to 360 in 50 steps of varying size. For the sound sources to be localized, we generated 500 ms broadband (0.2 – 20 kHz) noise bursts sampled at 44.1 kHz (as in ref. [105]). We then zero- padded these sounds to 2,000 ms, with the temporal offset of the noise burst randomly sampled from a uniform distribution such that it was fully contained in the middle second of audio. We generated 20 such exemplars using different samples of white noise. We then rendered each stimulus at ±20 and ±10° azimuths, and 0°, 10°, 20° and 30° elevation for all 45 sets of HRTFs as well as the standard set of HRTFs (that is, the one used for training the model). This led to a total of 14,720 stimuli (46 HRTFs × 4 azimuths × 4 elevations). The rendered locations were slightly different from those used in ref. [105] as we were constrained by the locations that were measured for the CIPIC dataset.

Use of ear-specific cues to elevation: analysis

The results graphs for this experiment (Fig. 2-4b–e) plot the judged source position for each of a set of rendered source positions, either for humans (Fig. 2-4b,c) or the model (Fig. 2-4d,e). For the model results, we first calculated the mean judged position for each network for all stimuli rendered at each source position. The graphs plot the mean of this quantity across networks. Error bars are the s.e.m., bootstrapped over networks. In Fig. 2-4d we plot model responses for stimuli rendered using the HRTFs used during network training. In Fig. 2-4e we plot the average model responses for stimuli rendered with 45 sets of HRTFs from the CIPIC database (none of which were used during network training). In Fig. 2-4f,g we plot the results separately for each alternative set of HRTFs, averaged across elevation or azimuth. The thickest bolded line denotes the mean performance across all HRTFs, and thinner bolded lines denote HRTFs at the 5th, 25th, 75th and 95th percentiles order by error. Each line plots the mean over the ten networks.

Limited spectral resolution of elevation cues: stimuli

We ran a modified version of the spectral smoothing experiment in ref. [137] on our model using the training HRTFs. The original experiment[137] measured the effect of spectral detail on human sound localization. The experimenters first measured HRTFs for each participant. Participants then sat in an anechoic chamber and were played broadband white noise bursts presented in one of two ways. The noise burst was either played directly from a speaker in the room or virtually rendered at the position of the speaker using the participant's HRTF and played from a set of open-backed earphones worn by the participant. The experimenters manipulated the spectral detail of the HRTFs as described below. On each trial, two noise bursts (one for each of the two presentation methods) were played in random order and participants judged which of the two noise bursts were played via earphones. In practice, this judgement was performed by noticing changes in the apparent sound position that occurred when the HRTFs were sufficiently degraded. The results of the experiment were expressed as the accuracy in discriminating between the two modes of presentation as a function of the amount of spectral detail removed (Fig. 2-4i). The experiment included four participants.

The HRTF is obtained from the Fourier transform of the HRIR, and thus can be expressed as:

$$\mathrm{H[k]} = \Sigma_{n=0}^{N-1} X_n e - \frac{i2\pi nk}{N}$$

where x is the HRIR, N is the number of samples in the HRTF and k = [0,N − 1]. To smooth the HRTF, we first compute the log-magnitude of $H[k]$. This log-magnitude HRTF can be decomposed into frequency components via the discrete cosine transform:

$$\log \mid H[k] \mid = \Sigma_{n=0}^{M} \; C(n) \; cos(2\pi nk/N)$$

where C(n) is the nth cosine coefficient of log|H[k]| and M = N/2.

As in the original experiment[137], we smoothed the HRTF by reconstructing it with M < N/2. In the most extreme case where M = 0, the magnitude spectrum was perfectly flat at the average value of the HRTF. Increasing M increases the number of cosines used for reconstruction, leading to more spectral detail (Fig. 2-4h). After smoothing, we calculated the minimum phase filter from the smoothed magnitude spectrum, adding a frequency-independent time delay consistent with the original HRIR. Our HRIRs consisted of 512 time points, corresponding to a maximum of 256 points in its cosine series.

We repeated this smoothing process for each left and right HRTF at each spatial position. We then generated 20 exemplars of broadband white noise (0.2 – 20 kHz, 2,000 ms in length) with a 10 ms cosine ramp at the beginning and end of the signal. The exemplars were rendered at elevations between 0° and 60° in 10° steps and a set of azimuths ranging from 0° to 355° , the spacing of which varied with elevation due to the locations in the original set of HRTFs. This yielded 74,340 stimuli (9 smoothed sets of HRTFS x 20 exemplars x 413 locations).

Limited spectral resolution of elevation cues: analysis

For the model, the effect of the smoothing was measured as the average absolute difference in degrees between the judged position and the rendered position for each stimulus. Figure 4j plots the mean error across networks for each smoothed set of HRTFs. Error bars are s.e.m., bootstrapped over networks. Figure 4k,l plot the mean judged azimuth (left) and elevation (right) versus the actual rendered azimuth and elevation, plotted separately for each smoothing level. Each line is the mean response pooled across networks. Error bars are shown as bands around the line and show s.e.m., bootstrapped over networks.

Dependence on high-frequency spectral cues to elevation: stimuli

In the original experiment[98], human participants were played high- and low-pass noise bursts. The high-pass cut-off frequencies took on one of six values: 3.8, 5.8, 7.5,

10.0, 13.2 and 15.3 kHz; low-pass cut-off frequencies took on one of seven values: 3.9, 6.0, 8.0, 10.3, 12.0, 14.5 and 16.0 kHz (imposed with an analogue Cauer–Chebychev filter). The sampling rate was 44.1 kHz. Each noise burst was 1,000 ms in duration, with a 5-ms squared-cosine ramp at the beginning and end. Each stimulus was presented from one of nine speakers spaced along the midline at 30° increments in elevation from 30° to 210°, with 0° being frontal horizontal. Participants judged which speaker the noise burst was played from, indicating their judgement with a keypress. The results graph (Fig. 2-4n) plots the proportion correct for each condition (error bars were not plotted in the original publication, and the raw data were no longer available). The experiment included ten participants.

Stimuli for the model experiment were similar to those from the human experiment apart from being presented from a subset of elevations used in the human experiment due to the constraints of the HRTF set in the model. We generated 50 exemplars of each cut-off frequency used in the human experiment, each with a different exemplar of white noise. Filtering was performed in the frequency domain by setting Fourier coefficients beyond the cut-off to zero. We then rendered all 650 noise bursts at one of six locations along the midline: 0°, 30°, 60°, 120°, 150° and 180°, with 0° being frontal horizontal. This led to 3,900 stimuli in total (650 noise bursts at each of six locations). All stimuli were rendered in an anechoic environment without any background noise using the virtual acoustic simulator.

Dependence on high-frequency spectral cues to elevation: analysis

We determined the model's response in the experiment to be the elevation in the stimulus set that was closest to the elevation of the softmax class bin with the maximum activation. Figure 4o plots the proportion of correct responses for each high-pass and low-pass cut-off frequency, averaged across the ten networks. Error bars are s.e.m., bootstrapped over networks.

Precedence effect: stimuli

For the basic demo of the precedence effect (Fig. 2-5b) we generated a click consisting of a single sample at +1 surrounded by zeros. We then rendered that click at ±45 azimuth and 0° elevation in an anechoic room without background noise using the virtual acoustic simulator. We added these two rendered signals together, temporally offsetting the 45° click behind the 45° click by an amount ranging from 1 to 50 ms. We then zero-padded the signal to 2,000 ms, sampled it at 44.1 kHz and randomly varied the temporal offset of the click sequence, constrained such that all non-zero samples occurred in the middle second of the stimulus. For each delay value, we created 100 exemplars with different start times.

To quantitatively compare the precedence effect in our model with that in human participants, we reproduced the stimuli from ref. [144]. In the original experiment, participants were played two broadband pink noise bursts from two different locations. The leading noise burst came from one of six locations (±20°, ±40° or ±60°) and the lagging noise burst came from 0°. The lagging noise burst was delayed relative to the leading noise burst by 5, 10, 25, 50 or 100 ms. For each pair of noise bursts, participants reported whether they perceived one or two sounds and the judged location for each perceived sound. The experimenters then calculated the mean localization error separately for the leading and lagging click for each time delay (Fig. 2-5c). The experiment included ten participants (all female) between the ages of 19 and 26 years old.

For both the human and model experiments, stimuli were 25-ms pink noise bursts, sampled at 44.1 kHz, with a 2-ms cosine ramp at the beginning and end of the burst. For the model experiment, we generated two stimuli for each pair of noise burst positions, one where the 0° noise burst was the lead click and another where it was the lag click. For each delay value, location and burst order, we created 100 exemplars with different start times. This was achieved by zero-padding the signal to 2,000 ms and randomly varying the temporal offset, constrained such that all non-zero samples occurred in the middle second of the stimulus.

Precedence effect: analysis

Because human experiments on the precedence effect typically query participants about positions in the frontal hemifield, we corrected for front–back confusions in the analysis of both the precedence effect demo and the Litovsky and Godar experiment by mirroring model responses of each stimulus across the coronal plane. Figure 5b plots the mean judged position at each interclick delay, averaged across the means of the ten individual networks. Error bars are s.e.m., bootstrapped over networks.

To generate Fig. 2-5d (plotting the results of the model version of the Litovsky and Godar experiment) we calculated errors for each stimulus between the model's judged position and the positions of the leading and lagging clicks. We calculated the average lead click error and average lag click error for each network at each delay. Figure 5d plots the mean of these quantities across the ten networks. Error bars are s.e.m., bootstrapped over networks.

Multi-source localization: stimuli

We reproduced stimuli from the original experiment[244], in which human participants were played between one and eight concurrent speech stimuli. Each stimulus was played from a different location (out of 12 possible, evenly spaced in azimuth). Participants judged the number of stimuli as well as the locations at which stimuli were presented in each trial. The experimenters then plotted the mean number of sources perceived versus the actual number of sources presented (Fig. 2-6b) and localization accuracy (proportion correct) versus the number of sources presented (Fig. 2-6d). The experiment included eight normal-hearing participants.

Stimuli were 10 s in duration and consisted of a concatenation of ten 1-s recordings of a person saying the name of a country (randomly drawn without replacement from a list of 24 countries). Each stimulus used recordings from a single talker (out of 12 possible talkers, six were female). Each stimulus was presented from one of 12 speakers at 0° elevation, spaced 30° apart in azimuth (Fig. 2-6a). On each trial, between one and eight stimuli were simultaneously presented, each spoken by a different talker

and presented from a different speaker.

The model experiment used the same 1-s recordings used in the original experiment (kindly provided by W. Yost), but presented a single 1-s recording (of a speaker saying a single country name, rather than the sequence of ten such recordings used in the human experiment) at each location, to accommodate the 1-s input length of the model. For each number of sources (one to eight) we computed each possible spatial source configuration and rendered 20 scenes for each configuration, randomly sampling talkers and country names for each trial (without replacement). All stimuli were rendered in an anechoic environment without any background noise using the virtual acoustic simulator. This led to 75,920 stimuli in total (20 exemplars in each of 3,796 spatial configurations).

Multi-source localization: output layer fine-tuning

To enable the model to perform the multi-source localization experiment, we altered the softmax output layer, which was designed to report one source at a time. We replaced the softmax function with independent sigmoid functions for each output unit. This allowed the model to independently report the probability of a source at each location. To allow our model to use this new output representation, we retrained this new final model stage. We froze all weights in each network except for those in the final fully connected layer, which we then trained using gradient descent for 10,000 steps ('fine-tuning'). The fine-tuning used a dataset consisting of auditory scenes generated and rendered in the same manner as the original training data (as described in Training data generation above), with two exceptions. First, each scene contained between one and eight natural sounds, each rendered at a different location. Second, the scenes did not contain background noise. This process was repeated for each network to allow the model to use its features on the multi-source localization task.

To measure accuracy after fine-tuning, we created a multi-source validation set using the natural sounds from the main model validation set. We measured the area under the curve for the receiver operator characteristic curve over the entire multi-

source validation set. The average area under the curve across fine-tuned networks after fine-tuning was 0.73.

Multi-source localization: analysis

The output layer of the multi-source model contained a unit for each location, as for the main single-source localization model, but differed in that the unit activation represented the judged probability that a source was present at that location. To enable the model to perform the multi-source experiment, we implemented a decision rule whereby the model would determine a source to be present at a location if the probability for that location exceeded a criterion. We set this criterion such that the model would correctly estimate the number of sources when a single source was present. We found empirically that the absolute activations resulting from the sigmoid output units varied considerably across sounds, presumably because the networks were trained with a softmax output layer that normalizes the output activations (which was no longer present in the multi-source decision layer). We thus adopted a criterion that was a proportion of the maximum probability across all output units and found that this yielded results that were stable across stimuli. Using all the experiment stimuli containing one source, we successively lowered the criterion from 1, each time running through the full set of scenes and estimating confidence intervals on the average predicted number of sources, until the 95% confidence interval for the predicted number of sources (after front–back folding) included 1. This yielded a decision criterion of 0.09 times the maximum probability across all output unit activations for the stimulus.

To perform a trial in the experiment, we first selected the model's location bins whose probability exceeded the criterion of 0.09 times the maximum probability across all output unit activations for the stimulus. We then mapped these locations to the 12 possible speaker locations in the experiment (for each output location bin, we selected the speaker location closest in azimuth). The number of sources was calculated as the number of these 12 speaker locations to which a localized source was mapped (Fig. 2-6c). The proportion correct was calculated as the hit rate: the fraction of the

12 speaker locations at which the model correctly judged there to be a source (Fig. 2-6e).

2.4.8 Evaluation of models trained in unnatural conditions

Once trained, each alternative model was run on each of the psychophysical experiments. The exception was the multi-source localization experiment, which was omitted because it was not clear how to incorporate the background noise training manipulation into the fine-tuning of the model output layer. The psychophysical experiments were identical for all training conditions.

2.4.9 Analysis of results of unnatural training conditions

Human–model dissimilarity

We assessed the effect of training condition on model behaviour by quantifying the extent of the dissimilarity between the model psychophysical results and the human results. For each results graph, we measured human–model dissimilarity as the r.m.s. error between corresponding y axis values in the human and model experiments. To compare results between experiments, before measuring this error, we min–max normalized the y axis to range from 0 to 1. For experiments with the same y axis for human and model results, we normalized the model and human data together (that is, taking the min and max values from the pooled results). For experiments where the y axes were different for human and model results (because the tasks were different, as in Figs. 3b,c and 4i,j), we normalized the data individually for human and model results.

The one exception was the ear alteration experiment (Fig. 2-4a–g), in which the result of primary interest was the change in judged location relative to the rendered location, and for which the locations were different in the human and model experiments (due to constraints of the HRTF sets that we used). To measure the human–model dissimilarity for this experiment, we calculated the error between the judged and rendered location for each point on the graph, for humans and the model.

We then calculated human–model dissimilarity between these error values, treating the two grids of locations as equivalent. This approach would fail to capture some patterns of errors but was sufficient to capture the main effects of preserved azimuthal localization along with the collapse of elevation localization.

This procedure yielded a dissimilarity measure that varied between zero and one for each experiment, where zero represents a perfect fit to the human results. For Fig. 2-7b, we then calculated the mean of this dissimilarity measure over the seven experiments. To generate error bars, we bootstrapped across the ten networks, recalculating all results graphs and the corresponding mean normalized error for each bootstrap sample. Error bars in Fig. 2-7b plot the s.d. of this distribution (that is, the standard error of the mean). Additionally, we plotted the mean normalized error individually for each of the ten networks (Extended Data Fig. 2-15).

Between-human dissimilarity

The dissimilarity that would result between different samples of human participants puts a lower bound on human–model dissimilarity, and would thus be useful to compare to the dissimilarity plotted in Fig. 2-7b. This between-human dissimilarity could be estimated using data from the original individual human participants. Unfortunately, the individual participant data were unavailable for nearly all of the experiments that we modelled, many of which were conducted several decades ago. Instead, we used the error bars in the published results figures to simulate different samples of human participants given the variability observed in the original experiments. Error bars were provided for only some of the original experiments (the exceptions being the experiments in Figs. 2 and 4n), so we were only able to estimate the between-human dissimilarity for this subset. We then compared the estimated between-human dissimilarity to the human–model dissimilarity for the same subset of experiments (Extended Data Fig. 2-16).

We assumed that human data for each experimental condition were independently normally distributed with a mean and variance given by the mean and error bars for that condition. Depending on the experiment, the error bars in the original graphs

plotted the standard deviation, the s.e.m., or the 95% confidence interval of the data. In each case we estimated the variance from the mean of the upper and lower error bar (for s.d. the square of the error bar; for s.e.m.: variance = $(\sqrt{N}\times\text{s.e.m.})^2$; for 95% confidence interval: variance = $(\sqrt{N} \times (\text{error bar width})/1.96)^2$, where N is the number of participants). To obtain behavioural data for one simulated human participant, we sampled from the Gaussian distribution for each condition. We sampled data for the number of participants run in the original experiment, and obtained mean results for this set of simulated participants. We then calculated the r.m.s. error (described in the previous section) between the simulated human data and actual human data (normalized as described in the previous section for the human– model dissimilarity). We repeated this process 10,000 times for each experiment, yielding a distribution of dissimilarities for each experiment. We then calculated the mean dissimilarity across experiments and samples. Extended Data Fig. 2-16 plots this estimated between-human dissimilarity (with confidence intervals obtained from the distribution of between-human dissimilarity) alongside the human– model dissimilarity for the same subset of experiments.

Models with internal noise

To test for the possibility that the noiseless training environments might have had effects that were specific to the lack of internal noise in the cochlear model used as input to our networks, we trained an alternative model with internal noise added to the output of the cochlear stage. This alternative model was identical to the main model used throughout the paper except that independent Gaussian noise was added to each frequency channel before the rectification stage of the cochlear model. The noise was sampled from a standard normal distribution and then scaled so that its power was on average 60.6 dB below the average power in the subbands of the input signal (intended to produce noise at 9.4 dB SPL assuming sources at 70 dB SPL[20]). In practice, we pregenerated 50,000 noise arrays, sampled one at random on each trial, and added it to the output of the cochlear filters at the desired SNR.

Cohen's d

To assess how training conditions affected individual psychophysical effects, we measured the effect size of the difference between human–model dissimilarity in the naturalistic and unnatural training conditions for each psychophysical effect. Specifically, we measured Cohen's d for each experiment:

$$d = \frac{\mu_{unnatural} - \mu_{naturalistic}}{s}$$

$$s = \sqrt{\frac{\sigma^2_{unnatural} + \sigma^2_{naturalistic}}{2}}$$

where μ and σ^2 are the mean and variance, respectively, of the human–model dissimilarity across our ten networks for the naturalistic or unnatural training condition. We calculated error bars on Cohen's d by bootstrapping across the ten networks, computing the effect size for each bootstrap sample. Figure 7c plots the mean and s.e.m. of this distribution.

2.4.10 Instrument note localization

Instrument note localization: stimuli

To assess the ability of the model to predict localization behaviour for natural sounds, we rendered a set of instruments playing notes at different spatial positions. Instruments were sourced from the Nsynth Dataset[59], which contains a large number of musical notes from a wide variety of instruments. We used the validation set component of the dataset, which contained 12,678 notes sampled from 53 instruments. For

each note, room in our virtual environment, and listener location within each room, we randomly rendered each of the 72 possible azimuthal positions (0° elevation, 0°–355° azimuth in 5° steps) with a probability $P = \frac{0.025 \times no.\ of\ locations\ in\ smallest\ room}{no.\ of\ locations\ in\ current\ room}$. We used a base probability of 2.5% to limit the overall size of the test set and normalized by the number of locations in the current room so that each room was represented equally in the test set. This yielded a total of 456,580 stimuli.

Instrument note localization: analysis

We anticipated performing a human instrument note localization experiment in an environment with speakers in the frontal hemifield, so we corrected for front–back confusions by mirroring model responses of each stimulus across the coronal plane. Different instruments in the dataset contained different subsets of pitches. To ensure that differences in localization accuracy would not be driven solely by the instrument's pitch range, we limited analysis to instruments for which the dataset contained all notes in the octave around middle C (MIDI note 55 to 66) and performed all analysis on notes in that range. This yielded 43 instruments and 1,860 unique notes. We calculated the mean localization error for each network judgement by calculating the absolute difference, in degrees, between the judged and rendered azimuthal location. We then averaged the error across networks and calculated the mean error for each of the 1,860 remaining notes from the original dataset. We plotted the distributions of the mean error over notes for each instrument (8 A) using letter-value plots[106].

To characterize the density of the spectrum we computed its spectral flatness. We first estimated the power spectrum x(n) using Welch's method (window size of 2,000 samples, 50% overlap). The spectral flatness was computed for each note of each instrument as:

$$SpectralFlatness = \frac{\sqrt[n]{\prod_{n=0}^{N-1} x(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)}$$

We averaged the spectral flatness across all notes of an instrument and then computed the Spearman correlation of this measure with the network's mean accuracy for that instrument.

2.4.11 Statistics

Real-world localization

For plots comparing real-world localization across models (Extended Data Fig. 2-14b,c), error bars are s.e.m., bootstrapped over stimuli (because there was only one version of the baseline models).

Psychophysical experiments

For plots assessing duplex theory (Fig. 2-2d), azimuth sensitivity (Fig. 2-3c), bandwidth sensitivity (Fig. 2-3f), ear alteration (Fig. 2-4d,e), spectral smoothing (Fig. 2-4j), sensitivity to low-pass and high-pass filtering (Fig. 2-4o), the precedence effect (Fig. 2-5b,d) and multi-source localization (Fig. 2-6c,e) error bars are s.e.m., bootstrapped across networks. In some cases, the graph of human results used s.d. rather than s.e.m. for error bars because that is what was used in the original paper, the results of which were scanned from the original figure. We opted to use s.e.m. error bars for all model results for the sake of consistency.

To assess the significance of the interaction between the stimulus frequency range and the magnitude of the ITD/ILD bias weights (Fig. 2-2d), we calculated the difference of differences in bias weights across the four stimulus or cue-type conditions:

$$\text{difference of differences} = (B_{ILD}^{highpass} - B_{ILD}^{lowpass}) - (B_{ITD}^{highpass} - B_{ITD}^{lowpass})$$

where B denotes the bias weight for each condition). We calculated the difference of differences bootstrapped across models with 10,000 samples, and compared it to 0. As this difference of differences exceeded 0 for all 10,000 bootstrap samples, we fit a Gaussian distribution to the histogram of values for the 10,000 bootstrap samples

and calculated the P value (two-tailed) for a value of 0 or smaller from the fitted Gaussian.

We assessed the significance of the low-pass ILD bias weight (Fig. 2-2d) by bootstrapping across networks, again fitting a Gaussian distribution to the histogram of bias weights from each bootstrap sample and calculating the P value (two-tailed) for a value of 0 or smaller from the fitted Gaussian.

Statistical significance of unnatural training conditions

We assessed the statistical significance of the effect of individual unnatural training conditions (Fig. 2-7b) by comparing the human–model dissimilarity for each unnatural training condition to a null distribution of the dissimilarity for the natural training condition. The null distribution was obtained by bootstrapping the human–model dissimilarity described above across networks. We fit a Gaussian distribution to the histogram of the dissimilarity for each bootstrap sample and calculated the P value (two-tailed) of obtaining the value of the dissimilarity measure (or smaller) for each unnatural training condition under the fitted Gaussian. The effect size of the difference in dissimilarity between training conditions was quantified as Cohen's d (calculated as described above for individual experiments, but with the dissimilarity aggregated across experiments, as is plotted in Fig. 2-7b).

We also assessed the statistical significance of the effect size of the change to individual experiment results (relative to other experiments) when training in alternative conditions (Fig. 2-7c). We first measured Cohen's d as described above for 10,000 bootstrap samples of the ten networks, leading to a distribution over Cohen's d for each experiment and each training condition. For each experiment of interest, we assessed the probability under its bootstrap distribution that a value at or below the mean Cohen's d of each other experiment could have occurred. The histogram of bootstrap samples was non-Gaussian so we calculated this probability by counting the number of values at or below the mean for each condition and reported the proportion of such values as the P value (two-tailed).

We assessed the statistical significance of the effect of training condition on real-

world localization performance (Fig. 2-7e) by bootstrapping the r.m.s. localization error across networks. We fit a Gaussian distribution to the histogram of the r.m.s. error for the normal training condition. The reported P value (two-tailed) is the probability that a value could have been drawn from that Gaussian at or above the mean r.m.s. error for each alternative training condition.

2.5 Extended Data Figures

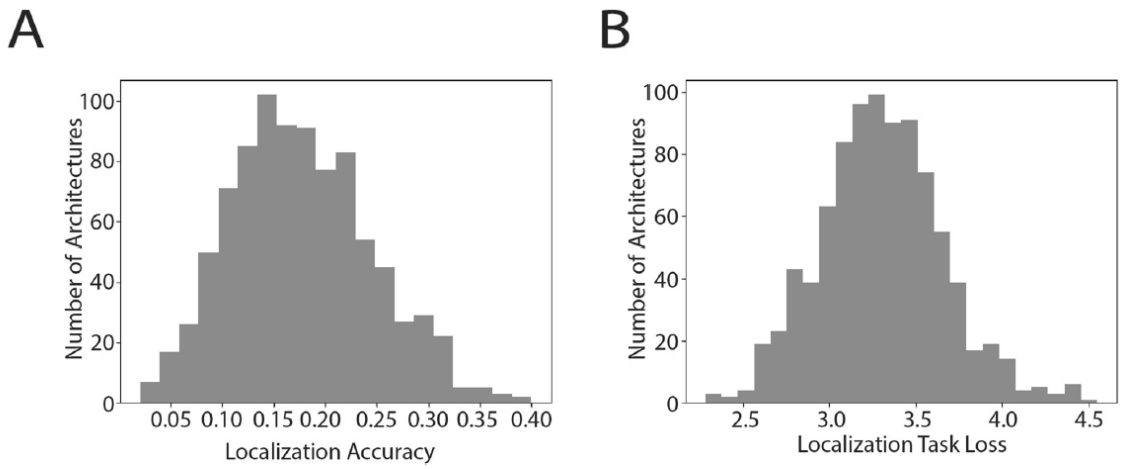


Figure 2-9: a. Histogram of validation set accuracies (proportion correct) for neural network architectures after 15k steps of training during architecture search. Here and in B, histograms include the 897 architectures that remained (out of the initial set of 1500) at this point in the architecture search. b. Histogram of validation set losses for neural network architectures after 15k steps of training during architecture search.

Network Layer	Convolutional Kernel Width	Convolutional Kernel Height	Pooling Kernel Width	Pooling Kernel Height
1	[4,8,16,32,64]	[1,2,3]	[2,4,8]	[1,2]
2	[4,8,16,32]	[1,2,3]	[2,4]	[1,2]
3	[2,4,8,16]	[1,2,3]	[2,4]	[1,2]
4	[2,4,8]	[1,2,3]	[1,2]	[1,2]
5	[2,4,8]	[1,2,3]	[1,2]	[1,1,2]
6	[2,3,4]	[1,2,3]	[1,1,2]	[1,1,2]
7	[2,3,4]	[1,2,3]	[1,1,1,2]	[1,1,1,2]
8	[2,3,4]	[1,2,3]	[1,1,1,2]	[1,1,1,2]

Figure 2-10: Discrete prior distributions used for architecture search. Pooling and convolutional kernel parameters at each layer were uniformly sampled from the lists of values.

	Network Architecture Numbers									
Operation	1	2	3	4	5	6	7	8	9	10
1	Conv[1,8,32]	Conv[2,8,32]	Conv[1,4,32]	Conv[3,8,32]	Conv[2,32,32]	Conv[1,64,32]	Conv[1,16,32]	Conv[1,64,32]	Conv[3,32,32]	Conv[2,4,32]
2	Relu	Relu	Relu	Relu	Pool[1,2]	Pool[1,8]	Relu	Relu	Relu	Pool[2,2]
3	Bn	Bn	Bn	Bn	Relu	Relu	Bn	Bn	Bn	Relu
4	Conv[1,64,32]	Conv[3,16,32]	Conv[3,32,32]	Conv[3,8,32]	Bn	Bn	Conv[1,8,32]	Conv[2,16,32]	Conv[2,16,32]	Bn
5	Relu	Relu	Pool[1,8]	Pool[1,2]	Conv[1,4,64]	Conv[2,4,64]	Pool[1,2]	Pool[1,8]	Pool[1,4]	Conv[2,4,32]
6	Bn	Bn	Relu	Relu	Pool[1,4]	Relu	Relu	Relu	Relu	Pool[1,4]
7	Conv[1,64,32]	Conv[2,4,32]	Bn	Bn	Relu	Bn	Bn	Bn	Bn	Relu
8	Pool[1,8]	Pool[1,8]	Conv[3,32,64]	Conv[1,32,64]	Bn	Conv[1,32,64]	Conv[2,4,64]	Conv[2,4,64]	Conv[2,32,64]	Bn
9	Relu	Relu	Relu	Relu	Conv[3,2,64]	Pool[2,4]	Relu	Relu	Relu	Conv[3,16,64]
10	Bn	Bn	Bn	Bn	Relu	Relu	Bn	Bn	Bn	Pool[1,2]
11	Conv[2,4,64]	Conv[3,16,64]	Conv[1,8,64]	Conv[3,8,64]	Bn	Bn	Conv[2,32,64]	Conv[2,16,64]	Conv[3,4,64]	Relu
12	Pool[2,4]	Relu	Pool[1,4]	Pool[2,4]	Conv[2,8,64]	Conv[3,4,128]	Pool[1,4]	Relu	Pool[1,4]	Bn
13	Relu	Bn	Relu	Relu	Relu	Relu	Relu	Bn	Relu	Conv[1,2,128]
14	Bn	Conv[1,8,64]	Bn	Bn	Bn	Bn	Bn	Conv[1,16,64]	Bn	Pool[1,2]
15	Conv[3,8,128]	Pool[1,4]	Conv[3,8,64]	Conv[2,2,128]	Conv[1,16,64]	Conv[2,16,128]	Conv[3,2,64]	Pool[1,2]	Conv[3,8,128]	Relu
16	Relu	Relu	Relu	Pool[1,4]	Pool[1,4]	Pool[1,2]	Relu	Relu	Pool[1,4]	Bn
17	Bn	Bn	Bn	Relu	Relu	Relu	Bn	Bn	Relu	Fc[512]
18	Conv[3,32,128]	Conv[3,8,128]	Conv[1,2,64]	Bn	Bn	Bn	Conv[1,2,64]	Conv[2,32,128]	Bn	Relu
19	Pool[1,4]	Pool[1,4]	Relu	Conv[1,4,256]	Conv[3,4,128]	Conv[1,2,256]	Pool[2,4]	Pool[1,4]	Conv[3,2,256]	Bn
20	Relu	Relu	Bn	Relu	Pool[1,2]	Relu	Relu	Relu	Pool[1,2]	Dropout
21	Bn	Bn	Conv[2,2,64]	Bn	Relu	Bn	Bn	Bn	Relu	Out
22	Conv[3,4,256]	Conv[2,2,128]	Pool[2,4]	Conv[3,2,256]	Bn	Conv[3,4,256]	Conv[1,8,128]	Conv[2,16,128]	Bn	
23	Relu	Pool[1,2]	Relu	Relu	Conv[3,4,256]	Pool[1,2]	Pool[1,1]	Relu	Conv[2,8,512]	
24	Bn	Relu	Bn	Bn	Relu	Relu	Relu	Bn	Relu	
25	Conv[3,8,256]	Bn	Conv[2,4,128]	Conv[2,2,256]	Bn	Bn	Bn	Conv[1,2,128]	Bn	
26	Pool[1,2]	Conv[3,2,256]	Relu	Pool[1,2]	Conv[3,4,256]	Fc[512]	Fc[512]	Relu	Conv[3,4,512]	
27	Relu	Pool[1,2]	Bn	Relu	Pool[1,1]	Relu	Relu	Bn	Pool[1,2]	
28	Bn	Relu	Conv[1,8,128]	Bn	Relu	Bn	Bn	Conv[3,16,128]	Relu	
29	Fc[512]	Bn	Relu	Fc[512]	Bn	Dropout	Dropout	Pool[1,4]	Bn	
30	Relu	Conv[1,8,512]	Bn	Relu	Conv[2,4,256]	Out	Out	Relu	Conv[1,3,512]	
31	Bn	Pool[1,2]	Conv[3,2,128]	Bn	Pool[1,2]			Bn	Pool[1,1]	
32	Dropout	Relu	Pool[1,4]	Dropout	Relu			Fc[512]	Relu	
33	Out	Bn	Relu	Out	Bn			Relu	Bn	
34		Fc[512]	Bn		Fc[512]			Bn	Fc[512]	
35		Relu	Fc[512]		Relu			Dropout	Relu	
36		Bn	Relu		Bn			Out	Bn	
37		Dropout	Bn		Dropout				Dropout	
38		Out	Dropout		Out				Out	
39			Out							

Architecture Layer Legend

Key	Description
Conv[X,Y,Z]	Convolutional Layer with Kernel Height X, Kernel Width Y, Z Number of Filters
Relu	Rectified Linear Unit Layer
Bn	Batch Normalization Layer
Pool[X,Y]	Max Pooling Layer with Kernel Height X and Kernel Width Y
Fc[X]	Fully Connected Layer with X Number of Units
Dropout	Dropout Layer
Out	Softmax Classification Layer with 504 Output Units

Figure 2-11: Summary of the 10 network architectures. These architectures performed best in the architecture search and were used as 'the model' in all experiments in this paper.

Air hockey	Chainsaw Cutting 2	Doorbell 4	Humming 1	Revving Engine 2	Tapdancing 1
Airplane	Chainsaw Revving	Door knocking	Humming 2	Ringing Phone 1	Tapdancing 2
Alarm 1	Chair Rolling	Drawer opening	Ice Cream Truck	Ringing Phone 2	Tapping Finger
Alarm 2	Cheering	Drilling screw	Insect chirping	Ringing Phone 3	Tapping Object
Alarm 3	Person clapping	Drilling into wood 1	Jackhammer 1	Ringing Phone 4	Tearing
Alarm clock	Chewing 1	Drilling into wood 2	Jackhammer 2	Road traffic	Telephone Ringing
Animal noises 1	Chewing 2	Drinking	Jackpot sound effect	Rocket Launch	Tennis Rally
Animal noises 2	Chicken Clucking	Driving sounds	Jumping rope 1	Rocking Chair	Thunder
Animal noises 3	Chimes 1	Drum Roll	Jumping rope 2	Rooster 1	Ticking Clock
Baby Crying	Chimes 2	Drums Beat	Kettle whistling	Rooster 2	Toothbrushing
Basketball Dribbling 1	Chimes 3	Duck quack 1	Person Laughing 1	Rooster 3	Train 1
Basketball Dribbling 2	Chimes 4	Duck quack 2	Person Laughing 2	Rotary Telephone Dialer	Train 2
Bear	Chopping Wood	Eating	Person Laughing 3	Rubbing Hands	Train 3
Bee 1	Chopping Food	Duck quack 3	Lawn mower 1	Running 1	Trainbell 1
Bee 2	Church Bells	Electric Hand Drill Starting	Lawn mower 2	Running 2	Trainbell 2
Beeping 1	Cicadas 1	Electric Shaver	Lawn mower 3	Running Up Stairs	Trainbell 3
Beeping 2	Cicadas 2	Elevator door	Lion 1	Running water faucet 1	Train Leaving Station
Beeping 3	Clanking	Engine 1	Lion 2	Running water faucet 2	Train Warning Bell
Bells Chiming 1	Clapping 1	Engine 2	Lion 3	Running water faucet 3	Train whistle 1
Bells Chiming 2	Clapping 2	Engine 3	Machine Running	Running water faucet 4	Train whistle 2
Bells Chiming 3	Clapping 3	Eruption	Marching	Sanding	Train whistle 3
Bells Chiming 4	Clashing Metal	Explosion 1	Metal Clinking 1	Hand saw 1	Trampoline
Bells Chiming 5	Clattering 1	Explosion 2	Metal Clinking 2	Hand saw 2	Treadmill
Bells Chiming 6	Clattering 2	Film Reel	Metal Clinking 3	School bell	Truck
Bike bell 1	Clinking Glasses	Finger Tapping	Monkey Scream	Scraping	Truck Backing Up 1
Bike bell 2	Clock ticking 1	House Fire	Morse code 1	Scratching	Truck Backing Up 2
Bird 1	Clock ticking 2	Fire Fighters	Morse code 2	Screwing Off Lid	Truck Backing Up 3
Bird 2	Clock Tower	Fire Alarm	Motor 1	Scrubbing	Truck horn
Bird 3	Coin Dropping 1	Fire Crackers	Motor 2	Seagull 1	Turkey
Bird 4	Coin Dropping 2	Fireworks	Motor 3	Seagull 2	Typewriter
Bird 5	Coloring	Flushing	Motor 4	Seal	Typing 1
Bird 6	Construction 1	Fountain	Motor 5	Sharpening knives	Typing 2
Blender	Construction 2	Cooking Bacon	Motorboat 1	Sheep	Vacuum
Boat	Cow Mooing 1	Gargling	Motorboat 2	Shopping Cart	Vegetable Peeler
Boat Horn	Cow Mooing 2	Gavel 1	Motorcycle Revving	Shower 1	Velcro
Boiling Water	Cow Mooing 3	Gavel 2	Music Box	Shower 2	Walking in Leaves 1
Bowling Pins Falling	Cracking	Geese 1	News Paper Rustling	Shuffling Cards	Walking in Leaves 2
Breaking Glass 1	Creaky Door	Geese 2	Opening Letter	Sink	Walking in Leaves 3
Breaking Glass 2	Crushing Can	Geese 3	Owl	Siren 1	Walking on Gravel
Brushing Hair	Crinkling paper 1	Glass Shattering	Pepper Grinder	Siren 2	Walking on Hard Surface
Brushing Teeth 1	Crinkling paper 2	Goats 1	Pig Oinking 1	Siren 3	Walking with Heels
Brushing Teeth 2	Crow	Goats 2	Pig Oinking 2	Siren 4	Water dripping
Busy Signal 1	Laughing	Goats 3	Pig snorting	Siren 5	Water Flowing
Busy Signal 2	Crumpling paper	Grandfather Clock 1	Ping-Pong 1	Siren 6	Water Splashing
Saw Cutting	Cuckoo clock	Grandfather Clock 2	Ping-Pong 2	Siren 7	Waves at Beach
Camera shutter 1	Cutting with scissors 1	Grating Food	Ping-Pong 3	Skateboarding 1	Weedwhacker
Camera shutter 2	Cutting with scissors 2	Growling 1	Plane crash	Skateboarding 2	Whales
Car crash	Dancing	Growling 2	Pool balls Colliding	Skateboarding 3	Whip 1
Car Accelerating	Dentist Drill	Gunfire	Popcorn	Slicing	Whip 2
Car Alarm	Dial Tone	Guns shooting 1	Pouring Liquid	Smashing Things	Whip 3
Car Driving 1	Dishes Clanking	Guns shooting 2	Pouring water 1	Smoke alarm 1	Whistle 1
Car Driving 2	DJ Record Scratching	Guns shooting 3	Pouring water 2	Smoke alarm 2	Whistle 2
Car Driving 3	Dog Lapping Water	Guns shooting 4	Pouring water 3	Songbird	Whistle 3
Car Driving 4	Dog panting 1	Guns shooting 5	Pouring water out of bottle	Splashing Water	Whistle 4
Car Driving 5	Dog panting 2	Guns shooting 6	Power tools	Sports Arena Buzzer	Windchimes
Car engine Starting 1	Dog panting 3	Hammering 1	Printing 1	Aerosol Can Shaking	Winding up device
Car engine Starting 2	Dog barking 1	Hammering 2	Printing 2	Spraying Aerosol can	Writing 1
Car Horn	Dog barking 2	Hawk	Printing 3	Stomach Growling	Writing 2
Car window rolling down	Dog barking 3	Heart Beat 1	Puppy whining	Stove	Writing on Chalkboard 1
Car Skidding	Dog barking 4	Heart Beat 2	Radio Tuning	Stream 1	Writing on Chalkboard 2
Car Sputtering	Dog barking 5	Heart Beat 3	Rain	Stream 2	
Cash Register	Dog barking 6	Horse neigh 1	Ratchet	Stream 3	
Castanets	Doorbell 1	Horse neigh 2	Rattling	Suitcase rolling	
Cell Phone Vibrating	Doorbell 2	Horse neigh 3	Reception Desk Bell	Swimming	
Chainsaw Cutting 1	Doorbell 3	Horse neigh 4	Revving Engine 1	Swords Clashing	

Figure 2-12: The set of sources contained multiple exemplars of some of the sound classes, denoted with the numeral at the end of the source name.

Room Geometry (Length, Width, Height)	Room Material (Walls, Floor, Ceiling)
9,9,10	Plaster on Concrete, Carpet on Concrete, Acoustic Tiles 0.625" Thick
5,4,2	Brick, Carpet on Foam Padding, Plaster
10,10,4	Wood Paneling, Audience in Upholstered Seats, Sound Dampening Panels 1" Thick
8,5,5	Heavyweight Drapery, Carpet on Concrete, Plaster on Lath
3,3,4	Grass, No Reflections, No Reflections

Figure 2-13: Room configurations used in virtual training environment. Dimensions of rooms are given in meters.

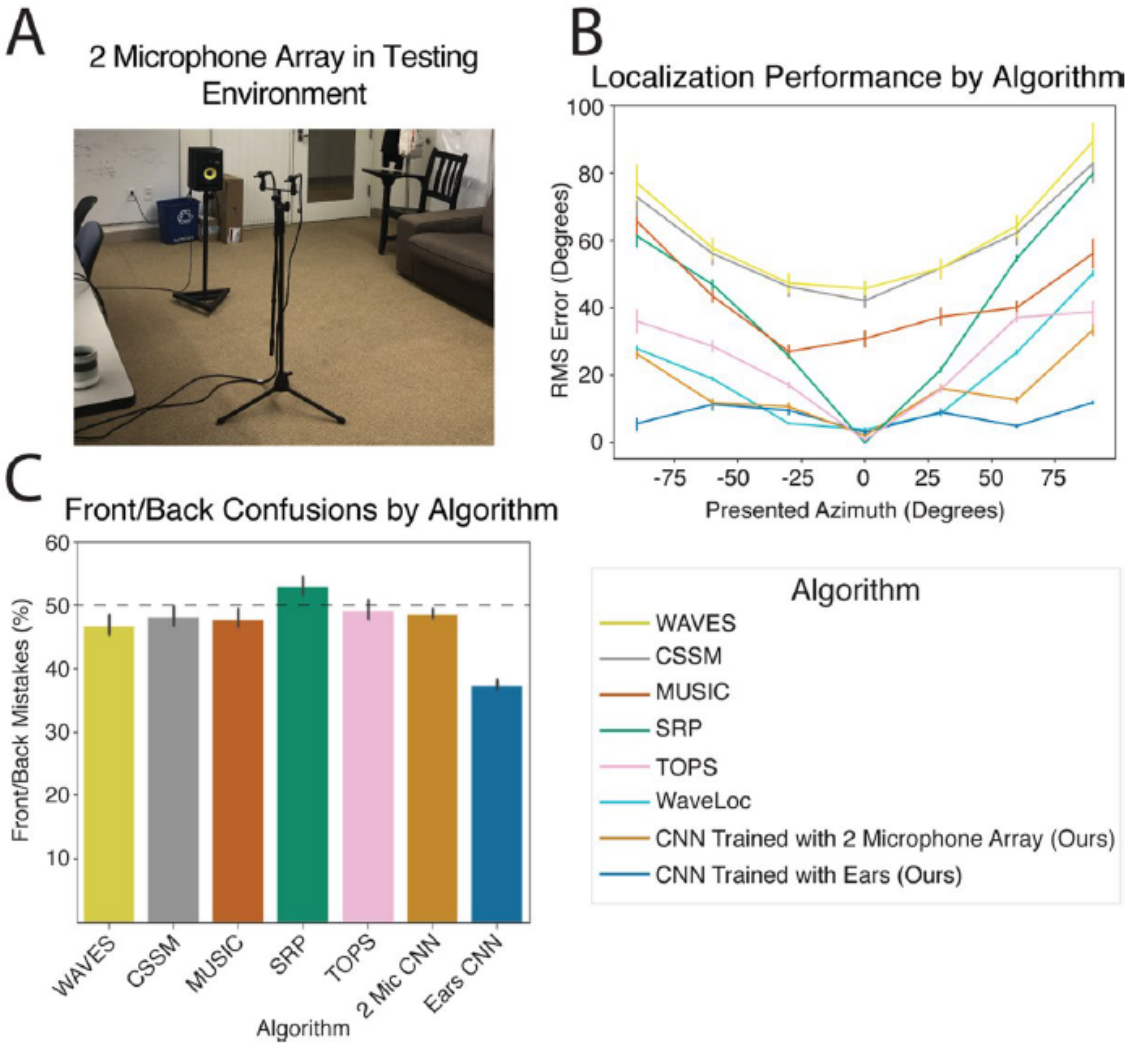


Figure 2-14: a. Photo of two-microphone array. Microphone spacing was the same as that in the KEMAR mannequin (shown in Fig. 2-1e) used to record our real-world test set, but the recordings lacked the acoustic effects of the pinnae, head, and torso. b. Localization accuracy of standard two-microphone localization algorithms, our neural network localization model trained with ear/head/torso filtering effects (same data as plotted in Fig. 2-1g,h), and neural networks trained instead with simulated input from the two-microphone array. Localization judgements are front-back folded. Error bars here and in C plot SEM, obtained by bootstrapping across stimuli. c. Front-back confusions by each of the algorithms from B. Chance level is 50%. Our main model (that is, the one trained with ears) is the only model whose front-back confusions are substantially below chance levels, confirming the utility of head-related transfer function cues for partially resolving front-back ambiguity.

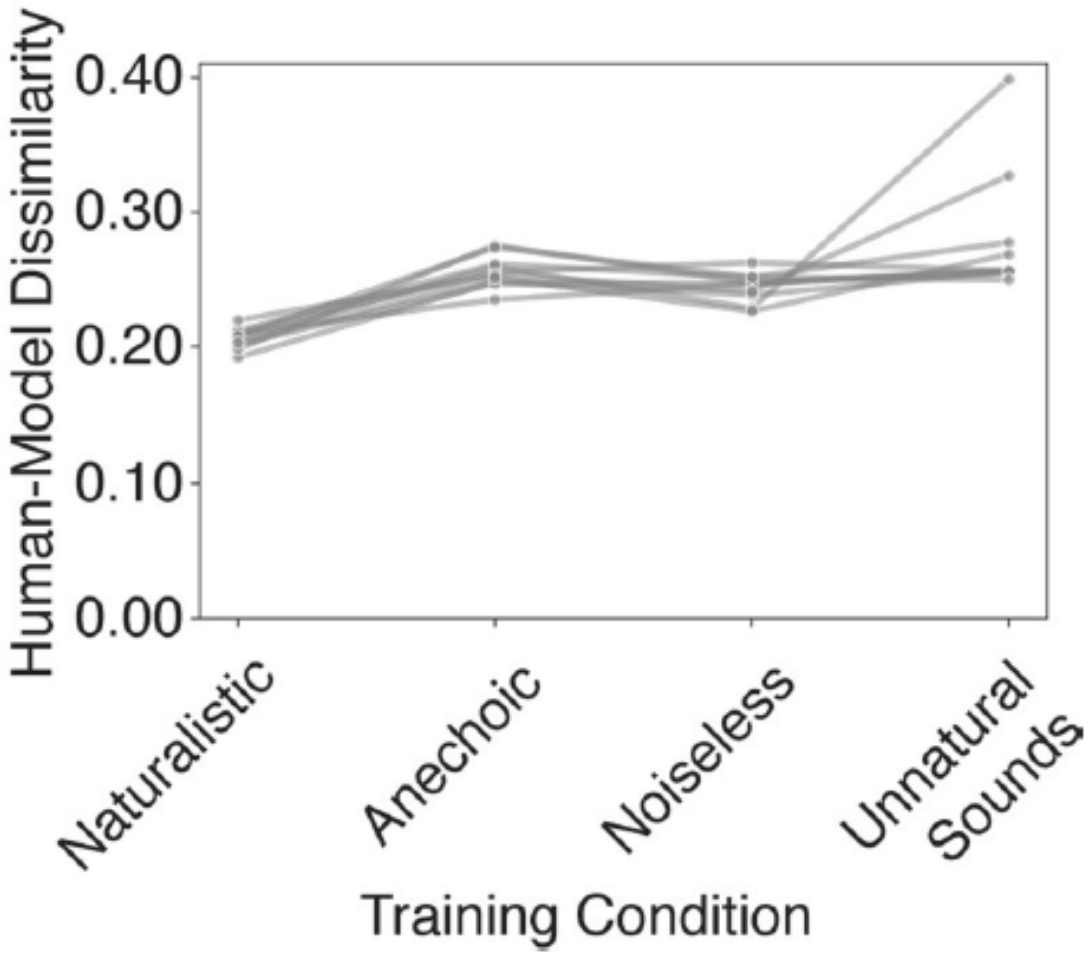


Figure 2-15: Human-model dissimilarity for natural and unnatural training conditions for each of the 10 individual neural networks

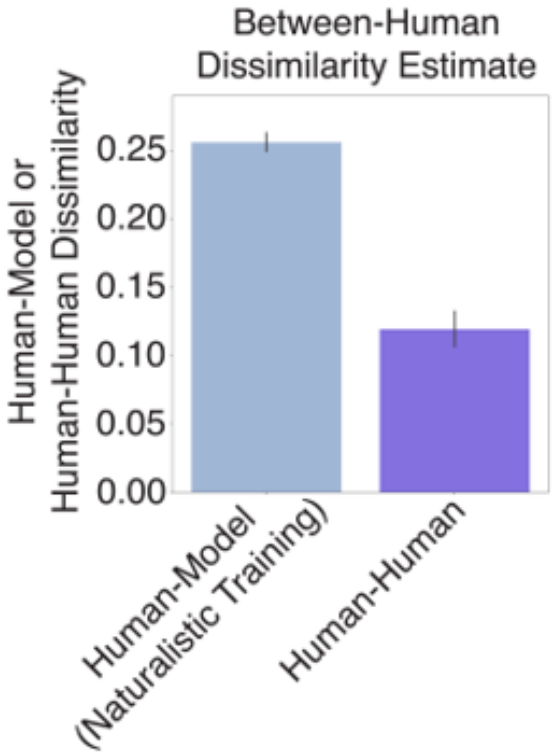


Figure 2-16: Human-model dissimilarity and human-human dissimilarity (root-mean-square error; RMSE) calculated over the subset of experiments for which across-participant variability could be estimated (typically from error bars in the original results graphs).

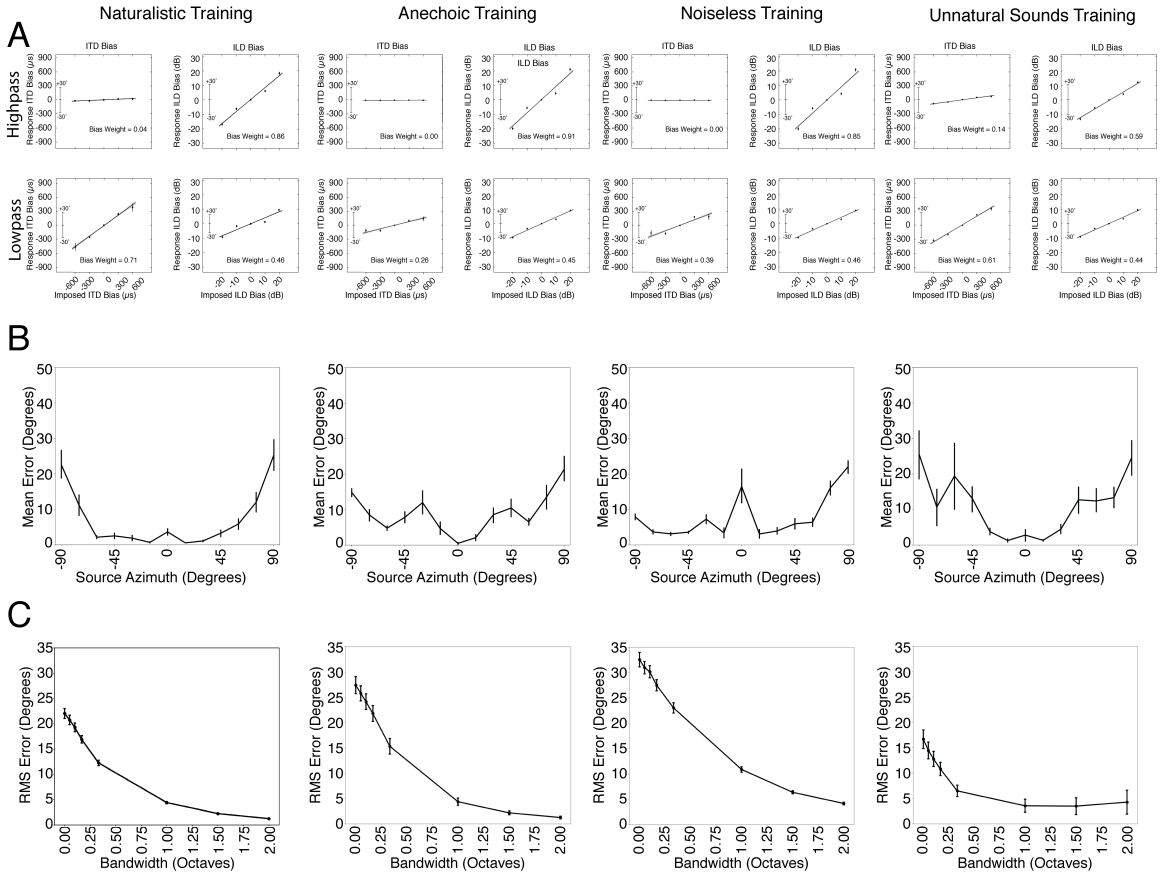


Figure 2-17: Model psychophysical results across training conditions for first three psychophysical experiments. a. Model sensitivity to interaural time and level differences (Fig. 2-2d). b. Model accuracy for broadband noise at different azimuthal positions (Fig. 2-3c). c. Effect of bandwidth on model localization of noise bursts (Fig. 2-3f). All plotting conventions are the same as in the corresponding figures in the main text.

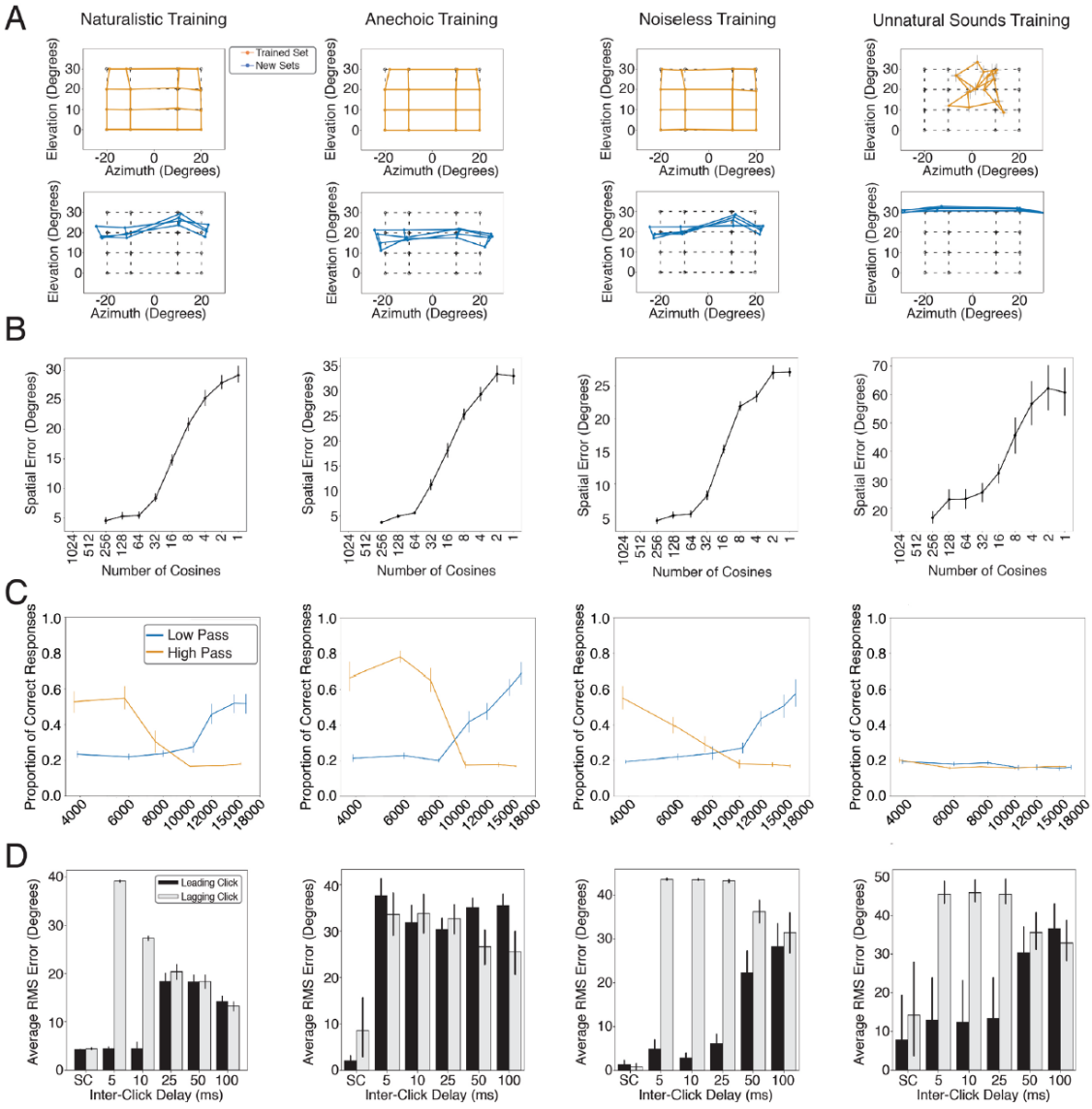


Figure 2-18: a. Sound localization by the model in azimuth and elevation before and after ear alteration (Fig. 2-4d,e). b. Effect of spectral smoothing on model localization accuracy (Fig. 2-4j). c. Effect of low-pass and high-pass cutoff on model localization accuracy for elevation (Fig. 2-4o). d. Model error in localization of the leading and lagging clicks in the precedence effect experiment, as a function of delay (Fig. 2-5d). All plotting conventions are the same as in the corresponding figures in the main text.

Chapter 3

Human Sound Localization with Natural Sounds

Abstract

Researchers often assess human sound localization using stimuli that deviate significantly from what humans experience on an everyday basis. Measurements of everyday sound localization have the potential to reveal new insights, and to provide an important benchmark for models of sound localization, which ultimately must explain real-world competence. In this work, we evaluated human sound localization in a naturalistic setting with natural sounds and quantified the accuracy of human localization. We also identified specific sounds that are difficult for humans to localize. Lastly, we assessed whether a neural network model of sound localization can predict the accuracy with which individual sounds are localized. We found that the model predicted human localization accuracy well above chance. However, the model biases were distinct from those evident in humans, suggesting room for future model improvements through refinements of the model constraints and datasets.

3.1 Introduction

Research on human sound localization has characterized human sensitivity to many features of an auditory stimulus. These include interaural time and level differences [182, 151], spectral cues to elevation [137], frequency bandwidth [241], and reverberation [221, 143]. One common approach is to carefully control stimulus presentation and all potential cues available to a listener by presenting stimuli over headphones.

While headphones ensure that subjects only have access to variables explicitly encoded in the stimuli by the researchers, it also means subjects do not have access to natural cues induced by the interaction of incoming sound with a person's head and torso. As a result, such studies are not measuring a natural human behavior but rather measure the sensitivity of the human auditory system to specific predefined cues.

A limited number of studies have investigated human localization in a "free-field" setting (i.e., in which sounds are played from speakers some distance from the listener), where subjects can use natural cues [152, 170, 227]. However, these studies are relatively challenging to set up, requiring speakers assembled in a spatial array or movable speaker mounts (in contrast to headphone studies that only need a quiet room and computer), and as a result are less common than experiments with headphones.

Many studies in free-field conditions were also forced to choose between maximizing the frequency range used in the experiment and the number of spatial positions that could be sampled. This tradeoff is due to the large speaker cones needed to produce a strong signal at low frequencies [154]. Large speaker cones require large and heavy speaker cabinets that need strong mounts, making them difficult to move around quickly. One solution to this problem is to limit the number of speakers to a few spatial locations [186, 98]. Another approach is to place subjects in a dark environment and use small speakers mounted on a movable frame [152, 170]. This second approach allows for experiments using many spatial locations, but suffers from two substantial drawbacks. The first is that moving the array takes time, which means that spatial sampling patterns either have to be blocked based on the speaker position or that the speaker has to be moved between most trials, slowing data collection. The second is that the speakers must be small enough to be easily movable, which means that they will be too small and lightweight to produce significant power at low frequencies. The lack of low frequencies means the experiment does not maintain all the cues available to a listener in a natural sound. In part because of these constraints, free-field experiments have rarely presented naturalistic stimuli apart from

some studies using speech [130, 5].

We built on this previous work by presenting natural sounds in a real-world environment across a large range of tightly spaced positions. To enable the experiment, we mounted 133 large speakers across a range of positions and supported the associated weight of the speakers using a custom-designed support truss. This approach allowed us to overcome several issues facing previous researchers, enabling us to accurately reproduce natural sounds, render sounds from a wide range of spatial locations, provide fine-grained reporting to allow accurate measurement of human errors, and control hundreds of speakers quickly enough to allow for high throughput data collection. In the sections that follow, we describe the design of the array, and the experiment that used it to collect a large set of human localization judgments with natural sounds.

We then use this new dataset of natural human behavior to further test the model described in chapter 1. We find that the correlation between the model and the human pattern of errors is significantly greater than chance and, in some cases, predicts about half of the explainable variance in the human data. However, we find our model falls short when predicting elevation biases in the human data, suggesting room for future improvement and the need to identify and incorporate other model constraints.

3.2 Results

3.2.1 Constructing the speaker array

We began by constructing a speaker array capable of reproducing natural sounds at a large number of spatial locations. The speaker array (Fig. 3-1) produced sounds from 133 speakers, which were arranged in a hemisphere with positions ranging from -90° to +90° in azimuth and -20° to +40° in elevation. Each speaker was exactly 2 meters from the head of a human listener sitting at the center of the hemisphere. We designed and built an aluminum support truss to mount the speakers. The truss was constructed of heavy-duty aluminum to support the combined weight of the speakers (which totaled one ton) while being light enough to transport and assemble without

the use of specialized equipment.

We used KRK Classic 5 speakers with a flat frequency response between 46Hz and 30 kHz, covering almost the entire range of human hearing. Each speaker was capable of playing out a unique audio channel, allowing any single or combination of speakers to play any combination of audio tracks. The speakers were all phase locked and routed using MOTU digital-to-audio converters. We labelled each speaker with an alphanumeric code, with the letters A through G representing elevations from +40° to -20° and numbers 1 through 19 representing azimuths from -90° to +90°. We created a custom keyboard with letters A through G and a standard number pad to allow for rapid response input.

Figure 3-1: Picture of the Speaker Array

3.2.2 Measuring Human Localization with Natural Sounds

We assembled a stimulus set of 160 natural sounds, each 1-second in duration, that were common in everyday life (Table 3-1). We crossed each sound with each speaker position. Due to practical constraints on participant time, we chose to split these 21,280 trials across 19 different subjects (i.e., presenting each combination of sound and location once, to a unique subject). We chose this number by estimating that each subject should be able to run about 600 trials/hour and targeted an experiment duration of less than 2 hours to prevent fatigue. Trials were randomly assigned to participants but constrained to be approximately uniform across position and sound identity for each participant. We recruited 19 subjects (8 female, ages 19-28) from the area around Cambridge, MA.

In the experiment, each subject fixated the speaker directly in front of them and was presented with a randomly selected sound from a randomly selected speaker. After the sound finished playing, the subject was allowed to move their head and was instructed to enter the label from the speaker where they believed the sound played from. Subjects would then reorient back to the speaker directly in front of them and begin the next trial. Trials were grouped into 4 blocks of 280 trials, with breaks between each block.

We separately analyzed the localization error in the azimuthal and elevation dimensions, as in previous studies. [152, 170, 227]. It is common to analyze these two dimensions separately because there are thought to be distinct the mechanisms to underlie localization in azimuth and elevation.

3.2.3 Measuring Model Localization with Natural Sounds

We measured the model's localization judgments for our set of 160 natural sounds rendered in an environment intended to mimic the one used with human subjects. We rendered each sound at the same speaker positions used for human subjects but excluded positions below 0° elevation because the model cannot report sources at those positions (limited by the set of head-related transfer functions we used in model

training). We collected a response from each of the 10 networks for each sound at each rendered position, and analyzed the responses in the same way as those of the human participants.

3.2.4 Accuracy vs. Azimuth

We began by measuring localization accuracy as a function of azimuthal position. First, we grouped the human behavioral responses by the ground truth speaker location. We then calculated the mean absolute error between the judged azimuth and true azimuth for sounds at each azimuth. The results (Fig. 3-2A) show that human localization accuracy is best at the midline, and becomes progressively less accurate as sources move to the left or right of the listener. This aligns with results from previous human studies with synthetic sounds [188, 165, 229], but shows that the result generalizes to natural sounds, and illustrates the overall accuracy for natural sound localization. One potential explanation for the effect of azimuth is that the first derivatives of ITD and ILD with respect to azimuthal location decrease as the source moves away from the midline and provide less information about location [16, 93]. We performed the same analysis for the model (Fig. 3-2B) and found that the model replicated the general effect but was less accurate than humans.

We also calculated the mean absolute error between the judged elevation and true elevation for sounds at each azimuth. The results (Fig. 3-2C) show that elevation accuracy was worst on the midline and improved somewhat as the source moves to the right or left of the midline, denoted by the roughly 1° decrease in absolute error. This effect was not observed in previous studies [152, 170, 227]. The same analysis of the model's responses showed a similar effect (Fig. 3-2D), but the model exhibited both higher overall error and a larger decrease in error for sources away from the midline than we saw in humans.

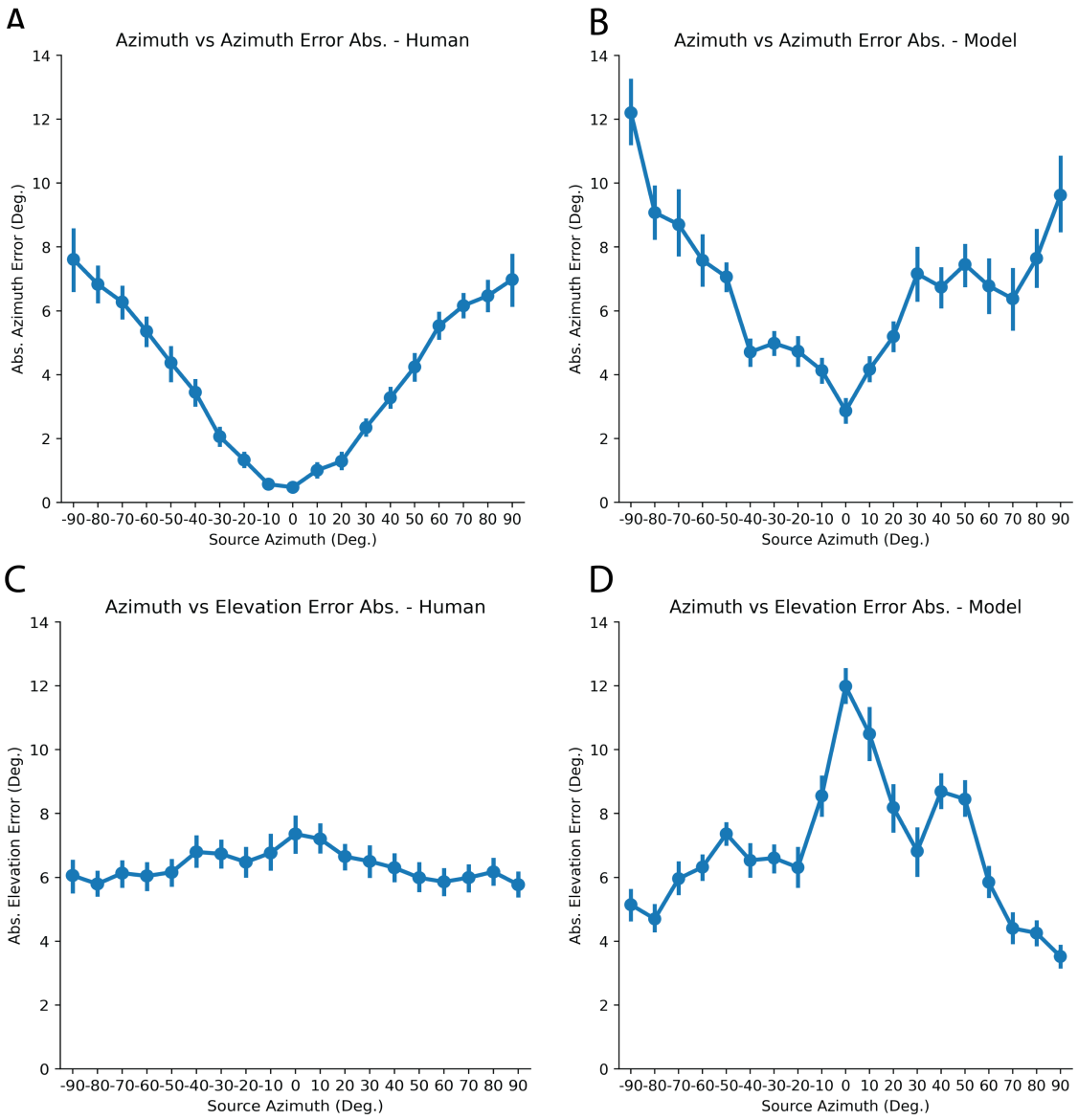


Figure 3-2: Azimuth localization accuracy at different azimuthal positions for human listeners. Here and in B, graph plots mean absolute azimuth localization error. Error bars plot SEM across listeners, obtained via bootstrap. B. Azimuth localization accuracy at different azimuthal positions for the model. Error bars plot SEM across ten networks. C. Elevation localization accuracy at different azimuthal positions for human listeners. Here and in D, graph plots mean absolute elevation localization error. Error bars plot SEM across listeners. D. Elevation localization accuracy at different azimuthal positions for the model. Error bars plot SEM across ten networks.

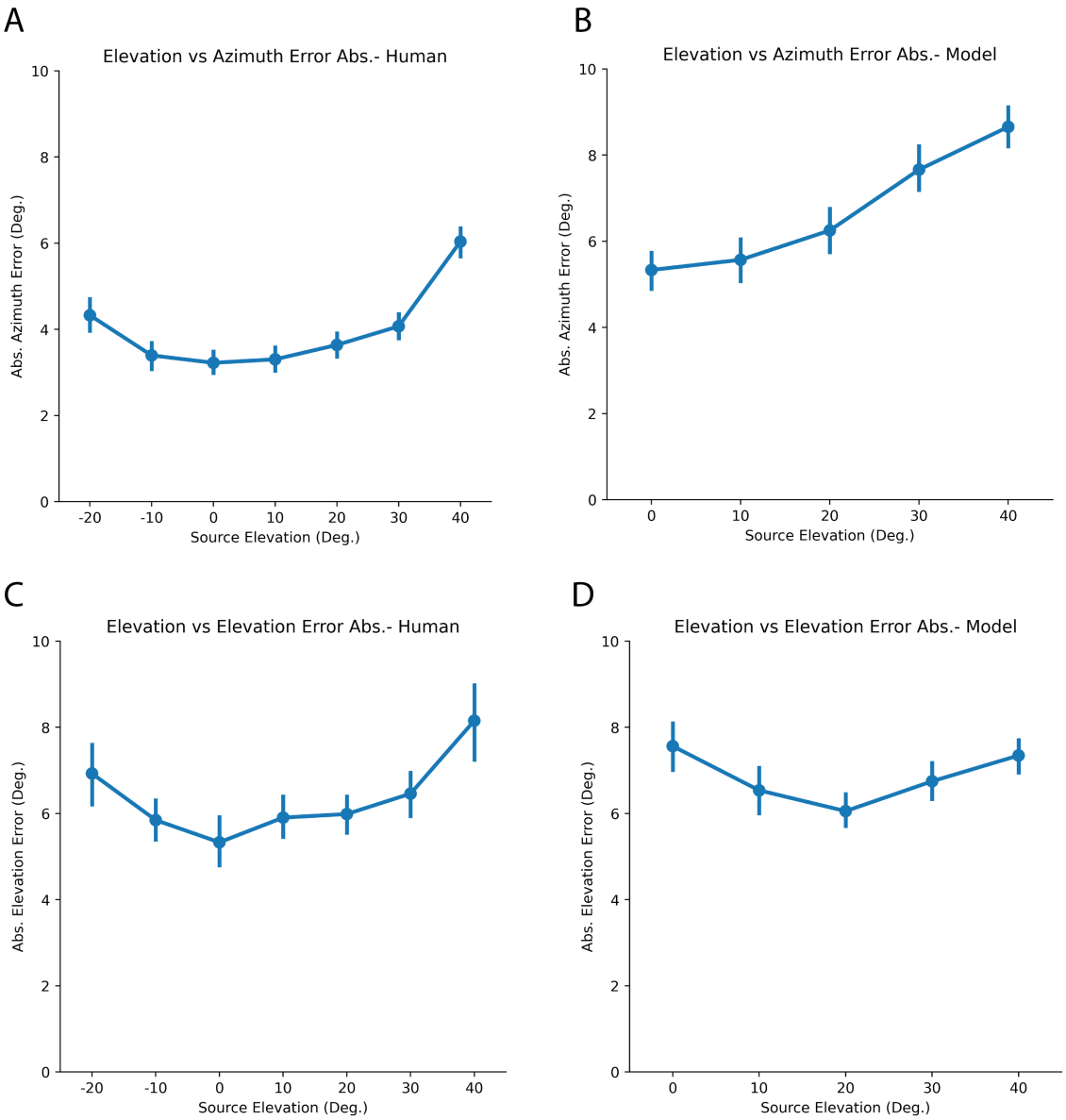


Figure 3-3: Azimuth localization accuracy at different elevation positions for human listeners. Here and in B, graph plots mean absolute azimuth localization error. Error bars plot SEM across listeners. B. Azimuth localization accuracy at different elevation positions for the model. Error bars plot SEM across ten networks. C. Elevation localization accuracy at different elevation positions for human listeners. Here and in D, graph plots mean absolute elevation localization error. Error bars plot SEM across listeners. D. Elevation localization accuracy at different elevation positions for the model. Error bars plot SEM across ten networks.

3.2.5 Accuracy vs. Elevation

We next measured accuracy for sounds presented at different elevations using the same approach we used for analyzing azimuth. We found that azimuthal accuracy was fairly consistent across elevations, with only slight advantages for sounds near the horizon (Fig. 3-3A). The model also localized most accurately at the horizon (Fig. 3-3B). One possible explanation may be that the first derivatives of ITD and ILD with respect to azimuthal location decrease for signals at higher elevations [220].

Next, we calculated the mean absolute error between the judged elevation and true elevation for sounds at each elevation. Overall, both humans and the model performed well across tested elevations. Humans were slightly more accurate when sources were positioned at 0° elevation, with absolute elevation error increasing as the source moved up or down (Fig. 3-3C). The model was slightly more accurate at 20° elevation than above or below (Fig. 3-3D). One potential explanation for the deviation between model and human results may be the idiosyncrasies of the specific head-related transfer function (HRTF) used to train the model, e.g. if the model's HRTF has a feature for sources at 20° that is particularly salient. However, the similar performance across the tested source elevation positions suggests that both humans and models are accurate across different source elevations.

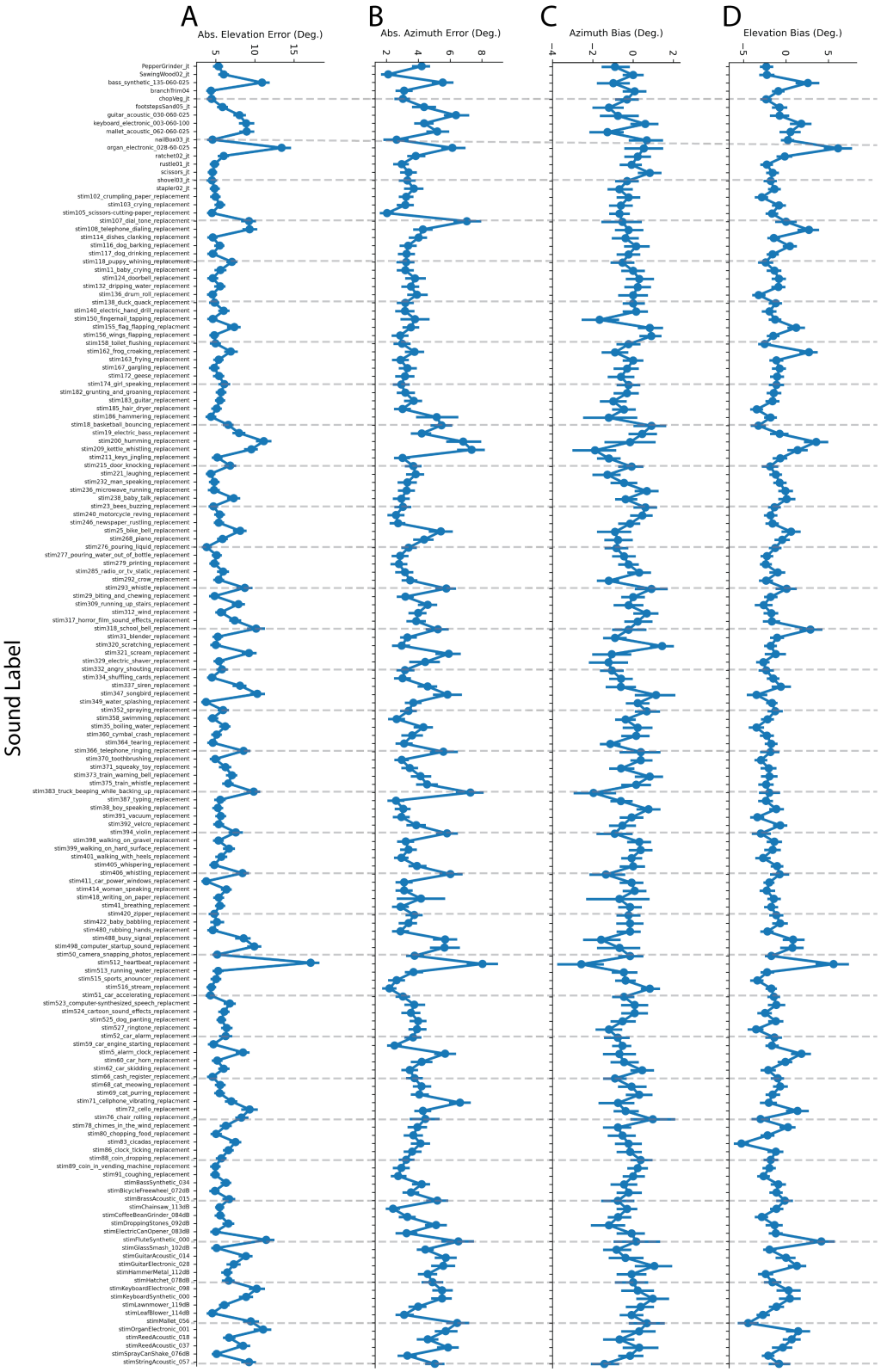
3.2.6 Sound Identity vs Accuracy

A primary question motivating the experiment was whether humans are better at localizing some natural sounds than others. We assessed this by calculating the mean absolute error in azimuth and elevation for each sound class (pooled across all positions, using responses from all 19 subjects). We plot the human mean absolute error for azimuth (Fig 4A) and elevation (Fig 4B) for each of the 160 sounds. Some sounds are much harder to localize than others. These differences are not specific to the source position or to idiosyncrasies of a single human subject's HRTF, because they are averaged across positions and participants. The variance across subjects is likely instead caused by the acoustic properties of the sounds.

We also measured localization bias, or mean signed error, for each sound, in azimuth and elevation (Fig. 3-4C&D). This analysis tests whether certain sounds are consistently judged as occurring above/below (elevation bias) or to the left/right (azimuth bias) of their actual position. We found the majority (131/160) of sounds showed a slight downward but not significant bias in elevation (mean bias of - 1.15°, $p > 0.05$) and little to no bias in azimuth (mean bias = - 0.24°, $p > 0.05$).

Together these results characterize human localization behavior for natural sounds and show that there is meaningful variance across sound classes.

Figure 3-4: A. Elevation localization accuracy by sound identity. Graph plots mean absolute elevation localization error. Here and elsewhere, error bars plot SEM across trials, via bootstrap. B. Azimuth localization accuracy by sound identity. Graph plots mean absolute azimuth localization error. C. Azimuth localization bias by sound identity. Graph plots mean azimuth error (the bias). D. Elevation localization bias by sound identity. Graph plots mean elevation error (the bias).


A
Abs. Elevation Error (Deg.)
B
Abs. Azimuth Error (Deg.)
C
Azimuth Bias (Deg.)
D
Elevation Bias (Deg.)
Sound Label

3.2.7 Comparison to Model Predictions

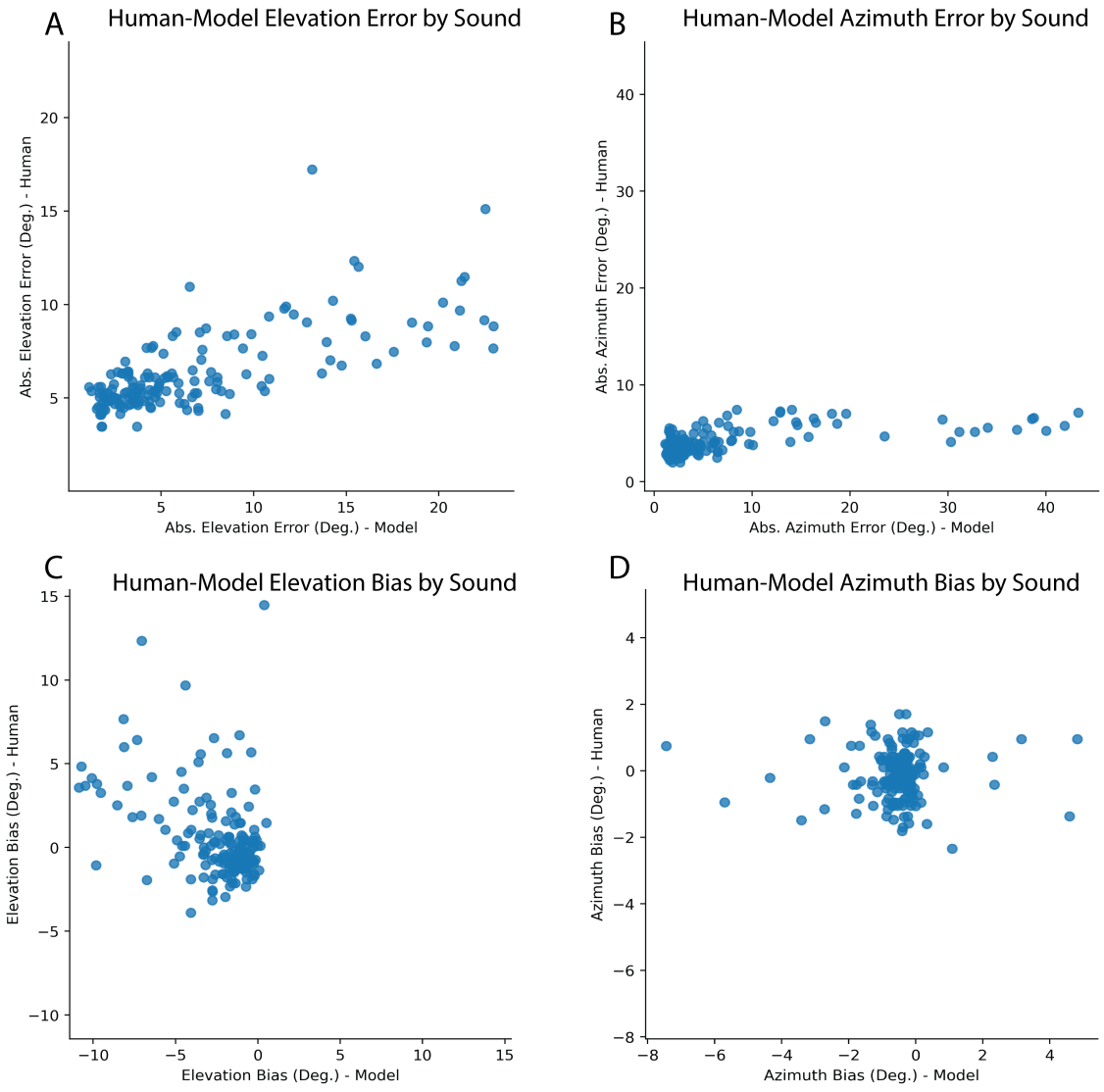
We next assessed whether our model could predict the pattern of human localization errors across natural sounds. We calculated the mean absolute error in azimuth and elevation for each sound class from the model's responses, and plotted the model error versus human error for each sound (Fig. 3-5). Next, we calculated the correlation between the human and models for each of the four types of error described in the previous section (Table 3-2). Lastly, we calculated the reliability of the pattern of human errors by measuring the split-half reliability of each dependent measure, as this determines the ceiling on the human-model correlation that is possible.

The model accurately predicted many of the sounds that human subjects struggled to localize (Fig 5 A&B). The human-model correlation was substantial for both elevation error (r=0.72; Fig. 3-5A) and azimuth error (r=0.6; Fig 5B). The model also tended to make larger errors than human subjects. In both cases, the split-half reliability of the human data was high, with split-half reliabilities greater than 0.75.

By contrast, the model biases were less closely related to human biases (Fig 5 C&D). There was in fact a negative correlation between human and model elevation bias (r=-0.4). The model and human azimuth biases were virtually uncorrelated, but the reliability of these biases was very low for humans, indicating that there is little reliable variance for a model to explain.

Lastly, we asked whether the model and human behavior could be accounted for by simple measures of the sparsity of a sound's spectrum, as is often argued to influence localization accuracy [204, 241]. We found that the spectral flatness of a sound (a common measure of the sparsity of the frequency spectrum) was fairly well correlated with human localization accuracy (r=0.68 for elevation error; r=0.56 for azimuth error; Table 3-3). This suggests that some of the difficulty in localizing sound is related to how power is distributed across frequencies in a signal, with sounds with denser spectra being easier to localize. To test whether the human-model correlation was driven by this acoustic factor, we computed the partial correlation between the human and model accuracy, controlling for the effect of spectral flatness.

This analysis revealed that a significant correlation remained between the model and human residuals for absolute errors in azimuth as well as elevation (Table 3-2), suggesting the model is capturing features beyond spectral sparsity alone, and indicating that the model has independent value for predicting human localization accuracy.



(Caption on next page.)

Figure 3-5: A. Elevation localization accuracy for model and human listeners. Each dot plots the mean absolute elevation error across all trials for one natural sound. B. Azimuth localization accuracy for model and human listeners. Each dot plots the mean absolute azimuth error across all trials for one natural sound. C. Elevation localization bias for model and human listeners. Each dot plots the mean elevation error (the bias) across all trials for one natural sound. D. Azimuth localization bias for model and human listeners. Each dot plots the mean absolute azimuth error (the bias) across all trials for one natural sound. All analysis of error and bias were performed across the same subset of positions for the human and model (i.e., at 0 degrees elevation and above).

3.3 Discussion

We designed and constructed a speaker array capable of producing natural sounds with high fidelity at many different locations. Building the array required us to overcome a number of technical challenges that constrained previous researchers. We addressed these challenges by physically supporting our speakers using a custom aluminum support truss and by using modern networking protocols to deliver sound rapidly to the array. This speaker array allowed us to measure human localization for a large set of natural sounds. This is the first experiment to document human localization behavior in a naturalistic setting with a wide range of natural sounds.

The human results reproduced some previously documented aspects of human sound localization, such as the decreasing accuracy of azimuthal localization (Fig. 3-2A) as a source moves away from the midline [152, 165, 229]. We also produced novel measurements of the overall accuracy of natural sound localization, which was overall fairly good (averaging between 1 and 8 degrees error in azimuth, depending on the location, and around 6 degrees in elevation). We also found that sounds varied substantially and reliably in the accuracy with which they were localized by humans (Fig. 3-4).

Next, we assessed how well the model from chapter 1 could predict human behavior. Specifically, we asked if the sounds that were easiest and hardest for human subjects to localize were also easiest and hardest for the model to localize. We found that the model predicted a majority of the explainable variance in elevation error and

49% (R2 in Table 3-4) of the explainable variance for azimuth error (Table 3-4).

Despite the similarities between the human and model results, discrepancies remain. The most notable difference is between the model and human elevation biases (Fig. 3-5C). Humans make systematic errors in elevation judgments for particular sounds (judging some sounds to be higher than they actually are, and some to be lower). The model also exhibited biases, but they tended to be in the opposite direction as humans. One possible explanation for the human-model difference is that human elevation biases are caused by the natural distribution of source positions in the world, which the model did not learn given that sources in its training data were rendered uniformly across space. Another potential explanation is that the model's training set was generated from a relatively small set of natural sound sources (due to limits on available recordings), with the model being somewhat overfit to its training set as a result. The model did generalize to novel sounds (all the sounds in the experiment were distinct from those in its training set), but the training set sound sources may nonetheless have influenced the learned localization strategy. Lastly, the model was trained to report sources located only between 0° and 60° elevation (due to constraints on available head-related transfer functions), whereas human listeners learn to localize over the full range of possible sound locations. This difference may have induced different biases in the model than naturally occur in humans.

Another salient discrepancy between the human and model data is the variability in accuracy across sounds and locations. This is visible in the graphs showing model error for source azimuth (Fig. 3-2 B&D), and in the size of the model errors in the human-model error scatter plots (Fig. 3-5). This discrepancy could again be due to the small set of sound sources in the training set, but could also reflect the small number of network architectures used to make the judgments and/or limits of the architectures and training due to the computational resources at the time of training (roughly four years ago). Training a larger set of networks and rerunning a larger-scale architecture search using modern computing resources (faster/larger memory GPUs, faster CPU preprocessing) and a more diverse training dataset may result in a better model with errors that more closely match those of human subjects.

We have presented just a single experiment run on the speaker array we built, but there are numerous promising future applications of the system. One such application would be to study concurrent localization with multiple natural sounds. While there have been some measurements of multi-source localization with noise bursts [**?**] or speech [244, 123], we know little about how spatial location and grouping cues interact in natural scenes to give rise to our perception of multiple sources. The current speaker array will allow researchers to better quantify this behavior in humans with more naturalistic stimuli and at a greater range of positions than was previously possible.

The fine timing control of the array also offers the chance to perform more sophisticated stimulus generation, including rendering full sound fields with ambisonics [64] or rendering moving sound sources. One use case could be to revisit work on the role of dynamic cues to source elevation. Hans Wallach noted that listeners experience smaller changes in the interaural time and level differences (ITD and ILD) for high elevation sources when they move their head [220]. He also showed that listeners report a source as originating from a higher elevation if the source is rendered with a smaller ITD and ILD than a listener would naturally experience for a given head rotation. Our array allows a more thorough exploration of this phenomena by rendering sources with smaller or larger interaural cues than would be expected for sources at different positions (by panning between speakers during participant head movements). Measuring how these changes affect human judgments could allow researchers to quantify how monaural cues and dynamic cues each contribute to elevation perception.

3.4 Methods

3.4.1 Room Impulse Response Measurement

We measured the room impulse response (RIR) using the method described in Traer & McDermott (2016)[212]. We recorded a noise signal produced by a speaker in the room. Because the noise signal was known and the speaker and microphone transfer

functions were flat, we could infer the RIR from the recording. The noise signal was played from a KRK Classic 5 speaker and was recorded using a Beyerdynamic MM-1 microphone. Both the broadcast and recording of the signal were managed using a Scarlet 8i8 sound card connected to a 2018 MacBook Pro running Audacity 2.2.1. The speaker was placed 1.5 meters from the microphone, and the microphone was placed at the listener's position in the room. Per the manufacturer's instructions, the microphone was oriented at 90 degrees to the speaker to maintain the flat frequency response from 30Hz to 20kHz. The noise signal consisted of interleaved repetitions of 6-second Golay sequences. The final noise signal was 3 minutes long and played in the room at 85 dBA. The recorded Golay sequences were cross-correlated with the broadcast Golay sequence to obtain a final measurement of the room-impulse response using the code from Traer and McDermott 2016, where a detailed description of the code and the mathematical basis for the approach is described.

3.4.2 Room Noise Level Measurement

We measured the background noise level in the room using a sound level meter. All measurements were made by a Svantek 979 Class 1 sound level meter with a G.R.A.S. 40AE microphone cartridge. The free-field compensation filter was enabled on the sound level meter, and frequency analysis was set to 1/3 octave bands during all measurements. The sound level meter was placed on a small foam pad on top of a small flat stand with the microphone extending off of the stand. The stand was placed in the same position occupied by subjects in the experiment. We made a total of eight recordings (two recordings at each of four separate microphone orientations). The microphone was rotated on its z-axis to 0°, +90°, -90°, and 180°, where 0° is the orientation of a forward-facing subject performing the experiment. For each measurement, we recorded the average dBA measurement and reported the average over the eight measurements (next section).

3.4.3 Noise Reduction Treatment – Room

We reduced the background noise in the experiment room by building two false walls along the walls adjacent to the room where a piece of mechanical equipment was generating the background noise. The first wall was 254 inches long and 98 inches tall. We constructed and installed a wood frame using 2x6 pine boards for the floor and ceiling plate and 2x4 vertical studs. The frame was placed 6 inches in depth away from the permanent room wall, and the studs were spaced 24 inches apart to decrease the mechanical coupling between the wood frame and the interior wall. The frame was secured to the back wall via 3 L brackets and masonry screws. Each of these contact points was insulated with rubber to reduce vibration transfer between the wall and frame. We installed Thermafiber Sound & Fire Guard mineral wool between each stud. The mineral wool was 5" thick and was installed flush with the front of the wall. We then mounted 5/8-in x 4-ft x 8-ft drywall to the front of the wood frame to further dampen the background noise. The edges of this wall were insulated with rubber weather stripping to minimize the airflow that could pass through the wall.

The second room wall where we installed noise reduction measures appeared to have less noise passing through it than the first wall. As a result, we used a more straightforward construction method where we simply leaned tightly joined drywall against the permanent wall. We spliced together 5/8-in x 4-ft x 8-ft drywall along the 4-ft edge such that the final drywall height made contact with the ceiling when the drywall was leaned against with wall, with the drywall base 9 inches away from the permanent wall at floor level. The splice between pieces of drywall was reinforced by backing the seam with 1/4-in x 4-ft x 4-ft oriented strand board. We lined each section of drywall with rubber weather stripping to minimize the airflow that could pass through the wall. The wall was assembled in this manner, section by section, until it spanned the entire length of the second wall. In addition to building two isolation walls, we also contracted a Heating, Ventilation, and Air Conditioning specialist to remove an exhaust vent from the ceiling and redirect it into the sub-ceiling to reduce the audible turbulent airflow around the vent. We also insulated three drain pipes to

attenuate the intermittent trickling sound.

The addition of these two walls reduced the overall noise level in the room from 53 dBA to 48 dBA and made the remaining background noise spatially diffuse.

3.4.4 Speaker Array Design

We designed the speaker array to have a speaker positioned every 10° in azimuth and 10° in elevation relative to the listener. Speaker positions ranged from +90° azimuth to -90° azimuth and °40 elevation to -20° elevation, with 133 speakers in total. The array was also set up across a hemisphere so that the distance between the front panel of the speakers and the listener was always 2 meters. A support truss was designed to allow us to mount the speakers at the chosen positions and support the weight of 133 speakers.

3.4.5 Speaker Array Truss Design

The support truss for the speaker array was designed to support the weight of over 133 speakers, weighing 15 pounds each. We decided to make the truss out of aluminum to ensure that the structure could withstand the physical strain of the speakers while being light enough to disassemble and move. To mount the speakers to the support truss, we designed a plate to support the speaker that would then be mounted to a metal tube using pipe clamps. This design choice provided future flexibility for the speaker's azimuthal position and the speaker mounting angle. The truss consisted of 7 rings of aluminum piping, one at each elevation where the speakers would be located. The truss was also designed to be separated into four pieces that could each fit through a standard door. The truss was assembled by bolting together the four pieces stabilized with additional metal plates that spanned the joints of the assembly.

To ensure the pipes could support the weight of all speakers, we performed a series of load tests to ensure that the pipe could withstand the weight of the speakers and selected a pipe with a 2-inch outer diameter and 1/8 wall thickness. When the pipe was bent to the geometry and length that would be used in the longest span

of the speaker array, the load test measured less than .3 inches of yield with 110 pounds of load. Additional static analysis suggested that the longest length of the tube could hold over 500 pounds before yielding beyond where it would return to its original shape. We estimated this provided a roughly 6x safety factor over the expected speaker load of 80 pounds (5 speakers at 15 lbs./speaker)

We designed vertical supports for the pipes on each end of the four sections of the support truss. These vertical supports were constructed from $\frac{1}{2}$-inch thick aluminum sheet stock, and the width varied from 11 inches at the bottom of the support to 6 inches across at the top of the support. Each vertical support was 98 inches in height. The base of the structure was constructed from $\frac{1}{2}$-inch thick aluminum sheet stock and was 11.5 inches in width.

The joints between the pipes, vertical supports, and truss base were welded together and ground down to simplify assembly and disassembly. Each seam between subsections was joined with steel bolts and ten steel bolts with $\frac{1}{2}$ inch diameter.

3.4.6 Speaker Array Construction

An external vendor cut all aluminum sheet stock components using a high-pressure computer-controlled waterjet to the specifications in the final design files. The aluminum tubes were then rolled to the proper curve and welded to the sheet stock pieces by MIT's central machine shop according to the specification in the final design files.

3.4.7 Speaker Mounting Mechanism

The speaker mount was designed to allow speakers to be moved to arbitrary azimuths to provide maximum flexibility for future experiments. The mount was constructed using a custom-cut aluminum plate that measured 11-in x 6-in (LxW) x 1/8-in thick. These plates were designed to support the speakers from below and to be clamped above the aluminum tube. Holes were drilled into each plate to secure the plates to the tube and secure the speaker to the plate. Four holes that were $\frac{1}{2}$ inch in diameter were drilled on the back of the plate with two sets of holes 1 inches from each of the

long edges and 0.5 inche from the wide edge. Each set of holes had a 2-inch intra-hole distance along the length axis. Each plate had four additional holes to secure the speaker, each 1/4 in diameter. These holes formed a rectangle with geometry 6.5 X 4.5 with 0.5 inches of margin from the front of the plate and 0.75 inches of margin from each side of the plate.

The mounting mechanism involved two pipe clamps per plate that secured each plate to the tube. We used Nickson 2 Inch Steel Exhaust Clamps with a 3/8 inch bolt. Prior to construction, we performed a load test using these clamps and determined that they provided enough clamping force to counteract the torque exerted by the speaker's weight pushing down on the speaker plate. We secured each speaker to the mounting plate using four medium-density fiberboard screws. The screws were $\frac{1}{2}$ inch long and did not penetrate the interior of the speaker cabinet (which might otherwise affect the sound output).

We inserted 1/16-inch thick rubber strips between the mounting plate and pipe clamps and between the pipe clamps and support truss tube. These dampened the vibrations transferred to the support truss from speakers when active and prevented any metal-to-metal contact that may have caused rattling sounds from the speaker vibration.

3.4.8 Speaker Array Construction – Speaker Calibration

We choose to use KRK Classic 5 speakers in the speaker array because they provide a flat frequency response between 56 Hz to 30 kHz. We verified the speaker response by measuring the transfer function of the speakers and found that the frequency response met the manufacturer's specifications. To measure this transfer function, we recorded the speaker producing a pink noise signal with frequencies between 20Hz and 24 kHz. We then calculated the power difference at each frequency between the recorded signal and the input signal. We recorded the signal played from the speaker with a Beyerdynamics MM-1 microphone and used a MOTU 16A as the audio interface for both the KRK Classic 5 speaker and MM-1 microphone.

3.4.9 Speaker Audio Interface and Routing

We used audio interfaces and network switches designed and built by Mark of The Unicorn (MOTU). These audio interfaces implement the IEEE Audio Visual Bridging (AVB) protocol for all-to-all routing network. The audio interfaces and network switches also ensured that all broadcast audio was phase-locked across speakers if multiple speakers played simultaneously. We constructed our routing network with two connected network switches. We connected four MOTU 24 Ao audio interfaces to one network switch and three 24Ao audio interfaces, and one MOTU 16A audio interface to the other network switch. We attached the computer controlling the routing network to the 16A interface via a thunderbolt cable, which allowed us to run up to 128 simultaneous output audio streams.

We controlled routing using the MOTU HTTP API, which requires an ethernet connection to the controlling computer, in addition to the thunderbolt connection used to send audio streams.

3.4.10 Controller Software

We built a software package to control the array and abstract the speaker routing and playback details from the experimental details. We route a speaker of interest by making an HTTP POST request from our controller computer to the 24Ao interface attached to the speaker of interest. The post request consists of a JSON string specifying the speaker to activate and the input stream for that speaker. We constructed a mapping between speaker position and the audio interface and channel serving that speaker. This approach allowed us to abstract away the networking details of the speaker routing problem and allow experiments to be described in terms of physical positions.

The controller software sets the level of playback by RMS-normalizing the incoming audio and then scaling the audio to be played at a specified volume. Sounds are normalized using the RMS calculated over the entire stimulus length and are scaled according to a linear function that we fit to measurements of the speaker's response to

yield a playback level in dBA. We obtained this function by measuring the sound level of white noise at the position of the listener played from the speaker at 0° azimuth and 0° elevation. We made the level measurements over 8 seconds using the Svan 979 sound level meter and GRAS 4AE microphone with the free field compensation filter active. We made sound level measurements ranging from 53.2 to 68.5 dBA for white noise with RMS values ranging from 0.1 to 1.0. We additionally implemented a level limiter function that prevents any speaker from playing sounds louder than 70 dBA. We chose this level to limit any danger to participants or experimenters if all speakers were to be simultaneously active at maximum volume. In this case, the total sound level would be 70 dBA + 10*log(133 speakers) = 91 dBA, which is loud (on par with a motorcycle) but not dangerous.

We play normalized sound signals through a routed speaker using the sound device library. The sound device library is a python wrapper that allows audio to be sent using the computer's native audio drivers. The library also supports simultaneous playback across an unbounded number of audio channels, which is critical for multi-source experiments and experiments rendering complex auditory scenes.

The controller software was implemented using the python programming language and ran using Microsoft Visual Studio on a Dell XPS 15 9500 computer.

3.4.11 Natural Sounds Set

We assembled a set of natural sounds intended to be representative of everyday life. We started with a list of sound types from Norman-Haignere et al. (2015) [169] that previous human subjects had rated as commonly heard in everyday life. We then searched freesound.org for a recording of each sound that had been recorded using a microphone that allowed power up to at least 15 kHz, did not have any significant reverberation, and only contained sound that could conceivably be emitted from a single object. This led to a core sound set of 121 sounds. We then supplemented this set with 24 recordings of natural objects made in a sound booth by Traer, Norman-Haignere, & McDermott (2021) [**?**] and samples from 15 different midi instruments playing middle C from the Nsynth dataset [60]. In total, our final stimulus set had

160 sounds.

We resampled all audio to 44.1 kHz and edited the sounds to 1-second. We edited the clips to maintain a natural sound onset wherever possible. In addition, we applied a 30 ms Hanning window to the beginning and end of the sound. The natural onsets in the recordings were delayed at least 25ms from the start of the recording to prevent the Hanning window from interfering with the natural onset.

3.4.12 Human Sound Localization Experiment – Experiment design and trial balancing

We aimed to present every sound in the stimulus set at every position within the speaker array. The stimulus set consisted of 160 sounds and the speaker array had 133 positions for a total of 21,280 sound/position pairs. Based on pilot experiments we estimated that each trial would take a subject approximately 5 seconds. We calculated that collecting a response at every location would take approximately 30 hours, which was impractical for a single subject. We chose instead to spread the set of sound/position pairs across 19 subjects. This had the additional advantage of averaging out effects of idiosyncrasies of a particular subject's head-related transfer functions, sound localization strategies, or biases. We split the number of trials evenly between subjects so that each subject had 1120 trials, which could be collected in approximately 2 hours, including breaks.

We assign trials to subjects by creating a list of all sound/position pairs and shuffling this list. We then assign the first 1,120 trials to subject 0, the next 1,120 trials to subject 1 and continue assigning trials this way through subject 18. We then searched across random shuffles until we found one that had an acceptable degree of uniformity across positions and sounds within the trials for each subject. Specifically, calculated the empirical marginal distributions for azimuth, elevation and sounds for a given subject and shuffle. We then calculated the probability that the empirical distributions were drawn from a uniform distribution by calculating the chi-square goodness of fit test between the empirical distribution and a uniform distribution

for positions and sounds. We accepted the shuffle if all the chi-square tests for all subjects had a probability of greater than 25% that the empirical distribution could have come from the uniform distribution. To make the acceptance criterion easier to satisfy, the distribution over azimuths was measured across nine groups of positions ranging from -90° to +90°, where eight groups spanned 20° each, and the center group spanned 30° from -10° to 10°.

3.4.13 Speaker Array – Speaker labeling and Subject Response Procedure

We labeled all speakers with a letter and number where each elevation corresponded to a letter and azimuth corresponded to a number. Elevations were labeled from top to bottom with 'A' corresponding to +40° and 'G' corresponding to -20°. Azimuths were labeled from left to right with -90° corresponding to '1' and +90° corresponding to '19'. The number was appended to the letter to form the code. For example, the speaker at -90° azimuth and +40° elevation had the code 'A1' while the speaker at 0° azimuth and 0° elevation had the code 'E10'.

We made a custom 23-key keyboard with three rows of six standard keys, one bottom row with four standard keys, and one double-width key. The first two columns of the keyboard contained the letters 'A' through 'G' in alphabetical order from top to bottom and left to right. The last key in the second column was the 'X' key, which indicates the subject did not hear any sound. We set up the next three columns as a number pad in the standard layout, with '0' located on its own in the bottom row. We placed a backspace key at the top of the rightmost column and placed a double-width enter key at the bottom right of the keyboard. The keyboard used a standard USB-C interface and had a cable connected to the computer that ran the controller software.

3.4.14 Human Sound Localization Experiment – Experimental procedure

Each participant was presented with 1,120 1-second clips of one of 160 natural sounds at one of 133 random positions at 65 dBA. For each trial, the participant fixated on the speaker at 0° elevation and 0° azimuth (the "center" speaker) and then pressed the enter key, indicating that they were ready for the next trial. The control software played a 1-second sound from one of the speakers while the subject continued to fixate. After the sound ended, the subject was allowed to orient to the position where they judged the sound to have originated. The subject then selected the label of the speaker they believed the sound to have originated from and entered it with the keyboard. The subject would then reorient their head to face the center speaker and press enter to begin the next trial.

If the subject believed that they did not hear any sound at all or they wanted to report a lapse in attention for that specific trial, they had the option of entering an 'X'. The experiment would progress to the next trial, and the marked trial would be played again at the end of the experiment. If the subject reported a response without a valid label, the experimenter would ask them what response they intended to enter and make a note of it. In practice, this was very rare, and fewer than 100 trials across all 21,280 were mistyped.

Subjects were instructed to face the center speaker before pressing the enter button, and the experimenter monitored the subject closely throughout the experiment to ensure they were properly orienting before beginning the next trial. In practice, all subjects readily understood and complied with the instructions.

Subjects performed trials in 4 blocks of 280 trials each. Each block took approximately 24 minutes, and subjects took breaks between trials that were unrestricted in time. In practice, most subjects spent around 5 minutes on each break. At the end of the experiment, trial responses were saved into a JSON file containing the metadata and responses for each trial.

3.4.15 Human Sound Localization Experiment – Analysis

Subject responses were loaded from the JSON files into a single pandas dataframe. The responses were then converted from the alphanumeric labels on the speakers into the speaker position, specified in azimuth and elevation. If the code did not correspond to a known label, the trial was discarded, although this occurred in fewer than 100 trials across all subjects.

We calculated absolute azimuth error for each trial by subtracting the true location from the judged location, both measured in degrees, and then taking the absolute value of the difference. We calculated the absolute elevation error using the same method and also calculated the signed elevation error by recording the difference in the true and judged locations without taking the absolute value.

We analyzed the patterns of absolute error in azimuth and elevation, grouped according to the ground truth position of the sound in azimuth, elevation, and the sound identity. For each graph, we grouped trials by the variable plotted on the x-axis (ground truth azimuth, elevation, or sound identity) and calculated the mean value for the error variable on the y-axis. Error bars plot the standard error of the mean for the error variable, obtained by bootstrapping over trials.

We also analyzed individual sounds for elevation bias by plotting the sound identity vs. the signed elevation error across all positions. Deviations above or below 0 denote an upward or downward elevation bias for that sound, respectively.

3.4.16 Model Comparison

We assessed the extent to which the model from Francl & McDermott (2022) could predict the pattern of errors across sound identities. First, we spatially rendered the same set of natural sounds used in the human experiments and converted them to binaural audio. Next, we used the model to make predictions about the locations of each sound. Last, we calculated the model localization error for each sound and compared that pattern of errors to human data.

3.4.17 Model Comparison – Stimulus Rendering

We used the virtual acoustic world renderer from Shinn-Cunningham et al. (2001) [198] and used in Francl & McDermott (2022) [63] to simulate binaural room impulse responses (BRIRs) for a room similar to the one for the speaker array. We simulated the room to have the same approximate geometry and materials as the actual room and rendered a set of binaural room impulse responses for a listener at the same approximate position in the room as the human subject in the localization experiment. The rendered room geometry is 6 by 4 by 3 meters (LxWxH) and simulated ceiling material with "Acoustic tiles, 0.625", 16" below ceiling", floor material "Linoleum", and wall material with "Plaster on Concrete". We simulated the listener position 2.4 meters down and 2.6 meters left from the upper left corner of the room. The renderer simulated elevations ranging from -20° to +40° in 10° increments and azimuths ranging from 0°to 355° in 5° increments. Specific details of the rendering procedure can be found in the methods section of Francl & McDermott (2022).

We used the simulated BRIRs to render the natural sounds used in the human experiment at the same position as the speakers in the speaker array for elevations 0° and higher. The current model is not capable of reporting elevations below this threshold, so all human-model comparisons were limited to the subset of speaker positions at 0° and higher.

3.4.18 Model Comparisons – Simulating Room Background Noise

The background noise of the room was simulated to align the model and human localization conditions more closely. We made recordings of the background noise at varying locations behind the lister in the room using a directional microphone. We then generated noise that was spectrally matched to that signal and spatially rendered the noise at the recorded locations to create a diffuse background field that approximated the background noise present when human subject performed their experiments.

We made recordings of the background noise using an AmiCV-2R Unidirectional microphone connected to a Scarlett 8i8 digital-to-analog converter. The digital signal was sent to a 2018 MacBook Pro running Audacity 2.2.1. We made these recordings with the microphone facing along the ray passing through the listener's head position and the position of the microphone. The microphone recordings were made 2 meters from the listener's head position and at ten source positions relative to the listener. The recording positions were (listed as (elevation, azimuth)): (40°,135°),(40°,180°), (0°,225°), (0°,135°),(0°,180°), (-20°,225°), (-20°,135°),(-20°,180°), (-20°,225°), (90°,0°).

We made spectrum-matched noise for each of these recordings by taking the Fourier transform of each recording, randomizing the phase information, and taking the inverse Fourier transform. We made 100 of these spectrally matched noise versions of each of the ten recordings. The rationale for the spectrum-matched noise was to avoid having the model's judgments influenced by details of the particular recordings we made.

We generated background noise by randomly selecting a spectrum-matched noise for each of the ten recordings and rendering each at the spatial position where it was recorded. We combined these spatially rendered spectrum-matched noises into one auditory scene and then repeated this rendering procedure 10,000 times, randomly selecting ten new spectrally matched corresponding to each of the ten recordings in each pass. These background scenes were randomly selected and combined with the natural sound stimuli used in the experiment at an SNR of 17 dB to match the level difference in the human experiment. Levels were measured using the root-mean-squared power of the audio signal.

3.4.19 Model Comparisons – Predictions

The final rendered stimuli were processed with the cochlear model described in Francl & McDermott 2022. The cochlear model had 39 frequency channels and a frequency response ranging from 30Hz to 20kHz. The model predicted the location of each source using the procedure described in Francl & McDermott 2022 (see the original paper for details).

3.4.20 Model Comparison – Analysis

Human subjects chose between positions spaced every 10 in azimuth. However, the model output layer had units corresponding to positions spaced every 5 in azimuth. We assumed that if a human judged a sound as being located exactly between two labeled locations, they would randomly select one of the two. We replicated the same strategy in our model by reassigning model judgement to the left or right bin with 50% chance if the model chose a location not available to human subjects. For example, if the model selected a position at 25 azimuth it would be assigned to either 20 azimuth or 30 azimuth with even chance between the two. In practice, this assignment means that 50% of the time the trial will have 10 error and 0 of error the other half of the time. The expected value of the error for each of these trails after reassigning the bins is 5, which is equivalent to the error we would observe if we did not reassign the stimuli to different bins.

Human Participants judged locations within the frontal hemifield so we front–back folded the model responses to enable a fair comparison. This consisted of treating each model response in the rear hemifield as though it was a response in the corresponding front hemifield. For example, the 10° and 170° azimuthal positions were considered equivalent.

We calculated error for each trial from each of the ten networks in the model using the same procedure outlined in "Human Sound Localization Experiment – Analysis". For model plots, error bars plot the standard error of the mean for the error variable bootstrapped over networks instead of trials.

3.4.21 Model Prediction of Human Behavior

We calculated the mean absolute elevation error for each of our 160 sounds and for both human subjects and the model. For human data, we calculated the average error for one sound by averaging across all subjects and across positions with elevations greater or equal to zero to match the set of positions used in the model. For the model, we calculated the average error for one sound by averaging across all rendered

positions and all networks. We generated a scatter plot with model error on the x-axis and human error on the y-axis. We calculated the Person and Spearman correlation coefficients between the vectors containing the human error and model error for each sound.

We repeated the procedure outline above for the other three error measures (absolute error in azimuth, and bias in elevation and azimuth).

3.4.22 Measuring the Reliability of Human and Model Judgments

We assessed the consistency between human subjects by measuring the split-half reliability of the human behavioral data. First, we first divided the human subject pool into two groups of eight subjects by randomly sampling from our nineteen subjects without replacement. Next, we separately calculated the absolute elevation error for each sound for the two groups. This resulted in two vectors containing the error from the first group and the error from the second group. Each vector contained one value for each of the 160 sounds in the dataset. We calculated the Person and Spearman correlation coefficients between the vectors. Last, we then converted the correlations between the split-halves into an estimate of reliability of the results from the full set of subjects by applying the Spearman-Brown correction.

We assessed the model consistency by measuring the split-half reliability across networks. First, we first divided our networks into two groups of five by randomly sampling from our ten networks without replacement. We used these two groups to calculate the reliability of the model data using the same procedure described above to calculate the reliability of the human data.

3.4.23 Spectral Flatness

We measured the spectral flatness of each natural sound using the following formula:

$$SpectralFlatness = \frac{\sqrt[n]{\prod_{n=0}^{N-1} x(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)}$$

Where x(n) is the FFT of the signal. The spectral flatness is highest when all frequencies have equal amplitudes.

3.4.24 Statistical Significance Testing

We assessed the statistical significance of Pearson and Spearman correlation coefficients by calculating the probability that they differed significantly from 0. For Pearson correlations, we compared to a null distribution of correlation coefficients for two normal and uncorrelated variables. This distribution is known as the exact distribution of r and is:

$$f(r) = \frac{(1-r^2)^{\frac{n}{2}-2}}{B(\frac{1}{2}, \frac{n}{2}-1)}$$

where B is the Beta distribution.

For Spearman correlations, we calculated a t value using the formula:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The t values from this function are distributed approximately normally with n-2 degrees of freedom under the null distribution that is calculated by permuting the ranks. The details of this approach are laid out in Kendall 1973[127]. In both cases we calculated the probability of drawing an r value from the null distribution with an absolute value at or greater than the r value we calculated for our data.

3.5 Acknowledgements

I thank Preston Hess for his help building the speaker array, writing the control software and running human participants and Ajani Stewart for his help constructing the speaker array. I would also like to thank Andy Gallant and the team at the MIT Central machine shop for their assistance designing and fabricating the speaker support truss.

3.6 Tables

Alarm Clock	Chair Rolling	Flag Flapping	Microwave Running	Spray Can Shake
Angry Shouting	Chimes In The Wind	Flute Synthetic	Motorcycle Reving	Spraying
Baby Babbling	Chop Vegtables	Footstepssand	Nail Box	Squeaky Toy
Baby Crying	Chopping Food	Frog Croaking	Newspaper Rustling	Stapler
Baby Talk	Cicadas	Frying	Organ Electronic	Stream
Basketball Bouncing	Clock Ticking	Gargling	Organ Electronic	String Acoustic
Bass Synthetic	Coffee Bean Grinder	Geese	Pepper Grinder	Swimming
Bass Synthetic	Coin Dropping	Girl Speaking	Piano	Tearing
Bees Buzzing	Coin In Vending Machine	Glass Smash	Pouring Liquid	Telephone Dialing
Bicycle Freewheeling	Computer Startup Sound	Grunting And Groaning	Pouring Water Out Of Bottle	Telephone Ringing
Bike Bell	Computer Synthesized Speech	Guitar Acoustic	Printing	Toilet Flushing
Biting And Chewing	Coughing	Guitar Acoustic	Puppy Whining	Toothbrushing
Blender	Crow	Guitar Electronic	Radio Or Tv Static	Train Warning Bell
Boiling Water	Crumpling Paper	Guitar	Ratchet	Train Whistle
Boy Speaking	Crying	Hair Dryer	Reed Acoustic	Trimming Branch
Brass Acoustic	Cymbal Crash	Hammer Metal	Ringtone	Truck Beeping While Backing Up
Breathing	Dial Tone	Hammering	Rubbing Hands	Typing
Busy Signal	Dishes Clanking	Hatchet	Running Up Stairs	Vacuum
Camera Snapping Photos	Dog Barking	Heartbeat	Running Water	Velcro
Car Accelerating	Dog Drinking	Horror Film Sound Effects	Rustle	Violin
Car Alarm	Dog Panting	Humming	Sawing Wood	Walking On Gravel
Car Engine Starting	Door Knocking	Kettle Whistling	Saxophone Acoustic	Walking On Hard Surface
Car Horn	Doorbell	Keyboard Electronic	School Bell	Walking With Heels
Car Power Windows	Dripping Water	Keyboard Electronic	Scissors	Water Splashing
Car Skidding	Dropping Stones	Keyboard Synthetic	Scissors Cutting Paper	Whispering
Cartoon Sound Effects	Drum Roll	Keys Jingling	Scratching	Whistle
Cash Register	Duck Quack	Laughing	Scream	Whistling
Cat Meowing	Electric Bass	Lawnmower	Shovel	Wind
Cat Purring	Electric Can Opener	Leaf Blower	Shuffling Cards	Wings Flapping
Cello	Electric Hand Drill	Mallet	Siren	Woman Speaking
Cellphone Vibrating	Electric Shaver	Mallet Acoustic	Songbird	Writing On Paper
Chainsaw	Fingernail Tapping	Man Speaking	Sports Anouncer	Zipper

Table 3.1: List of 160 natural sounds, ordered alphabetically

			Spectral Flatness Partial Correlations	
Metric	Pearson Correlation	Spearman Correlation	Pearson Correlation	Spearman Correlation
Abs. Elevation Error	0.722	0.678	0.372	0.351
Abs. Azimuth Error	0.603	0.542	0.310	0.298
Elevation Bias	-0.398	-0.306	0.204	0.032
Azimuth Bias	-0.006	0.021	-0.004	0.017

Table 3.2: The two leftmost columns display the correlation coefficients between the human and model for four different error metrics evaluated on each of the 160 natural sounds: absolute elevation error, absolute azimuth error, elevation bias, azimuth bias. The two rightmost columns display the human-model partial correlation coefficients when controlling for spectral flatness.

Metric	Person Correlation	Spearman Correlation
Abs. Elevation Error	-0.675	-0.689
Abs. Azimuth Error	-0.560	-0.536
Elevation Bias	-0.626	-0.542
Azimuth Bias	0.031	0.093

Table 3.3: The table displays the correlations between the spectral flatness of a sound and the value for that sound of one of four error metrics computed from human responses: absolute elevation error, absoluerror, elevation bias, azimuth bias.

	Human Split-Half Reliability		**Model Split-Half Reliability**		**Noise Corrected Correlation Coefficients**		**Noise Corrected R^2**	
Metric	Pearson Correlation	Spearman Correlation	Pearson Correlation	Spearman Correlation	Pearson Correlation	Spearman Correlation	Pearson Correlation	Spearman Correlation
Abs. Elevation Error	0.895	0.807	0.993	0.988	0.766	0.759	0.587	0.577
Abs. Azimuth Error	0.756	0.682	0.989	0.982	0.697	0.662	0.486	0.439
Elevation Bias	0.707	0.550	0.941	0.936	-0.490	-0.427	0.240	0.182
Azimuth Bias	0.03	0.048	-0.42	0.405	-0.051	0.15	.003	.023

(Caption on next page.)

Table 3.4: The first four columns display the split-half reliability of the human and model data, each of which determines the noise ceiling for its respective correlation. The reported correlations are the mean value from 10,000 bootstrap samples across subjects and are Spearman-Brown corrected. The last four columns use the split-half reliability to calculate the noise-corrected correlation coeffects, which adjust for the maximum correlation possible for that comparison (the correction for attenuation). This adjustment involves dividing the human-model correlation by the appropriate reliability for that comparison:

$$\mathrm{R}_{Corrected} = \frac{R_{Human-Model}}{\sqrt[2]{R_{Humansplit-half} \times R_{ModelSplit-half}}}$$

$$\mathrm{R}^2_{Corrected} = \frac{(R_{Human-Model})^2}{R_{Humansplit-half} \times R_{ModelSplit-half}}$$

Chapter 4

Self-Supervised Models of Human Sound Localization

4.1 Introduction

In humans, fine-grained sound localization does not appear to be innate. Although human newborns will correctly orient left or right if a sound is presented directly to their left or right [226, 168], they show a marked increase in sound localization acuity over the first 18 months of life [6, 43, 142, 167]. Infants have a smaller head size compared to adults, which makes binaural cues to localization less pronounced, but this difference is insufficient to explain the inability to localize sounds with high acuity [42]. These findings suggest that sound localization is not fully innate and is likely learned through experience. Experiments in adults have further demonstrated that pinna-specific cues of human sound localization can be relearned over a month or so [105], suggesting that humans continue to update sound localization strategies over the course of their life, integrating new experiences as they occur.

Recent deep-learning-based models of human perception [237, 126, 187, 136, 181], including recent models of human sound localization [63, 134], rely on supervised learning. Supervised learning relies on access to the ground truth labels for a large training data set. The models obtained via supervised learning provide examples of models optimized for a problem that may be fruitfully compared to human observers

[63]. But because humans do not have direct access to ground truth labels in the quantities typically used to train deep neural networks, this learning paradigm is likely implausible as a model of human learning in many domains, including sound localization.

We set out to evaluate a more biologically plausible learning mechanism for sound localization that relies exclusively on data available to humans. Vision could in principle provide a supervisory cue, providing the ground-truth location of a sound source if a listener could determine the sound identity and find the object visually. However, this would require listeners to be able to associate auditory and visual objects before learning how to localize. Additionally, congenitally blind individuals localize sounds with similar acuity to sighted individuals [10], which suggests that vision is not critical to learning to localize sounds. It occurred to us that one plausible alternative could involve head movements.

We hypothesized that head movements could provide a self-supervision signal that could be used to learn representations of sound location. Purkinje cells use cues from the vestibular system to calculate an internal representation of absolute head position [235], and it seemed plausible that knowledge of head orientation could be present early in development. This head orientation information could plausibly be used to learn representations of locations by leveraging the fact that sounds in the world tend to have stable locations over short timescales, such that the head-relative locations of sounds tend to change across head movements by an amount proportional to the head movement. We attempted to learn a representation of binaural audio for which representations from different excerpts of binaural audio were similar if the audio excerpts came from the same auditory scene heard at similar head positions. We learned this representation using a modified contrastive learning [37, 36, 81] approach with a loss function that tries to maximize the similarity between representations of binaural audio from the same scene and head position and otherwise minimize similarity. If successful, the examples from the same auditory scene and similar head positions will cluster in the representational space, potentially with separable representations of location. We assessed whether the learned representation could support sound

localization by fitting a linear decoder to the final learned representation. This final step requires classic supervision for fitting the decoder, but can be achieved with less labeled data than is needed to train a full deep neural network. If a biological system could learn a representation of location with self-supervision, it seems possible that it could then learn an analogous linear decoder with a modest amount of supervision, as might be obtained from interacting with the world. We found that the model performed well above chance, with an average of less than 10° of azimuthal localization error, which is approximately ± 6 inches for an object 6 feet away.

4.2 Results

We trained our model using a modified version of noise contrastive learning [37, 81]. In standard noise contrastive learning, a single image is augmented twice to create many pairs consisting of two augmented versions of the same image. The model then learns to maximize the similarity between the representations for these pairs of images [37, 36]. Our model used cochlear representations of binaural audio instead of images, with head motion replacing image augmentation. We also used many auditory scenes with different sets of sound in each scene. Our model was trained to maximize the similarity between representations of the same scene for similar head positions (Fig 4-1a). Critically, the model only had access to information about the head orientation associated with the binaural audio and did not have access to the locations of the sounds in the auditory scene.

We trained our model using a set of 5,000 auditory scenes, each containing between two and eight natural sounds. Each scene was rendered at 256 random head positions. We then used a linear decoder to map the learned representation to sound localization judgments (Fig 4-1b). Finally, we evaluated the model using a validation set of single sound sources not used in either training phase.

We quantified performance as the error in the azimuthal dimension, for which spatial resolution is well characterized in humans. The model localized sounds substantially better than chance (Fig 4-2) and showed an increase in azimuthal error away

from the midline, as is seen in humans. However, the model also showed a substantial increase in absolute azimuthal error as the source elevation increased. This could in principle be partially explained by the decreasing distance per degree azimuth at higher elevations (as sources move closer to 90 ° elevation, they converge to the same location). However, the observed errors exceed what would be expected from this effect. At 40° elevation, the radius of a circle would be cos(40°)=0.76 times the radius of the original circle. Thus, if errors were a constant size in cm, we would expect that the error at 40° elevation would be: (error at 0°)/0.76=(6.07° error)/0.76=7.92°. The actual azimuthal error at 40° elevation is closer to 16° error, suggesting that the results cannot be accounted for based on decreased distances between degrees at higher elevations alone.

To compare to previous work using supervised learning, we evaluated the model from Francl & McDermott (2022) [63] on the same validation set used for our self-supervised model. We found that the supervised model consistently outperformed the self-supervised model (Fig 4-2). The supervised model also showed a tendency to localize less accurately at higher elevations. However, this effect was much smaller than that in the self-supervised model (the supervised model error was 59% larger at 40° elevation than at 0° elevation, while the self-supervised model error was 272% larger at 40° than at 0° elevation). This suggests a limitation of the current self-supervised approach.

4.3 Discussion

Our current results provide a proof of concept of a biologically inspired self-supervised learning rule for sound localization. We found that providing information about head orientation with binaural audio was sufficient to learn a localization strategy that performed well above chance. However, this strategy still performed below the accuracy of supervised approaches.

This gap could likely be partially closed by systematically exploring variations across many aspects of the architecture, training procedure, and loss functions used

in this work. Prior work in contrastive learning has found that the accuracy of the final representation depends on many aspects of the architecture, embedding network, and update rules [37, 36, 81]. At the time of training, we did not have access to the computational resources necessary to pursue a systematic search over these hyper-parameters, and it is likely that the architecture we chose is sub-optimal in some respects. It is also likely that modifying the loss function could improve the model. In the current approach, we modified a discrete loss function that had been successfully used in past contrastive learning models [37, 36, 81]. To apply this type of loss function to our problem, we discretized our data by choosing a distance threshold and labeling pairs below the threshold as positive and above the threshold as negative. We also introduced a weighting for each positive pair that increased the loss for pairs with more similar head orientations to better align the continuous nature of the distance metric with the loss function. However, other more biologically plausible loss functions are possible [132]. For instance, a smoother loss function (such as squared error or an inverted gaussian) might be more appropriate for incorporating the distance between head orientations. Such loss functions could result in improved models of self-supervised sound localization, potentially eventually explaining human sound localization accuracy without extensive supervision.

4.4 Methods

4.4.1 Building a set of natural sounds

We started with the set of natural sound source recordings used in Francl & McDermott (2022) [63] and initially assembled in Norman-Haignere (2015) [169]. We excluded recordings shorter than 10 seconds to minimize the overlap between multiple 1-second excerpts of the sound that were to be used in training. This yielded 222 source recordings.

4.4.2 Rendering Binaural Room Impulse Responses

We used the virtual acoustic world simulator developed in Shinn-Cunningham et al. (2001) [198] and used by Francl McDermott (2022) to render the binaural room impulse responses (BRIRs) for the same five rooms, specified in Francl & McDermott (2022) [63]. Each pair (one impulse response per ear) of rendered BRIRs was associated with a specific location. To spatialize a sound, we used a pair of BRIRs to modify the sound so that the sound contained the cues a listener would experience with that sound at a particular location. We rendered BRIRs for positions every 5 degrees in azimuth from 0° to 355° and every 5° in elevation from -40° to 90°. This spacing was slightly finer in azimuth and elevation than the original set of head-related impulse responses (HRIRs); the renderer interpolated between positions to achieve this spacing.

4.4.3 Rendering Auditory Scenes

For model training, we rendered five thousand auditory scenes with 256 head positions for each. To render an auditory scene, we first sampled the number of sources from a uniform distribution, ranging between two and eight sources per scene. Next, we randomly sampled a sound file from each source, without replacement, from a set of 222 sounds (described above). We assigned each source to a position in space by drawing a random sample from a discrete uniform distribution corresponding to the set of possible source positions (i.e., those for which we had rendered BRIRs). We term the resulting auditory scenes the "base" auditory scenes. Our renderer supported spatializing sounds at locations between -40° and 90° elevation. We constrained the initial sound positions so that we could later make vertical head movements of up to 40° in elevation without sources moving to positions not supported by our renderer. This constraint meant that the set of valid sound positions in the base condition were locations at or above 0° elevation. We chose this limit in the base condition to provide a 40° margin between the sound and the lower vertical limit of positions that we could render, which is -40° elevation. This margin allowed us to make head

movements up to 40° up or down without sources moving to elevations that were not supported by our renderer. We used this sampling approach to generate a base auditory scene where a variable number of sources were rendered at random positions. We next simulated 256 head movements for each of the base auditory scenes. For each head movement, we sampled a random azimuth change between -180° and 180° and an elevation change between -40° and +40°. This limit on vertical head movement ensured that head motion would not cause a source to move to a position where we did not have a BRIR to render that source. We calculated the new positions of all sources after each head movement and rendered each source at its new position. For each movement, we also randomly sampled a one-second excerpt of each sound file that was distinct from that used in the base auditory scene. Auditory scenes from different head positions thus contained the same subset of sources with the same spatial relationship but did not use the exact same sound segment for a source.

4.4.4 Model Training – Overview

We trained the model to learn a representational space that maximized similarity between representations of the same auditory scene when the listener's head orientation was similar and that minimized similarity between representations of different auditory scenes or those for which the head orientation was very different. We trained the model on batches of binaural audio examples from multiple auditory scenes, with multiple head orientations for each auditory scene. We trained the model by optimizing the representation to increase the similarity between example pairs from the same auditory scene and similar head positions. The model was trained to decrease similarity in all other cases. We created a loss function for this purpose by calculating the pairwise similarity between the model representations of the examples (see below for the equation). The final model was trained for 207,000 steps of gradient descent, which took approximately three weeks using all available resources on an Nvidia DGX-1. We used a base learning rate of 0.015 and decayed the learning rate according to a cosine decay schedule as in Gordon (2020) [81].

4.4.5 Model Training – Contrastive Learning

We trained the model using noise contrastive estimation, similar to the procedure used in SimCLR [36] and MoCo [37], which seek to learn a representation of input data where augmented examples of the same object are close in the learned representational space, and all other examples are maximally separated. We used the same approach and learning rules but made two modifications. First, we augmented our examples through head rotation rather than modifying examples using unnatural transformations like cropping and color rotation (which is the standard approach in contrastive learning [36, 37, 84]). Second, we modified the loss function to identify similar examples using head orientation rather than image identity. We used momentum contrastive learning [37, 81] in which two separate networks with identical architectures each encode similarity for one example in each pair. This learning method is essentially a hack that approximates some of the benefits of larger batch sizes that are prohibitive for standard gradient descent learning given the memory constraints of current hardware. The two networks differ only in their weight update rule. One network is updated using standard gradient descent via backpropagation of error. We refer to this network as the encoding network. Because this network must compute gradients with respect to its parameters, it is constrained to a small batch size. The second network is updated via a momentum update rule. This network is known as the momentum encoder network. A momentum update linearly combines the the current weight from the momentum network with the equivalent weight from the encoder network using the following update rule:

$$\theta_m^{T+1} = w\theta_m^T + (1-w)\theta_e^T \tag{4.1}$$

Where θ_m is a network weight from the momentum encoder network, θ_e is a network weight from the encoder network, T is the training iteration time step, and m is the momentum weight. In this work we used a momentum weight of 0.999 as suggested in [96]. The momentum network does not involve explicit gradient updates. Pairwise comparisons within the loss function were always made between one example

from the encoder network and one example from the momentum encoder network, grouped into mini-batches. For the encoding network, each mini-batch contained 16 auditory scenes, each rendered at 8 different head positions. For the momentum network, each mini-batch contained the same 16 auditory scenes that were used for the encoding network, but each rendered at 64 head positions that were distinct from both each other and from the 8 head positions used in the encoding network examples. Each mini-batch thus contained 128 examples for the encoder network and 1024 examples for the momentum encoder network. This resulted in 512 intra-scene comparisons per mini-batch (8 x 64), and 7,680 comparisons across scenes (8 x 15 x 64). We processed more examples using the momentum encoder network to provide more pairwise comparisons per training step without significantly increasing the memory requirements to train the model. The momentum encoder requires less memory per example because it does not perform gradient updates and consequently does not have to save all network activations to calculate those gradients.

4.4.6 Model Training -Architecture

The model architecture used for both the encoder and momentum networks was based on a ResNet-50 architecture that was modified to have only two input channels instead of three input channels, corresponding to the cochlear representations of the left and right audio channels instead of the RGB channels of an image. The ResNet-50 transformed the binaural cochlear input into a vector with 2048 elements, which corresponded to the final fully connected layer in a standard ResNet-50 architecture. We appended a multi-layer perceptron with two hidden layers to the ResNet-50 architecture to project the 2048-dimensional vector representation to a 64-dimensional vector. This second embedding step has been shown to improve performance in recent contrastive learning approaches [36, 37], perhaps for the same reason that standard supervised network architectures often transform a higher-dimensional representation to something lower-dimensional when performing classification tasks. We used a 64-dimensional embedding to reduce memory requirements for calculating the loss and based on recent results [37] showing that output dimensionalities greater than 32

have little added benefit on final performance. After training, the 2048-dimensional vector (the "primary" representation) was evaluated for information about sound location (motivated by prior work using this approach, which found better read-out performance when using the higher dimensional representation).

4.4.7 Model Training – Model Input

The final rendered stimuli were processed with the cochlear model described in Francl & McDermott 2022. The cochlear model had 39 frequency channels and a frequency response ranging from 30Hz to 20kHz. The output of this cochlear model was the input to the ResNet-50 described above.

4.4.8 Model Training – Loss Function

The loss function was a modification of the standard noise contrastive loss function, known as InfoNCE [36, 215], computed on "positive" and "negative" pairs of training examples. This standard loss function is given by the following equation:

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)} \tag{4.2}$$

Here q is an encoder network representation, k+ is a momentum encoder representation of a positive (similar) example, and k are momentum encoder representations of negative (dissimilar) examples. The positive examples would typically be augmentations (e.g. croppings) of the same image, and negative examples would be augmentations of different images. is a temperature hyper-parameter. In essence, this function is a softmax cross-entropy across similarities between representations where each pair of vectors is given a binary label of similar or dissimilar.

We modified this standard loss function to weight positive examples (generated from the same auditory scene) based on the similarity of the head orientation with which they were rendered, with examples from closer head orientations receiving a

higher weighting. This forced the gradient updates to prioritize maximizing similarity between examples that had very close head orientations. We operationalized the distance between two head orientations by measuring the rotation angle θ from the rotation matrix between the head orientations in a pair. This was the shortest rotation along any axis that would move from one head position to another. We labeled all example pairs from the same auditory scene with a rotational distance of less than 15° as positive example pairs and all others as negative example pairs. We weighted positive example pairs based on the rotation angle between them so that example pairs with smaller rotational angles contributed more heavily to the loss function:

$$Loss = \sum_{i \in q,k^+} lossweight_i \times L_i \tag{4.3}$$

$$lossweight = global_weight(1 - (rotationaldistance)/threshold) \tag{4.4}$$

We added a global weight constant to scale the loss and preserve the average magnitude of the gradients during training. Without the global weight, the loss weight was always less than 1, which decreased the size of the final loss value and decreased the magnitude of the gradients, slowing training. We empirically derived the global weight by calculating the average loss over the first 100 batches with and without applying loss weights. We divided the average loss without the loss weights by the average loss with the loss weights, and used the resulting value as our global weight.

4.4.9 Model Evaluation – Overview

After model training was completed, we discarded the multi-layer perceptron embedding network and fit a linear classifier to the 2048-dimensional primary representation layer. This linear decoder used the primary representation layer to make judgments about the absolute location of a single source rendered in an auditory scene.

4.4.10 Linear Readout – Data generation

The linear readout classifier used to evaluate the model representations was fit using a separate set of training data. We used the same rendering procedure to generate this training data as that described for model training, with three notable differences. First, the dataset used to fit the linear classifier always contained only a single rendered source. Second, we used a new set of sounds. Specifically, we used a subset of the GISE-51 dataset [234], selected to ensure that the sound recordings contained power at high frequencies such that they would support the 3D localization cues that the model might have learned during training. We evaluated this by measuring the average power between 500-4000Hz and 8000-10000Hz using Welch's spectral density estimate. If the absolute value difference between the power in the two bands exceeded 25dB, we excluded the sound. This was designed to eliminate sounds that were recorded with a sampling rate of less than 16kHz. The final classifier training set had 2,163 total sounds, and the validation set had 390 total sounds. Third, we rendered examples ranging between 0° and 60° elevation (in 10° steps) and from 0° to 355° (in 5° steps) in azimuth. This resulted in 504 possible source locations. We rendered 500,000 training examples and 20,000 validation examples.

4.4.11 Linear Readout – Fitting and Evaluation

The linear readout classifier mapped the 2048-dimensional representation vector to 504 sound location output classes. We fit the linear readout classifier using stochastic gradient descent with batch sizes of 256 examples and performed 100 epochs of gradient descent. We used an initial learning rate of 30 and decayed the learning rate by a factor of 10 at 60 epochs and again at 80 epochs. To evaluate the trained representation, we fixed the weights of the linear readout and recorded predictions for each response from the validation set. To remove front/back confusions, we front–back folded the model responses as described in Francl McDermott (2022). This consisted of treating each model response in the rear hemifield as though it was a response in the corresponding front hemifield. For example, the 10° and 170° azimuthal positions

were considered equivalent.

4.4.12 Supervised Model Baseline

We calculated the error for the supervised model baseline using the model from Francl McDermott (2022) [63]. We recorded model predictions on the same validation set (with 20,000 examples) that was used to evaluate the primary learned representation above. We determined chance performance by calculating the average distance between two random points chosen from a uniform interval between 0 and 180 degrees. These points correspond to a randomly chosen azimuthal source location and model response.

4.5 Acknowledgements

I thank Cesar Duran for his help in implementing the model and refining the learning rule.

4.6 Figures

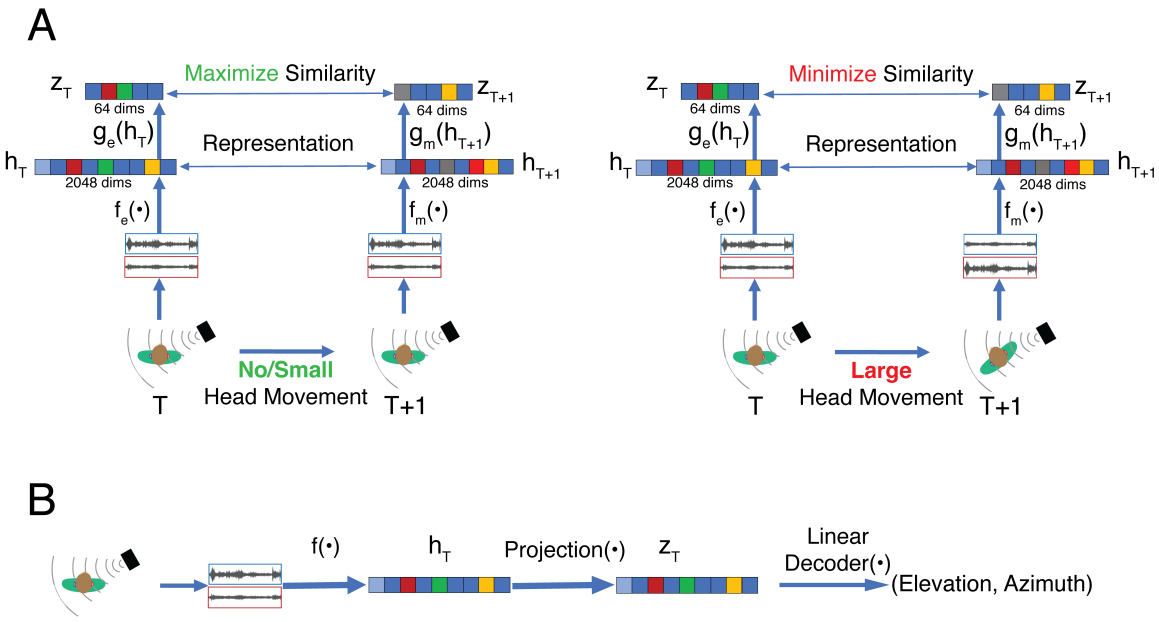
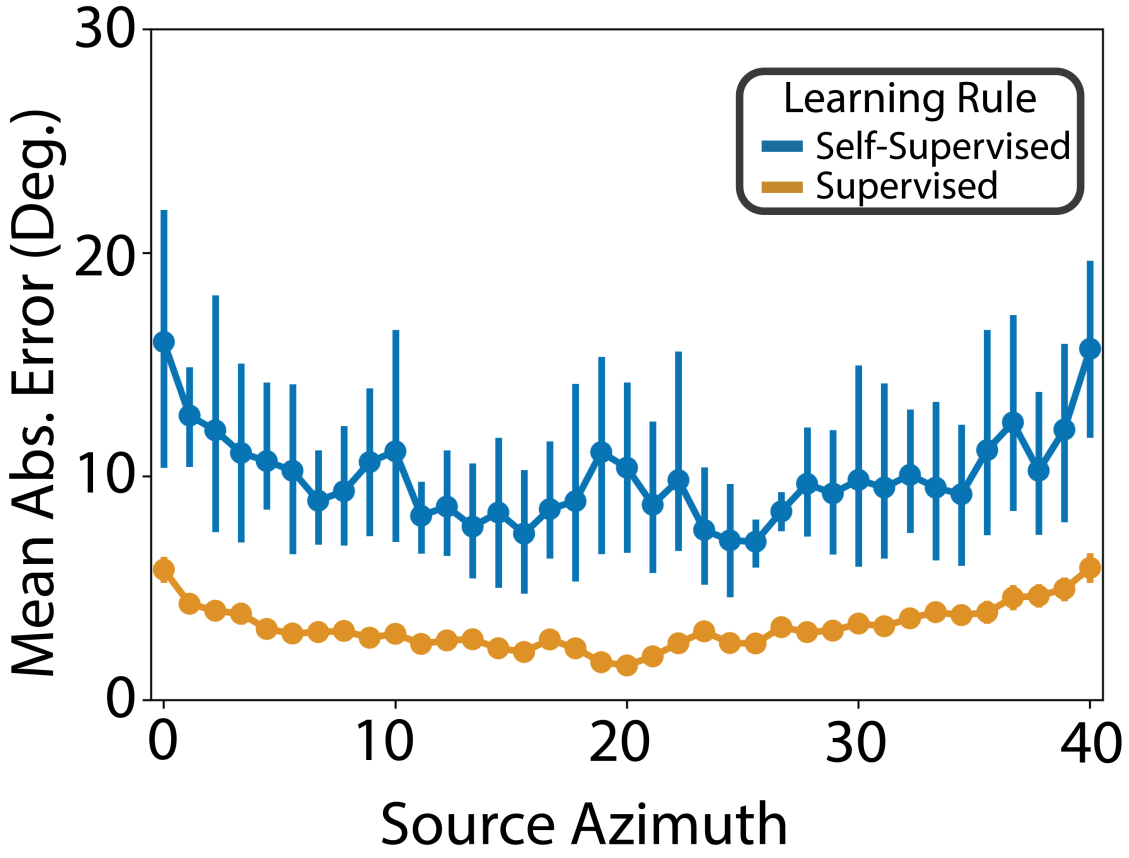
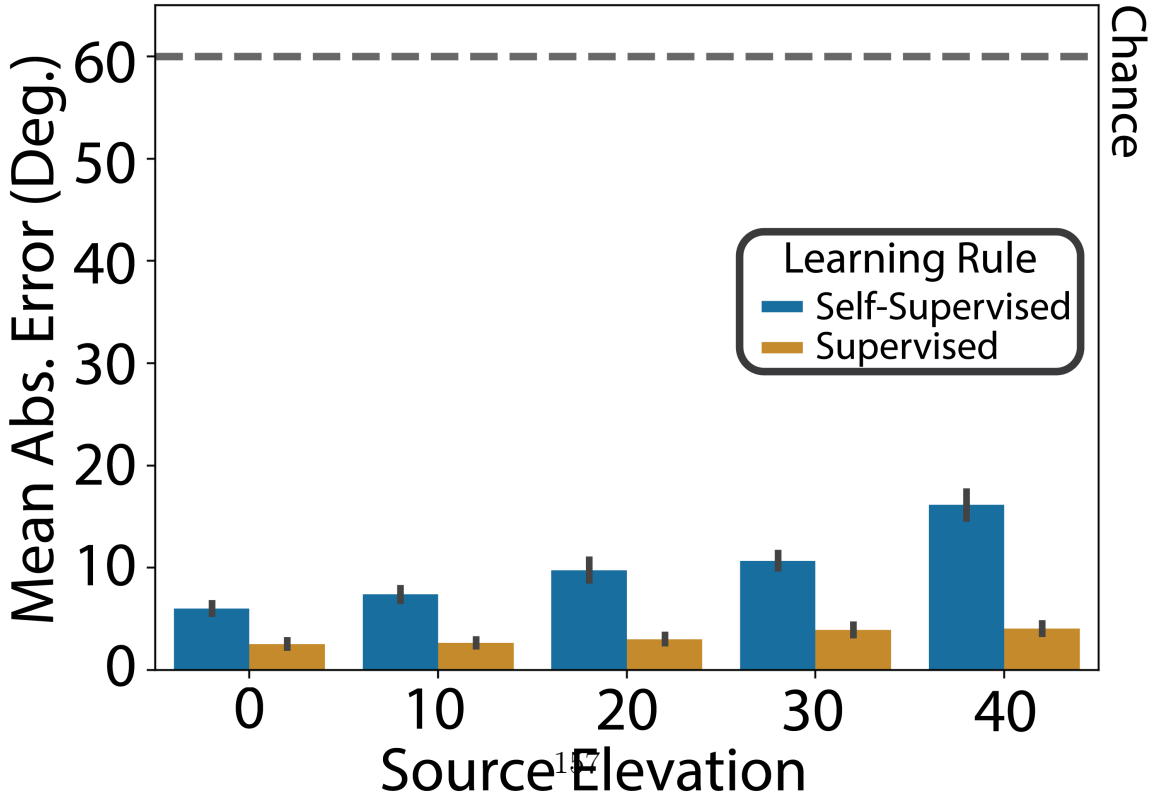


Figure 4-1: A. Schematic of Learning Procedure. The learning procedure tries to maximize the similarity between representations of binaural audio if a listener makes little or no head movement between time points. If the listener makes a large head movement, the learning procedure tried to minimizes similarity between representations. T and T+1 represent sequential points in time. f(•) is the network that transforms the binaural audio to the primary 2048-dimensional representation, labeled as h. g(h) is the embedding network that transforms the primary representation into the 64-dimensional similarity embedding, labeled as z. The e and m subscripts identify the encoding and momentum encoding networks, respectively. B. Schematic of Linear Decoder. The linear decoder maps the primary representation to a single position, specified by an elevation and azimuth.


Learning Rule
Self-Supervised
Supervised
Mean Abs. Error (Deg.)
0
10
20
30
0
10
20
30
40
Source Azimuth


Chance
Learning Rule
Self-Supervised
Supervised
Mean Abs. Error (Deg.)
0
10
20
30
40
50
60
0
10
20
30
40
Source Elevation

(Caption on next page.)

Figure 4-2: Localization Accuracy of the Self-Supervised Model. The graphs plots the absolute azimuthal error between the true and predicted sound location as a function of the true source position in either azimuth or elevation. The self-supervised results are derived from the predictions of a linear decoder operating on the learned primary representation. The supervised results are derived from the predictions from the model built in Francl McDermott (2022). Error bars in both conditions are SEM bootstrapped across trials. The dashed line at the top shows chance performance (the expected error between two positions randomly drawn from a uniform distribution).

Chapter 5

Speech Denoising

Abstract

Contemporary speech enhancement predominantly relies on audio transforms that are trained to reconstruct a clean speech waveform. The development of high-performing neural network sound recognition systems has raised the possibility of using deep feature representations as 'perceptual' losses with which to train denoising systems. We explored their utility by first training deep neural networks to classify either spoken words or environmental sounds from audio. We then trained an audio transform to map noisy speech to an audio waveform that minimized the difference in the deep feature representations between the output audio and the corresponding clean audio. The resulting transforms removed noise substantially better than baseline methods trained to reconstruct clean waveforms, and also outperformed previous methods using deep feature losses. However, a similar benefit was obtained simply by using losses derived from the filter bank inputs to the deep networks. The results show that deep features can guide speech enhancement, but suggest that they do not yet outperform simple alternatives that do not involve learned features.

Index Terms: speech enhancement, denoising, deep neural networks, cochlear model, perceptual metrics

5.1 Introduction

Recent advances in speech enhancement have been driven by neural network models trained to reconstruct speech sample-by-sample [173, 174, 222, 178, 185, 147, 172, 108]. These methods provide substantial benefits over previous approaches, but nonetheless leave room for improvement. The resulting processed speech usually contains audible

artifacts, and noise removal is usually incomplete at lower SNRs.

A parallel line of work has explored the use of deep artificial neural networks as models of sensory systems [236, 125]. Although substantial discrepancies remain [180, 61], such trained neural networks currently provide the best predictive models of brain responses and behavior in both the visual and auditory systems [236, 126]. The apparent similarities between deep supervised feature representations and representations in the brain raises the possibility that such representations could be used as perceptual metrics. Such metrics have been successfully employed in image processing [243], but are not widely used in audio applications.

Deep feature losses for denoising were previously proposed in [75, 177, 206, 109, 121], but were explored only for relatively high signal-to-noise ratios (SNRs), a single task and network, or were not compared to baseline methods using the same transform architecture. Additionally, direct comparisons have not been made to simpler losses derived from conventional filter banks. It was thus unclear the extent to which deep feature losses could improve on simpler approaches, and what choices in the feature training would produce the best results. The goal of this paper was to directly compare deep perceptual losses to alternative losses, and to explore the conditions in which benefits might be achieved. We found that deep feature losses produced more natural denoising compared to waveform losses, but that a similar benefit could be achieved using a loss derived from standard filter bank representations.

5.2 Methods

There were two components to our denoising approach (Figure 5-1). The first component was a recognition network trained to recognize either speech or environmental sounds. Once trained, this network was used to define deep feature losses. Speech recognition is a natural choice in this context, but it also seemed plausible that more general-purpose audio features learned for environmental sound recognition might help to achieve natural-sounding audio even in speech applications. The input to the network was the output of a filter bank modeled on the human cochlea.

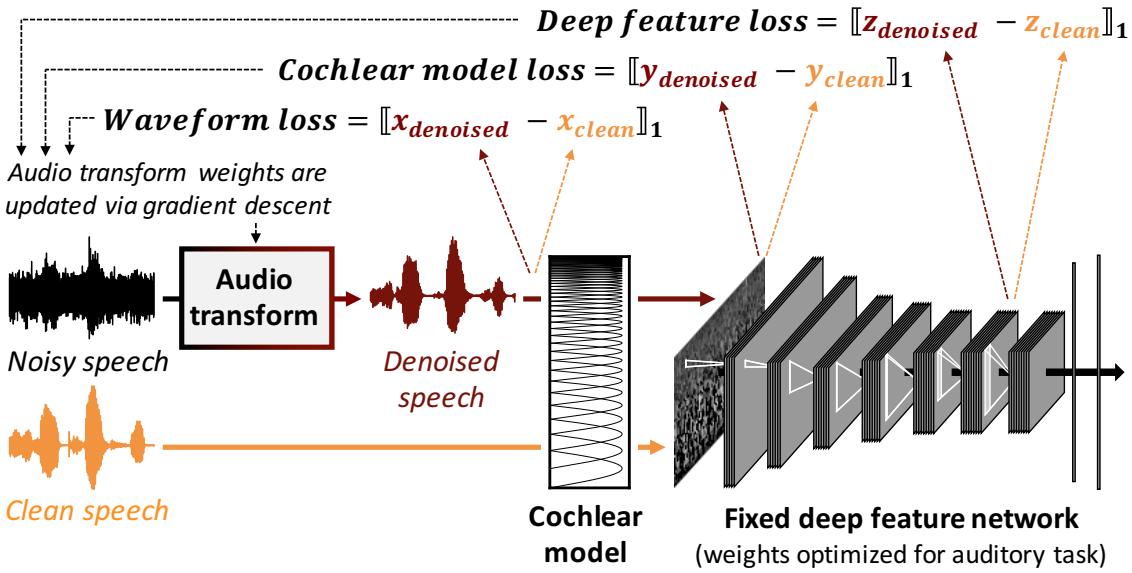


Figure 5-1: Schematic of audio transform training.

The second component was a waveform-to-waveform audio transform whose parameters were adjusted via gradient descent to minimize a loss function (evaluated on features of the recognition network, or the outputs of a filter bank, or on the waveform). We used a Wave-U-Net [205], which has been found to perform comparably to WaveNet [149] based on objective metrics of noise reduction, but which can be specified with many fewer parameters and run with a much lower memory footprint. Code, models, and audio examples are available at: http://mcdermottlab.mit.edu/denoising/demo.html.

5.2.1 Recognition Networks

The recognition networks took as input simulated cochlear representations of 2s sound clips (audio sampled at 20 kHz). The cochlear model consisted of a bank of 40 bandpass filters whose frequency tuning mimics that of the human ear (evenly spaced on an Equivalent Rectangular Bandwidth scale [78]), followed by half-wave rectification, downsampling to 10 kHz, and 0.3 power compression [159].

Recognition Network Architectures

We used three feed-forward CNN architectures for the recognition networks. Each consisted of stages of convolution, rectification, batch normalization, and weighted average pooling with a hanning kernel to minimize aliasing [100, 61]. The three architectures were selected based on word recognition task performance from 3097 randomly-generated architectures varying in number of convolutional layers (from 4 to 8), size and shape of convolutional kernels, and extent of pooling. The selected architectures had 6 (arch1) or 7 (arch2,3) convolutional layers.

Recognition Network Training

The recognition networks were trained to perform either word recognition or environmental sound recognition. For the speech task, each training example was a speech excerpt (from the Wall Street Journal [175] or Spoken Wikipedia Corpora [131]). The task was to recognize the word overlapping with the center of the clip [126, 61] (out of 793 word classes sourced from 432 unique speakers, with 230,357 unique clips in the training set and 40,651 segments in the validation set). For the environmental sound recognition task, each training example was a non-speech YouTube soundtrack excerpt (from a subset of 718,625 AudioSet examples [74]), and the task was to predict the AudioSet labels (spanning 516 categories in our dataset).

The three network architectures were trained on each task until performance on the validation set task plateaued. Word task classification accuracies for the three architectures were: arch1 = 90.4%, arch2 = 88.5%, and arch3 = 80.6%. AudioSet task AUC values were: arch1 = 0.845, arch2 = 0.861, and arch3 = 0.869.

5.2.2 Audio Transforms

Wave-U-Net Architecture

The Wave-U-Net architecture was the same as in [149]: 12 layers in the contracting path, a 1-layer bottleneck, and 12 layers in the expanding path. All layers utilized 1D convolutions with learned filters and LeakyReLU activation functions. There were 24

filters in the first layer, and the number of filters increased by a factor of 2 with each successive layer prior to the bottleneck.

Deep Feature Losses

The recognition networks were used to define a deep feature loss function as the $L1$ distance between network representations of noisy speech and clean speech. The total loss for a single recognition network and single training example was the sum of the $L1$ distances between the noisy speech and clean speech activations for each convolutional layer, weighted to approximately balance the contribution of each layer.

Cochlear Model Losses

We also trained transforms using losses derived from the cochlear model that provided input to the recognition network, as well as variants of the model that varied in i) the number of filters (5, 10, 20, 40, 80 and 160 filters, evenly spaced on an ERB-scale [78], with bandwidths scaled to tile the spectrum in all cases), ii) the dependence of filter bandwidth on frequency (linearly-spaced and 'reversed', with broad low-frequency filters and narrow high-frequency filters, opposite to what is found in the ear), and iii) in their phase invariance (subband envelopes computed by lowpass-filtering the rectified subbands; cutoff of 100 Hz).

Wave-U-Net Training

Out of concern that the audio transform might overfit to idiosyncrasies of any individual recognition network, we trained some transforms on losses computed simultaneously from an ensemble of three different networks (arch1,2,3), and some on just a single network (arch1).

In all cases the Wave-U-Net was trained on speech superimposed on non-speech AudioSet excerpts (the same corpora used to train the recognition networks) with SNR drawn uniformly from $[-20, +10]$ dB. AudioSet excerpts were used as the training 'noise' as they were highly varied and diverse. All Wave-U-Net models were trained with the ADAM optimizer for 600,000 steps (batch size=8, learning rate=10^{-4}).

Baselines

We used two baseline models, both trained to explicitly reconstruct clean speech waveforms from noisy speech waveforms drawn from the same training set described above. The first was a previously described WaveNet [185] and the second was the Wave-U-Net [149] used with the deep feature and filter losses.

We also compared our results to those of a previously published denoising transform trained with a deep feature loss [75], using both the pre-trained model made available by Germain et al. and a Wave-U-Net that we trained on our dataset using the feature loss from [75] (deep network features trained on the DCASE 2016 [163] environmental sound challenge).

5.2.3 Evaluation

We evaluated the trained models on 40 speech excerpts (from a separate validation set) superimposed separately on each of four types of noise signals: speech-shaped Gaussian noise, auditory scenes from the DCASE 2013 dataset [76], instrumental music from the Million Song Dataset [13], and 8-speaker babble made from public-domain audiobooks (librivox.org). These noise sources were chosen to be distinct from those in the training set, and to span a variety of noise types to assess the generality of the trained transforms.

Human Perceptual Evaluation and Objective Metrics

We evaluated the audio transforms by conducting perceptual experiments on Amazon Mechanical Turk. Participants first completed a screening task to help ensure that they were wearing headphones or earphones [232]. The participants who passed this screening task then rated the naturalness of a set of processed speech signals, presented seven at a time in a MUSHRA-like paradigm. Listeners could listen to each clip as many times as they wished and then gave each a numerical rating on a scale of 1-7. Listeners were provided with anchors corresponding to the ends of the rating scale (1 and 7). The anchor at the high end was always the original clean speech.

The low-end anchor was 4-bit-quantized speech (an example of very high distortion). To help ensure that participants were using the scale as instructed, each experiment included 3 catch trials where two of the stimuli were the two anchors. In order to be included in the analysis, participants had to rate all instances of the high and low anchors as 7 and 1, respectively.

We ran two identically structured experiments to evaluate all of our audio transforms. Experiment 1 compared various deep feature losses to baselines and contained all of the conditions listed in Table 5.1. Experiment 2 compared losses derived from different cochlear filter banks and contained all of the conditions listed in Table 5.2. 54 and 105 participants met the inclusion criteria for Experiments 1 and 2, respectively.

We also used three standard objective measures for evaluation: perceptual evaluation of speech quality (PESQ) [183], short-time objective intelligibility measure (STOI) [208], and the signal-to-distortion ratio (SDR) [219].

5.3 Results

5.3.1 Deep Feature Losses Yield Improved Denoising

The best-performing systems trained with deep perceptual feature losses outperformed both waveform-based baselines. The average objective and subjective evaluation results are shown in Table 5.1. Human listeners found the speech processed by the deep feature models to be more natural than the speech processed by the baseline models. We plot the naturalness results in more detail (Figure 5-2) for two of the best-performing models trained on each of AudioSet features (A123) and word recognition features (W123), as well as a model trained on random features (Random123), the two baselines, and the two versions of the denoising network from [75].

5.3.2 Learned vs. Random Deep Features

The benefit of deep feature losses was specific to models trained with learned features. Audio transforms trained to reconstruct random features did not produce better nat-

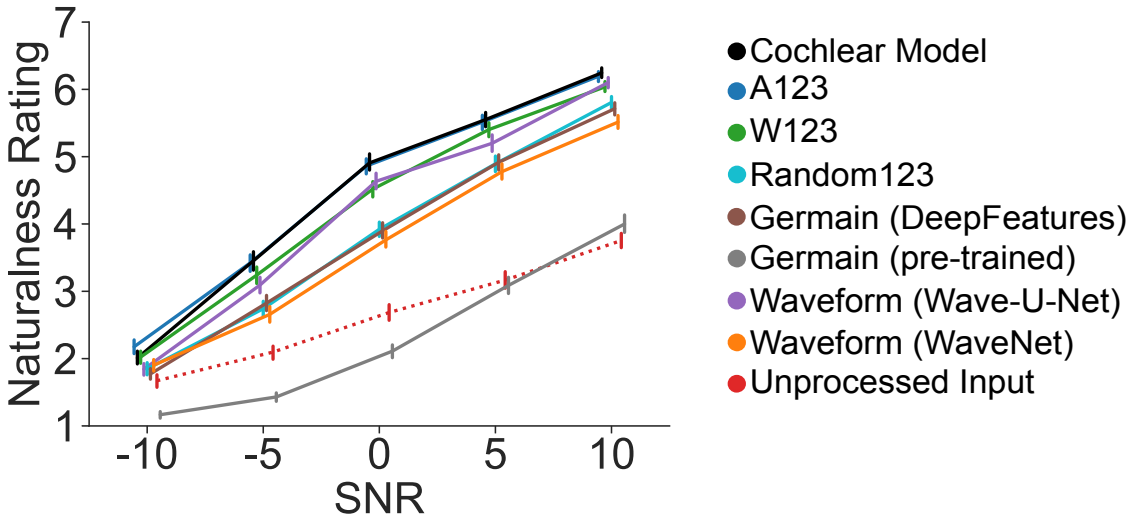


Figure 5-2: Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on deep feature losses, in addition to baseline models trained to reconstruct clean speech waveforms, and two versions of a related prior method [75]. Error bars plot SEM (across 54 participants).

uralism than the baseline WaveNet, and performed worse overall than the baseline Wave-U-Net (Figure 5-2; Table 5.1).

5.3.3 Comparison to Previous Deep Feature Systems

Our best-performing deep feature-based systems also outperformed previously published systems with deep feature losses. The pre-trained system from Germain et al. [75] generalized poorly to our test set. Furthermore, the Wave-U-Net we trained using the deep feature loss from [75] also performed worse than the baseline Wave-U-Net. These findings suggest that the features used for the perceptual loss are important, and that the DCASE task used in [75] may not have produced sufficiently general features.

5.3.4 Effect of Task Used to Train Deep Features

The best results occurred for features trained on the environmental sound recognition task – naturalism was consistently higher than for features trained on word recognition (Figure 5-2; Table 5.1). However, all of the models trained with feature losses from

our recognition networks produced more natural-sounding speech than the baselines, and than the systems trained with DCASE features based on [75]. There was no obvious benefit from training on features from three different networks.

5.3.5 Cochlear Model Losses Match Deep Feature Losses

Although deep features produced better performance than baselines trained using waveform losses, we found that we could reproduce their benefit using losses derived from the cochlear model inputs to the recognition networks. Based on rated naturalness, the transform trained with this 'cochlear' loss performed just as well as our best model trained with deep feature losses (Table 5.1).

5.3.6 Effect of Filter Bank Characteristics

The benefit of the cochlear loss depended to some extent on the filter characteristics (Table 5.2; Figure 5-3, left). Worse performance was obtained with a 'reversed' filter bank, with wide filters at low frequencies and narrower filters at high frequencies, opposite to that of the ear. Using the envelope of the filter outputs also produced worse performance (counter to the hypothesis that phase invariance might be critical). However, filters that were linearly spaced along the frequency axis worked about as well as those modeled on the ear.

Worse performance was also obtained using only five filters (scaled to cover the frequency spectrum), but good results were obtained provided at least 10 filters were used (Figure 5-3, right). This result suggests that splitting the audio up into multiple frequency channels is sufficient to replicate the benefit of deep features provided there are enough channels with reasonably sensible frequency tuning.

5.3.7 Objective Metrics

The models trained on deep recognition features also performed better than the baselines according to PESQ and STOI. Notably, this advantage was not evident when measured with SDR. The filter bank-trained models showed the opposite trend – bet-

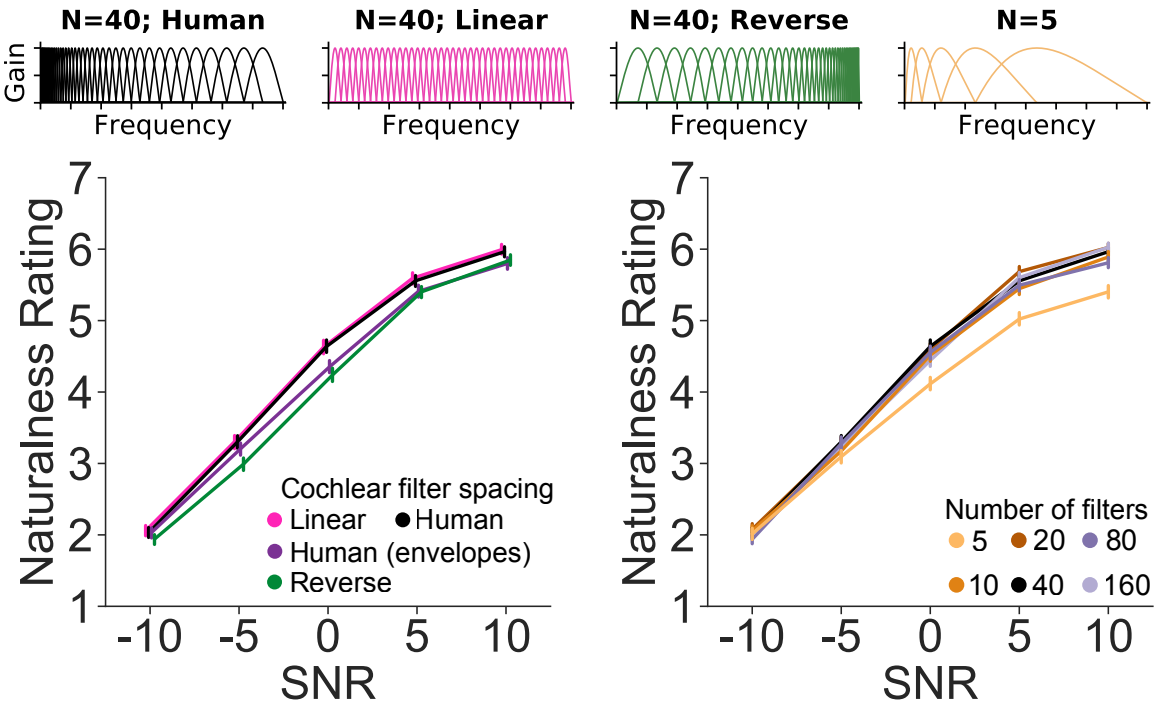


Figure 5-3: Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on cochlear model losses with different filter banks (select examples depicted above). Error bars plot SEM (across 105 participants).

ter performance as measured by SDR, and worse via PESQ and STOI (Table 5.2). These differences suggest that the filter bank and deep feature losses are not fully interchangeable despite having similar effects on overall naturalness. The results also underscore the limitations of objective metrics for capturing human perception of altered speech.

5.4 Discussion

Prior work has proposed denoising based on deep feature losses [75, 177, 206, 109, 121], but has not evaluated it relative to methods using simpler waveform- or subband-based losses. We found that deep recognition features could be used to train denoising systems that outperform waveform-based methods, but that their benefit could be matched using a standard one-layer auditory filter bank. The results thus provide no evidence that deep features provide a unique benefit for denoising.

Although deep neural networks yield the best current models of biological sensory systems [236, 125], our results indicate that these similarities are not yet sufficient to produce audio enhancement algorithms above and beyond what can be obtained from simple filter bank models. However, it is possible that building better models of human perceptual systems will also yield feature losses [4, 153] that would better transfer their perceptual benefits to humans, and produce benefits relative to simpler approaches. It also remains possible that the audio quality is limited more by the audio transform than the feature loss. More expressive transforms, or transforms with stronger generative constraints, might yield a clearer benefit of deep features.

The benefits of deep feature and cochlear model losses relative to waveform-based losses were clear from the ratings of human listeners, but were less evident in the objective metrics we tested (PESQ, STOI, SDR). This result indicates that optimizing for auditory model-based losses may provide perceptual benefits that conventional objective metrics are poorly suited to measuring, and suggests to the potential value of auditory model features as new objective metrics.

In sum, we found that audio transforms trained to modify noisy speech so as to reconstruct deep feature representations of clean speech produce better denoising performance than transforms trained to reconstruct clean speech waveforms, as measured by the ratings of human listeners. However, a similar benefit was obtained using one-layer auditory filter banks, suggesting the importance of multi-channel, overcomplete representations rather than learned features per se.

5.5 Acknowledgements

The authors thank Ray Gonzalez for developing the training dataset, John Cohn and the Oak Ridge National Laboratory for use of Summit, and the MIT-IBM Watson AI Lab for funding.

Table 5.1: Experiment 1 results. Reported metrics are averaged across the five tested SNR levels. Higher is better for all metrics.

Model name	**Loss function**	**Natural.**	**PESQ**	**STOI**	**SDR**
Cochlear model (N=40; human)	40 ERB-spaced subbands	**4.43**	1.55	0.75	7.16
A123	AudioSet features (arch123)	**4.43**	1.66	0.77	4.06
A1+W1	AudioSet + Word features (arch1)	4.36	**1.68**	0.79	6.18
A123+W123	AudioSet + Word features (arch123)	4.33	1.67	0.77	4.18
A1	AudioSet features (arch1)	4.33	1.65	0.78	3.63
W123	Word features (archs123)	4.24	1.67	**0.79**	6.64
W1	Word features (arch1)	4.22	1.63	0.77	3.30
Random1	Random features (arch1)	3.91	1.57	0.78	5.64
Random123	Random features (arch123)	3.84	1.57	0.77	5.08
Germain DeepFeatures	DCASE features from [75]	3.83	1.47	0.77	6.72
Germain (pre-trained)	DCASE features from [75]	2.36	1.14	0.64	0.93
Waveform (Wave-U-Net)	Waveform	4.17	1.51	0.76	**7.35**
Waveform (WaveNet)	Waveform	3.72	1.40	0.75	6.00
Unprocessed input		2.67	1.15	0.70	0.21

Table 5.2: Experiment 2 results. Reported metrics are averaged across the five tested SNR levels. Higher is better for all metrics.

Model name	**Loss function**	**Natural.**	**PESQ**	**STOI**	**SDR**
Cochlear model (N=20)	20 ERB-spaced subbands	**4.33**	1.54	0.77	**7.61**
Cochlear model (N=40; human)	40 ERB-spaced subbands	4.30	1.55	0.75	7.16
Cochlear model (N=160)	160 ERB-spaced subbands	4.26	1.60	0.77	7.51
Cochlear model (N=10)	10 ERB-spaced subbands	4.22	1.49	0.76	7.08
Cochlear model (N=80)	80 ERB-spaced subbands	4.21	1.53	0.74	6.69
Cochlear model (N=5)	5 ERB-spaced subbands	3.93	1.42	0.75	6.02
Cochlear model (N=40; linear)	40 linearly-spaced subbands	4.32	1.51	0.76	6.82
Cochlear model (N=40; env.)	Envelopes of 40 ERB subbands	4.16	1.59	0.75	6.94
Cochlear model (N=40; reverse)	40 reverse-ERB-spaced subbands	4.08	1.47	0.73	4.73
A123	AudioSet features (arch123)	4.27	**1.66**	**0.77**	4.06
Waveform (Wave-U-Net)	Waveform	4.17	1.51	0.76	7.35
Unprocessed input		2.47	1.15	0.70	0.21

Chapter 6

Conclusions

In this thesis, I have shown the potential value of considering the ecological perspective when studying human sound localization. The body of work presented here builds on previous modeling and behavioral research by extending the study of sound localization to natural sounds and environments. These extensions advanced our ability to account for human behavior using models. In one case, we were able to account for many psychophysical results in human hearing with a single model. In another, we demonstrated a simple local learning rule suffices to learn representations of sound location using only information available to human learners in a natural environment. The approach also led to a better understanding of factors that drive specific human behaviors in sound localization, such as the link between reverberation and the precedence effect or between spectral sparsity and natural sound localization.

We first hypothesized that an auditory task – in this case, localizing sounds in naturalistic conditions – provides substantial constraints on the methods that both models and humans can use to solve that task. In Chapter 2, we trained a model to localize sounds in naturalistic conditions and found that the resulting model reproduced a large and diverse array of human psychophysical judgments across various tasks. These comparisons, however, were limited to synthetic stimuli that had been run in prior psychoacoustics experiments, reflecting the historical focus of human sound localization research on tightly controlled stimuli such as tones and noise bursts. Chapter 3 sought to extend previous behavioral studies by measuring human local-

ization of natural sounds. We constructed a hemispheric speaker array, collected the first behavioral dataset of localization judgments of natural sounds, and found that human localization accuracy varied substantially across natural sounds. The model developed in Chapter 2 predicted these human responses to unseen natural stimuli well above chance. However, a substantial gap remains between the current model and ceiling performance. In particular, the correlation coefficients are still below the split-half reliability of the data. Additionally, the model makes errors in some cases that are much larger than the errors observed in humans. This modeling approach is a step toward models that can accurately predict human localization behavior in any scenario, which could have many practical uses, but the current model will need to be improved to realize the full potential of these benefits.

Next, we returned to the idea that human interaction with the natural environment constrains behavior and proposed a biologically inspired learning rule using self-motion to learn to localize sounds in complex environments. The model learned to localize sounds well above chance, suggesting the feasibility of such a learning rule (though the model's accuracy remained worse than that of humans, indicating a need for further refinement). Lastly, we demonstrated the value of models of human perception in applied settings where we used a model of human speech perception as a metric to improve speech denoising. This attempt demonstrates the potential applications of models of human auditory perception.

Throughout this thesis, I have worked to better align research questions to the situations and environments that humans experience in their daily lives. This approach has revealed that much of human behavior can be better understood by studying the natural environment. We hope the tools developed here will be useful in further advancing this understanding.

6.0.1 Future Directions

Our current model is a first step toward predicting human sound localization behavior in arbitrary auditory scenes. However, the discrepancies between model and human localization errors in Chapter 3 make it clear that we will need to improve the model to

provide a full account of human sound localization behavior. We began constructing the model 5 years ago, and were constrained by the availability of recorded audio and by the computing resources of that time. The technological advancements in machine learning and the proliferation of large datasets and data-sharing platforms over the past five years mean that a model built today could straightforwardly improve on our model in several respects.

Datasets and Renderer

The model's performance is likely limited by its training dataset. The dataset used to train the model in Chapter 2 contains only 385 natural sounds. While this dataset was sufficient to train a model that performs reasonably well in the real world, it seems likely that a future model could benefit from increasing the diversity of sounds in the dataset. In particular, our model makes azimuthal localization errors that are much larger than human errors for approximately twenty-five sounds in the dataset. The ten sounds localized least accurately in azimuth by the model are all musical instrument sounds gathered from the Nsynth musical instrument dataset [59]. The sounds in that dataset often contain a few dominant frequencies, which is not something well represented in our model's training dataset. The model would likely benefit from a training set that includes sounds dominated by a few specific frequencies, such a musical instruments.

We augmented the 385 training set sounds by bandpass filtering each stimulus with a bank of two-octave-wide filters, which yielded 2,492 sounds in total. The augmentation strategy used a fixed set of bandpass filters centered at predefined frequencies. This augmentation strategy meant that our final dataset had co-occurrence statistics between frequencies that likely deviated from natural sound statistics. Future models might benefit from randomizing the filter center weights.

We used a virtual acoustic world to simulate our auditory scenes and found that it captured important aspects of the natural environment. The simulator was sufficiently realistic to train a model that generalized to the real world, where it was able to accurately localize natural sounds in a real room. However, the simulator used the

image-source method [197],which assumes perfectly rigid walls. The simulator also is limited to rendering spatial audio in empty rectangular rooms using a fixed set of wall, floor, and ceiling materials. It is likely that a renderer that more directly models the physical propagation of waves would allow the model to learn to better localize in more complex environments, such as an auditory scene where a secondary object blocks the direct path between the source and listener. Speedups in modern computing hardware have enabled a new class of simulators, known as finite time difference simulators (FTDTs) [196, 210, 107, 87], that accurately model the propagation of sound waves in arbitrary environments.

Network Architecture

Identifying better architectures for the localization task may also help to build better models. Our model uses standard convolutional network networks, where each layer receives its input from only the previous layer. However, neural network architectures have diversified in recent years, including inception modules [207], residual connections [97], and transformers [216]. Many of these approaches have been successfully applied in the audio domain [103, 218, 145]. In one recent case, a transformer architecture was applied to binaural audio with promising results [134], and it seems possible that leveraging these types of architectures could yield better models of sound localization.

Our current input representations learn directly from a cochlear representation, which was roughly three times as large as a 256-by-256 image used in a standard neural network. Our architecture search was thus constrained by the size of the training examples and the limited memory (16GB) and training speed available with GPUs in 2017. These constraints induced biases in the architecture space by limiting the search to networks that needed less than 16GB of memory and that learned quickly in a small number of steps. Modern hardware has improved significantly, and a new architecture search would be able to explore a much larger space of network architectures, potentially allowing the search to find better architectures for sound localization.

Loss Functions

We chose to discretize our network output representation into bins, each corresponding to positions every 5 in azimuth and 10 in elevation. This representation may be suboptimal for localization in two respects. First, locations vary continuously in the physical world, and our discretization is likely to be too coarse to reflect the location differences that a human can resolve. Second, we used a softmax cross-entropy loss function to calculate the error between the true and predicted bins. This penalizes all incorrect location judgments equally, regardless of the magnitude of the error (i.e. the model is penalized equally for guessing 180 or 5 when the true sound position is 0). A future model may benefit from using a regression loss function which would allow some encoding of the spatial relationship between bins and represent data using a continuous metric.

Lastly, our work in Chapter 4 suggests a more biologically plausible approach to learning to localize sound that relies on head motion. However, the approach we employed used a loss function that discretizes the distance between head positions into two categories, close or far, and thus does not take full advantage of the continuous nature of head movement. Building a loss function that makes better use of the head orientation distance could yield a superior model. For example, the model might be improved with a loss function based on the goal of reporting the head motion that occurred between two excerpts of binaural audio.

6.0.2 Open Questions

Are the details of the neural network critical to explaining these results?

We have shown that the training environment has a large and somewhat interpretable effect on the behavioral phenotype of our models. However, it is unclear to what extent these results depend on the details of the model class. We chose to use neural networks because they have been effective for several tasks using natural data [133, 92, 126]. However, we cannot completely exclude the possibility that our model's behavior is a result of an interaction between the training environment and some

specific constraint imposed by convolutional neural networks or neural networks in general.

To test the specificity of these results to neural networks, new models of binaural sound localization could be constructed using other model classes, including support vector machines, gaussian processes, or probabilistic programs. If these models reproduce the same results, it would demonstrate that these results are not critical to the neural network model class. However, it remains unclear whether these alternative model classes can accurately perform a binaural localization task in complex environments. Previous work suggests that it may be possible [231, 124] and recently developed tools [68, 47] may increase the likelihood of successfully constructing such models.

Recent work has also raised the possibility that many traits observed in neural networks may be due to hyperparameter and implementation choices rather than details of the loss function or task [191]. In our work, we did not tune any network hyperparameters to achieve matches with human performance, instead relying on an architecture search to find network architectures based solely on task performance. In addition, we pooled network judgments over ten network architectures to reduce the likelihood that our results were specific to the idiosyncrasies of any one network architecture. It thus seems unlikely that the results depend critically on particular hyperparameter choices beyond the general model class that we used, But a broader architecture search enabled by current computing hardware could more definitively address this issue.

What other constraints are necessary to capture human behavior?

This thesis focused on the role of the environment and natural sounds as constraints on human sound localization behaviors. However, it is likely that some aspects of human sound localization behavior will be driven by factors specific to human physiology. These may include details of the sensory periphery, or the physical properties and limitations of neurons. The approach presented here, in which we optimized a model to perform a behavioral task under a defined set of constraints, is not limited

to investigating environmental constraints. The approach can be directly extended to any constraint that can be encoded into the model. This approach has already been fruitfully applied in models of visual scene recognition and audio word recognition, where more accurately modeling the sensory periphery resulted in models that were more closely aligned with human responses to adversarial examples [49, 48]. Additionally, a recent model of pitch perception has demonstrated how accurately replicating the cochlea in certain specific respects is critical for accurately modeling human behavior in pitch perception [187].

How can models of human sound localization be applied to improve human experiences?

The central goal of this work was to build models of human behavior to better understand the human mind and the conditions necessary to give rise to human behavior. However, it should be possible to fruitfully apply such models to auditory design problems and improve auditory experiences in immersive settings like virtual reality. For example, it may be possible to leverage model predictions to choose sounds that are likely to be well localized by human listeners. Using sounds that are well localized would provide users with a more complete sense of immersion and improve their experience.

In addition, the current work may be used to design better warning signals in cases where the signal needs to be quickly localized. For instance, the behavioral data we collected suggested that the beep associated with a truck backing up is very difficult to localize, which likely poses a safety risk. Our model could be used to screen a very large set of potential replacement sounds for the backing-up warning beep. This screening process would allow a researcher to efficiently identify candidate sounds before gathering human data, which is a much more expensive and time-consuming process.

Bibliography

[1] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. IEEE.

[2] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 99–102. IEEE.

[3] Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[4] Ishwarya Ananthabhotla, Sebastian Ewert, and Joseph A. Paradiso. Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France*, pages 1–8. ACM, 2019.

[5] Tanya L Arbogast, Christine R Mason, and Gerald Kidd Jr. The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, 112(5):2086–2098, 2002.

[6] Daniel H Ashmead, Rachel K Clifton, and Eve E Perris. Precision of auditory localization in human infants. *Developmental Psychology*, 23(5):641, 1987.

[7] Fred Attneave and Richard K Olson. Pitch as a medium: A new approach to psychophysical scaling. *The American journal of psychology*, pages 147–166, 1971.

[8] Deepak Baby, Arthur Van Den Broucke, and Sarah Verhulst. A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. *Nature machine intelligence*, 3(2):134–143, 2021.

[9] J Barker, M Cooke, S Cunningham, and X Shao. The grid audiovisual sentences corpus. *Sheffield University*, 2013.

[10] Ceren Battal, Valeria Occelli, Giorgia Bertonati, Federica Falagiarda, and Olivier Collignon. General enhancement of spatial hearing in congenitally blind people. *Psychological science*, 31(9):1129–1139, 2020.

[11] Dwight W Batteau. The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 168(1011):158–180, 1967.

[12] G von Békésy. Über die entstehung der entfernungsempfindung beim hören. *Akustische Zeitschrift*, 1938.

[13] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[14] Virginia Best, Simon Carlile, Craig Jin, and André van Schaik. The role of high frequencies in speech localization. *The Journal of the Acoustical Society of America*, 118(1):353–363, 2005.

[15] Jennifer K Bizley and Yale E Cohen. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10):693–707, 2013.

[16] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[17] Jens Blauert and W Cobben. Some consideration of binaural cross correlation analysis. *Acta Acustica united with Acustica*, 39(2):96–104, 1978.

[18] Markus Bodden and Jens Blauert. Separation of concurrent speech signals: A cocktail-party-processor for speech enhancement. In *Speech Processing in Adverse Conditions*.

[19] Léon Bottou. *Large-scale machine learning with stochastic gradient descent*, pages 177–186. Springer, 2010.

[20] Jeroen Breebaart, Steven Van De Par, and Armin Kohlrausch. Binaural processing model based on contralateral inhibition. i. model structure. *The Journal of the Acoustical Society of America*, 110(2):1074–1088, 2001.

[21] W Owen Brimijoin, Alan W Boyd, and Michael A Akeroyd. The contribution of head movement to the externalization and internalization of sounds. *PloS one*, 8(12):e83068, 2013.

[22] Adelbert W Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.

[23] Adelbert W Bronkhorst and Tammo Houtgast. Auditory distance perception in rooms. *Nature*, 397(6719):517–520, 1999.

[24] Andrew D Brown, G Christopher Stecker, and Daniel J Tollin. The precedence effect in sound localization. *Journal of the Association for Research in Otolaryngology*, 16(1):1–28, 2015.

[25] Ian C Bruce, Yousof Erfani, and Muhammad SA Zilany. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. *Hearing research*, 360:40–54, 2018.

[26] Andrew Brughera, Larisa Dunai, and William M Hartmann. Human interaural time difference thresholds for sine tones: The high-frequency limit. *The Journal of the Acoustical Society of America*, 133(5):2839–2855, 2013.

[27] Andrew Brughera, Jason Mikiel-Hunter, Mathias Dietz, and David McAlpine. Auditory brainstem models: adapting cochlear nuclei improve spatial encoding by the medial superior olive in reverberation. *Journal of the Association for Research in Otolaryngology*, 22(3):289–318, 2021.

[28] Johannes Burge. Image-computable ideal observers for tasks with natural stimuli. *Annual Review of Vision Science*, 6:491–517, 2020.

[29] Johannes Burge and Wilson S Geisler. Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, 108(40):16849–16854, 2011.

[30] Robert A. Butler. The bandwidth effect on monaural and binaural localization. *Hearing Research*, 21:67–73, 1986.

[31] Tingli Cai, Brad Rakerd, and William M Hartmann. Computing interaural differences through finite element modeling of idealized human heads. *The Journal of the Acoustical Society of America*, 138(3):1549–1560, 2015.

[32] Simon Carlile. *Virtual Auditory Space: Generation and*. Springer, 1996.

[33] Simon Carlile and Johahn Leung. The perception of auditory motion. *Trends in hearing*, 20:2331216516644254, 2016.

[34] Soumitro Chakrabarty and Emanuel A. P. Habets. Broadband doa estimation using convolutional neural networks trained with noise signals. IEEE.

[35] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[37] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[38] Brian Cheung, Eric Weiss, and Bruno Olshausen. Emergence of foveal image sampling from learning to attend in visual scenes. *arXiv preprint arXiv:1611.09430*, 2016.

[39] Taishih Chi, Powen Ru, and Shihab A Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

[40] Wing Chung, Simon Carlile, and Philip Leong. A performance adequate computational model for auditory localization. *The Journal of the Acoustical Society of America*, 107(1):432–445, 2000.

[41] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016.

[42] Rachel K Clifton, Jane Gwiazda, Joseph A Bauer, Marsha G Clarkson, and Richard M Held. Growth in head size during infancy: Implications for sound localization. *Developmental Psychology*, 24(4):477, 1988.

[43] Rachel K Clifton, Barbara A Morrongiello, John W Kulig, and John M Dowd. Newborns' orientation toward sound: Possible implications for cortical development. *Child development*, pages 833–838, 1981.

[44] H. Steven Colburn. Theory of binaural interaction based on auditory-nerve data. i. general strategy and preliminary results on interaural discrimination. 54(6):1458, 1973.

[45] M Coltheart. Visual feature-analyzers and the aftereffects of tilt and curvature. fhychol, 1971.

[46] John F Culling and Quentin Summerfield. Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *The Journal of the Acoustical Society of America*, 98(2):785–797, 1995.

[47] Marco Cusumano-Towner and Vikash K Mansinghka. A design proposal for gen: Probabilistic programming with fast custom inference via code generation. In *Proceedings of the 2nd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, pages 52–57, 2018.

[48] Joel Dapello, Jenelle Feather, Hang Le, Tiago Marques, David Cox, Josh McDermott, James J DiCarlo, and SueYeon Chung. Neural population geometry reveals the role of stochasticity in robust perception. *Advances in Neural Information Processing Systems*, 34:15595–15607, 2021.

[49] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns

improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

[50] CJ Darwin and RW Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *The Journal of the Acoustical Society of America*, 102(4):2316–2324, 1997.

[51] Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, 102(5):2892–2905, 1997.

[52] Bertrand Delgutte. Physiological mechanisms of psychophysical masking: observations from auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 87(2):791–809, 1990.

[53] Sasha Devore, Antje Ihlefeld, Kenneth Hancock, Barbara Shinn-Cunningham, and Bertrand Delgutte. Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain. *Neuron*, 62(1):123–134, 2009.

[54] Elio D Di Claudio and Raffaele Parisi. Waves: Weighted average of signal subspaces for robust wideband direction finding. *IEEE Transactions on Signal Processing*, 49(10):2179–2191, 2001.

[55] Joseph Hector DiBiase. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Thesis, 2000.

[56] Mathias Dietz, Le Wang, David Greenberg, and David McAlpine. Sensitivity to interaural time differences conveyed in the stimulus envelope: estimating inputs of binaural neurons through the temporal analysis of spike trains. *Journal of the Association for Research in Otolaryngology*, 17(4):313–330, 2016.

[57] Wenwei Zhang Chen Change Loy Shuai Yi Xuesen Zhang Wanli Ouyang Dongzhan Zhou, Xinchi Zhou. Econas: Finding proxies for economical neural architecture search. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[58] M. Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, 152:184–194, 2017.

[59] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1068–1077. JMLR. org.

[60] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

[61] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh H McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[62] Brian J. Fischer and José Luis Peña. Owl's behavior and neural representation predicted by bayesian inference. *Nature Neuroscience*, 14(8):1061–1066, 2011.

[63] 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.

[64] Roger K Furness. Ambisonics-an overview. In *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio*. Audio Engineering Society, 1990.

[65] Werner Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *The Journal of the Acoustical Society of America*, 94(1):98–110, 1993.

[66] Werner Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *The Journal of the Acoustical Society of America*, 94(1):98–110, 1993.

[67] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

[68] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

[69] William G Gardner and Keith D Martin. Hrtf measurements of a kemar. *The Journal of the Acoustical Society of America*, 97(6):3907–3908, 1995.

[70] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[71] Wilson S Geisler. Ideal observer analysis. *The visual neurosciences*, 10(7):12–12, 2003.

[72] Wilson S Geisler. Contributions of ideal observer theory to vision research. *Vision research*, 51(7):771–781, 2011.

[73] WS Geisler. Ideal observer analysis. the visual neurosciences, eds, chalupa l, werner j, 2003.

[74] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[75] François G. Germain, Qifeng Chen, and Vladlen Koltun. Speech Denoising with Deep Feature Losses. In *Proc. Interspeech 2019*, pages 2723–2727, 2019.

[76] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events: An ieee aasp challenge. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

[77] Ahna R Girshick, Michael S Landy, and Eero P Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, 14(7):926–932, 2011.

[78] Brian Glasberg and Brian C J Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[79] Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: pitting neural networks against each other as models of human recognition. *arXiv preprint arXiv:1911.09288*, 2019.

[80] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[81] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

[82] D Wesley Grantham and Frederic L Wightman. Detectability of varying interaural temporal differencesa. *The Journal of the Acoustical Society of America*, 63(2):511–523, 1978.

[83] David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[84] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[85] Benedikt Grothe, Michael Pecka, and David McAlpine. Mechanisms of sound localization in mammals. *Physiological reviews*, 90(3):983–1012, 2010.

[86] Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. Towards deep learning with segregated dendrites. *Elife*, 6:e22901, 2017.

[87] C Guiffaut and K Mahdjoubi. A parallel fdtd algorithm using the mpi library. *IEEE Antennas and Propagation Magazine*, 43(2):94–103, 2001.

[88] U. Güçlü and M.A.J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[89] ER Hafter, RH Dye Jr, and RH Gilkey. Lateralization of tonal signals which have neither onsets nor offsets. *The Journal of the Acoustical Society of America*, 65(2):471–477, 1979.

[90] Ervin R Hafter, Raymond H Dye, John M Neutzel, and Howard Aronow. Difference thresholds for interaural intensity. *The Journal of the Acoustical Society of America*, 61(3):829–834, 1977.

[91] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, and Adam Coates. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[92] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[93] Nicol S Harper and David McAlpine. Optimal neural population coding of an auditory spatial cue. *Nature*, 430(7000):682–686, 2004.

[94] William M Hartmann, Brad Rakerd, Zane D Crawford, and Peter Xinya Zhang. Transaural experiments and a revised duplex theory for the localization of low-frequency tones. *The Journal of the Acoustical Society of America*, 139(2):968–985, 2016.

[95] Monica L Hawley, Ruth Y Litovsky, and John F Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2):833–843, 2004.

[96] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[97] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[98] Jack Hebrank and Donald 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.

[99] Michael G Heinz, H Steven Colburn, and Laurel H Carney. Evaluating auditory performance limits: I. one-parameter discrimination using a computational model for the auditory nerve. *Neural computation*, 13(10):2273–2316, 2001.

[100] Olivier J Hénaff and Eero P Simoncelli. Geodesics of learned representations. *arXiv preprint arXiv:1511.06394*, 2015.

[101] G Bruce Henning. Detectability of interaural delay in high-frequency complex waveforms. *The Journal of the Acoustical Society of America*, 55(1):84–90, 1974.

[102] G Bruce Henning. Lateralization of low-frequency transients. *Hearing research*, 9(2):153–172, 1983.

[103] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[104] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[105] Paul M Hofman, Jos GA Van Riswick, and A John Van Opstal. Relearning sound localization with new ears. *Nature neuroscience*, 1(5):417–421, 1998.

[106] Heike Hofmann, Hadley Wickham, and Karen Kafadar. Value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, 26(3):469–477, 2017.

[107] Maarten Hornikx, Thomas Krijnen, and Louis van Harten. openpstd: The open source pseudospectral time-domain method for acoustic propagation. *Computer Physics Communications*, 203:298–308, 2016.

[108] Tsun-An Hsieh, Hsin-Min Wang, Xugang Lu, and Yu Tsao. WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement. *IEEE Signal Processing Letters*, 27:2149–2153, 2020.

[109] Tsun-An Hsieh, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao. Improving perceptual quality by phone-fortified perceptual loss for speech enhancement. *arXiv preprint arXiv:2010.15174*, 2020.

[110] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[111] Shinya Ito, Yufei Si, David A Feldheim, and Alan M Litke. Spectral cues are necessary to encode azimuthal auditory space in the mouse superior colliculus. *Nature communications*, 11(1):1–12, 2020.

[112] Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

[113] Nori Jacoby, Eduardo A Undurraga, Malinda J McPherson, Joaquín Valdés, Tomás Ossandón, and Josh H McDermott. Universal and non-universal features of musical pitch perception revealed by singing. *Current Biology*, 29(19):3229–3243, 2019.

[114] Eric Javel and John B Mott. Physiological and psychophysical correlates of temporal processes in hearing. *Hearing research*, 34(3):275–294, 1988.

[115] Lloyd A. Jeffress. A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1):35–39, 1948.

[116] Shilong Jiang, Lulu Wu, Peipei Yuan, Yongheng Sun, and Hong Liu. Deep and cnn fusion method for binaural sound source localisation. *The Journal of Engineering*, 2020(13):511–516, 2020.

[117] Dezhe Z Jin, Valentin Dragoi, Mriganka Sur, and H Sebastian Seung. Tilt aftereffect and adaptation-induced changes in orientation tuning in visual cortex. *Journal of Neurophysiology*, 94(6):4038–4050, 2005.

[118] James D Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on selected areas in communications*, 6(2):314–323, 1988.

[119] Philip X Joris, Philip H Smith, and Tom CT Yin. Coincidence detection in the auditory system: 50 years after jeffress. *Neuron*, 21(6):1235–1238, 1998.

[120] Oliver Kacelnik, Fernando R Nodal, Carl H Parsons, and Andrew J King. Training-induced plasticity of auditory localization in adult mammals. *PLoS biology*, 4(4):e71, 2006.

[121] Saurabh Kataria, Jesús Villalba, and Najim Dehak. Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models. *arXiv preprint arXiv:2010.11860*, 2020.

[122] Takayuki Kawashima and Takao Sato. Perceptual limits in a simulated "cocktail party". *Attention, Perception, Psychophysics*, 77(6):2108–2120, 2015.

[123] Takayuki Kawashima and Takao Sato. Perceptual limits in a simulated "cocktail party". *Attention, Perception, & Psychophysics*, 77(6):2108–2120, 2015.

[124] Hendrik Kayser and Jörn Anemüller. A discriminative learning approach to probabilistic acoustic source localization. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 99–103. IEEE, 2014.

[125] Alex Kell and Josh H. McDermott. Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55:121–132, 2019.

[126] 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.

[127] MG Kendall and A Stuart. The advanced theory of statistics volume two. *Charles Griffin and Co Ltd, London,*, 1973.

[128] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[129] Gerald Kidd Jr, Tanya L Arbogast, Christine R Mason, and Frederick J Gallun. The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, 118(6):3804–3815, 2005.

[130] Gerald Kidd Jr, Christine R Mason, Tanya L Rohtla, and Phalguni S Deliwala. Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America*, 104(1):422–431, 1998.

[131] Arne Köhn, Florian Stegen, and Timo Baumann. Mining the spoken wikipedia for speech data and beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.

[132] Konrad Paul Körding and Daniel M Wolpert. The loss function of sensorimotor learning. *Proceedings of the National Academy of Sciences*, 101(26):9839–9842, 2004.

[133] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[134] Sheng Kuang, Kiki van der Heijden, and Siamak Mehrkanoon. Bast: Binaural audio spectrogram transformer for binaural sound localization. *arXiv preprint arXiv:2207.03927*, 2022.

[135] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

[136] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018.

[137] Abhijit Kulkarni and H Steven Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, 1998.

[138] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[139] Erno HA Langendijk and Adelbert W Bronkhorst. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4):1583–1596, 2002.

[140] Michael S Lewicki. Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–363, 2002.

[141] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. ii. the law of the first wave front. *The Journal of the Acoustical Society of America*, 80(6):1623–1630, 1986.

[142] Ruth Y Litovsky. Developmental changes in the precedence effect: estimates of minimum audible angle. *The Journal of the Acoustical Society of America*, 102(3):1739–1745, 1997.

[143] Ruth Y. Litovsky, H. Steven Colburn, William A. Yost, and Sandra J. Guzman. The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654, 1999.

[144] Ruth Y Litovsky and Shelly P Godar. Difference in precedence effect between children and adults signifies development of sound localization abilities in complex listening tasks. *The Journal of the Acoustical Society of America*, 128(4):1979–1991, 2010.

[145] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE, 2020.

[146] Margaret S Livingstone and David H Hubel. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7(11):3416–3468, 1987.

[147] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.

[148] Ning Ma, Tobias May, and Guy J. Brown. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453, 2017.

[149] Craig Macartney and Tillman Weyde. Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*, 2018.

[150] Stephen L Macknik and Susana Martinez-Conde. The role of feedback in visual masking and visual processing. *Advances in cognitive psychology*, 3(1-2):125, 2007.

[151] Ewan A Macpherson and John C Middlebrooks. Listener weighting of cues for lateral angle: the duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236, 2002.

[152] James C Makous and John C Middlebrooks. Two-dimensional sound localization by human listeners. *The journal of the Acoustical Society of America*, 87(5):2188–2200, 1990.

[153] Pranay Manocha, Adam Finkelstein, Zeyu Jin, Nicholas J. Bryan, Richard Zhang, and Gautham J. Mysore. A differentiable perceptual audio metric learned from just noticeable differences. *arXiv preprint arXiv:2001.04460*, 2020.

[154] Warren P Mason. *Physical Acoustics V14: Principles and Methods*, volume 14. Elsevier, 2012.

[155] Margaret W Matlin and Hugh J Foley. *Sensation and perception*. Allyn & Bacon, 1992.

[156] Tobias May, Steven Van De Par, and Armin Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on audio, speech, and language processing*, 19(1):1–13, 2010.

[157] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[158] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493–498, 2013.

[159] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71:926–940, 2011.

[160] Richard McWalter and Josh H McDermott. Adaptive and selective time averaging of auditory scenes. *Current Biology*, 28(9):1405–1418, 2018.

[161] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):1–12, 2020.

[162] Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *bioRxiv*, page 2020.01.08.898288, 2020.

[163] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132. IEEE, 2016.

[164] John C Middlebrooks. Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, 92(5):2607–2624, 1992.

[165] Allen William Mills. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

[166] Wiktor Młynarski and Jürgen Jost. Statistics of natural binaural sounds. *PloS one*, 9(10):e108968, 2014.

[167] Barbara A Morrongiello, Kimberley D Fenwick, and Graham Chance. Sound localization acuity in very young infants: An observer-based testing procedure. *Developmental Psychology*, 26(1):75, 1990.

[168] Darwin Muir and Jeffery Field. Newborn infants orient to sounds. *Child development*, pages 431–436, 1979.

[169] Sam Norman-Haignere, Nancy G Kanwisher, and Josh H McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6):1281–1296, 2015.

[170] Simon R Oldfield and Simon PA Parker. Acuity of sound localisation: a topography of auditory space. i. normal hearing conditions. *Perception*, 13(5):581–600, 1984.

[171] Alan R Palmer and Ian J Russell. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing research*, 24(1):1–15, 1986.

[172] Ashutosh Pandey and DeLiang Wang. A new framework for cnn-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7):1179–1188, 2019.

[173] Se Rim Park and Jinwon Lee. A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*, 2016.

[174] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[175] Douglas B Paul and Janet M Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the Workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

[176] Patrick M. Peterson. Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *The Journal of the Acoustical Society of America*, 80(5):1527–1529, 1986.

[177] Rafal Pilarczyk and Władysław Skarbek. Multi-objective noisy-based deep feature loss for speech enhancement. In Ryszard S. Romaniuk and Maciej Linczuk, editors, *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019*, volume 11176, pages 858 – 865. International Society for Optics and Photonics, SPIE, 2019.

[178] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson. Speech enhancement using bayesian wavenet. In *Interspeech*, pages 2013–2017, 2017.

[179] J. Raatgever. *On the Binaural Processing of Stimuli with Different Interaural Phase Relations.* Doctoral dissertation, 1980.

[180] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

[181] Rishi Rajalingham, Kailyn Schmidt, and James J DiCarlo. Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, 35(35):12127–12136, 2015.

[182] Lord Rayleigh. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

[183] ITU-T Recommendation. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*, 2001.

[184] Stephan Gerlach Stefan Weinzierl Stefan Goetze Reinhild Roden, Niko Moritz. On sound source localization of speech signals using deep neural networks. *DAGA: Deutsche Gesellschaft für Akustik*, 2015.

[185] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[186] Suzanne K Roffler and Robert A Butler. Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America*, 43(6):1255–1259, 1968.

[187] Mark R Saddler, Ray Gonzalez, and Josh H McDermott. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, In press.

[188] TT Sandel, DC Teas, WE Feddersen, and LA Jeffress. Localization of sound from single and paired sources. *the Journal of the Acoustical Society of America*, 27(5):842–852, 1955.

[189] Olli Santala and Ville Pulkki. Directional perception of distributed sound sources. *The Journal of the Acoustical Society of America*, 129(3):1522–1530, 2011.

[190] Bruce McA Sayers and E Colin Cherry. Mechanism of binaural fusion in the hearing of speech. *The Journal of the Acoustical Society of America*, 29(9):973–987, 1957.

[191] Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *bioRxiv*, 2022.

[192] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355. IEEE.

[193] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

[194] Jan WH Schnupp and Catherine E Carr. On hearing with more than one ear: lessons from evolution. *Nature neuroscience*, 12(6):692–697, 2009.

[195] Andrew Schwartz, Josh H McDermott, and Barbara Shinn-Cunningham. Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America*, 132(1):357–368, 2012.

[196] Jonathan Sheaffer, Bruno Fazenda, et al. Wavecloud: an open source room acoustics simulator using the finite difference time domain method. *Acta Acustica united with Acustica*, 2014.

[197] B. G. Shinn-Cunningham, J. G. Desloge, and N. Kopco. Empirical and modeled acoustic transfer functions in a simple room: effects of distance and direction. IEEE.

[198] Barbara G Shinn-Cunningham, Joseph G Desloge, and Norbert Kopco. 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)*, pages 183–186. IEEE, 2001.

[199] William M Siebert. Frequency discrimination in the auditory system: Place or periodicity mechanisms? *Proceedings of the IEEE*, 58(5):723–730, 1970.

[200] G Christopher Stecker and Ervin R Hafter. Temporal weighting in sound localization. *The Journal of the Acoustical Society of America*, 112(3):1046–1057, 2002.

[201] G Christopher Stecker, Ian A Harrington, and John C Middlebrooks. Location coding by opponent neural populations in the auditory cortex. *PLoS biology*, 3(3):e78, 2005.

[202] Richard M. Stern. Lateralization of complex binaural stimuli: A weighted-image model. 84(1):156, 1988.

[203] Richard M. Stern, Wang DeLiang, and J. Brown Guy. *Binaural Sound Localization*, pages 147–185. IEEE, 2006.

[204] Stanley Smith Stevens and Edwin Broomell Newman. The localization of actual sources of sound. *The American journal of psychology*, 48(2):297–306, 1936.

[205] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[206] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. *arXiv preprint arXiv:2006.05694*, 2020.

[207] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[208] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[209] Ole Tange. *GNU parallel 2018*. Lulu. com, 2018.

[210] Matthew Reuben Thomas. *Wayverb: A Graphical Tool for Hybrid Room Acoustics Simulation*. PhD thesis, University of Huddersfield, 2017.

[211] Willard R Thurlow, John W Mangels, and Philip S Runge. Head movements during sound localization. *The Journal of the Acoustical society of America*, 42(2):489–493, 1967.

[212] James Traer and Josh H McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

[213] Constantine Trahiotis, Leslie R Bernstein, Richard M Stern, and Thomas N Buell. *Interaural correlation as the basis of a working model of binaural processing: an introduction*, pages 238–271. Springer, 2005.

[214] Fabian David Tschopp, Michael B Reiser, and Srinivas C Turaga. A connectome based hexagonal lattice convolutional network model of the drosophila visual system. *arXiv preprint arXiv:1806.04793*, 2018.

[215] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[216] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[217] Paolo Vecchiotti, Ning Ma, Stefano Squartini, and Guy J Brown. End-to-end binaural sound localisation from the raw waveform. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 451–455. IEEE.

[218] Prateek Verma and Jonathan Berger. Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions. *arXiv preprint arXiv:2105.00335*, 2021.

[219] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), 2006.

[220] Hans Wallach. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339, 1940.

[221] Hans Wallach, Edwin B Newman, and Mark R Rosenzweig. A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, 21(4):468–468, 1949.

[222] Deliang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *arXiv preprint arXiv:1708.07524*, 2017.

[223] Hong Wang and Mostafa Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):823–831, 1985.

[224] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604, 2002.

[225] Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993.

[226] Michael Wertheimer. Psychomotor coordination of auditory and visual space at birth. *Science*, 134(3491):1692–1692, 1961.

[227] Frederic L Wightman and Doris J Kistler. Headphone simulation of free-field listening. ii: Psychophysical validation. *The Journal of the Acoustical Society of America*, 85(2):868–878, 1989.

[228] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

[229] Katherine C Wood and Jennifer K Bizley. Relative sound localisation abilities in human listeners. *The Journal of the Acoustical Society of America*, 138(2):674–686, 2015.

[230] John Woodruff and DeLiang Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1503–1512, 2012.

[231] John Woodruff and DeLiang Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1503–1512, 2012.

[232] Kevin J.P. Woods, Max H. Siegel, James Traer, and Josh H. McDermott. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, 79:2064–2072, 2017.

[233] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L. Jones, Eng Siong Chng, and Haizhou Li. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. IEEE.

[234] Sarthak Yadav and Mary Ellen Foster. Gise-51: A scalable isolated sound events dataset. *arXiv preprint arXiv:2103.12306*, 2021.

[235] Tatyana A Yakusheva, Aasef G Shaikh, Andrea M Green, Pablo M Blazquez, J David Dickman, and Dora E Angelaki. Purkinje cells in posterior cerebellar vermis encode motion in an inertial reference frame. *Neuron*, 54(6):973–985, 2007.

[236] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

[237] 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.

[238] Yeo-Sun Yoon, Lance M Kaplan, and James H McClellan. Tops: New doa estimator for wideband signals. *IEEE Transactions on Signal processing*, 54(6):1977–1989, 2006.

[239] William A Yost and Raymond H Dye Jr. Discrimination of interaural differences of level as a function of frequency. *The Journal of the Acoustical Society of America*, 83(5):1846–1851, 1988.

[240] William A Yost, Louise Loiselle, Michael Dorman, Jason Burns, and Christopher A Brown. Sound source localization of filtered noises by listeners with normal hearing: A statistical analysis. *The Journal of the Acoustical Society of America*, 133(5):2876–2882, 2013.

[241] William A Yost and Xuan Zhong. Sound source localization identification accuracy: Bandwidth dependencies. *The Journal of the Acoustical Society of America*, 136(5):2737–2746, 2014.

[242] Paul Thomas Young. The role of head movements in auditory localization. *Journal of Experimental Psychology*, 14(2):95, 1931.

[243] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE CVPR 2018*, pages 586–595, 2018.

[244] Xuan Zhong and William A Yost. How many images are in an auditory scene? *The Journal of the Acoustical Society of America*, 141(4):2882–2892, 2017.

[245] Yi Zhou, Laurel H Carney, and H Steven Colburn. A model for interaural time difference sensitivity in the medial superior olive: interaction of excitatory and inhibitory synaptic inputs, channel dynamics, and cellular morphology. *Journal of Neuroscience*, 25(12):3046–3058, 2005.

[246] Muhammad SA Zilany, Ian C Bruce, and Laurel H Carney. Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America*, 135(1):283–286, 2014.

[247] Nathaniel J Zuk and Bertrand Delgutte. Neural coding and perception of auditory motion direction based on interaural time differences. *Journal of Neurophysiology*, 122(4):1821–1842, 2019.

[248] Patrick M Zurek. The precedence effect and its possible role in the avoidance of interaural ambiguities. *The Journal of the Acoustical Society of America*, 67(3):952–964, 1980.

[249] Jozef Zwislocki and RS Feldman. Just noticeable differences in dichotic phase. *The Journal of the Acoustical Society of America*, 28(5):860–864, 1956.